

Ph.D. Thesis

Architecting Physical Information Space: AI-Enabled Assistive Technology for Non-Visual Cooking

Franklin Mingzhe Li

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
mingzhe2@cs.cmu.edu

CMU-HCII-26-106

June, 2026

Thesis Committee:

Patrick Carrington, HCII, CMU, Chair
John Zimmerman, HCII, CMU
Chris Harrison, HCII, CMU
Shaun K. Kane, Google
Gregory Abowd, Northeastern University

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

Copyright © 2026 Franklin Mingzhe Li

Abstract

Many Activities of Daily Living (ADLs) and Instrumental Activities of Daily Living (IADLs), such as cooking, are highly visual and create barriers for people with vision impairments. With over 2.2 billion people globally affected by vision impairment, these challenges significantly impact autonomy, health, and overall quality of life. Cooking, in particular, involves navigating the kitchen, identifying ingredients, and following recipes, which often leads to reliance on unhealthy, costly pre-packaged meals. This contributes to a 150% higher obesity rate among people with vision impairments. Despite advancements in assistive technology, there remains a critical gap in providing solutions that address non-visual interaction with the physical information space required for complex tasks like cooking.

This thesis addresses that gap by asking how AI-enabled assistive technologies can support non-visual interaction with physical activities, using cooking as a concrete and demanding setting. For an AI-based system to be useful in a kitchen, it needs structured information about objects, environments, and user actions; a generalized text or vision model alone cannot supply the contextual information a blind cook actually needs at the moment they need it. I argue that progress requires first *architecting a physical information space* for non-visual cooking, a structured account of what to track, why it matters, and how it should be communicated, and then building AI-driven systems on top of that representation.

The thesis is structured around six research efforts. First, a content analysis of 122 YouTube videos and semi-structured interviews with 12 blind cooks established the practices and challenges of non-visual cooking, identifying eight key challenges that current technology fails to address. Second, an interview study with 20 blind cooks and four rehabilitation instructors produced a design framework for accessible recipes, characterizing what content, structure, and technological support a recipe needs in order to be usable non-visually. Third, an in-home contextual inquiry with 12 blind cooks produced a taxonomy of the contextual information about kitchen objects (five primary, two secondary, and one application-specific category) that is required to act safely and confidently in the kitchen. Fourth, OSCAR (Object Status Context Awareness for Recipes) operationalized one piece of that information space: it extracts object-status transitions from recipes, aligns them with cooking video, and uses a time-causal model to track recipe progress, improving step-prediction accuracy by more than 20% over text-only baselines on both YouCook2 and a newly collected dataset of 12 real-world non-visual cooking sessions.

Building on these four studies, the thesis introduces the Object Dependency Graph (ODG), a unifying representation of a recipe that integrates object status, prerequisites, safety, non-visual cues, measurements, progress tracking, and parallelism into a single, machine-readable structure. The ODG is decomposed into seven design rules (R1–R7), each explicitly traceable to one of the prior chapters, and is implemented as a voice-first cooking assistant paired with a researcher dashboard. I evaluated the ODG-based assistant in a within-subjects remote study with 10 blind and low-vision participants comparing it against a screen-reader baseline across nine recipes spanning multi-stage, linear, and high-parallelism cooking flows. The study used matched task probes, custom Likert measures mapped to specific ODG rules, and a closing semi-structured interview.

The evaluation shows that the ODG materially improves the dimensions of non-visual cooking that screen readers currently fail to support: hazard awareness, parallel-task identification, and measurement and technique clarity all improved substantially under the ODG condition, while progress orientation showed a mixed pattern that points to concrete onboarding and verbosity refinements. Participants also articulated when they would and would not adopt a voice-first assistant, surfacing populations the system does not yet serve well. Together, the chapters of this thesis show that non-visual cooking is best supported not by better screen readers or better vision models in isolation, but by an explicit information space, the ODG, that ties what a system can perceive to what a blind cook actually needs to know while cooking.

Acknowledgements

This thesis is based on research projects that I personally led. In the framing, synthesis, and evaluation chapters I write in the first person singular (“I”), since these reflect my own argument and contributions; the chapters adapted from collaboratively published studies retain the first person plural (“we”) out of respect to all of my collaborators who contributed to that research.

This thesis would not have been possible without my PhD advisor, Patrick Carrington, who took a chance on me when I arrived at CMU and gave me the freedom, trust, and resources to define a research agenda that I cared about. Patrick has been remarkably patient with me through every iteration of this work, from the earliest content analysis of cooking videos through OSCAR to the system that became the ODG, and has shown me, by example, what it means to do accessibility research that is rigorous, humane, and grounded in the lived experiences of the people we are designing with and for. The values I will carry forward from this PhD are, in large measure, the values I learned from him.

I am also deeply grateful to the rest of my thesis committee. John Zimmerman has consistently pushed me to think about my work at the level of the design argument it makes, not just the system it produces, and his questions have shaped this thesis more than he probably realizes. Chris Harrison helped me see what it means to make a research artifact *real*: to build it, ship it, and own its consequences. His feedback on the OSCAR and ODG systems was invaluable. Shaun K. Kane has been a generous mentor on accessibility research and a direct collaborator on several of the projects that fed into this thesis (the contextual inquiry, the recipe study, the audio-description authoring work), and his perspective has been a north star for mine. Gregory Abowd’s feedback on physical and ubiquitous information spaces is one of the conceptual foundations of this thesis, and I am honored that he agreed to join the committee from Northeastern.

I owe an enormous debt to the mentors who set me on this path before I arrived at CMU. Khai N. Truong took me on as an undergrad and a master’s student at the University of Toronto, taught me what HCI research looks like up close, and continued to collaborate with me long after I left his lab; the FMT, BrailleSketch, captioning-practices, face-recognition, and assistive-devices papers are all the product of that mentorship. Mingming Fan has been a collaborator, mentor, and friend at every stage of my career, from the very first papers we wrote together through to the VR-gesture work with Mingming Fan’s lab at HKUST. Michelle Annett, Matthew Lakier, Daniel Wigdor, Tovi Grossman, and George Fitzmaurice gave me my first taste of what a thriving research community could feel like during the Living Room project.

I was lucky to spend my industry internships with mentors who shaped how I think about deployment, design, and accessibility at scale. At Google, Cynthia L. Bennett and Shaun K. Kane hosted me on the audio-description authoring work that became the ADCanvas project. At Apple, Jeffrey P. Bigham and Xiaoyi Zhang gave me a window into accessibility engineering inside one of the largest assistive-technology platforms in the world. The thinking I did during those internships is woven through this thesis even where it is not visible on the surface.

I have been extraordinarily fortunate in the collaborators I have worked with across the projects that surround this thesis. Brad A. Myers and Aniket Kittur at the HCII brought rigorous

design and sensemaking lenses to the Selenite project that I learned a great deal from. Sarah E. Fox, Nikolas Martelaro, and Daragh Byrne welcomed me into the public-robot co-design work alongside Howard Ziyu Han and Alesandra Baca Vazquez. Toby Jia-Jun Li, together with Zheng Ning, Leyang Li, Daniel Killough, JooYoung Seo, Yapeng Tian, and Yuhang Zhao, built AROMA on top of the OSCAR object-status representation and gave this thesis one of its most direct pieces of external follow-up evidence.

Beyond CMU, my collaborators across accessibility research have shaped how I think about almost every chapter of this thesis. I am grateful to Leah Findlater, Jacob O. Wobbrock, Kristen Shinohara, Maryam Bandukda, Abigale Stangl, Lotus Zhang, and Zhuohao Jerry Zhang for the visual-arts and visual-content authoring projects; to Chieko Asakawa, Daisuke Sato, Akihiko Oharazawa, Chloe Qingyu Zhu, and Misty Fan for the more recent procedural-feedback work on non-visual makeup; to Cheng Zhang, Wei Sun, Benjamin Steeper, Songlin Xu, Congshu Huang, Zhenyu Lei, Songyun Tao, Hyunchul Lim, Nam Anh Dang, Dylan Lee, Tianhong Catherine Yu, Jane Lu, Yiqi Jin, Yan Ma, Xiaojun Bi, François Guimbretière, and Feng Tian for the wearable-sensing collaborations that taught me how to think about input for accessibility; and to Zhicong Lu and Cheng Lu for the captioning-practices work.

I am especially grateful to the collaborators who worked directly on pieces of this thesis. Michael Xieyang Liu has been a constant collaborator and friend throughout both of our PhDs, and his fingerprints are on the contextual inquiry, the recipe-access study, and the input-modality work. Jamie Dorst and Peter Cederberg co-authored the first non-visual cooking practices study with me and shaped the early framing of the entire research agenda. Ashley Wang co-led the recipe-access design framework. Kaitlyn Ng and Bin Zhu collaborated on the OSCAR object-status work, and Bin Zhu has continued to push that work forward through the OSCBench benchmarking project with Xianjing Han, Shiqi Hu, Roger Zimmermann, and Jingjing Chen. Francheska Spektor, Meng Xia, Mina Huh, and Yuqi Gong worked with me on the makeup-and-cosmetics study whose findings about non-visual sensory feedback continue to inform how I think about the ODG. Di Laura Chen, Mohammad Kianpisheh, Yang Zhang, and Yunzhi Li collaborated on the earlier assistive-technology projects whose threads I followed into the kitchen work. Tongshuang Wu and Tianying Chen were core to the sensemaking work that, in retrospect, gave me the vocabulary for thinking about *structured information* that became the ODG. Jingze Tian, Yingna Wang, Keye Yu, Liyi Xu, Junan Xie, and Yafeng Niu collaborated on the VR-gesture work with Mingming Fan’s lab. Mengjie Tang, Xinman Li, Juxiao Zhang, and Zhuying Li worked with me on the nature-engagement study with blind people. Zihe Ran, Xiyu Li, Qing Xiao, Xianzhe Fan, and Yanyun Wang collaborated on the mobile-games accessibility work. Tianyang Wen, Xucheng Zhang, Zhirong Wan, Jing Zhao, Yicheng Zhu, Ning Su, Xiaolan Peng, and Jin Huang collaborated on the PANDA driving-aid project.

I am grateful to the AXLE Lab community at CMU: Atieh Taheri, Yunzhi Li, Frank Elavsky, Joon Jang, Sanika Moharana, Ezra Awumey, Kaitlyn Ng, Eric Davidson, and the AXLE alumni. Beyond AXLE, the HCII at Carnegie Mellon has been an extraordinary place to do a PhD, and I am grateful to the many friends, colleagues, and labmates who made the past six years what they were. I want to thank everyone who shared offices, hallways, study sessions, and late-night discussions with me, and the HCII staff whose care holds the community together. I am also grateful to the broader accessibility-research community at ASSETS, CHI, UIST, and CSCW that has been my intellectual home throughout this PhD.

This work would not exist without the dozens of blind and visually impaired participants who invited me into their kitchens, shared their cooking practices with me, tolerated my prototypes, and corrected me when I was wrong. Every finding in this thesis is, at its root, something they taught me. I am also grateful to the rehabilitation instructors who shared their expertise, and to the staff at the vision-rehabilitation centers who helped make the studies possible. I thank the anonymous reviewers whose feedback shaped the published versions of this work, and the funding agencies, internship programs, and institutional partners that supported it.

Finally, to my family and friends: to my parents and grandma, whose belief in me has been unwavering across the distance and the time zones, and to the friends, both inside and outside academia, who reminded me to step out of the lab, eat a real meal, and laugh, thank you. None of this would have happened without you.

Contents

1	Introduction	1
1.1	Overview	2
1.2	Contributions	3
2	Background and Related Work	5
2.1	Cooking and Technology as a Research Domain	5
2.1.1	The Cultural and Everyday Importance of Cooking	5
2.1.2	Interactive Cooking Systems	6
2.1.3	Recipe Design and Instruction Structure	6
2.1.4	Assumptions and Gaps in Cooking Technology Design	7
2.2	Information Spaces for Physical Activity	7
2.2.1	Information Spaces as a Design Construct	7
2.2.2	Distributed Cognition and Embodied Interaction	8
2.2.3	From Theory to a Physical Information Space for Cooking	8
2.3	People with Vision Impairments and Cooking	8
2.3.1	Vision Impact on Cooking and Food-Related Activities	9
2.3.2	Challenges and Adaptations in the Kitchen	9
2.3.3	The Need for Real-Time Contextual Support	9
2.4	Assistive Technologies for People with Vision Impairments in Cooking	10
2.4.1	Tactile-Based Technologies	10
2.4.2	Vision-Based Technologies	10
2.4.3	Auditory and Voice-Controlled Technologies	10
2.4.4	Multimodal Technologies	11
2.4.5	Large Language Models, Vision-Language Models, and Conversational AI	11
2.5	Summary and Positioning of the Thesis	12
3	Exploration of Non-Visual Cooking Practices and Challenges	13
3.1	Overview	13
3.2	Technology Scaffolds for Cooking	14
3.3	Methodology	14
3.3.1	YouTube Video Analysis: Cooking Practices for Visually Impaired People	14
3.3.2	Interview Study: Challenges of Cooking for People with Visual Impairments	16
3.4	Findings	17
3.4.1	Cooking Practices of People with Visual Impairments	17
3.4.2	Challenges of Non-Visual Cooking	22

3.5	Discussion and Future Work	28
3.5.1	Zero-touch Interactions for Cooking	28
3.5.2	Status Tracking and Safety Monitoring	28
3.5.3	Collaborative Cooking	29
3.5.4	Contextual Inquiry and Learning Process	29
3.6	Summary	29
4	Design Framework for Learning and Following Accessible Recipes	31
4.1	Overview	31
4.2	Related Work on Recipe Structure and Access	32
4.3	Methodology	32
4.3.1	Semi-structured Interview with Blind Cooks	34
4.3.2	Semi-structured Interview with Adaptive Cooking Instructors	35
4.3.3	Data Analysis	35
4.4	Findings: Current Use of Recipes while Cooking (RQ1)	36
4.4.1	Searching for and Editing Recipe Information	36
4.4.2	Preparing Workspace and Tools	38
4.4.3	Following Recipes while Cooking	40
4.5	Findings: Recipe Content and Structure (RQ2)	42
4.5.1	Recipe Content Preferences	42
4.5.2	Recipe Structure Preferences	45
4.6	Findings: How Technology Can Improve Recipe Accessibility (RQ3)	47
4.6.1	Improving Recipe Organization, Navigation, and Level of Detail	47
4.6.2	Supporting Users in Learning to Cook	48
4.6.3	Integrating Technology into Kitchen Activities	50
4.7	Discussion and Future Work	51
4.7.1	Augmenting Recipes to Make Them More Accessible	51
4.7.2	Facilitating Information Extraction for Kitchen Tools and Ingredients	52
4.7.3	Supporting More Accessible Kitchen Work	53
4.8	Summary	54
5	Recognizing Contextual Information of Objects	56
5.1	Overview	56
5.2	Related Work on Accessing Contextual Information of Objects	58
5.3	Methodology	58
5.3.1	Participants	58
5.3.2	Study Procedure	60
5.3.3	Data Analysis	62
5.4	Findings: Contextual Information Needs of Objects (RQ1)	62
5.4.1	Five Fundamental Categories of Contextual Information for Objects	63
5.4.2	Secondary and Application-specific Information	65
5.5	Findings: Process of Obtaining Contextual Information of Objects (RQ2)	67
5.5.1	Multi-sensory Approaches to Obtaining Contextual Information	67
5.5.2	Altering Objects to Ease Identification	70
5.5.3	Optimizing Cooking Procedures	71

5.6	Findings: Contextual Information Communication and Deployment (RQ3)	73
5.6.1	Preferred Information Granularity and Level of Detail	73
5.6.2	Preferred Communication Modalities	74
5.6.3	Deploying New Technologies in the Kitchen	75
5.7	Discussion and Future Work	75
5.7.1	Importance and Opportunities for AI to Improve Contextual Awareness	76
5.7.2	Creating Smart Objects in the Kitchen	76
5.7.3	Augmenting the Kitchen or the User	76
5.7.4	Tracking Activities in the Background	77
5.7.5	Multimodal Interaction in the Kitchen	77
5.8	Summary	77
6	OSCAR: Object Status Context Awareness for Recipes	79
6.1	Overview	79
6.2	Background and Related Work	80
6.2.1	Recipe Structuring and Navigation	81
6.2.2	Dataset Creation for Accessibility Research	81
6.3	Technical Approach: Object Status Recognition for Recipe Progress Tracking	82
6.3.1	Recipe Formatting and Object Status Extraction	83
6.3.2	Visual Data Extraction and Recipe Step Alignment	83
6.3.3	Log of Similarity Metrics and Sequential Predictions	83
6.3.4	Time-Causal Model	84
6.3.5	Progress Tracking	84
6.4	Evaluation on Large-Scale Recipe Instruction Videos	85
6.4.1	Dataset Description	85
6.4.2	Methodology	85
6.4.3	Results	86
6.4.4	Why Object Status Improves Step Prediction in Large-scale Recipe Instruction Videos	87
6.4.5	Factors Affecting Prediction Performance	88
6.5	Evaluation with Real-World Non-Visual Cooking Dataset	91
6.5.1	Dataset Overview	91
6.5.2	Methodology	92
6.5.3	Results: Evaluating Step Prediction in Real-World Non-Visual Cooking	93
6.5.4	Why Object Status Improves Step Prediction in Non-Visual Cooking	93
6.5.5	Factors Affecting Prediction Performance and Design Implications	95
6.6	Discussion	98
6.6.1	Rethinking Progress Tracking in Real-World Non-Visual Activities	98
6.6.2	Vision-Language Models in Messy Real-World Conditions	99
6.6.3	Object Status as a Universal Design Primitive	99
6.6.4	Modeling Context, Memory, and Temporal Continuity	100
6.6.5	Toward Context-Aware Interaction with LLMs	100
6.6.6	Designing for Autonomy over Standardization	101
6.6.7	Interaction and Camera Design for Blind Cooks	101
6.6.8	Expanding Accessibility Datasets for Inclusive AI	101

6.7	Limitations and Future Work	102
6.8	Summary	102
7	ODG: An Object Dependency Graph for Non-Visual Cooking Assistance	103
7.1	Overview	103
7.2	Background and Related Work	104
7.2.1	Structured Representations of Cooking Procedures	105
7.2.2	Voice-First and Multimodal Cooking Assistants	105
7.2.3	Conversational and Mixed-Initiative Agents for Procedural Tasks	106
7.2.4	Accessibility-Focused Recipe and Cooking Support	107
7.3	Design of the Object Dependency Graph	107
7.3.1	Design Rationale: Four Prior Chapters, One Representation	107
7.3.2	The Seven ODG Rules	108
7.3.3	R1: Object Status and Step Narration	110
7.3.4	R2: Sensory Cues	110
7.3.5	R3: Safety Alerts	111
7.3.6	R4: Parallel Tasks	112
7.3.7	R5: Precise Measurements and Technique	113
7.3.8	R6: Progress Tracking and Recovery	114
7.3.9	R7: Automatic Timers	115
7.3.10	The Object Dependency Graph: A Common Substrate for the Seven Rules	116
7.4	Summary	118
8	ODG System Implementation and User Evaluation	120
8.1	Overview	120
8.2	System Implementation	121
8.2.1	Implementation Architecture	121
8.2.2	Participant-Facing Assistant (Condition B)	123
8.2.3	Researcher Dashboard (Wizard-of-Oz Rule Overlays)	126
8.3	User Study	129
8.3.1	Participants	129
8.3.2	Design, Conditions, and Recipes	129
8.3.3	Procedure	131
8.3.4	Data Sources and Analysis	135
8.4	Findings	136
8.4.1	Recipe Characteristics and Rule Coverage	137
8.4.2	Likert Ratings by Condition and Rule	137
8.4.3	Preferences for the ODG Rules	139
8.4.4	Preferences for Adopting the ODG-Aware Assistant in Cooking Scenarios	141
8.5	Discussion	143
8.5.1	The Gains Concentrate Where Mainstream Recipes Fail Blind Cooks . . .	143
8.5.2	Coverage and Recovery, Not Throughput	144
8.5.3	Personalization Is a First-Class Design Dimension for the Same Rule Set	144
8.6	Summary	145

9	Contributions, Discussion, and Future Work	146
9.1	Overview	146
9.2	Synthesis: From Understanding to Delivery	146
9.3	Contributions of This Thesis	148
9.3.1	Empirical Contributions	148
9.3.2	Conceptual Contributions	148
9.3.3	Technical and Artifact Contributions	149
9.3.4	Dataset Contributions	149
9.4	Design Guidelines for AI-Enabled Assistive Technology in Physical Activities . .	149
9.5	Limitations	151
9.6	Directions for the Research Community	153
9.6.1	Generalizing the Physical Information Space Beyond Cooking	153
9.6.2	Scaling the Substrate: From AI-Drafted ODGs to Community-Stewarded Authoring	153
9.6.3	From Perception-Augmented Assistance to Embodied Physical Intelligence	154
9.6.4	From Snapshots to Long-Term, Personalized Assistive Partnerships	155
9.6.5	The Relational Dimensions of AI Assistance for Blind Users	155
9.7	Reflection on the Research Process	156
	Bibliography	158

Chapter 1

Introduction

People with vision impairments face significant challenges in performing Activities of Daily Living (ADLs), such as eating and mobility, and Instrumental Activities of Daily Living (IADLs), including cooking [18]. Vision impairment, affecting over 2.2 billion people worldwide, has a profound impact on autonomy, health, and overall quality of life [191]. Cooking, in particular, poses unique challenges, including navigating the kitchen, identifying ingredients, and following recipes, often leading to a reliance on pre-packaged meals that are unhealthy and costly. This contributes to a 150% higher obesity rate among people with vision impairments compared to those without disabilities [32]. Beyond nutrition, these barriers also limit the ability of visually impaired individuals to engage in cooking as a social activity, further affecting their autonomy and social interactions [110]. Although assistive technologies have improved tasks such as navigation and discrete object identification, there remains a significant gap in their ability to support complex physical activities like cooking, which require real-time contextual information and step-by-step guidance grounded in the state of the physical world.

Cooking, as a dynamic physical activity, presents multiple challenges at various stages—before, during, and after meal preparation [68]. For those who have recently lost their vision, extensive rehabilitation is often required to relearn cooking skills, and even after this training, challenges persist when attempting unfamiliar dishes. These challenges include finding accessible recipe information, tracking ingredients, organizing the kitchen space, and understanding the status of objects (e.g., determining whether food is fully cooked or whether ingredients are at the right temperature) [110]. Mainstream AI vision apps can already help with discrete tasks such as recognizing the denomination of currency or the color of clothing [135, 144], but more complex tasks like cooking require a deeper understanding of physical context and ongoing changes in the environment [110]. An AI-based cooking assistant must detect, track, and interpret the status of objects in real time while also guiding users through multi-step processes that branch, overlap, and recover from interruptions.

Central to addressing these challenges is the creation of a *physical information space* for non-visual cooking—a structured representation that an AI system can use to understand the physical environment, know what each cook needs in the moment, and provide contextual, actionable support. This notion extends the established concept of an *information space* [57] from digital documents to the physical, continuously changing environment of a hands-on activity. Constructing such a physical information space requires answering three coupled questions: (1) *what* does the system need to track about the kitchen and the cooking process, (2) *why* does each piece of

that information matter for a blind cook, and (3) *how* should the system communicate it through a non-visual channel? Unlike tasks that can be addressed with static datasets, physical activities like cooking demand continuously updated information about the environment, objects, and user actions; an assistant must reason, for example, about whether a pan has reached the correct temperature or whether food has been properly prepared. **This thesis argues that progress in AI-enabled assistive technology for non-visual cooking depends on first *architecting* this physical information space as an explicit, structured object, and then building AI systems on top of that representation rather than directly on top of recipe prose or general-purpose vision-language models.** The Object Dependency Graph (ODG), the central technical contribution of this thesis, is one concrete instance of such a representation, and the voice-first cooking assistant built on top of it is one concrete demonstration of how that representation can be delivered to blind cooks.

The research efforts presented in this thesis trace this argument through six moves: **understand** the practices and challenges of cooking without vision; **design** accessible recipe systems for non-visual access; **inquire** into the contextual object information blind cooks rely on while cooking; **build** OSCAR (Object Status Context Awareness for Recipes), a vision-language pipeline that tracks object-status transitions and aligns them to recipe steps; **represent** the resulting information space as the Object Dependency Graph (ODG), decomposed into seven design rules (R1–R7) each traceable to a specific prior study; and **deliver and evaluate** the ODG as a voice-first cooking assistant, compared against a screen-reader workflow as a familiar reference condition under a within-subjects user study. The remainder of this chapter develops each of these moves in turn.

1.1 Overview

Given the limited research on non-visual cooking, my work began by conducting a content analysis of 122 YouTube videos and semi-structured interviews with 12 visually impaired individuals to understand their cooking practices and challenges (Chapter 3). I identified and categorized practices for 12 different cooking activities, such as measuring ingredients and testing food for doneness, and surfaced eight cross-cutting challenges, including accessing information while cooking and determining the status of objects. These findings frame what an assistive system for non-visual cooking must support if it is to deliver information that accurately reflects the needs of visually impaired cooks.

Building on this foundation, I focused on how visually impaired individuals access and follow recipes, since this is a critical bottleneck in non-visual cooking (Chapter 4). I conducted further interviews with 20 visually impaired cooks and four instructors from a vision rehabilitation center to explore how users interact with recipe information at various stages of cooking. This research uncovered challenges such as the need to wash hands frequently when interacting with recipe readers and difficulties in structuring recipe information, and led to a design framework for accessible recipe content, structure, and supporting technology.

In addition to improving recipe access, it is crucial to provide real-time object status information (e.g., temperature, shape, expiration date, and location) during cooking, since this directly impacts the confidence and independence of visually impaired cooks (Chapter 5). Existing vision apps often struggle to provide this type of contextual information in a cooking environment.

To explore this further, I conducted a contextual inquiry with 12 visually impaired participants in their own kitchens, identifying eight categories of contextual information they rely on while cooking, along with their preferences for how that information should be communicated.

To bridge the perception gap identified in the prior studies, I built OSCAR (Chapter 6), a vision-language system that extracts object status from recipes, aligns visual frames to recipe steps, and provides a context-aware progress tracker for following a recipe in real time. OSCAR demonstrated that reasoning about object *status transitions* substantially improves recipe step prediction over text-only baselines on both YouCook2 and a newly collected dataset of real-world non-visual cooking sessions, and made the gap between *perceiving* object status and *delivering* it to a blind cook visible as the next problem to solve.

Synthesizing the four prior studies, I then developed the Object Dependency Graph (ODG) (Chapter 7), a unifying representation of a recipe that integrates object status, prerequisites, safety, non-visual cues, measurements, progress tracking, and parallelism into a single machine-readable structure. The ODG decomposes into seven design rules (R1–R7), each grounded in a specific prior chapter, that govern the behavior of an ODG-aware assistant and that can be toggled independently at run time. I implemented the ODG as a voice-first cooking assistant with a paired researcher dashboard (Chapter 8) and evaluated it in a within-subjects remote study with 10 blind and low-vision participants, organized around three research questions on user experience, per-rule mechanism, and adoption in everyday cooking. The thesis closes (Chapter 9) with a synthesis of findings across all prior chapters, a consolidated list of empirical, conceptual, technical, and dataset contributions, seven concrete design guidelines distilled from the work, and a roadmap for future directions.

1.2 Contributions

The series of studies and systems presented in this thesis advances both the empirical understanding of non-visual cooking and the technical machinery for supporting it. The contributions are grouped below by type and consolidated in full in Chapter 9.

Empirical contributions.

- A characterization of non-visual cooking practices and challenges through content analysis of cooking videos and semi-structured interviews with visually impaired cooks (Chapter 3).
- A design framework and opportunities for non-visual recipe access, derived from interviews with both blind cooks and cooking instructors at a vision rehabilitation center (Chapter 4).
- A taxonomy of contextual object information needs for non-visual cooking, constructed through a contextual inquiry in blind participants’ own kitchens (Chapter 5).
- A within-subjects user study with 10 blind and low-vision participants that examines *how* blind cooks experience an ODG-aware assistant end-to-end, *how* each of the seven rules shapes what the assistant says and when, and *how* an ODG-aware assistant would fit into everyday cooking, using a screen-reader workflow as a familiar reference condition together with custom Likert measures mapped to specific rules and a closing semi-structured interview (Chapter 8).

Conceptual contributions.

- The articulation of a *physical information space* for non-visual cooking, framed as a structured account of what an AI assistant needs to track, why each piece matters, and how it should be communicated.
- The Object Dependency Graph (ODG), a unifying representation of a recipe that integrates object status, prerequisites, safety, non-visual cues, measurements, progress tracking, and parallelism into a single structure grounded in the four prior contributions of this thesis (Chapter 7).
- A decomposition of the ODG into seven design rules (R1–R7), each traceable to a specific prior chapter, that govern the behavior of an ODG-aware voice-first assistant and that can be independently enabled or disabled at run time (Chapter 7).
- A thesis-wide synthesis that distills seven concrete design guidelines (G1–G7) for AI-enabled assistive technology in non-visual physical activities and charts directions for future work (Chapter 9).

Technical and artifact contributions.

- OSCAR, a vision-language pipeline that extracts object status from recipes, aligns visual frames to recipe steps with a time-causal model, and substantially improves recipe step prediction over text-only baselines (Chapter 6).
- A runnable end-to-end voice-first cooking assistant built on the ODG, with a paired Wizard-of-Oz researcher dashboard and an input-agnostic reasoning engine that grounds every spoken reply in a structured ODG query rather than in recipe prose (Chapter 8).

Dataset contribution.

- A new dataset of real-world non-visual cooking sessions captured by blind cooks in their own homes, annotated for evaluating vision and language models in this setting (Chapter 6).

Chapter 2

Background and Related Work

This chapter situates the thesis within four bodies of prior work: (1) cooking as a research domain in HCI; (2) information spaces and embodied interaction as theoretical frames for AI-enabled assistive technology in physical activities; (3) the lived experience of people with vision impairments around food and cooking; and (4) the assistive-technology landscape currently available to non-visual cooks, including the recent generation of large language and vision-language models. Each section closes with a short pointer to what the present thesis adds, and Section 2.5 consolidates the gaps that the empirical chapters and the ODG (Chapters 3–8) take up. The narrower body of related work that touches directly on structured recipe representations, voice-first cooking assistants, conversational and mixed-initiative agents, and accessibility-focused recipe systems is positioned more precisely against the Object Dependency Graph itself in Section 7.2 of Chapter 7; here I keep the framing broad.

2.1 Cooking and Technology as a Research Domain

Cooking has long served HCI as a setting in which the physical, social, and cognitive dimensions of everyday life intersect, which is also what makes it a demanding test bed for assistive technology. This section reviews how the field has framed cooking as a research domain, the interactive systems that have been built around it, the way recipe instructions have been designed for those systems, and the assumptions about users that those systems have tended to encode.

2.1.1 The Cultural and Everyday Importance of Cooking

Cooking is more than a means of sustenance; it is tied to social, cultural, and personal identity, and shapes how individuals express themselves and connect with others [12]. Pollan describes cooking as a bridge between the natural world and human society: transforming raw ingredients into shared meals supports social bonds and personal satisfaction [149], and across many cultures cooking is also a vehicle for transmitting tradition and family practice [6, 25, 127]. These observations are what motivate the HCI community’s sustained interest in cooking as a domain in which interactive systems can support food-related activities ranging from meal planning and grocery shopping to cooking and social dining [41, 67, 68].

Recent decades have also seen a documented decline in everyday cooking skills, particularly among younger generations, attributed in part to the convenience of processed foods and the erosion of formal cooking education [11, 33, 53, 171]. This decline carries both health and cultural consequences and creates an opening for interactive systems that scaffold cooking skill: cooking apps, smart kitchen tools, and conversational assistants can offer step-by-step guidance, personalized meal plans, and real-time feedback, lowering the threshold to home cooking for users who would otherwise default to pre-prepared meals [165, 167]. For people with vision impairments, who already face elevated obstacles to home cooking [20, 32], this skill-scaffolding role is even more consequential.

2.1.2 Interactive Cooking Systems

A long line of HCI work has built interactive systems that operate inside the kitchen itself rather than around it. Early projection-based systems such as CounterActive [87] and PersonalChef [134] put recipe instructions directly onto the work surface or rendered a life-sized virtual chef into the room, removing the need to handle a phone or laptop while cooking. Kitchen of the Future [164] distributed instructions across multiple screens placed where the cook would need them, while CooksCollage [183] used a visual timeline of completed steps to help cooks resume after interruption. Voice- and gesture-based systems including eyeCOOK [26] explored hands-free recipe interaction with eye tracking and speech for moments when the cook's hands were occupied or dirty, and instrumented hardware such as Panavi [184] embedded sensors in the pan itself to give real-time feedback on cooking technique.

AI-driven kitchen technologies extend this line of work by monitoring kitchen activities and objects [107, 141], supporting multimodal control of appliances [21, 96, 188], embedding sensing in smart tools and utensils [99], and providing dynamic recipe guidance [35, 38, 55, 104]. Lei et al. [107], for example, used RGB-D cameras to recognize fine-grained cooking activities and objects, and König and Thongpull's Lab-on-Spoon [99] packaged temperature, color, and pH sensing into a multi-sensor utensil for food-quality inference. Across these systems, the consistent design move is to move some piece of the cook's cognitive load into a perception or guidance layer that the kitchen itself provides; what differs from system to system is which piece, and through which sensory modality.

2.1.3 Recipe Design and Instruction Structure

How a recipe is written and structured has a measurable effect on how effectively users can follow it. Traditional recipe formats, which separate the ingredient list from the method, force the cook to mentally integrate the two streams while cooking, which raises cognitive load and increases the chance of errors [181]. This burden is especially heavy on novice cooks, who are still building the working memory and procedural fluency to manage multiple tasks at once [59]. Interactive systems such as PersonalChef [134] and Panavi [184] have attempted to reduce this burden by interleaving ingredient quantities directly into the step-by-step method, so that the cook reads the relevant amount alongside the action that consumes it. Empirical work on instruction design more broadly recommends presenting steps in a clear chronological order, chunked into logical units, and supported by visual aids where appropriate [59, 168]. These principles also surface again in Chapter 4, which extends them to non-visual recipe access.

2.1.4 Assumptions and Gaps in Cooking Technology Design

Many interactive cooking systems are built on assumptions about what cooks need that are not always grounded in empirical observation. Researchers have routinely assumed that cooks want help with multitasking, with efficiency, or with confidence-building, but these assumptions do not always match how home cooks describe their own practice [68]. Systems like HappyCooking [76], which optimizes for cooking multiple dishes in parallel, presuppose that throughput is a primary user goal, and systems like CooksCollage [183] that focus on interruption recovery presuppose that interruptions are the dominant pain point. Bell and Kaye’s [13] ethnography of domestic cooking suggests that, for many cooks, the activity is closer to leisure or creative practice than to a task to be optimized.

A related gap is who these systems are designed for. The bulk of interactive cooking work has been built with younger, sighted, technologically fluent users in mind, with comparatively little attention to older adults, people with disabilities, or novice cooks. VERA [176], which uses visual cues to help individuals with aphasia follow recipes, is one of the few systems in this lineage that takes accessibility as its central design constraint. The empirical chapters of this thesis (Chapters 3–5) are partly motivated by this gap: they ground the design of an assistive cooking system in what blind cooks actually do, what they actually need to know, and where mainstream interactive cooking systems leave them unsupported.

2.2 Information Spaces for Physical Activity

The thesis is built around the claim that AI-enabled assistive technology for physical activity needs an explicit, structured representation of the physical environment, the user’s needs, and the action under way, what I call a *physical information space* (Chapter 1). This section reviews the theoretical foundations of that claim, drawing on three converging strands of HCI research: information spaces as a design construct, distributed cognition, and embodied interaction.

2.2.1 Information Spaces as a Design Construct

The notion of an *information space*—an environment of information that a person navigates, rather than a static document they read—has a long lineage in HCI. Early work framed interaction itself as navigation through an information space [16, 51], and information-foraging theory characterized how people move through such spaces under the cost of acquiring information, following *information scent* toward what they need [148]. This tradition shifts attention away from static, application-centered documents and toward dynamic, user-centered environments that respond to context. Building on it, Fox et al. [57] propose a multiscale personal information space that adapts to a user’s cognitive and contextual needs across tasks, and argue that future computing systems will need to move beyond document-based interfaces to accommodate richer, task-driven interactions. I take Fox et al.’s construct as the most direct starting point because it is explicitly personal, multiscale, and context-adaptive—the three properties an assistive cooking system most needs—while the foraging and navigation lineage supplies the underlying account of why structuring an information space, rather than serving a flat document, changes what a user can find and act on.

For physical activities such as cooking, this framing matters because the information a user needs is not primarily textual: it concerns where objects are, what state they are in, what is safe to do next, and which actions can be parallelized. A *physical* information space, in the sense used in this thesis, extends this construct from digital documents and screens to the physical environment of an ongoing activity: the objects, their changing states, and the cook’s position within the task become the space that the assistant helps navigate. It is the version of the information-space idea appropriate for AI-enabled assistive technology operating in a physical environment.

2.2.2 Distributed Cognition and Embodied Interaction

The information-space view dovetails with two long-standing HCI frameworks for thinking about physical-world tasks. Hollan, Hutchins, and Kirsh’s theory of *distributed cognition* [80] holds that cognitive processes are not confined to the head but are distributed across people, artifacts, and environments. For a blind cook, distributed cognition is not a metaphor: tactile labels on spice jars, the felt heat radiating from a pan, the angle of a cutting board against the counter, and the running spoken summary from an assistant are all part of the cognitive system that completes the recipe. Dourish’s *embodied interaction* framework [50] reinforces this view from a different direction: cognition and action are tightly coupled, and information becomes usable through the physical context in which a user encounters it. Together these two frameworks imply that assistive systems for physical activity must structure information so that it can be perceived and acted on in the body and in the environment where the action takes place, not retrieved from a separate textual surface.

2.2.3 From Theory to a Physical Information Space for Cooking

The bridge from these theoretical commitments to a runnable system is the question of *what* to put in the information space and *how* to expose it. Mainstream representations of cooking, whether recipe prose, screen-reader navigation through that prose, or the embeddings of a general-purpose vision-language model, do not by themselves answer this question. Recipe prose collapses the underlying procedural structure, the safety constraints, and the parallelism into linear text [117]; vision-language model embeddings, while expressive, do not enumerate the named fields a downstream assistant needs in order to deliver the right piece of information at the right moment. The Object Dependency Graph introduced in Chapter 7 is one concrete proposal for what a physical information space for cooking should look like: a typed structure with named fields for object status, prerequisites, safety, non-visual cues, measurements, progress tracking, and parallel groups, each of which is what makes a corresponding assistant behavior possible at run time.

2.3 People with Vision Impairments and Cooking

This section turns from cooking-in-HCI generally to the specific population this thesis serves. It reviews how vision impairments affect food-related activities, the challenges and adaptations blind cooks develop in the kitchen, and the gap that current technology leaves around real-time contextual support.

2.3.1 Vision Impact on Cooking and Food-Related Activities

Vision impairments profoundly affect the ability to engage in food-related activities such as shopping, meal preparation, and cooking. The loss of visual cues complicates these tasks and often requires alternative strategies or sighted assistance, with the severity of vision impairment tightly correlated with the difficulty of daily food-related activities [20]. In a survey of over 100 visually impaired individuals, Jones et al. [86] found that severe impairments substantially increase the difficulty of meal preparation and documented an association with malnutrition, with downstream effects on overall quality of life.

The downstream effects are also social and economic. Bilyk et al. [20] report that participants in their study consumed at least 40% of their meals in restaurants because of the complexity of cooking at home, which both adds financial cost and reduces control over meal choices. Kostyra et al. [100], in a survey of 250 respondents, found that only 49.6% prepared their own meals, with the remainder relying on sighted or fellow blind assistance, and that 57.6% leaned on ready-to-eat products to avoid the risks of cooking from scratch.

2.3.2 Challenges and Adaptations in the Kitchen

When blind cooks do prepare their own meals, certain tasks are reported as disproportionately difficult: peeling vegetables (82.1% of respondents in [100]), frying foods (72%), and more broadly, any task involving chopping, stirring, heat control, or sharp tools [20, 86]. These difficulties extend beyond food handling: organizing kitchen space and tracking ingredients are themselves cognitively expensive, and frequent mismatches between expected and actual object position can lead to frustration and prolonged cooking times [190].

To compensate, blind cooks adapt their practice in ways that rely on senses other than vision and on systematic spatial organization. They use tactile and auditory cues to gauge doneness and to monitor cooking progress [190], they reorganize the kitchen so that frequently used objects can be located by touch [20], and they develop personal routines that minimize the need for visual inspection. These adaptations work, but they impose a real cognitive cost and require time and rehearsal to internalize. Li et al. [110] characterize this in detail through interviews and video analysis, surfacing the specific pain points around measuring ingredients, inspecting food quality, and tracking objects during meal preparation, and arguing that assistive tools should be designed to reduce, rather than mirror, the cognitive load of these adaptations.

2.3.3 The Need for Real-Time Contextual Support

Even with adaptation, a clear class of needs persists around contextual information that is not available in standard recipes. Blind cooks struggle to navigate recipes efficiently because they must continuously track steps, manage workspace, and integrate ambient information that the recipe text leaves implicit [20, 117, 190]. The need to re-wash hands before each recipe consultation, documented in both [110] and [100], fragments the cooking flow and makes step-by-step interaction with a recipe disproportionately costly. Accessible recipe formats are also scarce: most digital and print recipes offer little non-visual structure, leaving the cook to reconstruct it on the fly [110, 190]. Real-time contextual systems, whether computer-vision-based or voice-guided, are well positioned to fill this gap by offloading some of the contextual reconstruction

onto a perception or guidance layer that the cook can query at the moment of need rather than between recipe consultations [117, 190].

2.4 Assistive Technologies for People with Vision Impairments in Cooking

Assistive technologies for non-visual cooking can be grouped by the modality through which they deliver information to the cook: tactile, vision-based, auditory and voice-controlled, multimodal, and (more recently) conversational systems built on large language and vision-language models. The grouping is convenient but not absolute; the most useful systems combine modalities, and the most recent generation of systems is starting to blur the boundary between perception, reasoning, and conversational delivery.

2.4.1 Tactile-Based Technologies

Tactile markers and overlays offer a low-friction way to make existing kitchen objects identifiable by touch. Guo et al.’s Facade [72] generates 3D-printed tactile overlays for the controls of household appliances such as microwaves and stoves, letting blind users locate buttons by touch where they would otherwise need sighted assistance. Suzuki et al.’s FluxMarker [173] uses magnetic and tactile markers to convey position and orientation on top of existing surfaces. These approaches are valuable because they are passive, durable, and require no power, but they only cover identification of static controls; the dynamic, time-varying state of food and cookware (the focus of much of this thesis) is outside their reach.

2.4.2 Vision-Based Technologies

Computer-vision systems capture visual information from a camera and translate it into a non-visual representation for the user. Guo et al.’s VizLens [71] combines computer vision with crowdsourcing to make appliance interfaces accessible by identifying buttons in real time. Kim et al. [95] use a YOLO-based pipeline to monitor induction cooktops, detecting cookware, cooking zones, and hand positions, and providing audio, voice, and light feedback. Fusco et al. [58] demonstrate vision-based assistance for ingredient recognition during food preparation, and Tekin et al. [177] target the LED and LCD displays of appliances so blind cooks can read settings such as cooking time and temperature. More recent AR-based systems such as CookAR [106] layer object identification, utensil recognition, and ingredient measurement into a heads-up overlay. These systems perceive the kitchen well, but they typically stop short of the next step in this thesis: tying perception to a structured recipe representation that knows what each perception means in the context of the recipe the cook is following.

2.4.3 Auditory and Voice-Controlled Technologies

Auditory feedback is the dominant delivery channel for non-visual cooking today. The simplest forms are ambient or alarm-style cues (e.g., beeping ovens, audible timers) [92, 157]; speech-synthesis tools [22, 70, 163] and modern text-to-speech systems such as those used in navigation

aids [160] deliver richer linguistic content. Voice-controlled assistants such as Amazon Alexa and Google Home have become standard fixtures in many blind cooks’ kitchens, used for setting timers, asking discrete questions, and (less successfully) for following recipes step by step [1, 29, 117]. The combination of smartphone-based screen readers, text-to-speech apps, and smart speakers covers a wide surface, and is the workflow against which the user study in Chapter 8 is framed as a familiar reference condition.

Existing voice-controlled cooking workflows have well-documented limitations. They tend to read the recipe linearly and so do not answer structural questions (“what can I do while the oven preheats?”, “which step did I skip?”); their verbosity and pacing are often poorly matched to blind users’ preferences [29, 40, 172]; and they bundle proactive behavior (timers, warnings, sensory cues) into a single fixed persona that cannot be configured. The ODG-based assistant of Chapters 7–8 addresses these specific gaps; the more granular comparison against voice-first cooking work is in Section 7.2.

2.4.4 Multimodal Technologies

Multimodal systems combine tactile, auditory, and visual feedback to provide more comprehensive support. CueSee [200] integrates auditory guidance with tactile cues to help users identify objects and understand their spatial relationships. Kuriakose et al. [103] pair auditory and tactile feedback for object recognition and spatial orientation, while Grussenmeyer et al. [69] explore haptic-plus-auditory feedback for device navigation. The promise of multimodal systems is that different cooking moments call for different delivery channels (a discreet tactile pulse during a conversation, an audible alert when the stove gets too hot, spoken guidance when the cook explicitly asks for it), and a single delivery surface rarely fits every moment. The ODG’s per-state `nonVisualCues` field (Chapter 7) is designed in this spirit: cues are typed by sensory channel so that a downstream assistant can choose the appropriate delivery surface rather than collapsing every cue into spoken text.

2.4.5 Large Language Models, Vision-Language Models, and Conversational AI

The last few years have seen a step change in the capability of large language models (LLMs) and vision-language models (VLMs), with direct implications for assistive technology for blind users. General-purpose LLMs such as GPT-4 [3] and the broader generation of conversational systems built on them [90] can interpret free-form questions, summarize long documents, and produce fluent natural-language explanations, removing several of the rigidity constraints that limited earlier voice assistants. VLMs such as CLIP [152] and SigLIP [199] extend this capability into the visual domain, supporting open-vocabulary recognition of objects and scenes without per-task retraining; the OSCAR pipeline in Chapter 6 is one example of how such models can be combined with object-status reasoning for a non-visual cooking task.

For accessibility specifically, this capability turn has begun to reshape research design. Recent systems use LLMs to mediate conversation between blind users and complex environments [10, 125], to coach blind makeup application through vision-grounded language [115], and to translate cooking video into accessible step-by-step coaching [83]. Independent work has begun to extend the object-status representation from OSCAR into mixed-initiative cooking assistance

through the AROMA system [140]. Benchmarks of LLM and VLM behavior on accessibility-relevant tasks [138] are also starting to characterize where these models succeed and where they hallucinate or omit safety-critical information.

At the same time, LLMs and VLMs are not by themselves sufficient for non-visual cooking. They hallucinate object states, miss safety-critical transitions, and lose track of which step is current when the conversation extends over many turns [138, 140]. They are also non-deterministic in ways that complicate audit and rehabilitation training. The position taken in this thesis is that LLMs and VLMs are most useful when paired with an explicit, structured representation of the recipe and the kitchen state, with the model providing fluency and the structure providing grounding. The ODG is this thesis’s proposal for that structured grounding layer; Chapter 8 demonstrates the pairing by routing every spoken assertion of the ODG-aware assistant through a structured engine call rather than through free-form generation.

2.5 Summary and Positioning of the Thesis

Read together, the four bodies of work surveyed in this chapter share a common shape. Cooking-in-HCI has produced a rich library of interactive systems but has tended to assume sighted, throughput-oriented users (Section 2.1). The theoretical frames of information spaces, distributed cognition, and embodied interaction give principled reasons to design assistive systems around the physical environment rather than around a textual document, but they do not by themselves prescribe what that environment-anchored representation should contain for a specific physical activity (Section 2.2). The empirical work on people with vision impairments and cooking has documented in detail what blind cooks find hard and what they adapt around, but the assistive technologies most easily available to them (linear screen-reader playback, smart speakers, tactile labels, ambient cues) cover only narrow slices of that need (Sections 2.3 and 2.4). The recent capability turn from large language and vision-language models opens a path to far more flexible non-visual cooking assistance, but, on its own, it lacks the structured grounding that would let a blind cook trust the assistant’s claims about the recipe and the kitchen (Section 2.4.5).

The empirical chapters of this thesis (Chapters 3–6) take up these gaps directly: a content analysis of non-visual cooking practices and challenges, a study of how blind cooks access and structure recipes, an in-home contextual inquiry into the contextual object information they rely on, and the OSCAR perception pipeline that turns object-status transitions into a tracked recipe progress signal. The ODG (Chapter 7) is the synthesizing construct: a structured physical information space for cooking, decomposed into seven design rules that a voice-first assistant can act on at run time. The implementation and user study in Chapter 8 then probe whether such an ODG-aware assistant, evaluated against the screen-reader workflow that today’s blind cooks actually use as a familiar reference condition, materially changes the dimensions of non-visual cooking that mainstream tools leave underserved.

Chapter 3

Exploration of Non-Visual Cooking Practices and Challenges

This chapter was adapted from my published paper:

Franklin Mingzhe Li, Jamie Dorst, Peter Cederberg, and Patrick Carrington. 2021. Non-Visual Cooking: Exploring Practices and Challenges of Meal Preparation by People with Visual Impairments. *In Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, 1–11.

3.1 Overview

Vision is an important sensory modality for humans. Many activities of daily living (ADLs), such as cooking and eating, can be difficult without visual support. Jones et al. [86] revealed that people with visual impairments tend to have poor nutritional status, which is often linked to problems with buying, preparing, and eating healthy food. People with visual impairments may have an aversion to cooking due to difficulty accessing visual information and cues during the cooking process [20, 100]. This has resulted in people with visual impairments more frequently eating out at restaurants or preparing frozen food that may be calorie-rich. According to the aforementioned Canadian study [20], eight out of nine participants stated they “disliked or hated cooking” because of the time it takes to cook without vision.

Christine Ha, the first blind contestant of MasterChef, won the third season of the show in 2012 and described the importance of cooking to people with visual impairments [74, 193]. To assist visually impaired individuals to cook independently, there have also been training guidelines released by blind communities [187] and cooking-related assistive technologies that are commercially available [78]. For example, people with visual impairments could use voice commands to set timers or use a speaking kitchen thermometer to check the temperature of a steak. In order to explore the physical information space, it is important to explore the practices and challenges of people with visual impairments in cooking and how they leverage different assistive devices for cooking. In this chapter, we explore the following research questions:

- RQ1: What are current cooking approaches and techniques employed by people with visual impairments?

- RQ2: What are the key challenges, concerns, and risks encountered while preparing meals?
- RQ3: What are potential opportunities for assistive technologies to support people with visual impairments to cook independently?

In this chapter, we present the findings for RQ1–RQ3 through a content analysis of 122 YouTube cooking videos by cooks with vision impairments and semi-structured interviews with 12 visually impaired people. We categorized 12 different activities essential to non-visual cooking and unique practices that people with vision impairments adopt while cooking. We further illuminate challenges encountered before, during, and after cooking including: utilizing tools, information access, touching and feeling, safety and consequence, precision and ambiguity, organizing and tracking, item and quality inspection, and collaborative cooking and communication. We then discuss potential opportunities to support people with visual impairments while cooking (e.g., status tracking and safety monitoring). This chapter contributes to the knowledge base of the physical information space of non-visual cooking with novel findings of cooking practices and challenges.

3.2 Technology Scaffolds for Cooking

The broad literature on how vision impairments affect food-related activities, on the assistive technologies (tactile markers, computer vision, and voice interaction) that make kitchen interfaces accessible, and on interactive cooking systems in HCI is consolidated in Chapter 2. This section focuses on the work most directly relevant to the present study—technology that scaffolds the *process* of cooking and skill acquisition—and the specific gap it leaves for people with visual impairments.

In terms of cooking processes, prior research has explored learning procedures of cooking [91] and different cooking techniques [75, 104]. For example, Kato and Hasegawa [91] introduced an interactive sauteed cooking simulator that could visualize different cooking states (e.g., temperature changes, browning from burns), helping users better manage skills such as how to cook medium-rare meat. Kusu et al. [104] further proposed a method to calculate a cooking recipe’s difficulty level during searching and recommend recipes that match the user’s cooking skills. Although this body of work shows how technology can scaffold cooking activities and skill-building, it is built almost entirely for sighted users: there is little research and understanding of what cooking-related practices and learning procedures people with visual impairments have adopted, and what challenges they face during these processes. In our work, we showed cooking practices of 12 different cooking procedures through a YouTube video analysis and uncovered eight themes of cooking challenges through interviews with people with visual impairments.

3.3 Methodology

3.3.1 YouTube Video Analysis: Cooking Practices for Visually Impaired People

To understand existing cooking practices and potential risks for people with visual impairments, we conducted a YouTube video analysis (searching, filtering, and analyzing YouTube videos related to cooking practices by people with visual impairments), inspired by prior research on lever-

aging the richness of YouTube video contents to understand accessibility needs [9]. Our video analysis consisted of two main steps: 1) searching for YouTube videos related to cooking practices for people with visual impairments; 2) analysis and coding procedures.

Table 3.1: Searching Keywords

Searching Keywords
Blind Cooking, Blind Person Cooking, Blind Chef, Legally Blind Cooking, Blind Cooking Food, Blind Cooking Dinner, Blind in the Kitchen, Visually Impaired Cooking, Visually Impaired Person Cooking, Visually Impaired Chef, Visual Impairment Cooking

3.3.1.1 Search Protocol

In the video searching process, we looked for videos focused on cooking practices for people with visual impairments. To search for relevant videos, three researchers independently combined visual-impairment-related keywords (e.g., blind, visually impaired, visual impairment) and cooking-related keywords (e.g., cook, cooking, chef, kitchen), forming the final set of search terms listed in Table 3.1. To come up with these, our researchers first started with basic searches (e.g., blind cooking) and gradually included other keyword combinations from candidate video titles or descriptions. Because each search may generate hundreds of results, we then followed the same approach as Komkaite et al. [98] by stopping our search for videos after the whole page of results started to be irrelevant.

In total, we initially created a video dataset of 136 relevant videos found by March 28, 2021. We then filtered out videos if: 1) the person who cooked in the video did not have visual impairments; 2) the person only heated frozen food; 3) it had poor audio and video quality or did not show the person cooking; 4) videos were duplicated. We then ended up filtering 14 videos and created the final video dataset of 122 videos (V1–V122). Among the 122 videos in our dataset, most videos were uploaded in 2020 (35), while others were uploaded in 2021 (34), 2019 (11), 2018 (10), 2013 (9), 2014 (7), 2017 (6), 2016 (5), 2015 (3), and 2012 (2). The average length of videos was 698 seconds (ranging from 77 seconds to 2694 seconds).

3.3.1.2 Video Content Analysis

To code the videos, three researchers first open-coded [37] the videos independently. Then, the coders met and discussed their codes. When there was a conflict, they explained their rationale for their code to each other and discussed to find a resolution. Eventually, they reached a consensus and consolidated the list of codes. Afterward, we performed affinity diagramming [77] to group the codes and identify emerging themes according to the Cooking Guidelines for people with visual impairments [187]. Overall, we established 12 themes and 37 codes. The following section describes the findings based on the 12 themes.

3.3.2 Interview Study: Challenges of Cooking for People with Visual Impairments

Table 3.2: Participants' demographic information

Participant	Age	Gender	Vision Impairment Description	Learned Cooking <i>Before or After</i> Vision Loss
P1	25	Male	Totally Blind, Congenital	After
P2	33	Male	Legally Blind, Congenital	After
P3	32	Non-binary	Legally Blind, Congenital	After
P4	19	Male	Totally Blind, Congenital	After
P5	36	Male	Totally Blind, Acquired (7 years)	Before
P6	24	Female	Legally Blind, Congenital	After
P7	22	Male	Totally Blind, Congenital	After
P8	35	Male	Legally Blind, Acquired (1 year)	Before
P9	28	Female	Totally Blind, Congenital	After
P10	31	Female	Legally Blind, Congenital	After
P11	55	Male	Legally Blind, Congenital	After
P12	48	Female	Legally Blind, Congenital	After

3.3.2.1 Study Procedure

To understand perceptions and challenges of cooking for people with visual impairments, we conducted semi-structured interviews with 12 visually impaired people who have experience cooking (Table 3.2). Our participants have an average age of 32, with a range from 19 to 55 years old. Five of them are totally blind, and the rest are legally blind. Furthermore, ten of our participants are congenitally blind and learned cooking after the vision loss, and the other two cooked before losing their vision. Participants were recruited through social platforms (e.g., Reddit, Twitter, Facebook). To participate in our interview, participants needed to 1) be 18 or above; 2) have visual impairments; 3) have experience cooking; 4) be able to communicate in English. The interviews were conducted through Zoom and took around 60–75 minutes for each participant. Participants who completed the interview were compensated with a \$20 Amazon gift card. The entire recruitment and study procedure was approved by the institutional review board (IRB).

In the interview, we first asked participants about their demographic information, general reasoning and perceptions about cooking, and any barriers they encountered when they started cooking. Based on our findings of different cooking procedures and practices in the YouTube video analysis, we asked our participants to discuss their perceptions and challenges around different cooking procedures, such as measuring, pouring, and cutting and chopping. Afterward, we asked participants to discuss their experiences and challenges with assistive cooking technologies, cooking with other people, and the modification of kitchen appliances for accessibility purposes.

3.3.2.2 Data Analysis

Two researchers independently performed open-coding [37] on the interview transcripts. Then, the coders met and discussed their codes. When there was a conflict, they explained their rationale for their code to each other and discussed to resolve the conflict. Eventually, they reached a consensus and consolidated the list of codes. Afterward, they performed affinity diagramming [77] to group the codes and identify emerging themes. Overall, we established eight themes and 26 codes. The results introduced in the findings are organized based on eight themes.

3.4 Findings

3.4.1 Cooking Practices of People with Visual Impairments

From the YouTube video analysis, we uncovered unique cooking practices and procedures of different people with visual impairments (Table 3.3).

Table 3.3: Cooking activities from YouTube video analysis for people with visual impairments.

Category	Activities
Preparation	General Safe Cooking Practices
	Cutting and Chopping
	Measuring
	Spreading
	Pouring
Cooking	Placing Pans on a Burner
	Baking
	Turning Food
	Testing Food for Doneness
Tools, Environment, and Recognition	Tools and Small Appliances
	Knowing the Kitchen Space
	Tell Things Apart

3.4.1.1 General Safe Cooking Practices

For general safe cooking tips and practices, we found that people with visual impairments usually **start everything with gentle actions**. This includes using a slow cooker rather than an open grill to avoid fire flare-ups (V18), always cooking on low to medium heat to avoid injury (V80), and putting all food and ingredients in a cold pan first before putting the pan on the burner (V4, V6). In V80, the person with visual impairments mentioned in his video:

“...For people with visual impairments like me, we might have to make different adjustments on the pan with our hands or utensils while cooking, I often start with using my hand to feel the temperature and cook at low to medium temperature on my pan...”

Beyond starting with gentle actions, we found that people with visual impairments tend to **use hearing and hand-feeling to compensate for vision needs during cooking**. In terms

of leveraging hands for feeling during cooking, we found that people with visual impairments hold their hands over the pan to gauge temperature (e.g., V11, V23) and use hands to guide food onto utensils (V15). For hearing, we found that people with visual impairments leveraged sound to tell if oil is hot or not (V90, V94). For example, one cook with visual impairments put a small droplet of water into a pan of oil to hear if it was hot (V90). Furthermore, we found that it is important for people with visual impairments to **memorize the environment in the kitchen**, including the kitchen layout (V1, V17, V46) and which knobs correspond to which burner on the stove (V9).



Figure 3.1: Different ways of cutting. (a) Cutting with a vegetable chopper. (b) Cutting practice of putting the whole hand on top of the knife to cut garlic. (c) Tucking the fingertips under and in toward the palm of the hand and cutting with the other hand. (d) Having fingers on both sides of the knife to position where to cut.

3.4.1.2 Cutting and Chopping

For cutting and chopping, we learned that many people with visual impairments tried **using alternative ways to avoid knives**, such as using a garlic masher, a vegetable chopper (Fig. 3.1(a)), a food processor (e.g., V103, V114), scissors (e.g., V84, V94), or even by buying pre-minced garlic (V9, V13). Moreover, we found that people with visual impairments **sort cut and uncut food into different specific locations on the cutting surface** (V1, V32). In V1, the person commented on her practices in the video: *“For me, I always place the cut food at the top right corner of my cutting board, so I can keep track of the food that I still need to cut.”* We also found that people with visual impairments have **different practices of using the knife to cut objects**. For example, we found people perform cutting by 1) having fingers on both sides of the knife to position where to cut (e.g., V23, V28) (Fig. 3.1(d)), 2) putting the whole hand on top of the knife

(e.g., V20, V27) (Fig. 3.1(b)), or 3) tucking the fingertips under and in toward the palm of the hand and cutting with the other hand (e.g., V16, V112) (Fig. 3.1(c)).

3.4.1.3 Measuring

To measure food, we first found that there is a common practice for people with visual impairments to **use hands to weigh meat and measure spices** (e.g., V12, V35). We further revealed that people with visual impairments highly **rely on using measuring cups with hand assistance to measure liquids** (Fig. 3.2). To accurately use measuring cups, people with visual impairments either have to memorize the measurements (V51) or use a Braille-labeled set (V10). Finally, visually impaired cooks also **prefer using existing containers or utensils to measure food**, such as using the same can to get a 1:1 ratio of water and soup (V89) and using the broad side of a knife to measure the size of meat (V96).



Figure 3.2: Using the hand to direct the oil and measuring cup to measure liquid (V17).

3.4.1.4 Pouring

For pouring, we found that people with visual impairments **use an extra-large bowl to reduce the difficulty of pouring** (V53, V56) and usually pour over the sink to make sure spills do not matter (V10, V26). To know how much people are pouring, we revealed that people tend to **use their finger over the spout and/or in containers to know the exact amount** (e.g., V16, V21). Another common practice we discovered is *“making sure you poured everything”* to ensure **all of the ingredients they want are added to their dish**. More specifically, we found that people first **add water to food processor or bowl to get any leftover ingredients** (V35) and then **wipe bowls with fingers to make sure they are empty after pouring** (e.g., V30, V59).

3.4.1.5 Placing Pans on a Burner

From the video analysis, we recognized that people with visual impairments **prefer using a traditional electric stovetop over a modern glass top or gas top**. The key reason was the ease of placing pans on the stovetop. As a practice, we found that people with visual impairments tend to **use either listening or feeling to know whether the pan is on the burner**. For



Figure 3.3: The person uses one hand to feel the food location and another to turn the food with a spatula

example, V19 showed that visually impaired individuals listen to sizzle to tell whether the pan is on the burner or not after turning on the fire (V19). As another option, we found people either use a hand to feel to make sure the pan is at the center of the stovetop (V96) or use the back of a spatula to feel if the pan is on the burner (V19).

3.4.1.6 Baking

In baking, we found that visually impaired people have to **use their hands to feel batter texture, check the batter readiness, and shape dough** (V41, V88). Specifically, we found that visually impaired cooks prefer to fold the dough over parchment paper (V92), use their hands to shape the dough (V121), and create barriers when rolling it out (V82). Furthermore, we revealed **the importance of keeping things clean during baking**. For example, V29 mentioned the importance of wiping edges of the pan before baking to prevent the burning of anything that accidentally got on the edges. Finally, we found people with visual impairments **use caution with the oven when baking**, such as by pulling the oven shelf out slightly to avoid burns (V40).

3.4.1.7 Turning Food

For turning food, we learned that people with visual impairments **have different preferences on choosing tools** to accomplish the task. This includes using a large two-prong fork (V55), using two utensils (V83), and using tongs and spatulas (V33, V78). While turning food, people with visual impairments have different practices of **keeping track of food locations**. For example, V22 showed that the person always memorizes how much food she put on the grill and roughly where it is. V16 further showed that people with visual impairments use one hand to feel for food locations and then use the other hand to turn the food with a spatula (Fig. 3.3).

3.4.1.8 Spreading

From the video analysis, we uncovered that **using hands to spread** butter or jam is a common practice for people with visual impairments to make sure everything is evenly distributed (e.g., V43, V63). V88 further showed the practices of using single-use plastic gloves for spreading. Beyond using hands, V45 showed the approach of **using utensils to spread** jam evenly on bread.

3.4.1.9 Tools and Small Appliances

In terms of tools and small appliances for cooking, we found that visually impaired individuals often **use a virtual assistant to set timers**. This includes using Siri (e.g., V9, V13), Alexa (V1), Google (V4), and analog timers (V44). In V44, the visually impaired person rotated the analog timer using her hand position to gauge the correct time. Furthermore, we uncovered that **speaking tools are widely used for different purposes**, such as using a barcode reader to know what is in a can (V5), Aira [43] for identifying cans, screen readers to read recipes off of a phone (V29), video magnifiers to read boxes (V17), talking scales (V41) and speaking thermometers (e.g., V42, V115). We further found that people with visual impairments tend to **use a lot of specific tools for specific tasks**. For instance, V36 showed that the cook used an automatic can opener and listened for when it was done. Other examples include using a tea strainer to rinse off rice (V84) and spatula and chopsticks for mixing (V95).

3.4.1.10 Testing Food for Doneness

To test whether the food is done, we found that people with visual impairments mostly use five different methods. First, they **feel to check food doneness**, such as feeling food texture with a spoon (e.g., V9, V86), feeling for fried food to float to the top with a utensil (V90), and feeling cooking food with hands (e.g., V16, V39). We also learned that visually impaired people **used measures for checking doneness**. For example, they initiated a timer (e.g., V19, V54) or used a thermometer to check whether the food is done (V19, V121). Moreover, we found people with visual impairments also **used listening to tell when food sounds done** (e.g., V11, V34) and **use smell to check for food doneness** (e.g., V67, V100). For example, V84 mentioned that visually impaired people could use sound to know whether water is boiling. Finally, we found that people who learned cooking before losing their vision **judge food doneness on their intuition and prior memories from when they were sighted** (V91).

3.4.1.11 Knowing the Kitchen Space

From the YouTube video analysis, we discovered that people with visual impairments put a lot of effort into **organizing the kitchen space to reduce cooking barriers**, such as having separate space for cooking and for ingredients (V10), keeping the container of utensils handle side down so they can feel the ends (V25), keeping measuring cups nested (V28) and always placing utensils back to their original places (V3). Furthermore, we realized visually impaired individuals have different practices to **feel the kitchen space**. For example, V22 showed that people use a spatula to feel around a hot grill to get to know the space.

3.4.1.12 Tell Things Apart

To tell ingredients or tools apart, we found a common practice for people with visual impairments is **adding labels on buttons or objects**. For instance, visually impaired individuals add Braille labels on spice jars (V27, V47) or bump dot stickers for recognizing buttons (V50). Another approach to tell things apart is by **recognizing the size or shape of containers**. For example, the person from V72 put each ingredient in a different type of bowl to keep track of different ingredients. Furthermore, people also **rely on smell to tell things apart**, such as smelling to tell the differences between spices (V35, V65). Finally, we found that some people with visual impairments **prefer putting their containers in certain orders**. For example, some people prefer to alphabetically organize spices in their drawers (V50).

3.4.2 Challenges of Non-Visual Cooking

In this section, we introduce the challenges of cooking derived from our interviews in eight themes: utilizing tools, information access, touching and feeling, safety and consequence, precision and ambiguity, organizing and tracking, item and quality inspection, and collaborative cooking and communication.

3.4.2.1 Utilizing Tools

From the interviews, we found that people with visual impairments encountered various challenges when utilizing tools (e.g., cooking utensils, appliances). In the video analysis, we found that people with visual impairments rely on either Braille markers or bump dots to interact with an interface (Section 3.4.1.12). In the interview, seven participants commented on the concerns of **messing up tools with dirty hands and food during cooking**. For example, P1 commented on the problems of messing up Braille tactile markers on kitchen appliances:

“...It is very common that I might accidentally have something on my hand, especially when I make a dough, then when I touch my Braille markers or bump dots, it just stays on the marker and affects me when I want to read the markers...”

From the video analysis, we learned that people with visual impairments usually utilize many tools for different purposes (Section 3.4.1.9). In the interview, we found that eight of our participants **struggled with having more tools for different purposes than necessary**. For example, P12 mentioned that she had to use a vegetable holder on one hand to hold the vegetable and use another hand to peel it. P4 further commented on his situation of even using syringes for cooking:

“...I have too many tools or cookware for different purposes. I even used syringes to measure and add soy sauce to my dishes. But having too many things really freak me out, it made me look for things all the time...”

During the interview, five participants mentioned that **existing tools lack confirmation and feedback**, especially for our participants who learned cooking before vision loss and were used to leveraging vision for confirmation and feedback. For example, P8 mentioned the challenges of just using his spatula to tell where food is and whether the food is done:

“...Feel through tools is hard, I cannot just use my spatula to do everything for me, it is really hard for me to use it to tell if the food is ready by using tools to feel the texture. That is why I said feel is difficult without hand, that is why we need more help with adding more sensors at the head of tools to tell me different status, and it can also accomplish basic needs...”

3.4.2.2 Information Access

In the interview, participants overall complained about information access problems with kitchen appliances’ guides and recipes. First, three participants mentioned the **lack of details for guide manuals and instructions of kitchen appliances**. For example, P11 commented on the need for more vocal descriptions about the interface layout:

“...I cannot understand existing manuals and instructions for kitchen appliances like pressure cookers, it definitely needs more vocal descriptions about the position of different parts and the interface layout. Especially that the current pressure cookers or rice cookers have so many buttons...”

Beyond the manuals, we found our participants complained that **recipe content does not correspond to how people with visual impairments cook**. This includes cookbooks that have too many pictures (P9), wording or cooking language varying among different recipes which confused visually impaired cooks (P7), online recipes containing many figures and ads (P1, P3), and the way current recipes explain different steps requiring lots of vision support (P8). P8 continued with his experiences with recipes like a guessing game:

“...The descriptions in many recipes are very hard to correspond to the ways that I cook now. I was reading a recipe, and it told me to cook the piece of chicken until it turns brown and then flip it over. It does not transfer to people with visual impairments, it is like a guessing game...”

Furthermore, we realized that **existing recipe structures are not friendly to people with visual impairments**. Participants commented that different recipes they found online tend to have different recipe structures, such as the cook’s notes, nutrition facts, ingredients and cooking steps. As a consequence, our participants complained that they always have to go back to check the recipe (P1) and do not know where to start the recipe (P1, P3). P6 then further talked about her frustration of starting over from the beginning of the recipe because of the non-standardized recipe structure:

“...I use voiceover for apple devices, it typically read the whole recipe. However, many recipes do not follow a specific structure which made me easily miss information. Once I lose my place, I have to use voiceover to read from the beginning again...”

Besides the need for better recipe structures, we also found that our participants **had a difficult time interacting with the recipe while cooking**. Challenges include problems interacting with phones or tablets while cooking (P1), irrelevant ads and pictures (P10), and worries of getting both physical recipes and electronic devices dirty while cooking (P10). P10 continued:

“...I just cannot find a better way to interact with both my braille recipes and electronic recipes by getting them dirty and have a better interactive interface while cooking. Because my hands often have oil or sauce when I am cooking my dishes...”

Therefore, it is necessary for future research to consider how to better support cooking-related information access and better interactive methods while cooking.

3.4.2.3 Touching and Feeling

In the YouTube video analysis, we mentioned that people with visual impairments leverage their hands’ feeling for measuring (Section 3.4.1.3), placing pans on a burner (Section 3.4.1.5), baking (Section 3.4.1.6), and testing food for “doneness” (Section 3.4.1.10). From the interview, we found that there exist various challenges and concerns about this from people with visual impairments in cooking. First, we found that **hand feeling is not sufficient to tell all the information about certain cooking procedures**. For example, P12 commented: *“it can be difficult to feel the liquid level if the liquid is room temperature”*. Therefore, P5 complained that he had to put oil in the fridge first and then use a finger to feel the level. P1 further mentioned his difficulties in using his hand to feel the pouring speed of liquid:

“...When I use my fingers to lead the liquid to pour into a container, I can tell something flows over my finger, but I cannot tell how fast that is, and it usually ended up with having too much or too little that got poured, because I do not have another hand to feel the weight of the container, I have to use one hand to pour and use one hand to lead the liquid...”

Another concern from our participants is that **using hand feeling during cooking might be inappropriate or unsafe**. As an explanation, P5 mentioned his experiences of messing up the decoration of a cake by using his hands. Moreover, P9 and P12 mentioned that using hands when cooking might cause them to burn themselves by accidentally touching the hot pan. P12 continued:

“...Using hands to feel the food on a hot pan is not always easy to do, you have to be very very slow and gentle, especially when you want to feel the doneness of the food and try to flip it. I have burned myself many times...”

Furthermore, we found that visually impaired individuals complained that **using hands to feel objects affects the efficiency of cooking**, that is, the person cannot use their hand, which is used for feeling and touching, to do other things. As an example we mentioned previously, the person has to use one hand to pour and one hand to lead the liquid to the container, which forces them to place the container at a stable position (P10). Therefore, future research should consider how to enable hand and finger sensing to help with knowing more information during different cooking procedures and explore alternative methods to substitute the need for hand touch and feel in cooking.

3.4.2.4 Safety and Consequence

From the interview, we found that all participants showed strong concerns of **getting burned during different cooking procedures**, which need heat detection assistance while cooking. For example, taking things out of the oven puts the visually impaired person in danger (P8), and the plate and dish for baking can be very hot even after taking it out of the oven (P7). P8 explained the danger of exposed heat sources:

“...A lot of kitchen appliances have exposed heat sources, it can be very dangerous, such as my toaster or the steamer. After I burned myself once, I always pay 200% attention when I am cooking...”

Beyond burning, we found participants reported the **difficulty of handling accidents while cooking**. This includes accidentally knocking things over (P3), it being hard to know if there is an oven fire or not (P10), and accidentally dropping things and making a mess (P5). P5 continued:

“...I had experiences making scrambled eggs and spilled the egg all over the stove-top. However, I did not notice the mess after cooking, and my partner told me this situation the next day. I need something to alert me or help me to be aware of these things happened...”

Finally, we found that seven participants mentioned that **other people walking by could potentially cause safety concerns**, especially for participants who learned cooking before their vision loss. For example, P8 mentioned: *“you may turn around, and someone might stay behind you, and you hold a knife...It is a factor that is out of your control with another person while cooking.”* Thus, future research should explore how to help people with visual impairments detect heat to prevent burns, track accidents, and notice people walking by.

3.4.2.5 Precision and Ambiguity

In the YouTube video analysis, we introduced the need for measuring (Section 3.4.1.3) while cooking. From the interview, we found that **making precise steps while cooking can be a huge barrier for people with visual impairments**, such as getting the right amount of baking soda while baking, because using a hand to feel the level in a measuring cup of powders is challenging (P1). P7 further mentioned the problem with always getting more during spreading:

“...Because it is hard to know how much butter I already spread on my bread, I always ended up with taking more than needed...”

P11 further continued with commenting on the difficulty of decorating the food:

“...Decorating the cake or dishes is extremely hard for me. It took a lot of practice to get used to frosting or adding decorations. And it is very easy to get messed up...”

To reduce cooking risks, we showed different safe cooking tips in Section 3.4.1.1. However, we found that our participants complained about the **long learning curve of new methods which makes people stick with their familiar way of cooking**. P7 mentioned: *“every person has their own of doing things.”* P3 further commented on the hardness of adopting advice from other people:

“...It is really difficult to get advice from other people when I am already used to doing things in a certain way. I know my current way of cutting and chopping might not be the recommended way by blind cooking guidelines, but I am already used to it, I tried using the recommended method of cutting and chopping, and I cut myself while learning it...”

According to participants’ responses, another challenge is **following cooking steps or procedures precisely according to a recipe or instructions**. P10 mentioned that a simple mistake during cooking could end up ruining the dish, such as measuring baking soda and tracking the leavening time during baking. All participants mentioned the correlation between the tedious work of following all cooking steps and procedures and why they do not like to cook. Therefore, future research should explore different tracking and measuring methods to help people with visual impairments follow steps more easily and should create various teaching techniques to reduce the learning curve of different safe cooking methods.

3.4.2.6 Organizing and Tracking

In Section 3.4.1.11, we showed that people put a lot of effort into organizing their kitchen space to reduce barriers. From the interviews, we uncovered various challenges in organizing the kitchen and tracking objects. First, we found that participants complained about the challenge of **finding previously used objects or cookware due to the mental load during cooking**. P1 commented: *“It is really easy for me to forget where my used knives are.”* P10 also mentioned that it is hard to find previously used ingredients. P3 then commented on the need of having a system that continuously tracks the kitchen environment:

“...Cooking itself is a very complicated task that requires people to do multi-tasking and keep tracking of different things. For me, I sometimes forgot where my used knives are, and I had to be very conscious about my actions to prevent being cut by touching my used knife. Therefore, I think I need a tracker that can track the positions of different utensils and cookware for convenience and safety...”

Furthermore, we found that cooking is not a simple binary task that will guarantee successful and tasty dishes if you follow the recipe. We also realized people with visual impairments **have a hard time tracking the status of tasks in progress**. Five participants commented that it is hard to know when things are done. They have to constantly check the food. P1 further commented on this problem and tended to overcook the food:

“...Knowing whether the food is done is such a complicated project for me without any vision. To make sure I do not get diarrhea afterward, I often overcooked the food, such as shrimp. Although I can make sure it is totally cooked, but it affected the texture and the taste of the food...”

We also found that participants **have challenges tracking and organizing parts of the food**, such as organizing the cut and uncut parts of a large volume of food, and memorizing which parts of the food on the pan need to be flipped (P10). Therefore, P1 said that she had to wash her hands a lot and touch everything to check the food. P6 continued with her challenges:

“...The pan is big. If only a portion of the food on the pan needs to be turned over, it would be very difficult. I usually had to use hand and smell to check if the food at the center of the pan is burned or not...”

Finally, participants mentioned that they often have a hard time **knowing and remembering where to clean while cooking**. Specifically, six participants complained about remembering where they need to clean afterward, or they have to clean things right after use. P4 further commented that it is also very common for him to not know or not be aware of where he has to clean. Overall, future work should explore how to help people with visual impairments to better organize and track things through cooking.

3.4.2.7 Item and Quality Inspection

From the interview, we realized that people with visual impairments often have concerns about inspecting an item or food. First of all, we found that participants have a hard time **checking food quality**, such as whether the bread is moldy. P5 commented on the difficulty of checking food quality just with smell and touch, even with computer vision assistance:

“...Knowing the food quality is hard just by smelling and touch. Such as knowing whether a piece of bread is moldy or whether bananas or pineapples are ripe. I even tried with some apps that could object recognition, but I realized it just told me the object it is...”

Furthermore, we found that our participants complained about the task of **telling whether things got peeled or cleaned thoroughly**. P8 commented on cleaning and peeling non-smooth vegetables:

“...Different vegetables have different difficulties in cleaning and peeling. Vegetables like carrots are very easy to peel. But non-smooth vegetables are hard to tell if they got peeled or cleaned completely. Ginger is a hard one, also spots on potatoes are hard too...”

In the interviews, nine participants reported having a hard time **inspecting an accidental mix of unwanted things**, such as eggshells and bugs in flour (P1). P1 continued: *“I had many times of having eggshells in my scrambled eggs without noticing, it was disgusting.”* Overall, we showed the existing challenges of items and quality inspection. Future research should explore how to bridge the gap of the challenges we introduced to people with visual impairments in cooking.

3.4.2.8 Collaborative Cooking and Communication

From interviewing people with visual impairments, we found that ten out of twelve participants mentioned that they do not like cooking with another person because of **continuous communication needs**. *“Cooking with another person needs me to constantly speak out things.”* said P4. P3 further complained about collaborative cooking processes:

“...I have to verbally confirm steps and actions with another person all the time. Sometimes they might just do something in a way that I do not like, and they might not inform me because cooking and talking is not common for any person...”

In the interview, we found that eight participants leveraged apps for remote assistance while cooking, such as Aira [43] or Be My Eyes [192]. However, our participants complained about **making sure the item is in the camera’s field of view**. P1 continued:

“...Remote assistance definitely needs more improvements, I often do not know whether the thing I am talking about is in the camera’s field of view. More importantly, my hands are often dirty while cooking, and I always tried not to touch my electronic device. This leads to more concerns about using remote assistance...”

Moreover, we found that our participants complained about **appliance modification barriers in a communal space**. P8 commented on this problem: *“the biggest thing about using a shared kitchen, even the simple one like adding bump dots or braille markers, it is hard for me to just add these markers on different appliances, because they are shared.”* Therefore, future research should explore how to improve the social acceptability of adding tactile markers on appliances or other alternative ways for people to interact with different appliances.

3.5 Discussion and Future Work

From the results of our interviews with people with visual impairments, we uncovered existing challenges and gaps for visually impaired people in cooking. In this section, we further discuss zero-touch interactions for cooking, status tracking and safety monitoring, collaborative cooking, and contextual inquiry and learning process in depth.

3.5.1 Zero-touch Interactions for Cooking

From the results, we showed that our participants often rely on touching and feeling when cooking (e.g., measuring, checking for food doneness, spreading), and interacting with touch-based electronic devices might be inconvenient because they often have food on their hands. As an input method, prior research has explored speech [2, 29, 40, 172] and mid-air gestures [46] as input options for people with visual impairments. However, many of the existing interaction paradigms still require people to physically touch the interfaces (e.g., [196]), especially to explore elements on the screen. Although visually impaired individuals can currently leverage voice control to input certain commands (setting a timer), future research should explore how to best integrate zero-touch interactions with electronic devices alongside existing practices utilizing touch to manipulate ingredients and cooking tools.

3.5.2 Status Tracking and Safety Monitoring

From our interviews, we highlighted that people with visual impairments have various difficulties in tracking the status of tasks in progress (Section 3.4.2.6), protecting themselves from getting burned during different cooking procedures (Section 3.4.2.4), and handling accidents while cooking (Section 3.4.2.4). Prior work has explored using thermography to detect heat sources for energy auditing (e.g., [132, 133]). A similar approach could be applied in the kitchen to detect heat hazards. Beyond heat detection, prior research also explored using computer vision approaches to opportunistically capture actions and provide proactive reminders to users with visual impairments [94]. Similar approaches using computer vision or audio [105] could be applied in the

kitchen to track object locations or track a user's actions and the status of cooking steps (e.g., boiling water).

3.5.3 Collaborative Cooking

The majority of our participants complained about cooking with other people because they have to constantly speak about their actions and ask about other people's actions. Cooking while constantly talking might not be a natural behavior for people both with and without visual impairments. However, three participants did mention the benefit of having another person to accomplish tasks that rely on vision, such as checking expiration dates and food doneness (P10). Prior research has explored collaborative behaviors for people with visual impairments for other purposes, such as gaming [61], use of tangible interfaces [39], music composition [142], and creating accessible home environments [28]. Therefore, future research should explore the collaborative cooking behaviors for people with visual impairments and further create guidelines and training methods to reduce safety concerns as well as both mental and physical effort.

3.5.4 Contextual Inquiry and Learning Process

In this chapter, we described our research approach using YouTube video analysis to understand more about the practices of cooking by people with visual impairments. In general, video analyses of this type have the benefit of allowing researchers to gain a broad understanding of the practices of assistive technology use for many different users. This has been exemplified in prior work in HCI and Accessibility (e.g., [9, 82]). In future research, a contextual inquiry study is well suited to gather a deeper understanding of individual users' practices and challenges (e.g., [48, 94, 101, 109]), which may then be aggregated to understand themes across those individuals. As it relates to the research opportunities we discussed above, a contextual inquiry study may enable future researchers to deeply explore the practices and challenges of zero-touch interactions, status and safety monitoring systems, and collaborative cooking. For example, conducting a contextual inquiry study at a participant's kitchen could identify more status and safety monitoring concerns that might not be recognized by the participants with visual impairments and are not included in the camera's field of view in YouTube videos. An example of this might be watching a visually impaired cook show how she cuts vegetables in front of the camera while, off-camera, cooking chicken in a pan. In this case, neither the audience nor researchers may recognize any potential safety concerns with the pan or other objects in the surroundings from the video. Beyond only interviewing visually impaired individuals, we realized it would also be beneficial for future research to conduct interviews with teachers or trainers who work at community centers and help people with visual impairments with their activities of daily living and meal preparation. This may generate more common practices of cooking by people with visual impairments.

3.6 Summary

In this chapter, we systematically categorize and identify the distinct cooking practices observed across 12 different kitchen activities. Through a detailed analysis, we aim to highlight the nuanced techniques and adaptations employed during non-visual cooking tasks. Following this, we uncover the specific challenges that people with vision impairments face in the kitchen, using insights gathered from semi-structured interviews with participants. These interviews reveal a

range of difficulties, from ingredient identification to task sequencing, all of which play a crucial role in shaping the non-visual cooking experience.

This chapter forms the knowledge foundation for understanding the complexities of non-visual cooking, serving as an essential precursor to exploring, understanding, and constructing the physical information space—a conceptual framework that represents how physical objects and actions are perceived, interpreted, and utilized in non-visual contexts. By identifying key patterns and gaps in existing practices, we provide a comprehensive scope that is vital for future research. The findings from this chapter will inform both the exploration of core challenges and the design of innovative assistive technologies. Additionally, it offers guidance for researchers and designers who aim to delve deeper into the design space of non-visual cooking, helping to shape tools and systems that enhance accessibility and autonomy for individuals with vision impairments in kitchen environments. Based on the findings of this chapter, we further dive into two core challenges that people with vision impairments face in cooking—Accessing Recipes (Chapter 4) and Accessing Contextual Information of Objects (Chapter 5).

Chapter 4

Design Framework for Learning and Following Accessible Recipes

This chapter was adapted from my published paper:

Franklin Mingzhe Li, Ashley Wang, Patrick Carrington, and Shaun K. Kane. 2024. A Recipe for Success? Exploring Strategies for Improving Non-Visual Access to Cooking Instructions. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*, October 27–30, 2024, St. John's, NL, Canada. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3663548.3675662>

4.1 Overview

Cooking often requires multitasking, as cooks must prepare food (often through time-sensitive tasks) while following an overall plan, either memorized or by following a video or recipe. Recipes, whether presented as text or video, contain important procedural and contextual details that are often essential to correctly preparing dishes [35, 136]. Increasingly, home cooks are looking beyond cookbooks and relying on online cooking resources such as blogs and videos as part of their home cooking toolkit [42].

Having the agency and autonomy to access recipes and instructions while cooking is important for cooking by people with vision impairments [110, 190]. However, recipes may be inaccessible when they include instructions that expect the cook to be sighted. For example, a recipe might include an instruction to cook an item until it is a particular color (e.g., cook the chicken until golden brown). The lack of non-visual alternatives and descriptions creates obstacles to cooking for people with vision impairments [20, 100]. Barriers to cooking one's own food can lead to over-reliance on takeout or prepared foods, which in turn can result in poor nutritional outcomes [86].

Inspired by our research on non-visual cooking practices and challenges [110], it is important to dive into recipe accessibility for people with vision impairments, as changes in how home cooks use recipes can perpetuate or increase inaccessibility. Furthermore, it is also important to explore and instantiate the design space of non-visual access of recipe content and enhance learning experiences for people with vision impairments. To do so, we need to explore the following research questions.

- RQ1: What methods, strategies, and challenges do people with vision impairments have when accessing recipe information while cooking?
- RQ2: What types of recipe information, descriptions, alternatives, and structures are most helpful during the cooking process?
- RQ3: How might improvements in technology improve the accessibility of cooking instructions?

To explore these questions, we conducted semi-structured interviews with 20 people with vision impairments who have experience with cooking, along with interviewing four cooking instructors from a vision rehabilitation center. In this chapter, we first present how people with vision impairments currently use recipes while cooking (Section 4.4), the preferred design of the recipe content and structure (Section 4.5), and design guidelines to create more accessible home cooking experiences (Section 4.6). Overall, this chapter contributes to the design framework that provides the foundational knowledge of non-visual access to recipes and design guidelines for accessible recipe content, structure, and user interactions.

4.2 Related Work on Recipe Structure and Access

This section concentrates on the work most directly relevant to this study—how recipes themselves are worded, structured, and accessed—and the gap it leaves for non-visual recipe use.

Recipes act as the main source of cooking activities [23], and prior research has explored recipe wording and construction. For example, Tasse and Smith [175] created CURD, a database of recipes annotated using the MILK language, a meaning representation language based on first-order logic, to break recipes and steps into pieces of a specific format and function. Other work has explored algorithms that provide more adaptive recipes based on individual needs [178], cooking difficulty [104], and comparison of similar recipes [35]. Teng et al. [178] measured the role of ingredients and the relationships between co-occurring ingredients, constructing networks that, together with user feedback, drive recipe recommendation algorithms. Given these approaches to extracting recipe information and providing adaptive recipes, it is unknown how they can be leveraged for visually impaired individuals, and more importantly, what procedural and contextual information people with vision impairments need and why.

Practically, recipe information is accessed through many channels—online resources (e.g., AllRecipes.com [7], FoodNetwork.com [139]), traditional cookbooks [44], cooking videos on YouTube [197], and smart speakers [8]. Because the majority of recipes are text-based, non-visual access has relied on the screen readers, speech synthesizers, and Braille displays surveyed in Chapter 2, while recipes stored in tactile, image, or video formats carry their own access barriers [126, 137, 146]. Building on these general access methods, it remains unknown how people with vision impairments currently access recipe information specifically, and what design features a system for non-visual access to recipes while cooking should have.

4.3 Methodology

We conducted semi-structured interviews with blind people who have experience in cooking, and with adaptive cooking instructors from a vision rehabilitation center. Participants were compen-

Participant	Age	Gender	Vision Description	Self-	Years at Current Vision Level	Years Cook-ing	Days per Week Cooking
P1	32	Female	Totally Blind		21	15	3
P2	61	Male	Totally Blind		22	53	4–5
P3	48	Female	Totally Blind		48	20	7
P4	57	Female	Totally Blind		42	40	2
P5	47	Male	Legally Blind with light perception		47	23	Several
P6	38	Male	Totally blind in one eye, Legally Blind in other		5	32	Several
P7	44	Female	Totally Blind, some light perception in one eye		44	20	4–5
P8	26	Male	Legally Blind		6	20	2–3
P9	31	Female	Totally blind in one eye, Legally Blind in other		4	18	3
P10	61	Female	Totally Blind		61	45	7
P11	38	Male	Totally Blind		10	30	7
P12	67	Male	Legally Blind		67	49	7
P13	36	Male	Legally Blind		5	about 20	7
P14	30	Female	Totally Blind		22	11	once per month
P15	70	Female	Totally Blind		54	58	7
P16	24	Female	Legally Blind		24	2	3–4
P17	39	Female	Totally Blind		13	26	7
P18	48	Female	Totally Blind with very little light perception		48	40	3–4
P19	46	Male	Totally Blind		46	29	7
P20	24	Male	Legally Blind		12	14	3

Table 4.1: Home cooks with vision impairments who participated in our study

sated with a \$20 gift card. The recruitment and study procedures were approved by our Institutional Review Board (IRB). The semi-structured interviews were conducted through Zoom [84], and all interviews were audio-recorded and transcribed verbatim.

4.3.1 Semi-structured Interview with Blind Cooks

4.3.1.1 Participants

We recruited 20 blind participants (Table 4.1) through social media (Twitter and Reddit), and mailing lists maintained by blindness organizations. All participants were 18 years or older, legally or totally blind, experienced in cooking, and able to communicate in English. Table 4.1 presents participants’ demographic information.

4.3.1.2 Pre-interview Survey

We asked participants to complete a pre-study questionnaire to confirm that they met the inclusion criteria and asked for demographic information, cooking frequency, recipe sources they use, and three dishes they would like to learn. The questionnaire was administered using a Google Form [63] and is included in the supplementary material.

4.3.1.3 Interview Procedure

Each interview session took approximately 90 minutes and covered the following topics:

1. **Demographics and Background [~ 5 Minutes]:** participants’ age, gender, descriptions of vision level, and cooking experience. (see Table 4.1);
2. **Current Cooking Practices [~ 20 Minutes]:** participants’ current ways of accessing recipes and cooking information, such as what technologies they use to search for cooking-related information, how they gather information about physical objects while cooking, and how they adapt information resources to make them accessible;
3. **Recipe Review and Editing [~ 30 Minutes]:** participants edited recipes as described in the next section (Section 4.3.1.4);
4. **Current & Desired Technology Use [~ 30 Minutes]** participants’ experiences with different technologies and services to access recipe information while cooking.

4.3.1.4 Recipe Review and Collaborative Editing

As part of the interview, we asked participants to read and reflect on recipes that they were unfamiliar with, but that reflected dishes they were interested in learning to cook. We used this activity as a probe to get concrete feedback about recipe content and structure, while choosing recipes that matched participants’ interest and abilities.

During the pre-interview survey, we asked participants to name three dishes that they would be interested in learning to cook. Before the interview session, we located recipes for each of the dishes through recommended recipe sources by participants, and edited them into a simplified text format. During the session, we walked through the recipes with participants and collected their feedback about each step. This process also informed our follow-up questions during the rest of the interview.

Participant	Age	Gender	Vision Description	Years at Current Vision Level	Instruction Experience
I1	58	Female	Legally Blind	4	9 months
I2	24	Female	Sighted	24	10 months
I3	58	Male	Legally Blind, can only see high contrast	3	4 months
I4	65	Female	Totally Blind	35	30 years

Table 4.2: Instructor Demographic Information

4.3.2 Semi-structured Interview with Adaptive Cooking Instructors

4.3.2.1 Participants

To complement feedback from cooks themselves, we interviewed four instructors (I1–I4) who teach blind cooking courses at a vision rehabilitation center (SAVVI, located in Arizona, USA); this included totally blind, legally blind, and sighted instructors. Their demographic information is shown in Table 4.2.

4.3.2.2 Adaptive Cooking Instructor Interview

Each interview session took approximately 75 minutes and covered the following topics:

- **Demographics and Background [~ 5 Minutes]:** participants’ age, gender, vision level, and length of teaching experience;
- **Experiences in Blind Cooking Instruction [~ 30 Minutes]:** experiences of instructing blind people in cooking classes, including instructional practices, challenges faced by their students, learning barriers encountered by their students, and preferences for recipe information and instruction;
- **Perceptions of Recipe Presentation Quality [~ 20 Minutes]:** perceptions of recipe information sources, quality, and quality of presentation for blind cooks; how information should be presented to blind cooks at different stages of their learning processes and cooking activities;
- **Instruction Information Access Technologies [~ 20 Minutes]:** technologies that participants used to teach blind people to cook, and suggestions for how technology could be designed to better support blind cooks.

4.3.3 Data Analysis

To analyze the interview data from blind cooks, two researchers performed open coding [37] on the same four transcripts to form and reconcile the initial codebook. Those researchers independently coded the remaining transcripts before discussing the codes and resolving conflicts (e.g., missing codes, disagreements). After the two researchers reached a consensus and reconciled

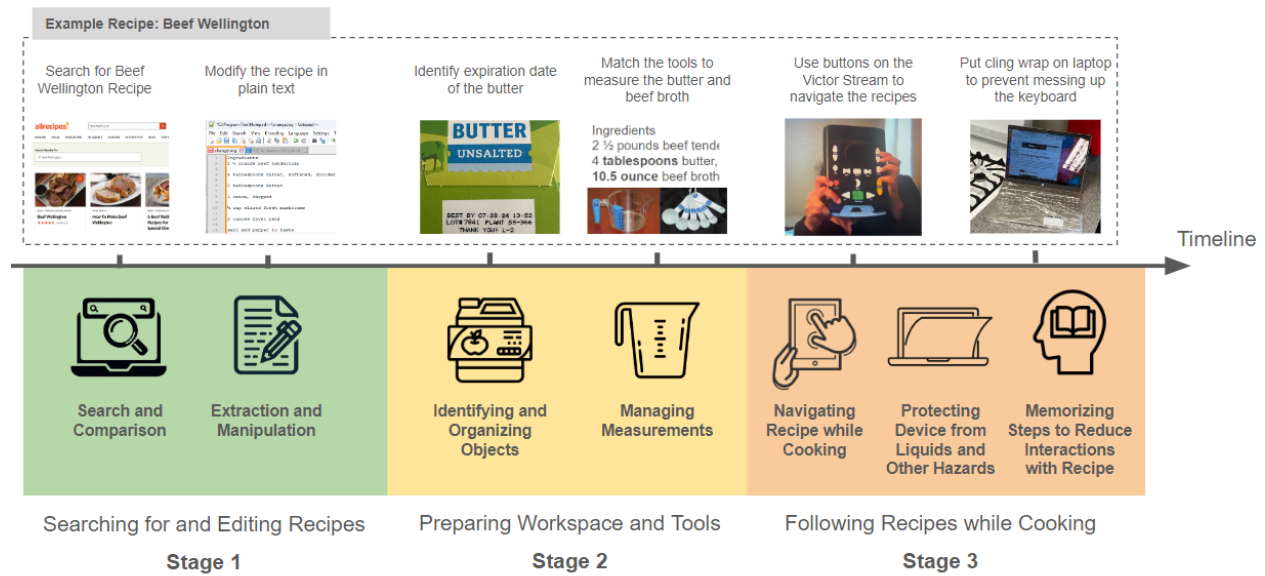


Figure 4.1: Illustration of cooking phases, using the example process of cooking Beef Wellington from a recipe.

the codebook, they performed affinity diagramming [77] using a Miro board [93] to cluster the codes and identify emergent themes. The interview data from the instructors' interviews were coded separately from the cook interviews but using a similar process of redundant coding for all four transcripts (I1–I4) by two researchers. This was followed by discussion, conflict resolution, reconciliation, and affinity diagramming to identify emergent themes.

4.4 Findings: Current Use of Recipes while Cooking (RQ1)

We identified current methods, strategies, and challenges of accessing recipes by cooks with vision impairments. These are loosely organized into three cooking phases (Figure 4.1): 1) searching for and editing recipes before cooking, 2) preparing their workspace and tools using recipe information, and 3) cooking while following recipe information.

4.4.1 Searching for and Editing Recipe Information

This stage of the process is characterized by actions taken to acquire information needed to perform specific cooking tasks, decide between different recipes, and manipulate the recipes themselves to be used in the kitchen.

4.4.1.1 Search and Comparison

Participants described their ways of searching for recipe information ahead of cooking. All participants mentioned that they use text-based online recipe information sources to learn non-visual cooking skills (e.g., measurement techniques) and new recipes and cuisines. Twenty participants described accessing recipes through general-purpose recipe websites (e.g., allrecipes.com, food-network.com), eight mentioned using specialized websites for blind cooks (e.g., visionaware.org,

acb.org), and six used Facebook groups (e.g., incredible recipes, the country cook). P17 described their experience of leveraging various online information for cooking with relative success:

“...The ads are really frustrating, but once you can get through those, I’ve had a lot of success. At Thanksgiving, I always forget how long to do a Turkey or whatever, so I get the weight of my Turkey and I go search and find all these recipes and things like that, and they’ve been super helpful and I’ve even gotten some new ideas that I didn’t necessarily know about from reading these recipes and things like that. Get a little more adventurous in my cooking. I’m a great basic cook, but trying new ideas and stuff, recipes are great for that...”

A common challenge was how **inaccessible and unpredictable visual content in online sources (e.g., ads, commercial videos) diminishes the accessibility of the resource itself**. P6 said:

“...When navigating the recipe websites, swipe right to scroll through them. Sometimes that would cause problems. If there was a video in the recipe, it would automatically start playing with sound. And even if I were to scroll past it, the video was already playing. Then the sound would just continue...I don’t know whether or not the video was still visually on the screen or not. But the audio still [played]. So that was a real hindrance. Especially when you’re relying on the audio of TalkBack at the same time, so I can’t remember if it was playing over TalkBack, or if TalkBack wasn’t working, but I do remember that it basically stopped me in my tracks...”

Participants often **search for multiple candidates and compare them based on ingredients, timing, and availability**. P7 explained:

“...I honestly just hop on Google...I’ll look to see how many people have rated it. I’ll look through three or four recipes about the same thing, and figure out what sounds good to me, what flavors sound right, what timing sounds right...”

In addition to accessing recipe information through websites, some participants have tried to use video and audio sources, often with limited success. Twelve participants mentioned that they have accessed recipes through YouTube, but do not rely on it due to **lack of audio description of actions in YouTube videos, which caused missing information**. Some still use YouTube for ideation and inspiration. Five participants mentioned using audio podcasts for cooking information, but did not generally follow recipes from those podcasts. When we asked participants what devices they used to access this information, most used their phones. Six participants also mentioned that they sometimes used smart speakers to access procedural and contextual information of cooking, but generally used them to answer questions or provide short directions rather than following a full recipe. When asked about challenges related to audio and video cooking information, participants mentioned the difficulty of following recipes with multiple steps (P1, P9), lack of consistency of the recipe sources (P5, P17), and difficulty in extracting recipe information into a format they could follow along with (P5, P12, P19).

4.4.1.2 Extraction and Manipulation

Once they selected a recipe, our participants applied a variety of strategies to extract and manipulate information to make the content more accessible and easy to navigate. Nine participants described **saving information either in a Word document or text file**, both for ease of navigation and to avoid ads. P1 commented on how she saved recipes from Facebook groups:

“...So I have joined a bunch of Facebook groups and I’ll typically get recipes from those groups, and if I see something I will save it, and typically I’ll not look at it at the moment. I’ll just sort of save and curate. And then when I have time, I’ll go look at the websites and copy everything out into easier-to-read plain text files, so that I don’t have to deal with ads and web formatting when I’m trying to cook...”

In addition to copying and pasting text into documents, participants **manually typed information from audio or video recipes into a text document**, although doing so required some tedious effort. P16 explained:

“...I’ll just play it a couple of seconds at a time. If she’s listing the ingredients, I’ll start it and pause it and then type it into my computer. It’s not convenient, but it’s nicer than if I miss an ingredient. I’d have to watch the whole video again...”

After converting the information into text, 13 participants described how they **edited ingredients and steps**, as well as adding personal notes. Participants usually saved and edited recipes on a desktop or laptop computer, then accessed those recipes on a mobile device while preparing and cooking. Participants often use their screen reader’s search function to **search for headers or specific text, such as an ingredient name or measurement unit**. P1 commented:

“...I’ll have to do a control F and find and I’ll search for a word of ingredients or I’ll search for the word ‘cup’ because I know that’s going to be in there somewhere. That is often why I’ll go ahead and copy it into Notepad, because especially on mobile, it’s a different experience...I’d rather do it when I’m sitting down and have the time to copy it out...”

4.4.2 Preparing Workspace and Tools

After finding the right recipe and converting it into an accessible format, participants often start the cooking process itself by preparing and laying out tools and ingredients; sometimes this included specific tasks to make objects and ingredients more accessible.

4.4.2.1 Identifying and Organizing Objects

Participants noted that preparing the workspace usually involves **visually identifying objects and food packaging** such as the name of the object, expiration date, nutritional information, and suggested cooking instructions. Participants use **visual recognition apps (e.g., SeeingAI [135], Lookout [64]) to scan objects**. Our participants discussed various challenges they encountered while using vision apps in the kitchen, such as accurately detecting text on different food packages, controlling when text is (or is not) read out, and navigating visual descriptions. P19 described the positive and negative aspects of using AI-based vision apps to scan food products:



Figure 4.2: Participants altered objects to make them more accessible, such as by adding NFC labels to spice jars.

“...If you buy canned fruit they might actually have a little recipe on the side of the can, and sometimes I try to read that and try to figure that out. And that’s kind of an adventure. It’s not very accurate at all times, but it’s something I like doing just because I can’t access such information from other sources...”

When using apps, participants sometimes struggle to get them **to only read their desired text, and to not just read any text that shows up in the camera’s field of view**. P9 explained:

“...So those are not the best if you want to get the ingredients honestly, because it would just start reading whatever is in front of you, but there’s no real order or you cannot explore using...So, it would be nice if there was a mode where you can really scan or take a picture of the product and it would give you a better interpretation of columns and better interpretation of tables on a package...”

P6 desires **more fine-grained control over what information is presented at what time**, for example describing what each object is when first organizing ingredients, but not to repeat that every time:

“...I don’t want the app to track other things. For example, I only want to know was it like a package of fish, or is it like a package of paddies. I only want to know this information while I’m turning on the system. ”

Beyond extracting information from physical packages, participants also faced challenges when identifying ingredients in similarly shaped containers, such as spices. Some participants choose to **apply tactile markers or NFC tags to distinguish similarly shaped objects** (e.g., the WayAround Tagging System, Figure 4.2) (P3, P6, P7, P18).

Recipe step (visual descriptions in **bold**)

Bake until baklava is **golden and crisp**.

Drop the covered ice cream balls into the hot oil and fry, turning occasionally so they **color evenly, until golden brown on all sides**.

Heat 2 tablespoons olive oil over high heat until **shimmering**.

Continue baking until **lightly browned on top**.

Fold the **sides, top, and bottom** of each husk in toward the center to enclose dough.

Add the onions to the skillet and sauté until **translucent**.

Slowly knead in a small amount of water at a time, using just enough to form a soft dough that is pliable and **even in color**.

Deep fry until **bright red in color** and crispy.

Reduce heat to 425 degrees F (220 degrees C) and continue baking until pastry is a rich, **golden brown**.

Table 4.3: Recipe steps gathered for the recipe editing task. Recipe steps often included inaccessible visual descriptions.

4.4.2.2 Managing Measurements

Using measuring cups or spoons presented accessibility challenges when the cook must measure something that partially fills the measuring cup (e.g., detecting when a measuring cup is half full). Participants **avored smaller measuring tools that could be filled completely, so as not to require reading markings on them**. P8 commented:

“...I frequently employ a variety of measuring tools for different recipes. Depending on whether I’m measuring liquids or solids, I use distinct measuring cups with different units. I tend to favor smaller measuring cups and measure multiple times, as it’s more convenient for me than attempting to use a larger measuring cup, which can be challenging to measure a portion of it...”

Participants found recipes challenging when they mixed units of measurement, as that required using and coordinating more measuring tools. Participants noted that **automatic unit conversion** could simplify this task (P2, P8, P11, P15).

4.4.3 Following Recipes while Cooking

Most participants listen to recipes in audio format while cooking. Participants use a variety of tools to access recipes while cooking, including their smartphone-based screen reader (9), laptop screen reader (8), audio players/text-to-speech apps (7), smart speakers (6), physical cookbooks (3), and smartwatches (1) (see Figure 4.3).

4.4.3.1 Navigating Recipes while Cooking

Participants generally (N=17) used **touch gestures to navigate while cooking, either via a screen reader or smartphone**. One challenge is that **screen readers often repeat unneces-**

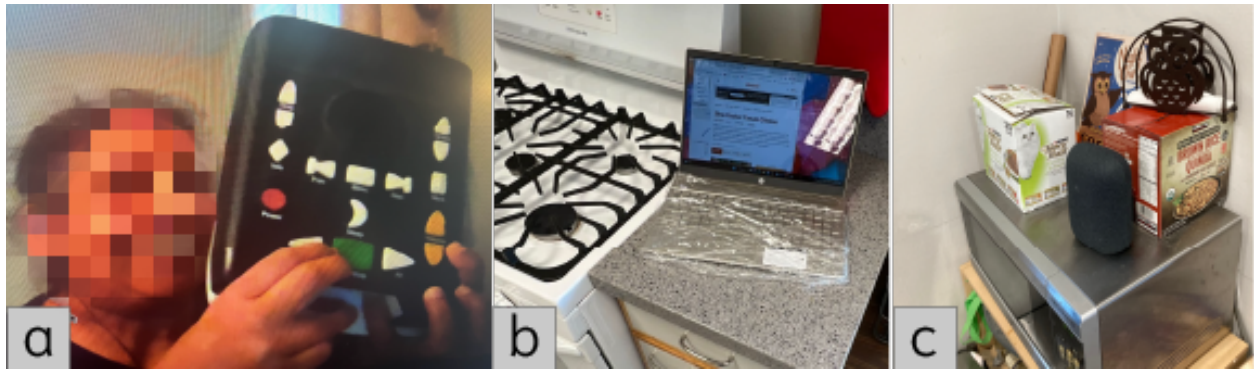


Figure 4.3: Technology used to access recipes. (a) P4 uses an accessible audio player to play back recipes; (b) P2 puts cling wrap on their laptop to prevent it from getting dirty; (c) P11 uses a smart speaker in the kitchen to access recipes and timers hands-free.

sary information, even when a cook needs to act quickly. This problem also occurs when recipes are not broken down into individual steps. P6 explained:

“...When I’m looking at the steps...I would listen to it. and then go back and complete a couple of actions, and then have to re-listen to the step. [When replaying the step] It often plays from the beginning of the step. And every time, I ended up having to listen to information that isn’t necessarily relevant to the actions I’m actually doing...”

Devices with **physical buttons, such as dedicated audio players, may be easier to manipulate while cooking** (see Figure 4.3 (a)). P2 commented:

“...I’d have it on my Victor Stream [audio player]. and I could just press the play/pause button and rewind it if I needed to, or fast-forward it...”

While some participants have **access to smart speakers, it can be difficult to navigate through a recipe using voice commands alone.** This difficulty is exacerbated by devices that repeat information or forget the current state of the recipe (P3, P15). P15 explained:

“...I typically do not use Alexa to guide me through the recipe while cooking. It is very common that it does not recognize your wording clearly and loses the current stage of information...”

4.4.3.2 Protecting Devices from Liquids and Other Hazards

One commonly described challenge was interacting with devices while in the process of cooking, especially when the workspace could be harmful to electronic devices (i.e., wet or messy) or when their hands are dirty. Participants described several strategies for overcoming these challenges: **positioning their devices far away from the cooking area, adding protective coverings** such as plastic wrap (see Figure 4.3 (b)), or **using voice or other hands-free interaction methods** (e.g., using a smart speaker (see Figure 4.3 (c))). P2 described his strategy:

“...When I was using my laptop, I would get like Saran Wrap or something, and I kind of wrap it over the keyboard. So when I was pressing the buttons, I knew it was kind of covered. So I could still press the buttons, but then the Saran Wrap would be covering it...”

4.4.3.3 Memorizing Steps to Reduce Interactions with Recipes

One strategy that applied regardless of the technology used to follow a recipe was to memorize several steps at a time to **reduce the number of times they consulted the recipe**. One participant (P7) shared her personal experiences on this matter:

“...And I look, kind of skim back and forth, and try to keep remembering. ‘Okay. So this much water for the rice. And no, this time when I make it, I’m not putting ginger in there, but I did last time for the slightly other recipes that they sent me.’ So yeah, looking back and forth to figure out what am I throwing in there and when and if my timing is right. So yeah, just scrolling back and forth as I went, as I would do steps...”

4.5 Findings: Recipe Content and Structure (RQ2)

During the study, participants read along with recipes that were chosen by the research team based on the three dishes they wanted to learn and suggested recipe sources. They provided feedback about their preferred structure, descriptions, and levels of detail for these recipes. Specifically, we engage the question: *What types of information, descriptions, alternatives, and structures might make the cooking process more effective?*

4.5.1 Recipe Content Preferences

Through reviewing recipes with the research team, participants shared their feedback on what information was necessary, unnecessary, or missing in each step.

4.5.1.1 Alternatives for Visual Descriptors

Recipes often include steps that require visually inspecting food as it is being prepared, such as checking doneness by examining the food’s color (see examples in Table 4.3). **These visual checks are inaccessible to many blind cooks**. One of P13’s sample recipes included the instruction: “Toast split bread under the broiler. Remove bread when it is **toasted golden brown in color**.” This accessibility problem was mentioned by most participants as an obstacle that they could not easily overcome. P7 commented:

“...It is really common to see ‘cook this until golden brown’ in recipes, but it is sometimes hard to standardize the conversion to a description in time or texture. By this time, I always need to further search online for a different recipe for the same cuisine and check if it has a different description...”

Many recipes include time-based instructions (e.g., “pan fry for 5 minutes”) along with visual descriptors. However, **timing information is not always sufficient, as cooking times and equipment vary**. P19 explained:

“...Yeah, so just knowing the time of cooking didn’t give you how well things cooked. So that’s the thing. Sometimes if your water evaporates faster or if your heat is too low, it will be raw on the inside. So, it’s not really telling you how you know if they’re cooked through. Even after an hour passes, if you have your heat too low, then they might still be raw...”

Besides timing, our participants highlighted the **importance of having texture descriptions (4), sound descriptions (6), as well as smell descriptions (3) to substitute the use of visual descriptions**. P4 commented on the importance of having smell descriptions:

“...Most of the recipes will not have the smell part of the description. They will say, when it looks reddish brown, caramelized onion. So caramelized onion, there will be a smell...Caramelized onion smells sweet and nice...You will know the smell of when it is cooked...Onions will smell differently than cooking potatoes or cooking peas or whatever you are cooking...”

4.5.1.2 Summaries and Outlines

Participants noted that, before starting cooking, they often read through the entire recipe to ensure that they had the needed ingredients and tools, and to make time to prepare the dish. Participants using screen readers found that **reading the entire recipe often took too much time, often because the recipe contained extraneous information** such as stories about the recipe. Participants **prefer having an overview or summary at the beginning of the recipe**. P9 explained her desire to know how the dish would turn out before reading the recipe:

“...And I also think getting a short overview of how the recipe is going to turn out. I guess or some of the ingredients, maybe the basic ingredients that you’re going to need without telling you the amounts necessarily. But at least that way you can figure out if, oh okay, I do have all the ingredients to make it, or oh, you know what, I’m missing something basic...”

Furthermore, beyond a high-level overview at the beginning of the recipe, our participants also described their **desire to have a short summary of the key task and estimated time for each step at the beginning of each instruction step to reduce the effort of navigation**. P12 mentioned:

“...I would rather [the summary] be part of the steps, and they shouldn’t go on too long. Like the scalding, it should be able to describe it pretty shortly. But I would like it to be right there in that step. You don’t want to have to go to someplace else to find it, because then you’d have to find your way back to where you were...”

4.5.1.3 Precise Descriptions of Components and Actions

Participants noted that some recipe descriptions were vague or left out important details, and noted that images within recipes often provide useful contextual information about the size, shape, and color of ingredients, and about how to physically complete cooking steps. Participants noted that **recipes should describe in text any contextual information that is presented visually, such as the size of an item or how a particular task should be performed**. For example, one of P16’s sample recipes stated “form into patties and place on the prepared broiler pan or baking sheet;” P16 commented:

“...I’d want to know an approximation of the size of the patty...”

Similarly, some recipes may omit **details about specific tools to use**, or might expect readers to infer this information from a photo. In response to the instruction “Heat oil in a large skillet or wok over moderate heat,” P9 commented:

“...[I don’t know] the size of the skillet, because large could mean different things to different people...”

Recipes with images might also omit **details about how to perform certain actions, such as distance, intensity, time, and repetition**. In response to the instruction “place foil around the edges of the crust to protect it from burning,” P3 asked:

“...Yep, how far on the edge? Is it over the edge into the pie? Is it the edge that is not covered by the egg mixture? I think I’d want a little more specific on that...”

4.5.1.4 Substitutions or Alternative Instructions

Recipes often make assumptions about the tools that the user has available. Our participants noted that they may sometimes perform tasks differently than others, such as by using specialized tools (e.g., accessible measuring cups) or cooking in a certain way for accessibility reasons. Twelve participants noted that they prefer recipes that **provide information about alternatives, such as substituting ingredients or using alternative cooking tools, and optional steps**. P9 commented on the importance of providing substitutions:

“...I guess it would also be cool to include in the recipe maybe at the end or in the ingredients list, if there’s any substitution for any ingredient that you can make if you don’t have that ingredient at home or if there’s something that you can skip, optional things. I think mostly recipes do tell you if something is optional, but they don’t give you any substitutions. For example, if you’re trying to make cookies and then you don’t have eggs, they don’t necessarily give you substitutions for, okay, if you don’t have eggs, we recommend you to use this...”

When given the instruction “Heat oil in a deep fryer to 400 degrees F (200 degrees C). Deep fry samosas until golden brown, 3 to 5 minutes each”, P15 noted that it would be helpful to suggest alternative cooking methods:

“...I don’t deep fry things because it is hard for me to track the food. What’s an alternative? If there’s a baking alternative, put it right there in the step. like: For deep frying, do this. For baking: do this. For air fryer: do this...”

4.5.1.5 Time Range and Tolerance Details

Participants preferred instructions that **provide a time range for each step, and note how these timings might affect the final product**. P6 commented on the uncertainty of timing included in recipes which results in over- or under-cooking food:

“...there’ll be a lot of times where the cooking time won’t be, it won’t necessarily reflect how long it takes for, say, chicken to cook. It’ll be either undercooked, or sometimes they’ll go in the opposite direction and in the recipe they’ll give instructions for times to cook a chicken that will end up being completely overcooked because they’re afraid that somebody’s going to undercook it and get sick...”

P4 commented on the need to have *time ranges* as descriptions: “*Five minutes timing is not enough, and sometimes they can say a range between five to seven minutes, not just five minutes.*” Furthermore, our participants also commented on the value of **describing time constraints around when certain steps must happen, and which steps can be combined**. P11 commented on their preferences about time management:

“...I need specifics, I need temperature, I need time, and things that keep time management going. So a recipe where it will give you the instructions to cook pasta and tell you while you’re cooking the pasta now you can do something, you can make the sauce. You can do something else within the timeframe, like the 11 minutes it takes to cook pasta...”

4.5.2 Recipe Structure Preferences

4.5.2.1 Separate Processes into Discrete Steps

All participants mentioned at some point that **existing recipes often have steps that are too long, or that combine multiple steps into a single instruction**. Participants suggested that **recipes should be broken into discrete steps for each ingredient and action**. P19 noted that it can be confusing when multiple ingredients are added in a single step:

“...when you were mixing the eggs white with the milk. that’s one step, then, when you fold in the egg yolks. I think that would be another step. That’s another that you know. That’s another agent entering. So anytime you have to add another ingredient. I would, but I consider that another step...”

Similarly, P11 noted that each action should be described in a separate step:

“...You have something, you’re putting a bunch of stuff in and cooking it for 25 minutes, that’s your first step. And then your next step is to put more stuff in, and cook that for six more minutes. The next step that changes or alters the cooking time or adds to the cooking time, makes that its own step. And then try to keep the activities in the step logically linked. So again, if you’re manipulating food, cooking, adjusting heat, timing, that’s one step...”

Recipes that did not follow these rules could be confusing. For example, P2 noted that a recipe was not clear about how often a repeated action should be performed:

“...No, I do not understand it all. Maybe because it was adding, adding the milk, you know, and then whisking it until it thickened, and then adding more milk and whisking it until thickened. You might wanna break that down. So because you’re adding milk. In fact, I can’t remember if it was 3 or 4 times. you might want to add each one of those. So if I was doing something new, I’m, you know, might say, okay, I’ve done that step. Okay, now I’m on step 2, and so the second time I add the milk, and then you know, step 3...”

4.5.2.2 Organizing Ingredients by Use or Type

P16 noted that **some recipes list ingredients in confusing order**, and that ingredients should be ordered by when they are used:

“...Yeah, I would say my ideal recipe is going to be my ingredients list first, grouped together based on usage. So, back to that banana bread, I want the baking soda, the flour, and the salt to be at the top of the list. And then I want it to be sugar, brown sugar, and bananas. And then I want it to be chocolate chips because those are the order that they’re added in. And then I like to go into the directions section...”

Alternatively, some participants preferred that **ingredients be listed by general category, so that they could be gathered at once and (when possible) mixed ahead of time**. As P3 explained:

“...For ingredients, I’d want...A lot of times in baking recipes, wet ingredients together, dry ingredients together...”

Participants emphasized the importance of **using consistent language to describe ingredients**, both in the ingredients list and later in the instructions. When navigating a recipe with a screen reader, it is often difficult to jump back and forth between the ingredients list and the current step, as explained by P3:

“...[is the garlic] dry? Is it dry garlic? Or was it garlic powder? Is it garlic salt? Whatever it is called for in the ingredient list, just call it by the same name all the way through, I’d say. Well, all the spices. Is it brown sugar? Is it white sugar? Is it ... I’m assuming it’s ground cinnamon and other things. Just keep consistency between the ingredient list and the step...”

4.5.2.3 Group Steps into Logical Chunks

Recipes comprise a list of steps in order with the expectation that users will complete all of the steps in order. However, there may be times when **blind cooks perform steps out of order, or break a recipe into several working sessions due to timing or accessibility constraints**. P6 mentioned his desire for **recipes that are broken into logical chunks** so that he can work on each part as time permits:

“...Making sure to you know if you’re putting a whole dish together, and you’re preparing a couple of different specific things that are gonna go together. Say, you’re cooking a sauce or a gravy, and then separately. You’re cooking, you know, meat like a steak, or and then, you know, maybe you’re also cooking, you know some onions and peppers to go on top of that clearly separating that into different sections is a big help for recipes...”

Participants also preferred when **recipes note which ingredients will be used multiple times**. P7 explained:

“...Oh, right. So you want two of them in this part and then one of them in that part. If more ingredients were that way, six cloves of garlic divided and then you’d know, Okay, let me just throw these in two separate areas so that I don’t accidentally put all of it in one space...”

4.6 Findings: How Technology Can Improve Recipe Accessibility (RQ3)

During our interviews, participants shared their current challenges and frustrations related to following recipes while cooking. These examples led to further discussion about how technology could be improved to mitigate these challenges. Here we consider the question: *How can technology make following recipes more accessible?*

4.6.1 Improving Recipe Organization, Navigation, and Level of Detail

4.6.1.1 Provide a Level of Detail that Matches the User

Participants may prefer different levels of detail depending on their skill or familiarity with the recipe. Since there may be no one-size-fits-all level of detail, **recipes can be adapted to the user's skills and experience**. P9 explained:

“...It has to be detailed, but not so much that you have a paragraph for each step. So I don't want to get all the information, but I do want to get the information that I need. And my background is in culinary arts, so maybe for me, it's easier to understand even with fewer words. But if I learned to cook or started cooking while I was blind, I would definitely want to get some non-visual pointers in my recipes...”

Participants noted that **each user might prefer more or less detail for a recipe, depending on their stage within the cooking process**. P1 said:

“...I'm thinking because I think there are different modes of viewing a recipe. So it could be that when someone is just looking at the recipe or when they're battle planning, they're preparing to cook, that information is annoying, it's too much. It is too much, it's cognitive overload to read all that in one step...”

4.6.1.2 Communicating Information via Voice Characteristics

Some participants suggested that a system could **communicate contextual information by changing the characteristics of text-to-speech voices**, such as using different voices to read different parts of the recipe (P12), or allowing users to adjust speed and volume during specific cooking stages (P2, P3, P4, P14, P17, P20). For example, P4 noted that she might use a fast voice when cooking alone, but slow the voice down if she was cooking with a friend.

4.6.1.3 Enable Non-Linear Navigation through Recipes

Participants described **the difficulty of tracking steps and managing timing, especially when multiple steps [must happen simultaneously]**. P11 explained:

“...I need time, and things that keep time management going. So a recipe where will give you the instructions to cook pasta and tell you while you're cooking the pasta now you can do something, you can make the sauce. You can do something else within the timeframe, like the 11 minutes it takes to cook pasta...”

One way to address this challenge is to **provide alternative paths for stepping through a recipe**. P9 presented an example of her desired navigation strategy:

“...Yeah. So how it works is I’ll ask for a recipe, it will give me the first option. If I like it, I can give it two different commands. The first command is to start with ingredients or start a recipe or something like that. I’m not getting the right wording, but it’s the concept of. So when you start gathering the ingredients, it will go one at a time and you just tell like the next ingredient. Of course you have to say her name and then next ingredient, next ingredient. And then it will tell you when you’re at the end of the ingredients, and it will ask you if you want to continue with the instructions and then it goes one step at a time...”

4.6.2 Supporting Users in Learning to Cook

We asked the adaptive cooking instructors about how their students work when taking classes, and how they could be better supported in their cooking activities.

4.6.2.1 Customizing Instructions for Skills and Ability

All instructors mentioned the importance of adapting instructions to the student, depending on their age, cooking experience, skill, and visual abilities. For example, I3 noted that some students knew how to cook before the vision loss and took cooking classes only to re-learn cooking skills non-visually, while others are learning to cook for the first time. We also found that people with fewer years of cooking experience or who have never cooked before blindness (e.g., P16) are more likely to prefer more detailed cooking descriptions and links to external resources through technology than people who had many years of cooking experience before blindness (e.g., P6). I4 emphasized the importance of **providing customized cooking instructions based on individual background**:

“...I found these newly blind people who show interest in cooking tend to have different understandings of cooking. Emmm, I remember there was a student who was a chef before becoming blind. His knowledge was much stronger than mine, so I only needed to teach him the basics of using tools non-visually, etc. But for many many other students who were blind and had never cooked independently before, I need to start from scratch to teach them everything...”

As noted in Section 4.4.1.2, more experienced cooks often edited recipes while transcribing them into a text file. Essentially, these cooks are customizing their own recipes, although **customizing and editing recipes could be made easier**. P8 described how and when he chooses to customize recipes:

“...Yeah, my first time when following the recipe, I just get to follow the recipe. And wherever I feel, okay, I’m not comfortable with this, I can actually make my own personal adjustments. Yeah. Yeah. That’s just my own recipe...”

The process of customizing a recipe also **creates an accessible version of the recipe that may be useful to other blind cooks**. P7 commented: *“I wish I could just see and use other people’s edits on the recipe who are also blind, which could save a lot of time. Other people could then see my edits too.”*

4.6.2.2 Structured Discovery and Professional Instructions

Instructors mentioned the teaching approach called Structured Discovery as a way to support students with varying skills and abilities. In Structured Discovery, instructors refrain from using hand-over-hand guidance and encourage students to experience cooking skills on their own with verbal guidance from instructors. I1 explained:

“...We’re a structured discovery center, so most of the time it’s students having to work through these things on their own...Typically, I don’t do a lot of hand-over-hand teaching. If I’m giving instruction or if students need any clarification or they’re learning something new, I always try to give verbal instructions first...”

We also found that our participants prefer **instructions that are intended for blind cooks specifically, rather than for the general population**. P11 commented on the usefulness of BeMyEyes [192] in his cooking:

“...So if they had some kind of connection to chefs or anyone that can note that they are cooks or they have good cooking knowledge, then when you go in and you specifically request. Hey, I need some help in the kitchen, then you can get someone who’s knowledgeable about cooking methods or just things that you need a little bit of guidance or things that you need to do something in the kitchen...”

4.6.2.3 Answering Questions and Providing Assurance

Our instructors (I1–I4) mentioned that during cooking classes, their students **constantly ask questions for confirmation and learning purposes**. They noted that **asking questions while practicing can be a distraction, and that students must learn to act without external confirmation**. I3 explained:

“...A lot of times it’s just looking for reassurance. Did I do this right? That’s the biggest one. I’m not sure if I’m doing this right or if this is the right piece of equipment. Because we try to get that out of the way ahead of time. I try to make them think through the process and ask most of the questions that they would’ve asked in the process ahead of time in conversation. It takes a lot of time to do that, but it saves time during the actual cooking process...”

Our participants mentioned that existing ways of asking smart speakers, searching through recipes, or using vision assistance apps either cannot get the expected answer or the whole process takes too long (Section 4.4.3). We found that cooking-related questions are often very context-specific, such as “how long should I cook this tomato again?” This points to the **importance of understanding the question’s intent and the current stage of the recipe**. P3 shared some common questions that she has while cooking:

“...Yeah. Standard questions like, what’s the cooking time? What’s the oven temperature? Things like that. That would be handy to know right away...”

4.6.3 Integrating Technology into Kitchen Activities

As discussed previously, participants often struggled with using their normal cooking devices in the kitchen, and invented workarounds for doing so, such as covering their laptop with cling film. During our interviews, we discussed how technology could better integrate into kitchen activities.

4.6.3.1 Using Multiple Devices Together

One problem noted by participants was the **difficulty in sharing information across multiple devices**, such as when they saved recipes on a computer and read them on a mobile device. P12 suggested that cooking information should propagate across his computing devices, including his smartphone and smartwatch:

“...[it should] notify me on my phone and my Apple Watch when it is done by vibration or sound whenever it is appropriate...”

4.6.3.2 Integrating At-School Learning and At-Home Practice

I1 noted that **recipes and tools used at school should also be available at home**, to support practicing skills:

“...Yeah, a huge, huge thing that I would love to support is having as many smart devices in the house. However, if a person is in training and they have the smart devices in the kitchen and they don’t have them at their house, then that’s going to be a challenge. So I also think meeting the students where they’re at is a good thing, too. So meaning that besides the smart devices, if they don’t have them at home, we’ll then also teach them without those smart devices...”

Instructors mentioned that their students may **struggle to reproduce work from class at home**, with different tools available and without feedback from instructors (I1, I2, I4). I4 explained:

“...I mean, many of my students do not have the same set of tools or devices at home compared with the kitchen at the rehab center, they sometimes struggle with the tools they have at home and there is not much I could do to help them...I usually have my phone available after hours and try to clarify and help as much as I can, but it is usually hard to do so through phone calls...”

4.6.3.3 Supporting Hands-Free Interaction

In a previous section, we noted the challenge of switching between cooking activities and interacting with technology. Participants suggested that technology could enable **hands-free interactions with their recipe devices**, which would reduce the need to rinse and dry their hands during cooking. P19 commented:

“...It would be nice to have either something that can be on some sort of a stand or hands-free something that you can ask your device to do, especially when you

have your hands full. So you got your potatoes in one thing, and you got your stuff, and it's really difficult to actually grab your phone and try to see. You know. Rinse your hands and then get back to other stuff. It'd be really nice to have something hands-free that can do that or even a wearable that concentrates on stuff like that cooking..."

Alternatively, computing devices could **provide simplified touch interfaces to prevent mistaken input**. P1 said:

"...So sometimes with apps, you can restrict the gestures that are available to make so that, what am I trying to say? When you're navigating with a screen reader on the phone, you have all the gestures available to you so you can accidentally go out of the file or you can accidentally lose your place. So if it was an app...that would restrict the gestures and only leave you a limited number of gestures available in that app. So you can't accidentally go out of the recipe..."

Our participants also emphasized the importance of **leveraging multiple interaction modalities while cooking** (P2, P8, P11, P17). P17 commented:

"...I only have one option to interact with my recipe reader which is through touch, I had to wash my hands multiple times and I wish I could simply use voice or wave my hand to move the recipe to the next step..."

4.7 Discussion and Future Work

In this section, we will discuss opportunities and considerations to augment recipe information and interactive technologies to better support non-visual access to cooking instructions and related information.

4.7.1 Augmenting Recipes to Make Them More Accessible

The first strategy we suggest is to augment the information content within the recipe itself. This would include modifying recipes to add the suggested information additions or to make them more widely adopted. Referring back to the previous section, this would include support for non-visual alternative descriptions for objects (including sound, smell, or touch) and especially changes in ingredient or cooking status (Section 4.5.1), incorporating both summaries and additional precision where appropriate. Today's AI tools (e.g., Large Language Models (LLMs)) may be used to summarize short pieces of the recipe to achieve the vision of concise step descriptions based on key actions without the need to re-write the full recipe [125, 147, 185]. By gathering further information from blind cooks, it may also be possible to identify unique substitutions for actions, in addition to utilizing resources for ingredient substitutions. Furthermore, leveraging the expert knowledge of chefs and food safety experts we may also be able to identify resources to build a repository of tolerance information for various ingredients.

4.7.1.1 Non-Visual Database for Common Cooking References and Substitutions

In our study, we showed that our participants prefer having non-visual interpretations to substitute visual descriptors (e.g., cook until golden brown) to reduce the barriers to learning and following cooking instructions effectively. We found that our participants gradually supplemented their cognition by either converting the visual description of a certain action or process non-visually before starting to cook, or utilizing strategies directly learned from blind cooking classes. This approach varies based on different ingredients, ways of cooking, size of the food, shape, and many other variables, which makes the process very experimental. We believe that it could be beneficial to create a database of non-visual descriptors for common cooking tasks and references to allow simple and automatic substitutions of relevant information. This database of recipe descriptions could include attributes such as [Original Visual Description], [Ingredient], [Cuisine Name], [Cooking Method], and [Non-Visual Description]. We believe this would reduce the learning curve for people with vision impairments, and having such a database could also contribute to recipe customization models (e.g., [35, 104, 178]) to automatically transform recipes with non-visual equivalence.

4.7.2 Facilitating Information Extraction for Kitchen Tools and Ingredients

4.7.2.1 Automatic Extraction of Visual Content from Photos and Video

From our study, we found existing online recipes contain various visual content that is inaccessible to blind people, such as having pictures or videos attached to the text to show the ingredients, steps, or size of objects. For example, in a video recipe, the person might verbally say, “Cook this until golden brown,” however, we can leverage data extraction tools [194] to identify any text shown in the video that indicates the timing or even extract relevant or unique sounds that are generated through the process (e.g., sizzling). This approach could potentially contribute to the non-visual reference database for visually impaired people to leverage while following a typical recipe.

4.7.2.2 Supporting Extraction of Static Information from Physical Objects

In this chapter, we found that our participants leveraged vision apps to identify information on physical packages, and we showed various challenges that they encountered through the process (Section 4.4.2). This is similar to prior research that explored package fetching by people with vision impairments [108]. Here, we propose future research to consider providing universal designs of packaging for visually impaired people, which includes considerations of the absolute position of certain information on the physical package (e.g., nutrition list, ingredients list, cooking instructions), support recognition of unpackaged items (e.g., a bunch of parsley from the store), as well as tactile readings [155]. Due to the existing difficulties of using vision apps to identify visualizations (Section 4.4.2), we also recommend that package designers utilize plain text structure and provide high-level summaries for easier reading through vision apps.

4.7.2.3 Supporting Extraction of Dynamic Information from Physical Objects

Furthermore, objects used in the kitchen often change or move, thus necessitating a constant re-evaluation of space and the objects that are being used. Future technology tools need to be

able to support identifying these kinds of changes and communicate such information at appropriate stages during the cooking process [109]. Our participants mentioned how the granularity of recipe information may change through different cooking stages (Section 4.5), and the same applies to details about physical objects, especially while learning (Section 4.6). For instance, consider a simple example of a knife, a versatile tool in the kitchen. At the start of a cooking task, people might need information about its location to ensure it’s safely stowed in its designated spot. However, as the cooking progresses, the context for this object evolves. People may need updates on whether the knife has been used, requiring information on its cleanliness and any food residues. Later in the process, people might require data on the knife’s orientation, particularly if it has been set aside with the blade exposed. This need for dynamic information motivates technology to assist blind cooks by reducing cognitive load and increasing environmental awareness.

4.7.3 Supporting More Accessible Kitchen Work

4.7.3.1 Multimodal Opportunities for Blind People in the Kitchen

We also learned that our participants prefer having the option to have multiple devices (e.g., connect the smartphone with the smart speaker) and modalities (e.g., sound, haptic) in the kitchen to interact with recipes while cooking for convenience and to reduce false positives (Section 4.6.3.1). This would provide the freedom to choose the modality to interact with the system and receive output in preferred ways [112]. This provides opportunities to consider: 1) How the input and output modalities should be designed? 2) What are the preferred gestures for the multimodal systems? 3) What are the design considerations of having multimodal systems in the kitchen for blind people (e.g., deployment, activation)? 4) What are the broader perceived benefits beyond interacting with recipes by using multimodal systems in the kitchen?

4.7.3.2 Information Modality for Cognition

In this chapter, we showed that our participants leveraged text and video for different purposes (e.g., manually transcribing video content for details to text). Specifically, we discussed how our participants prefer leveraging sound and sayings from the cook to gain motivation for cooking (Section 4.4.1.1). Beyond cooking, our findings correlate to prior research on leveraging multimodal guides or interaction to support blind people in everyday activities, such as navigation [120], social interaction [94], gaming [158], doing makeup [116], and art experiences [34]. Therefore, to enhance the adaptation of information, further research should also consider how to enable information in different modalities (e.g., sound, smell, tactile) to better motivate and assist people with vision impairments in accessing information and interacting with devices in cooking.

4.7.3.3 Conversational User Interfaces for Kitchen-AI Systems

In our findings, we showed that our participants commonly leveraged communications with their instructors in the cooking class to reassure themselves about their process and ask for clarifications on specific steps (Section 4.6.2.3). This provides opportunities for leveraging tools such as LLMs [90] to provide prompts to answer questions that are highly contextual based on specific recipes to support learning experiences and agency to cook without instructors. For example,

this includes providing details of specific steps that require high memory load (e.g., quantity, time, ingredients), having detailed guidance on how to perform certain actions without vision (e.g., suggestions on measuring food doneness), having summarization of a recipe, and describing alternative ways to do something. To do so, future research should conduct contextual inquiries to observe existing Q&A conversations between blind cooking students and their instructors in cooking classes to further explore the proper prompts to provide accurate and ethical responses [90, 201] through the conversational user interfaces.

4.7.3.4 Customization Based on Individual Backgrounds and Conditions

In our findings, we showed that our instructors tailored their instructions based on people's diverse backgrounds (Section 4.6.2.1), such as people with or without cooking experiences before becoming blind or people of different ages [102]. In our study with blind cooks, we had participants who were totally blind as well as legally blind who were capable of finger counting (Table 4.1). By analyzing their behaviors, we found that our participants who are legally blind and can perform finger counting or exhibited some degree of tunnel vision (e.g., P6, P9, P12) also employed vision-assistive technologies during cooking procedures non-visually, such as P6 used BeMyEyes to read physical packages, P9 used SeeingAI to read labels, and P12 used VoiceOver to walk through recipes on mobile devices, which is identical to other participants who are totally blind. On the other hand, we found there are slight differences regarding how they incorporated vision-assistive technologies with their remaining visual capabilities (e.g., light perception). For instance, while using SeeingAI to identify object locations, P12 utilized his light perceptions to estimate the facing direction based on the daylight from the window and the surrounding area that might contain certain objects (e.g., milk in the fridge). Drawing insights from both blind cooks and instructors, our recommendation for future systems is to offer users comprehensive control options. This includes preferences for spoken information, customizable levels of navigational assistance, and targeted guidance on specific cooking skills. Empowering end-users with detailed controls can potentially enhance their individual user experience and adaptability to diverse needs.

4.8 Summary

In this chapter, we presented the design framework of recipe accessibility that includes recipe content accessibility, recipe structure accessibility, and the technological adaptations necessary to support non-visual cooking. The design framework introduced in this chapter highlights key elements needed for accessible recipe support:

- **Accessible Recipe Content:** Recipes should provide non-visual alternatives to visual descriptions (e.g., substituting color-based instructions with sound or texture cues). This helps blind cooks navigate tasks more effectively.
- **Structured and Modular Recipe Format:** Recipes should be broken down into discrete, manageable steps, with clear, consistent language throughout. This ensures that each action and ingredient is clearly defined and easy to follow.
- **Technological Interventions:** Enhancements in multimodal interaction, voice guidance, and context-aware systems can improve navigation and time management, helping blind

cooks engage more independently with recipes. Future AI systems should enable personalized levels of detail and support non-linear navigation through the recipe.

These findings offer a foundation for improving the design of recipe accessibility systems and supporting tools, contributing to greater autonomy and agency for people with vision impairments in the kitchen.

Chapter 5

Design Framework for Recognizing Contextual Information of Objects

This chapter was adapted from my published paper:

Franklin Mingzhe Li, Michael Xieyang Liu, Shaun K. Kane, and Patrick Carrington. 2024. A Contextual Inquiry of People with Vision Impairments in Cooking. *In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 38, 1–14. <https://doi.org/10.1145/3613904.3642233>

5.1 Overview

The culinary process leans heavily on visual cues, relying extensively on the contextual information of objects nestled within the kitchen (e.g., location, status), which remains elusive to people with vision impairments [110, 180]. This creates substantial barriers for them to cook independently and, ultimately, negatively impacts their quality of life [86]. For example, seemingly simple tasks like finding an item in a cluttered refrigerator or gauging whether a dish is fully cooked are challenging for visually impaired individuals, particularly for beginners in rehabilitation training programs [110, 190]. Prior research has highlighted the importance of recognizing contextual information [110, 180], such as the color and placement of objects in everyday tasks. However, there exists a notable gap in knowledge regarding the specific contextual cues that visually impaired people rely on and the rationale behind these preferences when cooking.

Although assistive vision systems like SeeingAI [135] and TapTapSee [174] have proven effective in helping visually impaired people identify objects under various settings, these systems frequently fall short in the kitchen due to their inability to extract contextual information the end-user prefers [110]. For instance, past studies have highlighted their lackluster performance in recognizing clusters of objects or retrieving precise information about an object, such as the expiration date of products [14, 49]. Additionally, these systems often require holding a smartphone to capture information, which is inconvenient and impractical during cooking tasks that are time-sensitive [110]. Therefore, it is crucial to understand the design space of how visually impaired individuals obtain contextual information about kitchen objects as well as explore effective ways to communicate this information while cooking, thereby shedding light on opportunities to increase culinary independence and enhance their overall quality of life.

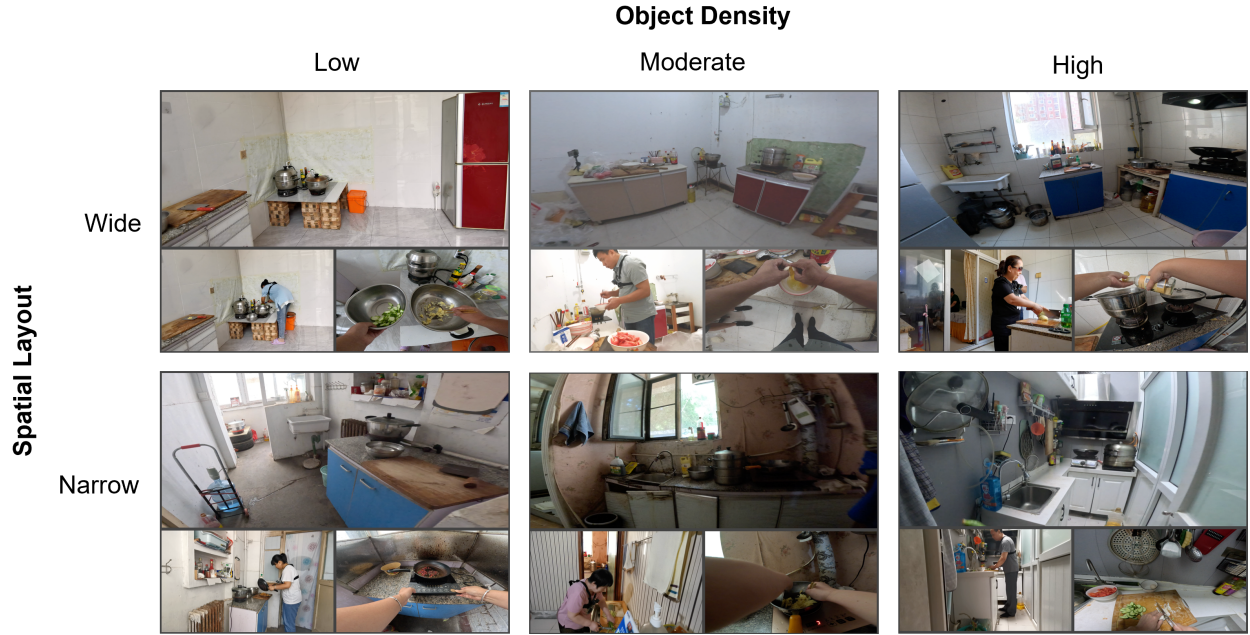


Figure 5.1: Contextual Inquiry: sample views of participants' home kitchens with varying spatial layouts (from wide to narrow), as well as object density.

To summarize, in this work, we aim to investigate:

- RQ1 What contextual information about objects is important to people with vision impairments during cooking?
- RQ2 What techniques and strategies do blind cooks use to non-visually gather and utilize information about objects during cooking and food preparation tasks?
- RQ3 What are the most effective methods for conveying necessary contextual information about objects at the right times and in the right ways?

To answer the research questions, we conducted a comprehensive contextual inquiry study with 12 visually impaired individuals experienced in cooking (Figure 5.1). This inquiry took place in participants' own kitchens, where they were asked to naturally prepare meals using ingredients available in their refrigerators (as detailed in Section 5.3.2.2). Subsequently, they were instructed to perform specific kitchen-related tasks related to object recognition, such as identifying ingredients and organizing their kitchen space. In the end, we conducted semi-structured interviews to gain further insights into the motivations and rationales behind their actions, understand their needs for object-related information, and explore their views on integrating AI-powered assistive technologies in the kitchen (Section 5.3.2.3).

In this chapter, we contribute to the design framework of recognizing contextual information of objects for people with vision impairments. It includes:

- A comprehensive analysis of the contextual information requirements for objects, comprising five primary, two secondary, and one application-specific category (Section 5.4).

- The nuanced process of gathering contextual information about objects for cooking activities, which involves establishing intentional contextual associations with objects (Section 5.5).
- Design guidelines for AI-powered assistive technologies tailored for visually impaired individuals in kitchen environments (Section 5.7).

5.2 Related Work on Accessing Contextual Information of Objects

Chapter 2 consolidates the literature on cooking and food-related activities for people with vision impairments, on the assistive technologies (tactile markers, crowdsourcing, computer vision, and multimodal systems) used to identify objects non-visually, and on the AI-powered kitchen technologies that monitor activities and appliances. This section focuses on the work most directly relevant to this study and the specific gap it leaves.

Prior research has documented a wide range of practices and challenges that people with vision impairments face in cooking [20,110,190]; for example, Chapter 3 [110] surfaced difficulties such as measuring, organizing space, tracking objects, and quality inspection. Many of these tasks hinge on identifying *contextual* information about objects (e.g., shape, color, location, internal state), and a substantial body of assistive technology—tactile markers [72, 173], crowdsourced visual question answering [19,73], computer-vision object identification [58,71], presentation and guidance toward target objects [200], and AI-powered kitchen sensing [99, 107]—has been built to surface it (reviewed in Chapter 2). Yet these systems largely stop at naming objects: it remains unknown what contextual information of objects people with vision impairments actually need across different cooking procedures, how they currently obtain it, and how a system should be designed to deliver it for cooking-related tasks. Our contextual inquiry addresses this gap.

5.3 Methodology

Contextual inquiry has been used in field research to understand end users’ experiences, preferences, and challenges during their everyday activities [17,81,88,122,123,153]. Contextual inquiry research can support deeper understandings of everyday human behaviors through surfacing facts, details, constraints, and structures [17]. Contextual inquiry has been widely adopted as a research method to better understand marginalized groups, cultural learning, and accessibility practices (e.g., [45, 48, 129, 166]). To investigate our research inquiries comprehensively, we initiated a **contextual inquiry involving 12 participants with vision impairments naturally cooking in their kitchens**. Our study unfolds in three distinct phases: the pre-study interview, the contextual inquiry, and the semi-structured interview.

5.3.1 Participants

We recruited 12 people with vision impairments from the mailing list of the China Disabled Persons’ Federation (Table 5.1). To participate in our study, participants were required to be 18 years or older, legally or totally blind, and have prior experience with cooking. Among the 12 participants we recruited, six of them are female and six are male (Table 5.1). The average age of our

participants was 42.5 (SD = 6.1). Eight of them are totally blind and four are legally blind (Table 5.1). Regarding the four participants who are legally blind, P1 has no vision in the left eye and light perception in the right eye. P3, P8, and P12 have some light perception in both eyes. They had an average of 21.7 years of cooking experience (SD = 9.0). As per their self-reports, seven of them cook every day, one cooks three or four times a week, two of them cook once per week, one cooks two or three times a month, and one cooks once per month (Table 5.1). Regarding living arrangements, eight resided with their families, one with roommates, and the remainder lived independently (Table 5.1). Participants were compensated in local currency, which is equivalent to \$40 USD. The recruitment and study procedure was approved by our organization's Institutional Review Board (IRB). Each participant's engagement took approximately 90 to 105 minutes.

PID	Age	Gender	Vision Condition	Cooking Experience and Frequency	Living Condition
P1	41	Male	Legally Blind	16 Years, Three or Four Times a Week	Living with Family
P2	40	Male	Totally Blind	18 Years, Once per Month	Living with Family
P3	32	Female	Legally Blind	12 Years, Everyday	Living with Family
P4	46	Male	Totally Blind	24 Years, Once per Week	Living with Family
P5	52	Male	Totally Blind	29 Years, Once per Week	Living with Family
P6	44	Female	Totally Blind	15 Years, Everyday	Living with Family
P7	49	Male	Totally Blind	26 Years, Everyday	Living Independently
P8	36	Female	Legally Blind	10 Years, Everyday	Living with Family
P9	39	Male	Totally Blind	18 Years, Everyday	Living Independently
P10	42	Female	Totally Blind	36 Years, Everyday	Living Independently
P11	51	Female	Totally Blind	38 Years, Everyday	Living with Friends
P12	38	Female	Legally Blind	18 Years, Two or Three Times a Month	Living with Family

Table 5.1: Demographic information of our study participants



Figure 5.2: Contextual inquiry settings. Three camera views captured during the study to show the overall kitchen environment, egocentric view, and a 3rd-person camera view of the person cooking. The top image is the overall still image of the kitchen setting. The bottom left image is the view from the stationary camera that shows our participant making a meal. The bottom right image is the egocentric view captured from the chest-mounted camera.

5.3.2 Study Procedure

Our study unfolds in three distinct phases: a pre-study interview, a contextual inquiry, and a semi-structured interview.

5.3.2.1 Pre-study Survey [5 Minutes]

In the initial pre-study interview, we gathered demographic information from our participants. This included details such as age, gender, vision condition, cooking experience, cooking frequency, living arrangements, preferred cooking activities, and any challenges they encountered in the culinary domain.

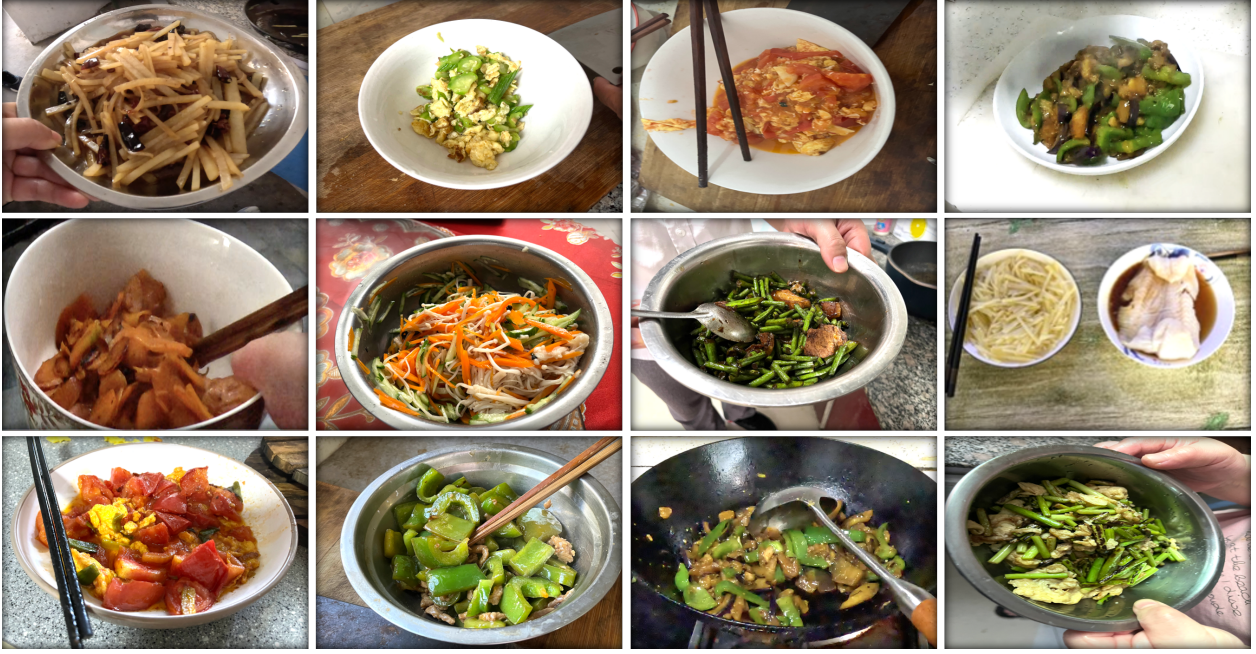


Figure 5.3: Dishes made by our participants during the contextual inquiry in their kitchens.

5.3.2.2 Contextual Inquiry [75 Minutes]

During the contextual inquiry phase, participants engaged in various cooking activities within their personal kitchens (as depicted in Figure 5.2). To record their culinary experiences comprehensively, participants were equipped with a GoPro 11 camera [65] attached to their chests, allowing us to document their actions and behaviors (as shown in Figure 5.2). To provide a comprehensive view of their activities, we installed a stationary camera within their kitchen environments (Figure 5.2). In the contextual inquiry, there were two main tasks:

[TASK 1: Self-Directed Cooking]: To observe a full experience of cooking, our participants were asked to first explore the food and ingredients that they had in their own kitchen, and then make a dish based on the availability of ingredients (See Figure 5.3 for dishes made by our participants). After making the dish, participants were asked to serve the dish and clean the kitchen. During the cooking process, participants were encouraged to vocalize their thoughts while performing tasks, following the think-aloud protocols throughout the contextual inquiry [186].

[TASK 2: Specific Cooking-related Activities]: Following the completion of Task 1, participants were further engaged in specific cooking-related activities to gain a deeper insight into their procedural methods and information requirements. Drawing from prior research findings [110], we chose eight key activities that epitomize the essential cooking tasks for individuals with vision impairments that are relevant to object identification. These activities encompassed: 1) identification of ingredients and food items, 2) recognition of cookware and utensils, 3) precise measurement, 4) monitoring of cooking progress, 5) the process of serving food, 6) ensuring safety measures, 7) maintaining kitchen organization, and 8) executing grocery shopping. Participants were requested to demonstrate their typical approach to completing these tasks and to vocalize their thought processes throughout, adhering to the think-aloud protocols [186].

5.3.2.3 Semi-structured Interview [25 Minutes]

Following the completion of the contextual inquiry, we proceeded to conduct a semi-structured interview. This interview served as a platform for a comprehensive debriefing, allowing us to delve into our observations and findings from the contextual inquiry. Additionally, it provided an opportunity to explore various facets of future design considerations related to the communication of visual information and the form factors of assistive technologies within the kitchen environment.

[Debriefing of the Contextual Inquiry]: In this segment of the interview, we initiated discussions concerning the behaviors and processes observed during the contextual inquiry (refer to Section 5.3.2.2). We delved deeper into the challenges encountered by participants, as well as their specific visual information needs. This encompassed topics such as object localization, contextual information regarding objects, and the dynamics of visual information requirements throughout different stages of the cooking process.

[Design of Information Communication and Deployment]: Subsequently, our conversation shifted toward the design aspects related to information communication and deployment within the kitchen setting. We inquired about participants' preferences for how information should be conveyed to them while cooking, their information needs during culinary activities, their opinions on the form factors of assistive technologies within the kitchen, and any other concerns or considerations they wished to share.

5.3.3 Data Analysis

The contextual inquiries were meticulously documented using both video recording methods. High-resolution footage was captured at 5.3k resolution with a frame rate of 30Hz utilizing the GoPro 11 camera in HyperView [65]. Additionally, a stationary camera was employed to supplement the recordings. Meanwhile, the semi-structured interviews were recorded in audio format. We leveraged the video resources captured by the body-worn camera and the stationary camera for the **contextual inquiry** analysis [97]. For the analysis of the contextual inquiry videos, we used thematic analysis [30]. Two researchers independently annotated the video and open-coded the observations. Our analysis centered on two key aspects: the contextual information needs related to objects and the processes involved in acquiring contextual information about these objects (as detailed in Section 5.3.2.2). As for the analysis of the **semi-structured interviews**, a similar thematic analysis approach was employed [30]. This analysis revolved around themes related to the processes of acquiring contextual information, expected methods of information communication for objects, visual information requirements, and considerations regarding form factors (refer to Section 5.3.2.3).

5.4 Findings: Contextual Information Needs of Objects (RQ1)

In this section, we first present the five fundamental contextual information needs of objects that people with vision impairments prefer while cooking: position information, orientation information, proximity and grouping information, similarity and duplicate information, and internal state information. We further present three secondary and application-specific information needs

that are relevant to kitchen activities: safety-related information, health-related information, and plating and serving information. We illustrate this in Table 5.2.

Attribute	Description	Priority
Position	The relative location to a reference point, or an “anchor,” to indicate the object’s position	Primary
Orientation	Information about how an object is currently oriented (e.g., vertical, horizontal, or tilted) relative to a reference point (e.g., human, object)	Primary
Proximity and Grouping	Information about groupings of objects relative to others and the environment	Primary
Similarity and Duplicates	Information about similar or duplicate objects, which includes differentiation between similar objects, relative positions between objects that are similar, and the overall quantity of similar objects	Primary
Internal State	Information about internal state of objects, such as cleanliness, freshness, boiling water, the amount of solid or liquid inside the container, and the doneness of food	Primary
Safety-related Information	Information to monitor anything that might be harmful (e.g., knocked-over objects, open flame)	Secondary
Health-related Information	Information to track objects with health hazards after consumption (e.g., expired food, overcooked food)	Secondary
Plating and Serving	Information about the final appearance and presentation of objects upon finishing (e.g., color distribution)	Application-Specific

Table 5.2: Contextual information needs revealed during our study. Primary attributes were required to interact with objects, while secondary and application-specific attributes were related to understanding the state of objects and manipulating them.

5.4.1 Five Fundamental Categories of Contextual Information for Objects

5.4.1.1 Position Information

During our contextual inquiry conducted in the participants’ kitchens, we uncovered a unanimous consensus among our participants regarding the paramount importance of knowing the precise position of objects within their culinary domains, which could reduce the time and effort of finding objects while cooking (9), and support autonomy and agency in the kitchen (3). We found over half of our participants encountered difficulties when attempting to locate specific items while cooking, due to the mental load of memorizing the positions (P1, P8, P11) or misplacement of items by their family members or friends (P2, P4, P6, P10). This challenge increased

when dealing with less frequently used items. To remember object locations, our participants typically advocated for the **use of a reference point in the kitchen, or an “anchor,” to indicate the object’s position**. For instance, they preferred describing the oyster sauce as being “on the windowsill” rather than specifying its coordinates in 3D space (e.g., x, y, z). P5 elucidated:

“To determine the location of the item I’m searching for, I simply require a relative position in relation to a reference point in my kitchen, such as near my gas range or on my fridge.”

5.4.1.2 Orientation Information

Our inquiry also shed light on another critical facet—the need to ascertain the orientation of various kitchen objects, signifying their **current alignment or positioning relative to a specific reference point or another object**. We found that our participants encountered challenges in discerning the orientation of objects during different culinary tasks (P2, P4, P5, P6, P9, P11). For example, both P2 and P5 grappled with slicing pork belly with the correct orientation, often leading to the unintended separation of fat and lean meat portions. Similarly, our participants faced recurrent tribulations when endeavoring to align a wok precisely with plates or bowls during the serving process (P4, P6, P9, P11). This issue invariably resulted in unwanted food spillage. P11 provided illuminating insights into the complexities of this matter:

“Perfectly aligning the wok with the plate during serving can be quite challenging, and as a consequence, I frequently find myself grappling with food spillage, necessitating subsequent cleanup efforts.”

5.4.1.3 Proximity and Grouping Information

In addition to object orientation, our participants placed significant emphasis on contextual information related to proximity and grouping—essentially, the **arrangement and organization of objects in relation to one another and the kitchen environment**. Knowing the proximity and grouping information of certain object groups can support maintaining the kitchen, as well as navigating the kitchen space. This aspect encompassed the pressing need to determine if objects were correctly placed, if any misplacements or disarray had occurred (P4, P11), and whether alterations or rearrangements to the kitchen space had transpired during or after culinary activities (P5, P9). For instance, we observed that P5 faced difficulties locating sauces and ingredients following his daughter’s cooking session, as she had inadvertently rearranged various items. P5 articulated his predicament:

“I usually...have my sauces like soy sauce and vinegar arranged on the second shelf and other solid ingredients such as sugar and salt carefully positioned on the third shelf. However, my daughter cooked a meal yesterday, and the displacement of items made it exceedingly challenging to locate things today.”

5.4.1.4 Similarity and Duplicates Information

Our research also illuminated the critical need to acquire **information about similar or duplicate objects**, encompassing differentiation between similar items, determining the relative positions of objects, and assessing the overall quantity of similar objects. Understanding the

quantity of specific items, such as tomatoes within the fridge, emerges as vital for monitoring food supplies, particularly when preparing for grocery shopping. During our study, we observed that P3 had multiple tomatoes located at various spots within her refrigerator. When asked about the tomatoes' whereabouts, she discovered two with holes and remarked:

"I wasn't aware that I had these tomatoes tucked away in the corner of my fridge, and I can't recall how long they've been there. It's possible my son placed them there."

Moreover, we noted that providing information regarding the differentiation between similar objects can significantly assist individuals with vision impairments in comprehending their kitchen environments. For example, P1 and P8 both expressed the desire to know how different plates were stacked together on a shelf and whether all the plates were identical or if any variations existed during the serving process.

5.4.1.5 Internal State Information

Kitchen objects may possess various **internal states, including temperature, freshness, cleanliness, the condition of solid or liquid contents, and the degree of doneness for food that is being cooked**. Our participants noted that ascertaining these internal states can be a challenge, primarily because many of these assessments rely on visual cues. For instance, our participants expressed difficulties in determining the cleanliness of vegetables or meat (P2, P9). Other challenging tasks include monitoring the water temperature to determine if it has reached boiling point (P5) or gauging the readiness of food (P6). P9 explained this issue:

"It's impossible for me to determine if vegetables are clean or not during the washing process. Consequently, I often find myself repeatedly washing them to ensure their absolute cleanliness."

5.4.2 Secondary and Application-specific Information

In addition to the five fundamental contextual information categories, we identified three types of information that people with vision impairments are acutely aware of while interacting with objects in the kitchen.

5.4.2.1 Safety-related Information

The first secondary information is safety-related information, which encompasses monitoring objects that have the potential to cause harm. This includes objects that could be accidentally knocked over during kitchen tasks (P4, P9, P10) and safety-related concerns, such as monitoring the temperature of cooking equipment to prevent accidents (P4, P5, P6, P10, P11). Our observations revealed instances where P4 accidentally knocked over a salt bottle while searching for a plate, and P9 tipped over a nearby water cup. P9 emphasized the significance of being aware of potential obstacles:

"Being aware of potential obstructions that I might knock over could greatly benefit me throughout the preparation and cooking process. It could also help reduce the anxiety associated with interacting with objects while learning."

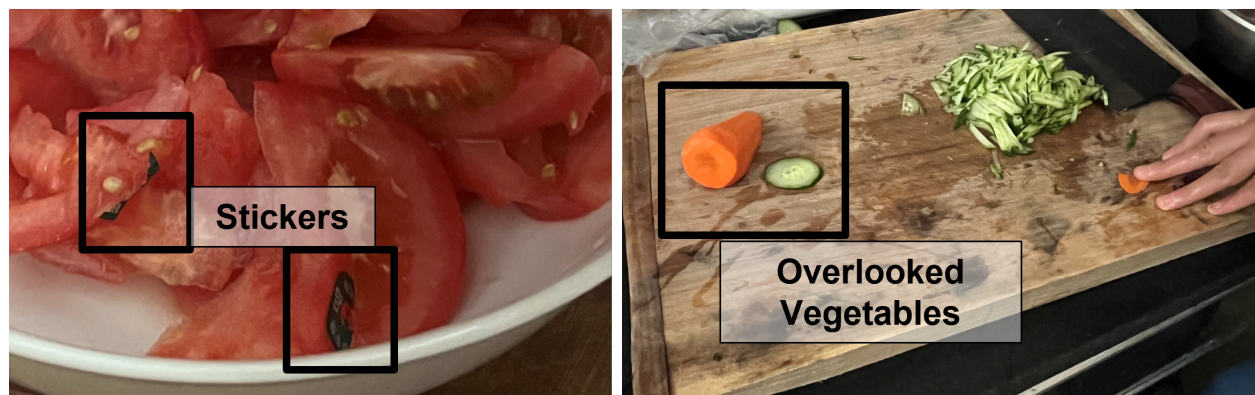


Figure 5.4: Some visual information can have health implications. Left: Sliced tomatoes still have stickers attached (highlighted); Right: Some vegetables were overlooked and left on the cutting board (highlighted).

5.4.2.2 Health-related Information

Our contextual inquiry highlighted the critical importance of knowing contextual information about objects for health-related considerations. This encompasses factors like checking expiration dates (P1, P10, P12), identifying food with potential health risks, such as overcooked items (P2, P5), and recognizing visual cues on food items, such as stickers on vegetables (P6, P8). P2 articulated the challenges he faced in this regard:

“To ensure all of the meat is fully cooked, I usually cook it for a longer period, which sometimes had some of the food got overcooked or even burned.”

We also observed instances where participants unintentionally left behind or missed certain food items, especially when dealing with round-shaped objects like green beans, or during the transfer of items from the cooking vessel to the plate. This oversight could lead to health-related concerns, such as consuming spoiled food or creating conditions favorable to pests like cockroaches (P2, P4, P5, P6, P11) (Figure 5.4). For example, P5 inadvertently included a food sticker while slicing a tomato, and it was subsequently cooked in the dish (Figure 5.4). P5 expressed:

“I had no idea there were stickers on the tomato, and it’s challenging to use my hands to feel the entire tomato to detect the sticker. Wouldn’t it be better if they stopped using stickers altogether?”

5.4.2.3 Plating and Serving Information

Upon completing the cooking process, we uncovered our participants’ preferences for being informed about the presentation and final appearance of the dishes they prepared. This encompassed details such as the arrangement of items on the plate (position), the degree of cooking (internal state), the spatial relationship with other items on the plate (proximity), and the orientation of food items (orientation). P5 emphasized the value of being aware of the visual presentation of their culinary creations:

“It’s valuable for me to be aware of how appealing my food looks once it’s served, or if I should consider adding additional vegetables or meat to enhance the overall visual presentation.”



Figure 5.5: P9 using his left hand to touch the side of the wok to gauge the temperature of the wok.

5.5 Findings: Process of Obtaining Contextual Information of Objects (RQ2)

In this section, we begin by discussing how individuals with vision impairments rely on multiple sensory inputs to acquire various types of contextual information [indicated in brackets] concerning objects during cooking activities. These sensory inputs encompass touch, sound, and smell. Subsequently, we explore two distinct approaches employed by our participants to streamline the process of identifying objects of interest: the creation of supplementary contextual information and the simplification of the contextual information acquisition process related to objects.

5.5.1 Multi-sensory Approaches to Obtaining Contextual Information

5.5.1.1 Touch

Our contextual inquiry revealed that our participants relied on their sense of touch for tracking the status of objects in their kitchen, including dual-purpose scanning and memorization, manipulating objects, and performing safety inspections.

Dual-purpose Scanning and Memorization: We found that our participants used touch for multiple purposes, such as localizing objects and checking the internal state of objects. This included locating items like vegetables, fruits, and meat within the refrigerator, as well as identifying sauces, ingredients, or containers within the kitchen [**Position**]. Touching food also communicated information about the freshness of the food [**Internal State**]. P4 commented: “Using

touch is my primary way of finding things in the kitchen, while I explore objects, I also feel the object to know if it is fresh through the stiffness or if there are holes on the skin.”

During this process, they engaged in pre-organizing objects of interest and often memorized the positions of other similar or nearby objects as they scanned the space [**Proximity and Grouping**][**Similarity and Duplicates**]. For instance, P1 searched for eggplants in the fridge, scanning through it while simultaneously committing the location of garlic to memory. Later, when he needed minced garlic, he easily found it, remarking, *“I memorized the position of the garlic last time when I was scanning through the fridge!”* However, we found this approach can be slow, and that the complexity of the space sometimes caused people to miss objects while scanning and to form an inaccurate memory of objects and space (P3, P7, P11). P3 explained:

“It often takes me a while to find the vegetable that I want to get. And it is easy for me to miss some of it, because kitchen shelves and refrigerator storage are complex, such as my tomatoes were placed at multiple positions in the kitchen. Once I did not find it, then it might just stay in a corner for many days.”

Precision and Manipulation: While manipulating and interacting with objects, we found our participants also leveraged touch to ensure objects are organized and aligned, which maintained order and reduced the risk of spillage, [**Orientation**] as well as to count and measure objects (e.g., sugar, vinegar) [**Internal State**]. To align objects, or when transferring materials between containers (e.g., adding sauces to salad, serving food from a wok) our participants usually used one hand to hold the object then used another hand to find and secure the other object. To determine the quantity of dry ingredients such as salt and sugar, they relied on touch, using their hands to feel and specify the exact amount (P1, P4, P5, P8, P9). For liquids, they often placed a finger beneath the lid, allowing them to feel the liquid passing through their finger to gauge the quantity (P5, P7, P10).

Safety Inspection: Furthermore, we discovered that access to safety-related information could significantly reduce the risks associated with cooking for individuals with vision impairments. For example, both P9 and P10 routinely performed thorough inspections of flammable objects that are close to the gas range before cooking [**Proximity and Grouping**][**Safety**]. P10 expressed a desire for pre-cooking safety checks: *“I would appreciate having some form of support for conducting safety checks before cooking to ensure there are no objects in close proximity to the range during cooking.”* Additionally, we found that gauging the temperature of the wok [**Internal State**][**Safety**] presented one of the most formidable challenges for individuals with vision impairments, as this information was traditionally obtained through tactile means, such as direct touch (see Figure 5.5). This practice, although effective, often resulted in burns and blisters, as expressed by P9:

“I use my hand to feel the temperature of the wok; you can see my arm has many blisters and burns, but I have to use this method as there is no other way for me to gauge the temperature or balance the wok correctly.”

5.5.1.2 Sound

Recognizing and tracking sounds played a pivotal role in our participants’ ability to assess the status of objects [**Internal State**] in cooking, such as temperature. As an illustrative example, P11 described a method involving the addition of a small amount of egg yolk to hot oil in a wok



Figure 5.6: Unique containers differentiate contents. Participants use and reuse unique containers to help identify objects.

to listen to the resulting sound to estimate the oil's temperature. While sounds were usually helpful, it was sometimes difficult to follow sounds because of excessive background noise (9), such as kitchen exhaust fans or conversations with others in the room. P11 provided further insights into this issue:

"I prefer using sound-based assessments, such as checking if water has boiled, but sometimes the differences in sound characteristics can be quite subtle...which is difficult to specify with my exhaust fan on."

In addition to assessing object status, participants utilized sound to estimate the quantity or volume of objects [**Internal State**]. An interesting example involved the use of containers with narrow nozzles. When pouring liquids into a pan or wok, the air inside the container compresses as the liquid flows, generating a distinct sound, often described as a "burp." This auditory cue allowed individuals to approximate the amount of liquid dispensed. P8 detailed the practicality of this method for estimating quantities through sound while acknowledging its inherent limitations, such as reduced precision, particularly when the liquid level in the container was low: *"I use sound to estimate how much oil I've poured into the pan. It's challenging to discern through touch alone. The first drop of oil hitting the wok generates a small sound, and since the oil bottle I purchased has only one nozzle, it doesn't continuously pour. Instead, it dispenses intermittently, producing a 'burp' sound. I rely on this auditory feedback to gauge the amount of oil in the pan. However, it's not always precise, as the sound may become less noticeable when the liquid level in the bottle is low."*

5.5.1.3 Smell

The sense of smell plays a critical role in helping individuals with vision impairments determine the internal state of objects and health-related information [**Internal State**][**Health**], such as the freshness or the doneness of the food. In cases where touch exploration was not feasible or practical, participants relied on their olfactory senses to detect signs of spoilage. P3 illustrated this practice, stating, *"It is common for us to have leftovers of main dishes as well as rice and bread."*

Sometimes, touching is often not feasible to assess the condition of the food. So, I use my sense of smell to determine if the food has gone bad. Spoiled food often emits a sharp and unpleasant smell due to fermentation." This reliance on smell allowed them to make informed decisions about whether it was safe to consume leftover food items.

Participants also utilized their sense of smell as a means to determine the readiness of certain dishes [**Internal State**]. Specific foods emitted distinctive aromas when they were close to being fully cooked. For example, P4 and P9 mentioned that particular dishes, such as those containing green peppers and meat, would release a savory aroma, signaling that they were nearly done. This olfactory cue served as an indicator of cooking progress. However, participants emphasized the importance of swift action once these aromatic cues were detected, as there was little room for delay between sensing the enticing aroma and preventing the food from becoming overcooked or burned. As P4 humorously put it, *"I like to use smell to gauge the readiness of my food. When you catch that aroma, it feels like a culinary achievement. But don't celebrate too long; you need to promptly remove the food from the wok to prevent it from burning."*

5.5.2 Altering Objects to Ease Identification

In addition to their multi-sensory strategies, our participants demonstrated the approach of actively creating additional contextual information for objects, thereby facilitating their recognition and organization. Through contextual inquiry, we uncovered a pervasive practice among all participants: the deliberate customization of objects to imbue them with supplementary contextual information. This approach involved introducing distinctive attributes, such as unique container shapes or deliberate organizational strategies, to enhance object identification and ensure secure organization within the kitchen.

5.5.2.1 Using Containers with Unique Shapes

We found that all of our participants typically chose to use different sizes or shapes of containers to indicate the difference between sauces, oil, or seasonings [**Similarity and Duplicates**] (Figure 5.6), which corresponds to prior research showing that bartenders used different glasses to remember orders [54]. From the observation of the contextual inquiry, we found that they easily spotted the sauce that they wanted, and we found that for P1, he used a thin-headed, mid-sized glass to store oyster sauce, a large, rectangle bottle to store vinegar, and a wide-headed, large plastic bottle to store soy sauce. P5 also refills oils to the bottle that he purchased a long time ago (Figure 5.6). P6 further explained:

"I personally use different shapes of containers to indicate the difference of oyster sauce, vinegar, and soy sauce. There was one time that my daughter purchased a new bottle of oyster sauce, which confused me for many days."

5.5.2.2 Securing Objects for Organization

Participants favored using bowls or plates to secure objects, promoting better organization, rather than letting them roll freely on the counter [**Orientation**][**Position**]. For example, P4, P5, P11, and P12 placed cut vegetables inside a bowl for ease of management. Similarly, P4, P6, and P10 employed a knife as a makeshift "container" to secure objects before cooking, such as scallions or minced garlic (Figure 5.7). This additional layer of context minimized the risk of accidents and lightened the cognitive load associated with memorizing object locations.



Figure 5.7: P4 used a knife as a “container” for scallions or minced garlic.

5.5.2.3 Spatial Separation and Grouping of Objects

In addition to enhancing the identification of individual objects, our study revealed that participants deliberately arranged objects in separate spatial groupings based on their purposes [**Proximity and Grouping**], aiming to facilitate easier identification and access (P2, P3, P6, P8, P9, P11). During interviews, participants elaborated on the advantages of spatially separating objects, highlighting how it added another layer of context and improved the spatial recognition of object groupings. P8 provided insights into this behavior:

“I prefer to keep my kitchen essentials minimal and uncomplicated. Simultaneously, I adopt a practice of spatially separating different objects, aiding me in distinguishing between various items more effectively and reducing the chances of encountering unwanted objects due to spatial constraints.”

Furthermore, participants mentioned their practice of organizing objects within dedicated mini-spaces, each containing a specific group of items. For instance, P11 stored all sauces in the same location, while P9 demonstrated how he arranged daily-use bowls within easily accessible spaces and placed less frequently used ones in cabinets:

“Locating and reaching certain spaces can be quite demanding, especially when I have to stoop down to access bowls near ground level. Therefore, I keep utensils I use daily on the counter, within easy reach, and store the others in cabinets.”

5.5.3 Optimizing Cooking Procedures

In their pursuit of greater convenience and efficiency, our participants devised a range of strategies to simplify their cooking procedures, reducing the need for extensive contextual information gathering. These innovative methods encompassed pre-assigning orders, sequential organization, strategic seasoning placement according to usage frequency, and preserving spatial arrangements

during the cleaning process. Through these tactical approaches, they were able to optimize their cooking routines while minimizing the effort required to access contextual information about objects.

5.5.3.1 Pre-assigned Orders and Sequential Organization

Our study revealed that participants frequently prepared objects in a **specific order based on their sequential requirements** (P2, P5) [**Position**]. This approach involved completing preparation tasks before embarking on the actual cooking process, thus optimizing their time and effort. P2 exemplified this by arranging the bowl of oil in front of the onions, followed by the meat, to denote the order of preparing the dish: *“Pre-arranging all the containers would save me time and effort during execution by minimizing the need to recheck the objects in the bowl.”* Furthermore, beyond arranging ingredients in the order they would be used, participants also **organized seasonings according to their frequency** of use (P9, P11) [**Similarity and Duplicates**]. P9 elaborated:

“I typically position the seasonings I use daily closer to the gas range, while placing others in cabinets or higher shelves.”

Additionally, participants maintained the **spatial arrangement of objects** even during the post-cooking cleaning process (P4, P5, P7, P12) [**Proximity and Grouping**]. This practice served a dual purpose, as explained by P5:

“The cleaning process not only ensures cleanliness but also guarantees that all objects are returned to their designated positions, preventing difficulties in locating ingredients during subsequent cooking sessions.”

5.5.3.2 Step and Purpose Combination

In pursuit of greater cooking efficiency, our participants skillfully combined specific cooking steps and harnessed objects for multiple purposes, simplifying their culinary endeavors, which reduced the need to obtain contextual information of multiple positions as well as differentiating different objects [**Position**][**Similarity and Duplicates**]. These strategies included merging cooking steps and employing objects with dual functionalities. Participants deliberately **integrated particular cooking steps and repurposed objects, effectively streamlining their culinary processes**. For instance, P3 favored using a pot as a multi-purpose container, consolidating ingredients in the pot before cooking. P3 elucidated: *“I adopt this approach to reduce the complexity and hassle associated with employing multiple containers.”* Similarly, P4 opted to place pepper and other seasonings inside the wok with oil before igniting the flame, effectively condensing multiple steps into a single action, which was typically divided in conventional recipes. Furthermore, our participants exhibited resourcefulness by **utilizing objects with dual functionalities to diminish the need for additional tools**. For instance, P11 employed a large bowl both as a wok lid and as a receptacle for washing vegetables. P11 explained the practicality of this approach: *“Utilizing the large bowl as a wok lid not only eliminates the necessity for a separate lid but also serves as a convenient vessel for washing various vegetables.”*

5.5.3.3 Reducing the Need for Precise Movement

In their quest for enhanced convenience and reduced demand for precision and accuracy, our participants expressed a preference for **minimizing complex 3D spatial actions [Orientation]**. They identified strategies that streamlined their actions, such as adjusting their knife-handling technique and adopting a pragmatic approach to discarding waste. Several participants indicated a desire for simplified knife handling, opting for a 2D movement approach by gripping the knife's back edge rather than the blade's tip (P4). This technique allowed for more straightforward and manageable motion when manipulating the knife during culinary tasks.

To further simplify their kitchen activities, our participants highlighted the practice of initially disposing of waste items in the sink (P6) **[Orientation]**. This approach was especially advantageous for items that required precise disposal, as participants found it challenging to accurately target a conventional waste bin placed at ground level. By contrast, the sink offered a more accessible, waist-level receptacle, reducing the need for precision and diminishing concerns about transferring waste to a traditional trash bin after cooking. P6 explained: *"This would reduce my effort of accurately throwing the garbage inside the bin...The water sink is big enough and at my waist level so I can easily throw things in it without much effort and worry about transferring them into the trash bin later after cooking."*

5.6 Findings: Contextual Information Communication and Deployment (RQ3)

In this section, we show our findings about contextual information presentation and deployment considerations by people with vision impairments during cooking activities. We present participants' preferences for information granularity in communication (Section 5.6.1), communication modality (Section 5.6.2), and form factor considerations for future technologies (Section 5.6.3).

5.6.1 Preferred Information Granularity and Level of Detail

5.6.1.1 Providing Precise Spatial Descriptions

Participants preferred specific, stationary spatial references to describe object positions, using objects in the kitchen as landmarks. P4 elaborated:

"I typically say that the salt is on the left-hand side of the gas range, or the oyster sauce is on the windowsill. I prefer not to know if my oyster sauce is close to my sugar, because other people might place it in different spots."

Moreover, participants requested detailed spatial references. P1, for instance, suggested describing an object as being on the third shelf of the fridge door, rather than simply mentioning that it is on a shelf. Best practices here depend on layout: a kitchen with only one window has objects "on the windowsill," while more detail is needed for a kitchen with multiple windows. In some cases, participants combined multiple objects to describe a location, such as *"the bowl is inside the cabinet at the left-hand side of my gas range."* (P1).

5.6.1.2 Limiting Verbosity

Participants expressed a dislike for systems that chatter constantly, as it could distract them from tasks that require focused auditory attention. P2 conveyed this sentiment:

“I don’t want the system to constantly provide verbal updates for every new object it encounters. It should provide contextual information thoughtfully.”

Participants requested that notifications be provided when most needed, and taking into account the user’s current location (P4, P7, P10). P7 elaborated:

“I don’t require precise navigation to a specific reference point in my kitchen, such as the position of my sink. I can manage that on my own. What I need is a brief notification to confirm that I’ve reached the desired position.”

5.6.2 Preferred Communication Modalities

5.6.2.1 Combining Speech and Non-Speech Audio Feedback

During the post-interviews, we asked participants about their preferred modes of receiving information in the kitchen. A majority of participants preferred non-speech audio notifications, such as a familiar “beep-beep-beep” sound, as these would be less distracting. This preference was primarily motivated by the desire to minimize cognitive load during cooking, as noted in Section 5.5.1.2. P4, P8, and P10 mentioned the advantage of less distracting audio feedback, particularly when engaged in conversation or consuming other auditory content. P10 pointed out:

“I prefer non-verbal sounds over human voice notifications as they are less disruptive while I’m cooking. If I’m in the middle of cooking and listening to a lecture, having a human voice as a notification can disrupt my experience.”

Participants noted that speech was still needed to convey details, but they requested control over when speech feedback was offered. As articulated by P11:

“I wouldn’t want the system to speak to me in a human voice unless I specifically request it. Typically, not during active cooking, but perhaps I might ask for more detailed information before or after I perform certain actions, such as inquiring about the quality of the dishes I’ve prepared.”

Overall, participants hoped for a balance between less distracting audio cues and on-demand speech feedback.

5.6.2.2 Offering Proactive Reminders

Furthermore, our participants desired proactive reminders around key topics such as safety- and health-related information (P8, P11), as noted in Sections 5.4.2.1 and 5.4.2.2. Participants stressed the importance of being able to edit and customize reminders. Reminders could be triggered by events in the kitchen or based on the user’s location. P12 elaborated on this concept, likening it to an indoor navigation system:

“This would resemble an indoor navigation system that can automatically provide me with the references I’ve designated as relevant to my current location.”

5.6.2.3 Increasing Environmental Awareness

When discussing the benefits of using a wearable camera to track the user's activities in the kitchen, nine participants emphasized that such a system would be tolerable, but that they also might need awareness of events happening outside the view of a wearable camera. P5 articulated this need:

“I’m unsure if I can always have the tracking system facing the exact objects or reference points. Therefore, I want the system to be capable of providing me with contextual information even when I’m not directly facing it.”

5.6.3 Deploying New Technologies in the Kitchen

5.6.3.1 Kitchen Deployment Considerations

During the contextual inquiry, we inquired about our participants' interest and concerns related to technology that could track contextual information. Our participants raised concerns related to introducing new technology into cooking, including the risk of water splatter, exposure to fire, oil-proofing, battery life, and dirt resistance. P9 also expressed concerns about the potential high cost of such a system, suggesting that additional support could be provided in existing mobile device hardware.

5.6.3.2 Device Form Factors and Location

Assuming new technology became available to support activities in the kitchen, we asked participants whether they would prefer a stationary system installed in the kitchen or a body-worn system. Out of the 12 participants, 11 expressed a preference for body-worn systems due to factors such as ease of installation, extensibility, and cost efficiency. P4 raised questions about the challenges of installing technology themselves. P7 mentioned that new technology might also be used for other purposes, such as indoor navigation. One participant favored a stationary camera due to privacy concerns associated with wearing such a system during other activities (P12).

In addition to this feedback, we conducted an evaluation in which participants wore the camera in various body positions, including the chest, head, and wrist, and asked about their preferences for device location. Eight participants indicated a preference for a chest-worn camera, while four favored a head-worn camera. Those in favor of the chest-worn position cited its non-obtrusive nature, expressing concern that a head- or hand-worn device might collide with obstacles (P5). Participants who preferred the head-worn position emphasized the advantages of greater freedom of movement when engaged in cooking tasks.

5.7 Discussion and Future Work

In this section, we explore how the contextual needs identified in this study might be integrated into future technology such as AI-powered kitchen assistants.

5.7.1 Importance and Opportunities for AI to Improve Contextual Awareness

Prior research has explored various ways of adopting AI-powered systems to identify objects, such as using computer vision with overhead RGB-D cameras [169] or recognizing objects in mobile device images [15, 135]. While these systems typically detect objects, our study notes the importance of both identifying objects and describing the contextual information of an object. For example, existing systems mostly identify the object name (e.g., milk bottle) or a group name of objects (e.g., vegetables), instead of providing more contextual information to people (e.g., distance to the object, expiration date of food). Based on our findings on what and how contextual information should be presented, there is also an opportunity for these AI-powered systems to provide more precise spatial descriptions that are customized to the user's workspace. Beyond kitchen contexts specifically, we recommend that future research also explore other scenarios in which embedding specific contextual information of objects can support people with vision impairments with agency and autonomy (e.g., art museum [118], grocery store [200], makeup [116]).

5.7.2 Creating Smart Objects in the Kitchen

In our study, we found that visually impaired participants often substituted touch for visual information, and did additional work to make recognition by touch easier, such as using different bottle shapes for different ingredients (Section 5.5.2). If users are already choosing or augmenting the shapes of objects, it might be feasible to expect users to attach tags to objects that could improve recognition [72, 110]. Using more complex tagging methods such as 3D printed models [162] or sensor-enabled tags [173] could provide additional contextual information. This leads to future fabrication research to consider exploring: 1) tagging methods that are sustainable and deformable so that they can be attached to different kitchen objects, 2) platforms that support people with vision impairments to create customized 3D objects for their home, 3) low-cost embedded systems to track and report back about item status.

5.7.3 Augmenting the Kitchen or the User

Beyond customizing contextual information of objects, we also uncovered the importance of managing the space (Section 5.4.1.3) and knowing the internal state of objects (Section 5.4.1.5), which often require them to leverage touch to scan through the space (Section 5.5.1.1). Prior research has explored approaches to **augment the kitchen space**, such as by instrumenting a kitchen with cameras, microphones, and motion tracking [141, 198]. These approaches should also work for people with vision impairments, and may provide even more benefit as they may help address accessibility challenges in the kitchen. Creating and deploying such systems would need to acknowledge that visually impaired users may move and act differently within the kitchen space, and may have particular concerns around issues such as spilling ingredients or tracking cooking status.

Along with augmenting the kitchen space, most of our participants were open to the idea of using **wearable devices** (e.g., body-worn cameras or smartwatches) on themselves to track their activities and provide contextually relevant suggestions (Section 5.6.3). For example, we found that tracking the internal state of the space and objects can be highly visual and it brings

opportunities to create systems to visually check the internal state of objects through user definitions (e.g., dumplings floating indicate doneness). To understand visual information, devices with worn cameras could leverage pre-trained Vision-Language Models (VLMs) [36,52,60,202] to track the status of objects and answer user questions (VQA), such as safety-related questions as well as relative positions and state of objects (e.g., material, quantity) [60]. Such a system could also provide proactive notifications, such as noting if an object has moved or if the space has been rearranged by another user of that kitchen.

5.7.4 Tracking Activities in the Background

Participants noted that they would be interested in knowing the status of the kitchen even when they had moved outside of that space (Section 5.6.2.3). Implicit tracking systems [66,79,89] could discreetly and continuously monitor contextual information, such as the location of objects, while the user is engaged in various activities within the kitchen [109]. This approach would alleviate the need for users to consciously track objects in the kitchen at all times. Future research should consider 1) achieving comprehensive coverage within diverse kitchen spaces, 2) the practicality of addressing challenges related to power consumption and storage capacity [109], 3) the interactive interface for people with vision impairments to pre-assign objects of interest.

5.7.5 Multimodal Interaction in the Kitchen

Participants' activities in the kitchen often leveraged multiple sensory modes at once and in concert (e.g., touch, sound, and smell) (Section 5.5.1). Participants also experienced overload and related challenges during these tasks (Section 5.5.1). This type of multimodal interaction is known in HCI to support users with diverse abilities [24, 121, 124, 143, 151, 154]. As users with vision impairments already interact multimodally in the kitchen, technology that supports these users should also be multimodal and adapted to their existing ways of performing tasks (Section 5.5.1). We noted that participants' abilities to engage the environment were sometimes affected by context, for example, using touch is not always feasible when the user's hands are dirty or greasy, which suggests that future systems should be context-aware and should adapt to the user's activities and current state.

5.8 Summary

This chapter presents a design framework for recognizing and delivering contextual information about kitchen objects to support people with vision impairments during cooking. The framework is based on a study that identified key information needs, strategies for gathering that information, and ways to integrate it into assistive technologies.

The framework consists of three main components:

- **Contextual Information Categories:**
 - Five primary information types are critical for cooking:
 - * *Position Information:* The object's location relative to familiar reference points in the kitchen.

- * *Orientation Information*: How objects are aligned, important for tasks like serving or pouring.
 - * *Proximity and Grouping Information*: Spatial relationships between objects, aiding in kitchen organization and navigation.
 - * *Similarity and Duplicates Information*: Differentiating between similar objects and tracking duplicates to prevent confusion.
 - * *Internal State Information*: Understanding the status of objects (e.g., food doneness, freshness, liquid levels).
- These categories are complemented by secondary needs such as safety, health, and presentation-related information.

- **Strategies for Gathering Contextual Information:**

- Users rely on multi-sensory approaches (touch, sound, and smell) to gather information about objects. For instance, touch helps identify positions, while sound helps monitor cooking progress.
- Organizational strategies, like grouping frequently used objects, streamline information gathering and reduce cognitive load.
- Assistive technologies should replicate or enhance these sensory strategies to provide relevant information efficiently.

- **Guidelines for Communicating Contextual Information:**

- Information should be presented in a non-intrusive way, using audio cues for routine updates and detailed verbal feedback on-demand.
- Clear spatial references based on familiar landmarks improve users' spatial understanding.
- Proactive reminders for safety and health-related information should be customizable and context-aware, adapting to the user's tasks and location.

This design framework emphasizes reducing cognitive load and improving efficiency, allowing AI-powered systems to deliver relevant, timely, and accessible information to support independence and autonomy for visually impaired users in kitchen environments. Having presented design frameworks for both accessing recipes (Chapter 4) and recognizing the contextual information of objects (Chapter 5), it is important to create systems that put these frameworks into practice to support people with vision impairments in cooking, and to evaluate and validate their usefulness.

Chapter 6

OSCAR: Object Status Context Awareness for Recipes

This chapter was adapted from the following two papers:

Franklin Mingzhe Li, Kaitlyn Ng, Bin Zhu, and Patrick Carrington. 2025. Exploring Object Status Recognition for Recipe Progress Tracking in Non-Visual Cooking. *In Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)*. Association for Computing Machinery, New York, NY, USA, 1–15.

Franklin Mingzhe Li, Kaitlyn Ng, Bin Zhu, and Patrick Carrington. 2025. OSCAR: Object Status and Contextual Awareness for Recipes to Support Non-Visual Cooking. *In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–6.

6.1 Overview

Following recipes while cooking is an important yet challenging task for people with vision impairments, particularly due to limited support for tracking progress and receiving contextual feedback [117]. Prior research has shown that blind and low vision (BLV) individuals often rely on embodied strategies—such as spatial memory, tactile exploration, and environmental routines—to navigate the kitchen without vision [110, 117, 190]. These strategies support independence but are often brittle in unfamiliar or complex recipes, where confirming task completion, anticipating next steps, and recovering from disruptions are critical for both usability and safety [86, 100, 111].

While tools like screen readers and smart speakers (e.g., Alexa, Google Home) enable audio-based access to recipes, they operate linearly and lack awareness of what is happening in the physical cooking space [1, 29]. These systems cannot verify whether a step has been completed, adapt to changes in task flow, or reason about what remains to be done. As a result, many BLV users report uncertainty about their current progress, whether they have missed a step, or when it is safe to move on—especially when cooking alone or without pre-memorized routines [110, 117]. Prior work has emphasized the need for cooking technologies that provide context-aware feedback, rather than just access to static instructions [111].

In parallel, computer vision and assistive technologies have enabled object and environment recognition to support non-visual interaction [19, 71, 95]. However, most systems focus on object

presence, not object status—the condition or transformation of ingredients and tools. In cooking, this status is constantly changing: onions get chopped, sauces thicken, meat browns. Recent research shows that such transformations offer rich signals for tracking procedural progress [195], but their potential to support accessibility remains underexplored. To investigate this gap, we ask:

- **RQ1:** How useful is object status information in improving the performance of recipe step prediction using vision-language models?
- **RQ2:** How can object status information support reasoning about the current context of a cooking task, especially in real-world non-visual cooking scenarios?

To address these questions, we developed OSCAR (Object Status Context Awareness for Recipes), a technical pipeline (rather than a deployed assistive system) that explores the use of object status recognition to support recipe progress tracking in non-visual cooking. OSCAR integrates recipe parsing, object status extraction, visual alignment with cooking steps using vision-language models (VLMs) [152, 199], and a time-causal model for structured prediction. Unlike tools that rely solely on recipe text or voice interaction, OSCAR reasons about the real-time cooking context through the evolving visual state of ingredients and tools—enabling both progress tracking and contextual feedback.

We emphasize that our focus is on evaluating the feasibility and performance of this technical pipeline through controlled dataset-based analysis, rather than assessing a complete assistive system in user-facing scenarios. We evaluate OSCAR across two datasets: (1) 173 instructional cooking videos from the YouCook2 dataset [203], and (2) a real-world dataset of 12 non-visual cooking sessions recorded by BLV individuals in their homes. Our results show that object status improves step-prediction accuracy by over 20% across multiple VLMs. We also identify key factors affecting performance in real-world conditions, including implicit tasks, lighting, occlusion, and camera placement. Our work contributes:

- A technical pipeline (OSCAR) for object status-driven recipe progress tracking, combining recipe modeling, vision-language alignment, and time-causal prediction.
- A real-world annotated dataset of 12 cooking sessions by people with vision impairments, reflecting diverse non-visual practices and challenges, which we publicly release to support future accessibility and AI research.
- A technical evaluation demonstrating the feasibility and limitations of object status recognition for progress tracking in both instructional and real-world cooking contexts.
- An analysis of real-world performance factors and design implications for future context-aware assistive cooking technologies.

6.2 Background and Related Work

The general landscape of kitchen technology, of cooking practices and assistive tools for people with vision impairments, and of the large language and vision-language models that OSCAR builds on is reviewed in Chapter 2. Building on that grounding—in particular the finding that

the compensatory strategies blind cooks rely on (tactile organization, spatial memorization, and auditory cues such as a sizzling pan) are effective but brittle, breaking down in unfamiliar environments and on complex, multi-step dishes [110, 117]—this section focuses on the two lines of work most directly relevant to OSCAR: structured recipe modeling with non-visual progress tracking, and dataset creation for accessibility research.

6.2.1 Recipe Structuring and Navigation

Structured recipes are central to guiding cooking tasks and have been widely digitized for applications ranging from personalized meal planning to AI-driven cooking support. Prior research has explored recipe parsing, search, and recommendation through online platforms, voice interfaces, and annotated datasets [7, 114, 130, 139, 175, 178]. More recent efforts have focused on aligning recipe steps with instructional videos to track procedural content, enabling systems to provide step-by-step prompts or generate timelines for cooking actions [31, 47, 62, 119].

Yet most of these systems are built under idealized assumptions—clear visual data, fixed action-object mappings, and known recipes. In real-world cooking, ingredients and tools often transform over time: raw ingredients become chopped, cooked, or mixed, making it difficult for vision-based systems to track progress through static labels alone. Xue et al. [195] argue that recognizing object status changes (e.g., from raw to chopped to sautéed) offers a more accurate reflection of procedural progress than traditional object-verb detection. Despite this insight, few systems explicitly model these transformations, and even fewer consider how they might support non-visual or assistive applications.

For people with vision impairments, recipe access typically relies on screen readers, smart speakers, or Braille displays, which translate static text into audio [5, 22, 156, 163]. While tools like Google Home and Amazon Alexa support basic recipe navigation through commands like “next step,” they lack awareness of user progress or confirmation of completed actions [1, 29]. This limitation becomes problematic when users lose track of their place, face interruptions, or attempt to manage complex, multi-step processes without visual feedback.

Recent work has shown that BLV users often rely on internal strategies such as mental models, tactile routines, and step memorization to stay oriented during cooking [117]. However, these approaches are vulnerable to breakdowns, especially when switching tools, transitioning between stages, or reorienting after a distraction. While static access to recipe content is increasingly supported, real-time, context-sensitive feedback that reflects the procedural nature of cooking remains underexplored. Bridging structured recipe technologies with non-visual progress tracking presents a promising opportunity for more inclusive and supportive cooking systems.

6.2.2 Dataset Creation for Accessibility Research

Recent advances in accessibility research have increasingly embraced the creation and use of specialized datasets to evaluate technical feasibility, benchmark system performance, and train models tailored to the needs of people with disabilities [73]. These datasets are not only vital for developing robust machine-learning systems but also serve as a foundation for understanding the everyday contexts, preferences, and constraints of disabled users.

One influential example is the ORBIT dataset, which provides a real-world benchmark for few-shot object recognition. It was constructed by blind users who captured their own personal items across varied home settings, resulting in a dataset that reflects authentic use cases and user

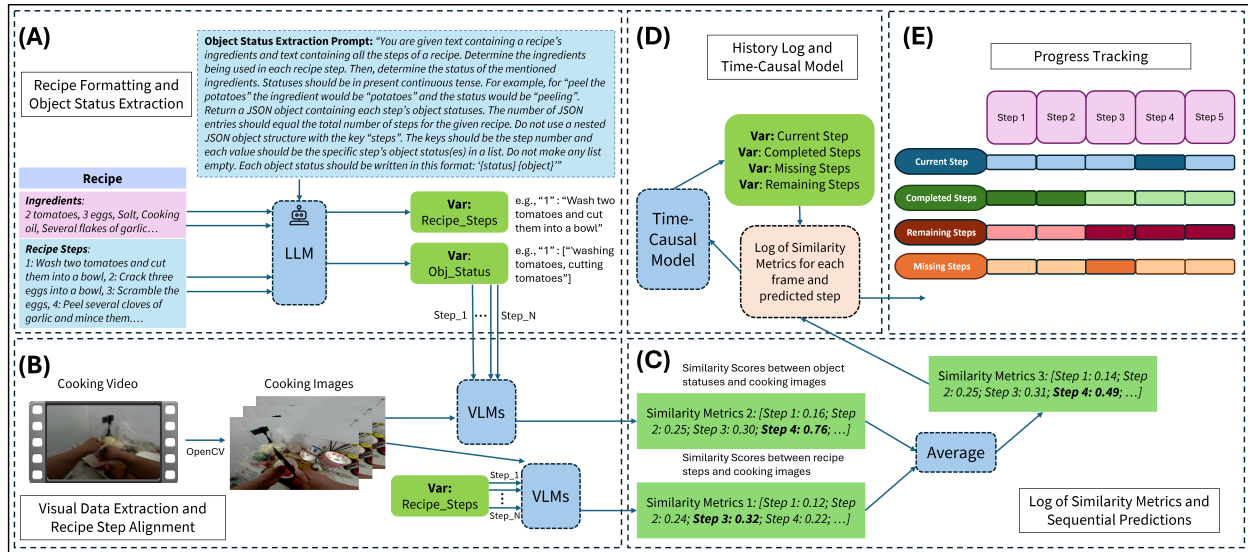


Figure 6.1: Illustration of OSCAR. (A) Recipe Formatting and Object Status Extraction, (B) Visual Data Extraction and Recipe Step Alignment, (C) Log of Similarity Metrics and Sequential Predictions, (D) History Log and Time-Causal Model, (E) Progress Tracking.

environments [131]. Other work, such as that by Theodorou et al., has called for a disability-first approach to dataset design, emphasizing the importance of representing the practices, tools, and goals of disabled individuals rather than retrofitting general-purpose datasets to accessibility tasks [179].

Additional datasets have targeted domains such as assistive navigation, scene description, and input technologies. For instance, SANPO captures real-world pedestrian navigation tasks with visual impairments [189], while the VizWiz dataset focuses on visual question answering using images taken by blind users in natural environments [73]. Others have explored datasets for technology-mediated object search, text entry, and gesture-based interaction [56], reinforcing the need for multimodal, real-world data that reflects disabled users' information needs and interaction contexts.

Collectively, these efforts illustrate a growing recognition that dataset creation is a critical site of access work—not just a technical step, but a design choice that shapes what problems are solvable and for whom. While existing datasets often center on static scenes or discrete recognition tasks, our work builds on this disability-centered trajectory by capturing real-world, user-generated cooking practices that reflect the procedural, tactile, and adaptive nature of non-visual cooking. By documenting these practices and aligning them with structured recipe steps, we contribute a novel resource for evaluating accessible AI systems in the kitchen—one grounded in authentic behaviors rather than idealized laboratory conditions.

6.3 Technical Approach: Object Status Recognition for Recipe Progress Tracking

To explore the feasibility of using object status for recipe progress tracking, we developed OSCAR (Object Status Context Awareness for Recipes)—a technical pipeline that integrates recipe parsing, object status extraction, visual alignment, and progress prediction. While OSCAR is not

a deployed assistive system, it serves as a technical probe to evaluate how object status information can enhance vision-language models for cooking progress tracking. In this section, we describe the design of OSCAR and its components.

6.3.1 Recipe Formatting and Object Status Extraction

From prior research, we learned that recipes contain formatting differences to present recipe steps information [117]. For example, they might indicate recipe step numbers in various ways as ‘Step 1:’, ‘1.’ or ‘Step 1.’ In OSCAR, we first utilized GPT-4o [3] to re-format the recipe steps (Figure 6.1(A)). We then save the recipe steps as a variable inside OSCAR (Figure 6.1(A)).

OSCAR then extracts object status information of a recipe by analyzing the ingredient list in conjunction with the associated cooking actions [195]. The extracted object statuses are represented in a structured format, typically as a combination of [verb] + [noun], such as ‘chopping carrots’ or ‘sautéing mushrooms’ [195]. This format allows OSCAR to clearly define the actions that are performed on each ingredient [195]. For example, a recipe that includes steps such as ‘whisking eggs’ or ‘peeling potatoes’ will have these actions identified and recorded to maintain a clear and actionable understanding of the object’s status. We then saved these object statuses along with the recipe steps that they belong to as a variable inside a JSON file, and a single step can contain multiple object status pairs, such as ‘peeling carrots,’ ‘chopping carrots,’ and ‘storing carrots.’

6.3.2 Visual Data Extraction and Recipe Step Alignment

After formatting recipe steps and extracting object statuses, OSCAR processes visual data by directly using the image input or extracting relevant frames from video clips using OpenCV (Figure 6.1(B)) [27]. Once the visual content has been processed, OSCAR separately aligns recipe steps and object status with the corresponding visual frame (Figure 6.1(B)). This alignment can be achieved through different Vision-Language Models (VLMs). In this chapter, we applied two state-of-the-art VLMs—CLIP (Contrastive Language-Image Pre-Training) [152] and SigLIP (Sigmoid Loss for Language Image Pre-Training) [199]—in OSCAR to show the performance differences and evaluate the approach of using object status to track recipe progress. These models are adept at associating textual descriptions with visual data, allowing OSCAR to compare the similarity scores between visual frames and the object statuses across different recipe steps. For example, when a recipe step involves ‘chopping tomatoes,’ OSCAR uses VLMs to identify the frame that best matches this action by comparing it to other potential matches within the video. OSCAR then predicts the recipe step to which the frame corresponds based on the highest similarity score. This process ensures that visual content is accurately aligned with the progression of the recipe, reducing the chances of errors or confusion during cooking. Through the alignment, OSCAR will generate two lists of similarity scores, one for recipe steps and one for object status. OSCAR then averaged the similarity metrics to evaluate the enhancement of using the object status for the prediction of the recipe step [128].

6.3.3 Log of Similarity Metrics and Sequential Predictions

Based on the similarity metrics generated for the alignment of the recipe steps, OSCAR generates a detailed prediction log for each frame (Figure 6.1(C)). The history log of each image frame is

contained in a JSON file, and each entity includes the similarity scores with all steps, the predicted step info, the predicted step number, the completed steps, any missing steps, and the remaining steps based on previous predictions. The missing step is determined by having a gap between the last completed step and the current step. This log serves as a record of the cooking process, which can be used to generate a progress tracker and support context-sensitive queries (Figure 6.1(C)). For example, a user might ask, ‘What step am I in?’ or ‘Have I already sautéed the mushrooms?’ OSCAR uses the prediction log to provide accurate and context-sensitive responses, guiding the user through the recipe. This feature can be beneficial in complex recipes where multiple steps or actions might overlap or occur simultaneously.

6.3.4 Time-Causal Model

After updating the log of similarity metrics and sequential predictions each time with a new image frame, we applied a time-causal model [145, 150, 170], to improve the temporal coherence of predictions, ensuring that OSCAR’s output aligns with the natural flow of cooking tasks (Figure 6.1(D)). By enforcing sequential logic, the model is designed to reduce false positives and prevent the misprediction of duplicate or out-of-order steps. Specifically, if a user completes Step N in the recipe, the model prevents the prediction of Step N-X (i.e., an earlier step in the sequence) unless that earlier step was never predicted in the first place. This ensures that the system adheres to the expected order of operations and avoids unnecessary confusion or redundant prompts for the user. OSCAR then updates the progress tracking information back to the history log based on the time-causal model. This time-causal approach not only enhances the accuracy of recognizing completed steps but also strengthens the model’s ability to track and predict the appropriate next action, thereby supporting a smoother and more intuitive cooking process for individuals.

6.3.5 Progress Tracking

To support reasoning about recipe state over time, OSCAR maintains a prediction log that records similarity scores between each visual frame and recipe steps. This log includes both raw similarity outputs from vision-language models and derived metadata about likely step completion, ordering, and gaps. Using this information, we apply a time-causal model to enforce temporal coherence—filtering out improbable jumps, duplicates, or regressions in step predictions based on the logical structure of recipes. From this history log, OSCAR computes a real-time snapshot of the user’s position within the recipe. Specifically, it identifies:

- **Current Step:** The most likely step being performed in the current visual frame.
- **Completed Steps:** All steps with high prediction confidence and valid temporal alignment.
- **Remaining Steps:** Steps not yet observed or predicted, and appearing after the current step.
- **Missing Steps:** Steps that were skipped or not recognized in the expected sequence.

This structured tracking enables a time-aware understanding of progress and allows downstream applications to reason about what has happened before, what is happening now, and what comes next. An illustration of this output is shown in Figure 6.1(E).

6.4 Evaluation on Large-Scale Recipe Instruction Videos

To evaluate the feasibility and performance of OSCAR in tracking cooking progress through object status recognition, we conducted a large-scale evaluation using YouCook2, a widely used instructional cooking video dataset annotated with temporal recipe steps. This experiment aims to quantify how incorporating object status and time-causal modeling improves the alignment of recipe steps with visual frames, in comparison to baseline vision-language model (VLM) approaches.

6.4.1 Dataset Description

6.4.1.1 YouCook2 Dataset.

To validate OSCAR’s ability to use object statuses to predict recipe steps in visual frames, we used the YouCook2 dataset [203]. The YouCook2 dataset is a comprehensive collection of cooking videos, totaling 176 hours, with each video averaging 5.27 minutes long [203]. All the videos are untrimmed, and the length can be up to 10 minutes; they were recorded using handheld or stationary cameras [203]. Each video segment is temporally localized with start and end times for each procedural step. The dataset includes diverse metadata, such as video duration, the number of recipe steps per video, segment duration, and the number of words per sentence [203]. This diversity in content and metadata makes YouCook2 an ideal dataset to assess OSCAR’s ability to handle varied and complex cooking scenarios [203]. The dataset is publicly available for download at (<http://youcook2.eecs.umich.edu/>).

6.4.1.2 Dataset Processing and Curation.

The original YouCook2 dataset did not include ingredient information in its metadata [203]. To address this, we manually reviewed each cooking video to extract ingredient details, sourcing information directly from the video content or the publisher’s notes on YouTube. In addition, we filtered out videos that were irrelevant to the cooking process, such as those featuring chef discussions or promotional content that is not related to cooking. After this curation process, we compiled a refined dataset consisting of 173 YouTube videos (Y1–Y173) with comprehensive ingredient lists and step-by-step instructions. These videos have an average step number of 7.7, which matches the average step of the original dataset, with a minimum of 5 recipe steps and a maximum of 15.

6.4.2 Methodology

6.4.2.1 Baseline Prediction.

Each video is annotated with timestamps of the start and end of each recipe step. For each step, we divide the video clip belonging to that step into five equal segments and randomly select a frame from each segment [182]. We use this approach to reduce the impact of noise in the data. We applied a blur filter to choose the frame with the least blur adjacent to the selected frame in each segment [27]. To evaluate and understand the generalizability of how object status information can enhance the performance of VLMs, we used two state-of-the-art VLMs in the evaluation, CLIP [152] and SigLIP [199]. As baseline methods, we then calculated the similarity score between each frame and all the recipe steps; each frame would have a list of similarity

VLM Model	Baseline Accuracy	Baseline Standard Deviation	OSCAR Accuracy	OSCAR Standard Deviation	Δ Accuracy (OSCAR – Baseline)
CLIP	41.7%	17.5%	68.0%	19.0%	26.3%
SigLIP	62.2%	18.0%	82.8%	14.7%	20.6%

Table 6.1: Evaluation results with recipe instruction videos.

scores [204]. Each time, one frame from each segment is processed, which includes five at a time for the step. The average of these similarity scores was used to determine the predicted step with the highest overall similarity [130, 204]. If the predicted step aligns with the ground truth, we will mark 100% for this step; otherwise, we will mark 0%. This process was repeated three times per step to ensure consistency in the results, and we used the average accuracy of the three times as the final prediction accuracy for each step [130, 204]. We then repeat the same process for the rest of the recipe steps, as well as all other recipe instruction videos. Repeating the prediction process three times per step helps mitigate variability across frames within dynamic cooking scenes, where key objects may be temporarily obscured by motion or occlusion. Averaging the results across these repetitions yields a more stable accuracy estimate by reducing the influence of transient visual noise such as blur or obstruction [161].

6.4.2.2 OSCAR with Object Status and Time-Causal Modeling.

To evaluate the usefulness of object statuses, the status information of the object was extracted from the ingredient list and the corresponding cooking steps (Section 6.3.1). Using the same video frames selected in the baseline method, we further calculated the similarity score between the frames and the object statuses using both CLIP [152] and SigLIP [199] (Section 6.3.2). We then utilized the time-causal model to further enhance the performance (Section 6.3.4). The process of averaging the similarity scores within the five frames per step and repetition three times is the same as the baseline method. The results were then averaged with the baseline similarity scores to obtain the final prediction accuracy to understand the usefulness of the object statuses compared to only using the recipe step text as a prediction [128].

6.4.3 Results

6.4.3.1 Baseline Performance.

Among 173 annotated instructional videos (Y1–Y173), the baseline results revealed an average accuracy of 41.7% (SD = 17.5%) for recipe step predictions using CLIP (Table 6.1). SigLIP, on the other hand, achieved an average accuracy of 62.2% (SD = 18.0%) (Table 6.1). These results highlight the challenges of accurately predicting the visual data with recipe steps individually, particularly in cases where the visual content is ambiguous or lacks clear cues for specific actions.

6.4.3.2 OSCAR Performance.

By incorporating object status information and utilizing a time-causal algorithm, OSCAR outperforms the baselines with a large margin in terms of prediction accuracy. Specifically, the average



Figure 6.2: Image frame of Y19 that shows a top-down view with all ingredients of making the burrito that impact the prediction accuracy through object status.

accuracy for CLIP increased from 41.7% of the baseline to 68.0% (SD = 19.0%), while SigLIP’s performance rose from 62.2% to 82.8% (SD = 14.7%) (Table 6.1). These improvements demonstrate the effectiveness of using object status to enhance the ability to accurately track and predict recipe steps, especially in complex cooking scenarios.

6.4.4 Why Object Status Improves Step Prediction in Large-scale Recipe Instruction Videos

6.4.4.1 Disambiguating Incomplete or Occluded Visual Scenes.

By analyzing OSCAR’s performance improvements compared to the baseline with 26.3% for CLIP and 20.6% for SigLIP (Table 6.1), we found that the performance where OSCAR outperformed was because recipe text may reference objects or tools not directly visible in the frame. For example, some recipes instruct the use of specific tools (e.g., spatula, spoon), but these tools can be difficult to detect when obscured by other objects (e.g., soup) or when the user’s hand position occludes the tool’s shape. With the use of object states, visual information can complement missing details that are not directly visible with the recipe context (e.g., diced beef in the soup). As a result, this can improve the accuracy of aligning visual frames with the corresponding recipe steps.

6.4.4.2 Differentiating Visually Similar or Repetitive Actions.

Additionally, we found that implementing a time-causal model also improved performance, since some cooking steps involve similar or duplicate actions or objects at different stages of the recipe. For instance, in a stir-fry recipe for steak and vegetables, minced garlic is first added to the pan along with the steak. After a few minutes, the cook sets aside the partially cooked steak, adds more minced garlic to cook the vegetables, and then returns the steak to the pan once the vegetables are nearly done. This method ensures the steak is not overcooked since vegetables take longer to cook. In this scenario, minced garlic is added twice, which appears visually identical. However, by incorporating a time-causal algorithm, we can distinguish between the two actions, even though they involve duplicate steps.

6.4.4.3 Disambiguating Visually Cluttered Frames with Overlapping Ingredients.

In visually cluttered scenes—such as top-down views where all ingredients appear simultaneously—baseline models often struggle to determine the correct recipe step. In Y19, where rice, refried beans, salsa, sour cream, and guacamole were visible at once (Figure 6.2), SigLIP achieved only 25% accuracy, misclassifying most steps due to overlapping visual cues. OSCAR significantly improved performance by incorporating object status. For example, in step 3, baseline similarity scores were too close across steps (e.g., 0.734, 0.064, 0.612, 0.575), leading to misalignment. In contrast, OSCAR’s object status-based scores created clearer distinctions (0.381, 0.050, 0.575, 0.288) (Figure 6.3), enabling more accurate prediction. Even when object status alone was not enough to resolve repeated steps—such as “repeat the steps for each tortilla and serve”—adding the time-causal model corrected the prediction. This case illustrates how object status provides critical disambiguating cues in cluttered frames, and how temporal modeling helps resolve step repetition. The analysis demonstrates that both object status and time-causal model enhance the prediction performance for both CLIP and SigLIP models significantly with more than 20% enhancement.

6.4.5 Factors Affecting Prediction Performance

While OSCAR significantly outperforms baseline models, our analysis revealed several factors that limit the effectiveness of vision-language models (VLMs) when aligning video frames with recipe steps. We conducted a detailed error analysis to identify common challenges, comparing object status descriptions against visual content across underperforming videos.

6.4.5.1 Exocentric vs. Egocentric View.

From the evaluation, we observed that the performance of our model using both CLIP and SigLIP dropped for some videos because certain frames in the cooking videos captured an exocentric view, which presented a broad, external perspective of the kitchen environment. For example, in Y28 (Figure 6.4 left), the performance of OSCAR with CLIP was only 22.2% compared to the average of 68.0%. In these frames, all ingredients were displayed on the kitchen table rather than focusing closely on the specific cooking actions being performed. This wide-angle view included not only the ingredients but also the surrounding kitchen space, which made it challenging for VLMs to accurately associate these frames with the correct steps in the recipe. As a result, there

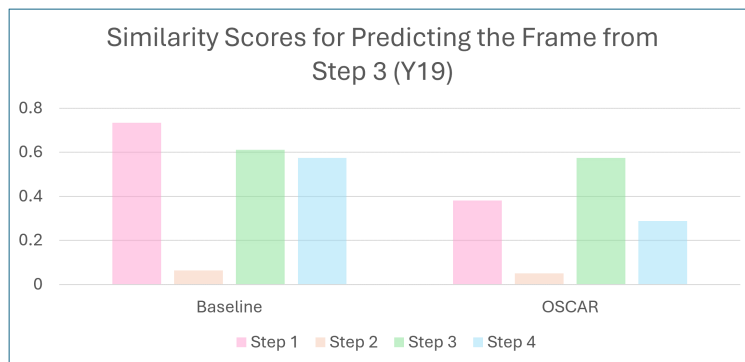


Figure 6.3: Similarity scores for predicting the frame from step 3 (Y19); it shows step 3 is correctly predicted by OSCAR, but not the baseline.



Figure 6.4: Image frames of cooking video that shows an exocentric view of the cooking steps. Left: This figure shows one frame in Y28, in which the exocentric view fully shows the ingredients clearly. Right: This figure shows one frame in Y33 that the person is trying to cook food in a slow cooker, and a side view cannot show what is being cooked.

was an increase in false-positive rates, where these frames were mistakenly linked to different recipe steps.

Besides creating false positives from all ingredients, the exocentric views often resulted in cooking tools, such as pots, pans, and utensils, partially or completely obscuring the ingredients being cooked. For example, Y33 shows a performance of 55.6% with OSCAR and SigLIP while the average was 82.8%. This obstruction made it even more difficult for VLMs to correctly identify the status of the ingredients and the corresponding cooking actions. The lack of a clear view of what was happening in the pots or pans due to these obstructions compounded the challenge of accurately predicting the sequence of cooking steps, leading to a decrease in model performance (Figure 6.4, right).

6.4.5.2 Subtle Visual Changes Across Steps.

We found that the type of dish being prepared had an impact on the performance, particularly when there were only subtle visual differences between steps. For certain types of dishes, the lack of clear and distinguishable visual cues made it difficult for VLMs to identify and track the status of objects accurately. For example, in dishes such as soups (Y5), which typically involve mixing or simmering ingredients together in a pot, the changes between steps can be very gradual and subtle (Figure 6.5). Its baseline performance with SigLIP was 54.2%, which is below the average. As ingredients are mixed and cooked, the appearance of the contents in the pot often becomes homogeneous, making it difficult for VLMs to discern individual ingredients or specific object statuses (Figure 6.5). The lack of distinct, observable changes in the visual state of the ingredients



Figure 6.5: Image frames of cooking video that show similar visual frames of different steps that caused misprediction.



Figure 6.6: Image frame of cooking video that shows misprediction because of the descriptions against the norm. This figure shows the recipe of making Korean watermelon rind kimchi.

leads to fewer visual cues that VLMs can use to accurately align the cooking steps with the recipe instructions (Figure 6.5).

6.4.5.3 Uncommon Object States or Descriptions.

Our analysis reveals that existing vision-language models are often trained with generic descriptions and standard norms, which limits their ability to accurately identify and interpret objects and their states in real-world cooking scenarios. For instance, in a Korean Watermelon Rind Kimchi recipe video (Y58), the CLIP model demonstrated limitations (Figure 6.6). It only achieved a 20% accuracy rate through OSCAR with CLIP when identifying the watermelon rind as an ingredient, which is starkly lower compared to its average accuracy of 68% across other, more commonly recognized ingredients. This discrepancy is likely due to the model’s difficulty in recognizing the watermelon rind, which is white and visually distinct from the green watermelon skin typically associated with watermelons in most training datasets (Figure 6.6). Instead of identifying the rind based on its unique characteristics and contextual use in the recipe, the model’s training on standard images led it to misidentify or overlook it. These models typically rely on large datasets that contain images of common objects and ingredients in their most typical forms [138], focusing on standard visual characteristics. As a result, when these models are confronted with less common object states or atypical features—such as variations in color, texture, or shape—they tend to perform poorly.

6.4.5.4 Low-Quality Video and Cooking Artifacts.

The video dataset of YouCook2 included cooking recipe videos uploaded in a large span of time-frames [203]. In our evaluation, we found that low-resolution data can impact the performance of predicting object statuses through vision-language models. For example, Figure 6.7 presents one of the frames for making BBQ pork fried rice (Y48) and the video is an older recording presented in 480P. In the evaluation, OSCAR with SigLIP achieved only 42.9% while the average was 82.8%. Furthermore, the visual impact of cooking can affect the performance of tracking recipe steps (e.g., hand occlusion, cooking steam (Figure 6.7)). Figure 6.7 showed the process of making BBQ pork fried rice, and it generates water vapor when the cook is stir-frying the rice, which causes

the camera lens to become blurry. This results in a loss of visual clarity and detail, making it difficult to accurately align the cooking video with the recipe steps. The steam obscures important visual cues and object statuses, such as the texture and color changes of the ingredients, which are crucial for understanding the cooking process and ensuring the recipe is followed correctly.

6.5 Evaluation with Real-World Non-Visual Cooking Dataset

To assess OSCAR’s effectiveness beyond curated online videos, we further evaluated its performance in real-world cooking scenarios by people with vision impairments. Unlike sighted cooking practices, non-visual cooking often involves unique behaviors—such as tactile exploration, prolonged object localization, or personalized tool use—that introduce new challenges for vision-based step tracking [110]. This evaluation aims to examine not only OSCAR’s prediction accuracy in natural settings but also to surface design considerations for future assistive systems that support non-visual cooking practices.

6.5.1 Dataset Overview

To provide the benchmark for evaluating different VLMs for non-visual cooking, we collected a new dataset of 12 cooking videos (V1–V12) featuring people with vision impairments (6 male, 6 female, average age of 42.5) preparing meals in their own kitchens (Figure 6.8). Four of them are legally blind and eight are totally blind. Each participant cooked independently, using their preferred techniques and tools to make one everyday dish. Videos ranged from 14 to 51 minutes, with an average of 8.1 recipe steps per video. Videos were recorded using a GoPro 11 chest-mounted camera in 5.3K resolution for high visual fidelity and later downsampled to 1080P for evaluation. Following the annotation protocol used in YouCook2 [203], we extracted ingredients and recipe steps from the egocentric video, then annotated the start and end timestamps for each step to support precise step alignment to evaluate OSCAR. For example, one video details washing, chopping, and assembling a salad. Participants were compensated with \$40 USD for completing the cooking session. The data collection was approved by our Institutional Re-



Figure 6.7: Image frames of cooking video that shows cooking BBQ pork fried rice and the cooking steam blurred the camera’s view.



Figure 6.8: Thumbnail of the non-visual cooking dataset of 12 videos by people with vision impairments.

view Board (IRB). The dataset is available at https://huggingface.co/datasets/limingz3/Non-visual_Cooking_Video_Dataset.

6.5.2 Methodology

Our evaluation followed the same experimental pipeline as the YouCook2 evaluation. We processed each video to extract object status information from ingredients and recipe steps and compared OSCAR’s performance to baseline VLM predictions.

6.5.2.1 Baseline Prediction.

For each annotated recipe step, we divided the corresponding video segment into five equal parts and sampled one frame per segment [182]. A blur filter was applied to select the clearest frame [27]. Using CLIP [152] and SigLIP [199] as baseline models, we computed the similarity between each sampled frame and all recipe steps. The highest average similarity score across the five frames determined the predicted step. Predictions were repeated three times per step for consistency, and the final accuracy was calculated by averaging the results.

6.5.2.2 OSCAR with Object Status and Time-Causal Modeling.

To evaluate the contribution of object status information, we incorporated extracted object states (Section 6.3.1) and recalculated frame-step similarity using the same video frames as the baseline. We further applied OSCAR’s time-causal model (Section 6.3.4) to refine predictions, especially for repetitive or ambiguous actions. The final accuracy for each step was obtained by averaging the baseline similarity scores with the object status-enhanced scores, maintaining the same evaluation protocol.

VLM Model	Baseline Accuracy	Baseline Standard Deviation	OSCAR Accuracy	OSCAR Standard Deviation	Δ Accuracy (OSCAR – Baseline)
CLIP	33.7%	21.8%	58.4%	16.5%	24.7%
SigLIP	41.9%	14.0%	66.7%	18.7%	24.8%

Table 6.2: Evaluation results with non-visual cooking dataset.

6.5.3 Results: Evaluating Step Prediction in Real-World Non-Visual Cooking

6.5.3.1 Baseline Performance.

Baseline results revealed substantial limitations in using vision-language models (VLMs) alone to predict cooking progress in non-visual cooking contexts. On average, CLIP achieved an accuracy of only 33.7% (SD = 21.8%), while SigLIP reached 41.9% (SD = 14.0%) across the 12 annotated cooking videos (Table 6.2). These figures are noticeably lower than performance on the YouCook2 dataset, emphasizing the complexity and unpredictability of real-world environments where people with vision impairments cook in ways that diverge from the assumptions encoded in conventional training datasets.

These low scores reflect multiple compounding factors: longer durations spent locating tools, non-standard tool usage, repetitive motions such as tactile verification, and variable framing and lighting due to user-generated chest-mounted camera views. Together, these introduce significant ambiguity in visual input that makes it difficult for general-purpose VLMs to determine what cooking step is currently being performed.

6.5.3.2 OSCAR Performance and Improvement.

Incorporating object status representations and a time-causal prediction model, OSCAR significantly outperformed the baseline models. With OSCAR, CLIP’s average accuracy improved from 33.7% to 58.4% (SD = 16.5%), and SigLIP’s from 41.9% to 66.7% (SD = 18.7%) (Table 6.2). This represents a performance boost of +24.7% for CLIP and +24.8% for SigLIP. These gains were consistent across all 12 videos, and in several cases, such as video V12, OSCAR achieved a perfect 100% prediction accuracy—60 percentage points above the baseline.

These results not only reinforce the feasibility of using OSCAR for procedural tracking in real-world cooking tasks but also highlight its strength in adapting to user-specific routines and unconventional kitchen setups. Whereas baseline models often misinterpret lengthy or non-standard interactions, OSCAR’s attention to the current state of objects and the temporal history of actions enables it to maintain alignment with the true recipe step.

6.5.4 Why Object Status Improves Step Prediction in Non-Visual Cooking

Our analysis of OSCAR’s performance improvements—achieving a +24.7% gain for CLIP and +24.8% for SigLIP (Table 6.2)—highlights the critical role of object status recognition in supporting cooking progress tracking, particularly in real-world non-visual cooking scenarios. Unlike



Figure 6.9: These three frames were captured for V4 during step 1: ‘Crack an egg and scramble it.’ The middle frame showed the data where the blind cook was looking for a garbage bin to throw the eggshell, which got captured and impacted the performance of predicting the step.

controlled video datasets, the non-visual cooking dataset reveals rich variations in cooking behaviors, tool use, and object interaction patterns by people with vision impairments. We identify several core reasons why object status improves prediction robustness and system adaptability.

6.5.4.1 Reducing False Positives from Prolonged and Exploratory Interactions.

In non-visual cooking, people with vision impairments often engage in longer interactions with objects—feeling, exploring, or verifying their location and state through touch. These extended or repeated actions can lead to false positives when vision-language models rely only on recipe steps to infer progress, as the system may mistakenly interpret prolonged handling as task completion. By incorporating object status recognition, OSCAR provides an intermediate representation that filters out incidental or exploratory interactions unless a change in object state is detected (e.g., uncut vs. chopped tomatoes). This mechanism helps disambiguate real progress from slow or repeated exploration, improving step-prediction accuracy in environments with non-linear and variable cooking workflows.

6.5.4.2 Accommodating Personalized Tools and Cooking Strategies.

People with vision impairments frequently use personalized or adapted tools that differ from standard recipe expectations [110, 117]. These substitutions—such as using a butter knife instead of a spatula—often lead to mismatches between the tools referenced in the recipe text and those visible in the video frame. Recipe-step-only models struggle in these scenarios because they anchor predictions on canonical tool appearances. In contrast, object status recognition focuses on the transformation of ingredients (e.g., spread beans, washed cucumber) rather than the tool performing the action. This design flexibility enables OSCAR to generalize across different user behaviors and tool choices, supporting accessibility without requiring users to conform to pre-defined cooking procedures.

6.5.4.3 Supporting Inclusive and Autonomy-preserving Design.

Tracking cooking progress through object status fosters a more inclusive interaction paradigm—one that adapts to user routines rather than forcing behavior changes. People with vision impairments often develop efficient but idiosyncratic cooking strategies shaped by their skills, preferences, and environmental constraints. OSCAR’s ability to align progress tracking with these individualized practices—without penalizing tool substitution, workspace variation, or non-standard sequencing—preserves user autonomy and supports natural cooking workflows. This is critical

for designing assistive systems that respect user expertise rather than enforcing rigid procedural structures.

6.5.4.4 Consistency Across Real-world and Instructional Datasets.

The performance gains of OSCAR were consistent across both the real-world non-visual cooking dataset and the large-scale YouCook2 instructional dataset. This consistency highlights that object status modeling is a transferable and generalizable approach, benefiting both controlled and naturally occurring cooking scenarios. This finding underscores the importance of moving beyond recipe text alignment alone. While recipe steps provide useful guidance, the variability in real-world cooking—especially among people with vision impairments—demands systems that can reason about the physical state of the environment. Object status, combined with time-causal modeling, provides a scalable foundation for robust cooking progress prediction in diverse settings.

6.5.5 Factors Affecting Prediction Performance and Design Implications

While OSCAR demonstrated substantial improvements over baseline models, our evaluation on the real-world non-visual cooking dataset revealed several unique factors that challenged the performance of vision-language models (VLMs) in non-visual cooking settings. These findings offer critical insights for the design of future vision-based assistive systems.

6.5.5.1 Implicit and Preparatory Tasks Beyond Recipe Instructions.

In real-world cooking, people with vision impairments often perform implicit or preparatory tasks that are not explicitly described in recipe steps—such as cleaning, organizing tools, or discarding waste. These actions are essential for maintaining safety, hygiene, and workflow readiness, yet they create gaps between recipe expectations and real-world behavior. For example, in V4, the participant paused during the step “crack an egg and scramble it” to search for a garbage bin and dispose of the eggshell (Figure 6.9). Since this action was not annotated as part of the recipe step, OSCAR mispredicted the progress as 0%. These moments introduce ambiguity in vision-based step tracking, particularly when object interaction occurs without a corresponding transformation in object status.

Design Consideration #1: The presence of frequent implicit and preparatory tasks in real-world cooking—such as cleaning, organizing tools, or discarding waste—suggests that assistive systems should be designed to recognize and accommodate these actions rather than treat them as noise. Distinguishing between core procedural steps and incidental yet essential behaviors can help models interpret natural deviations from scripted instructions without misclassifying user intent. Such consideration is critical for ensuring that systems remain aligned with the practical, adaptive workflows common in non-visual cooking.

6.5.5.2 Rechecking and Securing Tools and Ingredients.

A notable pattern in the non-visual cooking dataset was participants’ frequent rechecking of ingredients and tools—a behavior deeply rooted in strategies for ensuring safety, spatial awareness, and task confidence. Unlike sighted cooking routines, where visual scanning often suffices, people with vision impairments rely on tactile feedback to confirm object identity, location, and



Figure 6.10: Illustration of objects not captured in the center of the field of view. Left: V8 is showing ‘Cut the cucumber into pieces and store them in a bowl.’ Right: V6 is showing ‘Wash the enoki mushroom, carrot, and cucumber. Shred the carrot and cucumber.’

readiness. This leads to a recurring process of touching, repositioning, or verifying objects before proceeding. For example, in video V4, after the participant chopped tomatoes and placed them in a bowl, they revisited the bowl multiple times—gently touching the tomatoes and confirming their position on the countertop—before continuing to the next step. These interactions were not part of the functional transformation of the tomatoes, yet from the model’s perspective, the repeated contact resembled the type of object manipulation seen in active steps.

These exploratory or confirming actions often occurred not only immediately after processing ingredients but also as part of users’ adaptive strategies to stay prepared and oriented for future steps. For instance, some participants proactively ensured that utensils like spoons or knives were in the expected place before they were needed. While such behaviors helped users maintain workflow continuity, vision-language models frequently misclassified these moments as step transitions, resulting in inflated or inaccurate step completion predictions.

Design Consideration #2: The frequent use of exploratory and safety-checking behaviors—such as rechecking object positions or preparing tools in advance—indicates that assistive systems should distinguish these non-transformative actions from actual step progression. When users interact with objects without changing their state, such as lightly touching or repositioning already-prepared ingredients, these moments should not be treated as evidence of step completion. Instead, systems need to integrate object status recognition with temporal reasoning and patterns of interaction that reflect common non-visual strategies. This approach can help reduce false positives in step tracking, improve alignment with user intent, and make the system more resilient to the natural variability in non-visual cooking practices.

6.5.5.3 Lighting Conditions and Visibility Challenges.

Non-visual cooking environments often involve variable and uncontrolled lighting, which can significantly degrade vision model performance. For instance, in V6, suboptimal lighting reduced OSCAR’s accuracy to 40%, well below the 66.7% average. Two lighting issues emerged frequently: (1) strong backlighting caused models to fixate on irrelevant background features, and (2) dim environments led to poor contrast, making ingredients hard to distinguish from shadows. These conditions made it difficult for VLMs to detect object transitions or ingredient changes, undermining step prediction.

Design Consideration #3: The frequent impact of poor or uneven lighting on model performance highlights the need for systems to account for real-world visibility constraints. Inconsistent lighting—such as strong backlighting or dim conditions—can obscure ingredient changes

and lead to misclassification. These findings point to the importance of incorporating lighting-aware feedback mechanisms and designing for robustness under suboptimal visual conditions. Strategies such as real-time lighting quality checks, user guidance for repositioning light sources or the camera, and multimodal sensing (e.g., audio or haptics) may help ensure more consistent and reliable step tracking in varied kitchen environments.

6.5.5.4 Impact of Camera Angles and Field of View.

Inconsistent camera angles and limited field of view significantly impaired prediction accuracy in real-world non-visual cooking videos. In several cases, the camera failed to center on the active work area, leading to partial or missing views of ingredients and tools. For example, in V6, the camera was angled too high during the step “Wash the enoki mushroom, carrot, and cucumber. Shred the carrot and cucumber”, resulting in only a partial view of the cutting board (Figure 6.10). This contributed to V6 showing the lowest accuracy in the dataset: 40% with OSCAR and just 16.7% with the baseline model. For users with vision impairments, adjusting the camera angle or position during cooking is often impractical, especially when hands are occupied. Misaligned views can distort objects or exclude them entirely, reducing the model’s ability to detect status changes and step transitions.

Design Consideration #4: The frequent occurrence of misaligned or incomplete camera views underscores the need to design for hands-free, robust framing support. When the field of view fails to capture the active workspace, vision models struggle to detect object status changes and task progression, leading to substantial drops in prediction accuracy. These challenges highlight the importance of incorporating mechanisms that either assist users in maintaining appropriate framing or compensate for imperfect views—such as real-time framing feedback, multi-camera setups, or adaptive cropping techniques. Such support is especially critical in non-visual contexts, where users cannot easily verify what the camera is seeing during dynamic, hands-busy tasks like cooking.

6.5.5.5 Challenges with Recognizing Pre-prepared Ingredients.

People with vision impairments often adopt efficient cooking strategies, such as preparing and storing ingredients ahead of time in separate containers. While this helps streamline their workflow and improve task flow, it introduces ambiguity for vision-based systems like OSCAR, which



Figure 6.11: Illustration of multiple objects stored before cooking.

rely on visual cues to align recipe steps with observed actions. For instance, in V4, the cook followed two steps in sequence: “Cut the tomatoes into pieces and store them in a bowl” and “Take a stalk of green onion and cut it into pieces”. However, during the second step, the bowl of pre-cut tomatoes remained visible in the frame (Figure 6.11). This led OSCAR to misclassify the current step as a continuation of the earlier one, as the presence of tomatoes was interpreted as an ongoing action rather than a previously completed one.

This overlap between past and current ingredients illustrates the need for more sophisticated context modeling. Simply detecting object presence or status is insufficient—models must also reason about temporal sequencing and distinguish between active manipulation and background presence.

Design Consideration #5: The presence of pre-prepared ingredients in the visual frame—even when they are no longer being manipulated—reveals a critical gap in how current systems interpret context. Users’ strategies, such as batch-preparing ingredients and keeping them nearby for later use, result in scenes where previously completed steps remain visually present. This makes it difficult for models to distinguish between ongoing actions and residual context. These observations point to the need for assistive systems to move beyond static object detection and incorporate reasoning about temporal context and interaction history. By understanding when ingredients have been processed and set aside, systems can more accurately interpret which actions are in progress, improving the fidelity of step alignment in real-world cooking scenarios.

6.5.5.6 Summary.

Taken together, these findings highlight the unique complexities of supporting real-world non-visual cooking through vision-language models. Unlike instructional videos, non-visual cooking involves adaptive behaviors—such as exploratory touching, pre-preparation, and implicit tasks—that do not always align neatly with recipe instructions or visual cues. Performance was also sensitive to environmental variables like lighting, camera positioning, and field of view, which are difficult for users with vision impairments to control. These insights underscore the importance of designing assistive systems that go beyond frame-by-frame recognition: systems must model temporal context, support user autonomy, and adapt to the idiosyncrasies of real-world practice. By integrating object status tracking with reasoning about intent, memory, and environment, future systems can better align with the actual workflows and needs of blind and visually impaired cooks.

6.6 Discussion

This work presents OSCAR, a technical pipeline that leverages object status recognition to support progress tracking in non-visual cooking. Our findings offer broader implications for assistive AI—particularly around modeling real-world variability, supporting user autonomy, and designing systems that adapt to the embodied and procedural nature of non-visual tasks.

6.6.1 Rethinking Progress Tracking in Real-World Non-Visual Activities

Our findings challenge the dominant paradigm of recipe progress tracking that assumes clear task boundaries, discrete action-object mappings, and linear step completion. While these as-

sumptions hold in curated instructional videos (e.g., YouCook2), they break down in real-world non-visual cooking, where people with vision impairments often engage in exploratory, repetitive, and non-linear behaviors. For example, participants frequently performed extended tactile searching, rechecking ingredient locations, or organizing tools for future use, none of which directly transformed object status but were essential for task success and safety. These behaviors led to consistent mispredictions in baseline models that relied solely on aligning static recipe steps to video frames (Table 6.2). This finding echoes prior accessibility research that emphasizes the non-linear, adaptive nature of non-visual activities [94, 110, 116, 117, 118], where progress is often embodied, situational, and opportunistic rather than strictly sequential.

To support real-world assistive applications, future systems must move beyond rigid step-alignment and instead adopt dynamic progress inference models that accommodate the fluidity of non-visual workflows. This requires combining object status recognition with temporal reasoning (Section 6.3.4) and possibly integrating interaction modeling (e.g., recognizing exploratory touch vs. transformative actions). Such an approach would allow systems to reason about “task readiness” or “step proximity” rather than enforcing binary step completion. Ultimately, rethinking progress tracking in this way aligns with inclusive design principles that respect user autonomy, adapt to diverse routines, and support the messy realities of everyday non-visual activities.

6.6.2 Vision-Language Models in Messy Real-World Conditions

Our evaluation revealed a stark contrast between OSCAR’s performance on curated instructional data (YouCook2) and its performance in real-world non-visual cooking settings. While vision-language models (VLMs) like CLIP and SigLIP achieved recipe step-prediction accuracies above 60% in the controlled, well-lit videos of YouCook2 (Table 6.1), their baseline performance dropped significantly—to 33.7% and 41.9% respectively—in user-generated videos recorded by blind and low vision (BLV) individuals in their own kitchens (Table 6.2).

This performance gap exposes key fragilities in current VLMs when deployed outside of clean, curated datasets. In our real-world dataset, videos featured challenging visual conditions: inconsistent or dim lighting, cluttered backgrounds, and occluded ingredients or tools due to variable camera angles. More critically, BLV users engaged in adaptive behaviors—such as tactile searching, using alternative tools, or pre-preparing ingredients—that led to object states or visual frames not typically encountered in training corpora. As a result, models often misclassified steps or failed to recognize when progress had occurred.

These findings underscore a core challenge in accessibility AI: models trained on polished, idealized data rarely generalize to the messy, situated realities of lived experience. As Theodorou et al. argue [179], disability-centered technologies must be evaluated and optimized for the contexts in which they are used—not just the environments in which they are easiest to test. To build truly inclusive and dependable systems, VLM research must treat robustness under user-authored, imperfect, and unpredictable conditions as a core design goal—not a downstream deployment concern. Models must be built with the variability of real-world usage in mind, especially when intended to support autonomy in daily life.

6.6.3 Object Status as a Universal Design Primitive

Our findings highlight object status recognition as a transferable design primitive for assistive technologies supporting non-visual activities. In our non-visual cooking evaluation, participants

frequently substituted tools—using fingers to spread beans (V3) or butter knives to lift food (V5)—which led to baseline model errors (Table 6.2) that assumed fixed action-tool mappings and canonical tool usage. These errors expose a limitation in systems that treat progress as a function of specific object-action pairs rather than the outcome of a transformation.

In contrast, recognizing the evolving state of objects—such as ingredients becoming chopped, bowls becoming filled, or pans becoming empty—enabled OSCAR to track progress independent of the tools or gestures used. By grounding progress in the status of objects, rather than how the status was achieved, OSCAR offers greater flexibility and better accommodates diverse user routines. This abstraction aligns with inclusive design principles [110, 117], supporting user autonomy, customization, and adaptation—critical values in accessibility technology.

Importantly, object status recognition is not limited to cooking. Tasks like makeup application, cleaning, crafting, or home repair similarly involve the transformation of materials and tools through personalized strategies [83, 112, 115, 116]. In these domains, users may mix techniques, skip steps, or use improvised tools, yet the meaningful unit of progress remains the changing state of the task environment. Designing assistive systems around object status—rather than predefined actions or tools—offers a more resilient and generalizable foundation for context-aware support across varied non-visual, hands-on tasks.

6.6.4 Modeling Context, Memory, and Temporal Continuity

While our real-world evaluation exposed limitations of VLMs in non-visual cooking, these issues stem not only from data quality but from fundamental modeling constraints. Current VLMs rely heavily on per-frame classification, lacking mechanisms to reason about what has already occurred or how scenes evolve over time—a shortcoming echoed in recent work on temporal grounding and task tracking [195]. For example, in V4, previously prepared ingredients (e.g., chopped tomatoes) remained in view during later steps, causing baseline models to misclassify the user’s current action. Without temporal memory or an understanding of scene history, these systems cannot distinguish between residual context and active manipulation. To better support non-visual, multi-step tasks with high ambiguity, future systems must move beyond isolated image-text alignment. Techniques such as short-term memory buffers, lightweight scene graphs, or temporal event segmentation could help systems maintain continuity, disambiguate repeated actions, and better track progress through non-linear or interleaved workflows.

6.6.5 Toward Context-Aware Interaction with LLMs

While OSCAR’s primary contribution lies in its object status-based progress tracking, our architecture also opens up promising opportunities for context-aware, conversational interaction powered by large language models (LLMs). By maintaining a structured history log—including step predictions, visual context, and object transformations over time—OSCAR can dynamically construct prompts that reflect the current cooking state. This enables LLMs to reason not only about static recipe text but also about the user’s actual progress through the task. For instance, users could ask situational questions such as “What step am I on?” or “Did I already add the garlic?” and receive responses grounded in their cooking history rather than generic recipe flow. This integration represents a shift from linear recipe readers toward procedurally aware dialogue systems, where interactions are informed by both real-time observations and past actions. As LLMs continue to improve in temporal reasoning and contextual understanding, such a pipeline

could support richer, personalized guidance—especially for users who rely on non-visual cues to navigate complex tasks. Future work may explore the design space of such interactions, including how to present uncertainty, verify state interpretations, and align system responses with user mental models.

6.6.6 Designing for Autonomy over Standardization

Our findings underscore that non-visual cooking is shaped not by deviation from norms, but by adaptation grounded in embodied expertise and personal strategies. Participants frequently modified workflows—substituting tools, reordering steps, or combining actions—not out of error, but to optimize for safety, efficiency, or tactile feedback. These adaptations often confounded baseline models that assumed fixed action-tool mappings or linear task flows. In contrast, OSCAR’s object-status-driven tracking—focusing on the evolving state of objects rather than prescriptive actions—was more resilient to user-specific variations. This finding aligns with broader accessibility research cautioning against assistive technologies that enforce normative or idealized procedures [110, 117]. Future systems should foreground adaptability and respect user autonomy, allowing people to work with their own tools, sequences, and strategies—rather than requiring conformance to model assumptions.

6.6.7 Interaction and Camera Design for Blind Cooks

A recurring challenge in our real-world evaluation was camera misalignment—with chest-mounted cameras often angled too high, too low, or failing to capture the primary workspace. In V6, for example, the camera missed the cutting board entirely during ingredient preparation, contributing to a low prediction accuracy of 40% despite adequate lighting. For BLV cooks, this issue is not simply about camera placement—it reflects a fundamental interaction gap. Without visual feedback, users cannot easily verify what the camera captures during dynamic tasks like cooking. This challenge echoes prior research in blind photography, where real-time framing support—such as haptic cues, audio tones, or spoken guidance—has been shown to help users position cameras effectively [4, 85]. For systems like OSCAR to operate reliably in deployment, camera-aware interaction must become a core design consideration. Future systems should provide real-time feedback to guide framing, suggest re-alignment when needed, or explore multi-camera setups to increase robustness—all while minimizing user effort during hands-busy tasks like cooking.

6.6.8 Expanding Accessibility Datasets for Inclusive AI

We release our 12-video non-visual cooking dataset to support evaluating procedural AI in real-world accessibility contexts (Section 6.5). To more fully validate systems like OSCAR, future datasets must reflect a broader range of kitchen environments, cultural cuisines, and non-visual cooking practices. As with prior datasets like ORBIT [131], the true value of such resources lies in their diversity and realism. Future efforts should broaden this foundation by capturing data across varied kitchen setups, lighting conditions, cultural cooking practices, and user expertise levels. In particular, collecting multi-session recordings from the same participants could support modeling of longitudinal learning, routine formation, and adaptation over time. Echoing principles from disability-centered dataset research [179], inclusive AI development requires datasets that

reflect the messiness, variability, and situated practices of real users—not idealized or sanitized tasks. Expanding datasets along these dimensions is critical for advancing robust, user-centered assistive technologies.

6.7 Limitations and Future Work

While this chapter demonstrates the feasibility and performance of OSCAR for recipe tracking using object status recognition, several limitations remain. Our current evaluation was conducted entirely on pre-recorded video datasets, without direct user interaction with the system. As a result, we were not able to observe how blind users might engage with OSCAR in real-time—particularly how they interpret contextual feedback, interact with step tracking, or formulate conversational queries during the cooking process. Future work should focus on developing an accessible mobile prototype and conducting in-situ studies in users’ own kitchens. These studies will examine how users naturally interact with the system, including their input strategies, preferences for conversational support, and ways of integrating real-time feedback into non-visual workflows. Insights from these investigations will help refine OSCAR’s interactive capabilities and guide the design of context-aware assistive systems grounded in real-world use. Finally, our evaluation focused on two state-of-the-art VLMs (CLIP and SigLIP). As multimodal models continue to advance (e.g., GPT-4o), evaluating OSCAR with a wider range of models will help uncover its broader applicability. In particular, future work should explore what levels of model performance are “good enough” for practical use and define benchmarks that center user experience, not just accuracy.

6.8 Summary

This chapter presented OSCAR (Object Status Context Awareness for Recipes), a technical pipeline that explores object status recognition as a foundation for context-aware cooking assistance for people with vision impairments. By integrating recipe parsing, object status extraction, vision-language alignment, and time-causal modeling, OSCAR enables real-time recipe progress tracking that adapts to diverse user practices and non-visual workflows. Through large-scale evaluation on instructional videos and a real-world non-visual cooking dataset recorded by blind and low vision individuals in their homes, we demonstrated that object status recognition consistently improves step-prediction accuracy over baseline models, even in challenging, uncontrolled environments. Our findings highlight the limitations of current vision-language models in real-world accessibility contexts and underscore the importance of modeling temporal continuity, user autonomy, and task adaptability. We contribute both a technical pipeline for object-status-driven progress tracking and a real-world cooking dataset to support future research for non-visual cooks. Looking forward, we envision expanding OSCAR into interactive, user-facing prototypes and conducting in-situ studies to explore how BLV cooks engage with real-time feedback and context-aware guidance. Our work provides design insights for building robust, inclusive assistive technologies that align with the complexities of non-visual everyday practices.

Chapter 7

ODG: An Object Dependency Graph for Non-Visual Cooking Assistance

7.1 Overview

The preceding chapters of this thesis each addressed one piece of the non-visual cooking puzzle. Chapter 3 characterized the day-to-day cooking practices and challenges of people with vision impairments through a content analysis of 122 YouTube videos and 12 semi-structured interviews [110]. Chapter 4 examined how blind cooks access and structure recipes through a follow-up study with 20 blind cooks and 4 rehabilitation instructors, and articulated design strategies for non-visual recipe access [117].

Chapter 5 moved from the recipe to the kitchen itself, conducting a contextual inquiry with 12 blind cooks in their own homes to identify the eight categories of contextual object information (five primary: position, orientation, proximity and grouping, similarity and duplicates, and internal state; plus two secondary and one application-specific) that non-visual cooks rely on [111]. Chapter 6 built OSCAR, a vision-language system that extracts object status from recipes, aligns visual frames to recipe steps, and provides a context-aware progress tracker for following a recipe in real time.

Each of these contributions, however, was studied in isolation. Blind cooks in the preceding studies consistently described three failure modes that cut across all of them: (1) *recipe texts remain linear*, since tactile audio readers and screen readers force sequential playback even when the underlying cooking graph has many independent paths [117]; (2) *object status, safety warnings, and sensory cues are locked inside the recipe's prose*, so information that could be delivered proactively at the moment it is needed instead has to be re-read, remembered, and mentally re-planned [110, 111]; and (3) *progress, skipped steps, and re-orientation are left to the cook's working memory*. OSCAR's contribution was to track progress, but blind cooks also need to *ask* about and *recover* from where they are in the cooking process at any point [117].

This chapter introduces the Object Dependency Graph (ODG), a unifying representation of a recipe that operationalizes the physical information space introduced in Chapter 1. An ODG explicitly encodes each cooking state as a grounded [verb] + [noun] object transition (Chapter 6), the prerequisites and safety constraints between states (Chapter 3), the non-visual cues that signal completion (Chapter 5), exact quantities and techniques (Chapter 4), and the parallel groups

in which independent states can run concurrently. An ODG-aware voice assistant answers questions the underlying recipe was never written to answer (“What do I need before I sear the beef?”, “What can I do while the phyllo bakes?”, “What did I skip?”, “Where am I right now?”) while delivering safety alerts, sensory cues, and automatic timers proactively at the moment they become relevant.

The ODG is not a new recipe format; it is a common substrate on top of the four prior contributions of this thesis. It is *grounded* in the object-status representation from OSCAR [113,114], *populated* with the design strategies identified in the recipe work [117] and the contextual-information types from the kitchen contextual inquiry [111], and *informed* by the practices and challenges catalogued in the original non-visual cooking study [110]. To probe which of these ingredients actually matter for a blind cook in practice, the ODG decomposes into seven **ODG Rules** (R1–R7), each backed by one of the prior chapters, that can be individually toggled on or off at run time. These rules serve as both the system’s demonstration surface and as the independent variable of the within-subjects evaluation reported in Chapter 8.

The contributions of this chapter are scoped to the ODG itself: its representation, its decomposition into rules, and the running system that realizes it. The empirical evaluation of those rules with blind cooks is the contribution of Chapter 8 and is not claimed here.

- **Representation.** A formal definition of the Object Dependency Graph as a tuple $G = (S, E, P, T)$ that integrates object status, prerequisites, safety, non-visual cues, measurements, progress tracking, and parallelism into a single machine-readable structure, grounded in the four prior empirical studies of this thesis.
- **Rules.** A decomposition of the ODG into seven design rules (R1–R7), each rule motivated by one specific prior chapter and consuming one or more specific ODG fields, with explicit traceability between the empirical findings of this thesis and the behaviors of an ODG-aware assistant.
- **System.** A voice-first ODG-aware cooking assistant with a paired Wizard-of-Oz researcher dashboard, built around an input-agnostic reasoning engine that grounds every spoken reply in a structured ODG query rather than in recipe prose. The same engine drives the participant interface and the dashboard, so rule toggles made by the researcher take effect on the next assistant turn without restarting the session, making each of the seven rules an independently controllable run-time variable. The implementation is described in Chapter 8.

These three contributions form the substrate that the next chapter evaluates: the representation, the rules it expresses, and the running system that lets blind participants interact with the rules in practice.

7.2 Background and Related Work

The broad background on cooking technology, information spaces, and assistive technologies for people with vision impairments is covered in Chapter 2. This section focuses on the work most directly relevant to the ODG and positions it against four neighboring lines of research: structured

representations of cooking procedures, voice-first and multimodal cooking assistants, conversational and mixed-initiative agents for procedural tasks, and accessibility-focused approaches to recipe support for people with vision impairments.

7.2.1 Structured Representations of Cooking Procedures

A long line of work has treated cooking as a structured procedural task rather than as linear prose. Graph-based and workflow-based representations have been used to align instruction texts with video demonstrations [31, 47, 119], to compute recipe difficulty and workflow complexity [104], to analyze recipe corpora at scale [35], and to perform semantic processing of ingredient and step structure [130, 175]. These representations make explicit the dependencies between preparation steps, ingredients, and cooking actions, structure that is otherwise implicit in the recipe text. Beyond graph-based formalisms, researchers have also explored flow-chart-style and tree-structured recipe encodings [104], frame-semantic action representations [175], and large-scale paired text/image datasets that enable end-to-end learning of recipe structure [130]. Across this body of work, the recurring design move is to externalize what the recipe text leaves implicit (ingredient identity, action sequence, intermediate states, and tool-to-action bindings) into a representation that a downstream system can reason over.

The ODG builds on this tradition but differs in three important ways. First, existing representations are designed primarily to support retrieval, alignment, or analysis tasks (e.g., aligning narration to video, predicting the next step, retrieving similar recipes) rather than to drive a non-visual assistant in real time. As a result, they optimize for static, batch-processing settings and rarely consider how the representation will be *spoken* into a kitchen. Second, they typically model *what* needs to be done but not the non-visual cues that indicate *when* a state has been reached, the safety constraints that gate dangerous transitions, or the parallel structure that allows a cook to overlap independent work, all of which are fields whose absence the prior chapters of this thesis showed to be most consequential for blind cooks [110, 111, 117]. Third, where prior representations are designed for machine consumption, the ODG is explicitly designed so that each of its fields has a corresponding human-facing assistant behavior, a property that the seven rules in Section 7.3 make concrete. The ODG therefore extends the classical step-dependency view to include sensory cues, edge-level safety annotations, parallel groups, an explicit preparation track, and a measurement/technique layer, while preserving the underlying directed acyclic graph (DAG) structure that earlier representations established.

7.2.2 Voice-First and Multimodal Cooking Assistants

Interactive cooking assistants have been studied for two decades, spanning projector-based guidance [134, 159, 184], instrumented pans and cooking simulators [91, 184], gaze- and voice-controlled recipe systems [26], and early multitasking coordinators [75, 76]. These systems explored a broad design space, from situated displays that anchor instructions to the work surface, to ambient sensing of tools and ingredients, to conversational interaction during high-load moments such as searing or plating. More recent work has examined how commercial voice assistants (Amazon Alexa, Google Assistant, Siri) support (and fail to support) everyday cooking and other household activities, particularly for people with vision impairments [29, 40, 172]. This recent literature is more diagnostic than constructive: it documents what mainstream

assistants get wrong rather than proposing alternative architectures, which is one of the gaps that motivates the constructive contribution of the present chapter.

Three limitations are consistently reported across this literature. First, voice assistants that read the recipe linearly do not answer structural questions (“what can I do while the oven pre-heats?”, “which step did I skip?”, “what do I need before I can sear the beef?”); the structure that would support such questions exists in the underlying procedure but is collapsed into prose at delivery time. Second, their pacing and verbosity are often poorly matched to blind users’ preferences [29, 40]: information is either too dense to absorb in a single utterance or too sparse to act on without follow-up. Third, they bundle proactive behavior (timers, safety warnings, sensory cues, parallel-task suggestions) into a single fixed voice persona that cannot be configured: a cook who wants safety alerts but not parallel-task prompts, or who wants sensory cues but not measurement reminders, has no granular control. The ODG is designed to address all three: it exposes structural queries as first-class engine methods that traverse the DAG directly, it stores measurement, sensory, and safety content as separate fields so the assistant can adapt verbosity on demand, and it decomposes proactive delivery into seven independently toggleable rules rather than a single bundled persona.

7.2.3 Conversational and Mixed-Initiative Agents for Procedural Tasks

A complementary line of work studies conversational agents and mixed-initiative interaction for procedural tasks more broadly, including assembly, do-it-yourself projects, repair, and physical-therapy routines, alongside cooking. This work is relevant to the ODG in two ways. First, it has produced a vocabulary for describing when an assistant should speak proactively versus reactively, when it should ask clarifying questions, and how it should handle errors and re-orientation. The seven ODG Rules draw on this vocabulary, particularly R3 (proactive safety alerts), R4 (proactive parallel-task suggestions), and R6 (re-orientation queries). Second, it has shown that procedural assistants which expose an explicit task model to the user (via a visible list of steps, a progress bar, or an interactive transcript) tend to be more trusted than assistants that treat the task as a black box. The ODG is designed to be such an explicit model: the researcher dashboard described in Chapter 8 mirrors the live state of the graph back to the moderator, and the same surface could be exposed to the cook in a future deployment to support trust and recovery.

Recent work on large language model (LLM) agents has further sharpened the question of *what* a task representation should expose to a conversational interface. LLMs can flexibly answer free-form questions about a recipe, but they hallucinate object states, miss safety-critical transitions, and lose track of which step is current when the conversation extends over many turns. Representations such as the ODG act as a structured grounding layer that constrains what the LLM can claim to be true at any moment: the assistant can phrase its outputs naturally, but the facts it asserts (which steps are done, which are pending, which are concurrent, what quantity to use) are read from the graph rather than generated. Mixed-initiative work that grounds language outputs in a structured task model, including AROMA [140] (which builds on the object-status representation from Chapter 6), demonstrates that this combination of LLM fluency with structured grounding is feasible and useful in cooking specifically. The ODG can be read as a proposal for what such a grounding structure should look like when the user is blind.

7.2.4 Accessibility-Focused Recipe and Cooking Support

A smaller set of recent work has looked specifically at recipe and cooking support for people with vision impairments. This includes vision-based assistance on the stove [95], studies of cooking training and everyday practice for blind and low-vision cooks [190], and the four prior chapters of this thesis, which together characterized non-visual cooking practice [110], recipe access strategies [117], in-kitchen contextual information needs [111], and real-time object-status extraction via OSCAR. Beyond academic research, the everyday tooling that blind cooks actually use today is a mix of mainstream screen readers (VoiceOver, TalkBack, NVDA, JAWS) layered on top of recipe websites that were not designed with non-visual access in mind; mainstream smart speakers used for timers and limited step-by-step playback; and human-mediated remote sighted assistance through services such as Aira and Be My Eyes. Each of these tools fills a narrow slice of the non-visual cooking workflow, but none of them models the recipe as a structured object that can be queried, re-entered after interruption, or interrogated for safety, parallelism, or measurement detail. The result is a fragmented experience in which a cook may switch between three or four different tools to complete a single recipe, manually carrying state across them in working memory.

The ODG is designed to operationalize the findings from the prior chapters of this thesis into a single substrate that the everyday tools above cannot currently express. It is grounded in OSCAR’s object-status representation, populated with the recipe design strategies and contextual information categories identified in the two design studies, and informed by the practices and challenges catalogued in the original non-visual cooking study. Crucially, the ODG is not intended as a replacement for screen readers, smart speakers, or remote-sighted services. A blind cook in 2026 still benefits from each of those tools in different moments. Rather, the ODG provides the missing structured layer that an assistive cooking system can read from in order to deliver the right information at the right moment through whatever delivery surface the cook prefers. The empirical chapters of this thesis identified the gap; this chapter contributes the representation that closes it; and the next chapter (Chapter 8) evaluates the resulting assistant with blind participants.

7.3 Design of the Object Dependency Graph

7.3.1 Design Rationale: Four Prior Chapters, One Representation

Viewed together, the four prior chapters of this thesis tell a consistent story about what non-visual cooking support is missing. Table 7.1 summarizes the specific gap that each prior chapter surfaced but could not itself close, and the ODG element that addresses it.

Prior Chapter	Gap Surfaced but Not Closed	ODG Element That Addresses It
Ch. 3 Non-Visual Cooking Practices [110]	Blind cooks juggle many simultaneous sub-tasks; safety incidents cluster at transitions.	Parallel groups; per-edge safety annotations; transition warnings.
Ch. 4 Recipe Access [117]	Linear TTS/screen-reader playback hides structure; cooks rewash hands to re-scan; measurements and techniques get lost.	Non-linear state graph queryable by voice; explicit quantity and technique fields per state.
Ch. 5 Contextual Inquiry in the Kitchen [111]	Cooks need contextual object information (position, orientation, temperature, status, ...) that is not in the recipe text.	Per-state <code>contextCategories</code> and typed <code>nonVisualCues</code> (tactile / auditory / olfactory / thermal / timing).
Ch. 6 OSCAR	Real-time tracking works, but queries like “what did I skip?” and “where am I?” need an explicit step graph, not only a time-causal log.	Grounded <code>[verb]+[noun]</code> state identity; prerequisite graph; skip detection; <code>queryWhereAmI()</code> .

Table 7.1: The ODG as a unifying representation of the four prior chapters: each prior study identified a specific gap whose shape implied a structural requirement, and the ODG satisfies all of them in one representation.

Each of these gaps describes something a blind cook lacks at run time, not something the recipe text lacks at authoring time. Translating them into concrete assistant behaviors (behaviors that can be turned on or off, narrated aloud, and individually evaluated) yields the seven **ODG Rules** introduced in the next subsection. The rules come first because they are the empirical statement of what a non-visual cooking assistant must do, derived directly from the prior chapters of this thesis. The representation that makes those behaviors implementable, the Object Dependency Graph itself, is described after the rules in Section 7.3.10, as the substrate that all seven rules read from.

7.3.2 The Seven ODG Rules

I refer to the seven behaviors distilled from the four prior chapters as the **ODG Rules** (R1–R7). Each rule is a small, self-contained behavior that a non-visual cooking assistant should exhibit, and each is directly traceable to one of the prior empirical studies of this thesis. They are listed in Table 7.2, together with the chapter that motivates each rule and the ODG field each rule consumes once the underlying representation is introduced in Section 7.3.10.

Rule	Behavior	Motivated by	ODG Field Consumed
R1	Object Status & Step Narration: refer to everything as [verb]+[noun] and break complex steps into micro-steps.	Ch. 6, Ch. 5 [111, 113, 114]	state.verb, state.object, state.description
R2	Sensory Cues: replace visual doneness checks with sound, smell, touch, and timing.	Ch. 3, Ch. 4, Ch. 5 [110, 111, 117]	state.nonVisualCues
R3	Safety Alerts: warn before dangerous transitions (hot oil, open flame, sharp tools).	Ch. 3, Ch. 4 [110, 117]	edge.safety
R4	Parallel Tasks: proactively suggest independent work during wait times.	Ch. 4 [117]	parallelGroups, state.time.passive
R5	Precise Measurements & Technique: surface exact quantities, tool sizes, time ranges, and techniques on demand.	Ch. 4, Ch. 5 [111, 117]	state.quantity, prepTrack.items
R6	Progress Tracking & Recovery: give summaries, detect skipped steps, and support re-orientation.	Ch. 6, Ch. 4 [113, 114, 117]	completed, inProgress, detectSkips(), queryWhereAmI()
R7	Automatic Timers: set and track timers automatically whenever a state has a time duration.	Ch. 4, Ch. 5 [111, 117]	state.time.active, state.time.passive

Table 7.2: The seven ODG Rules, each motivated by a prior chapter of this thesis and each consuming one or more ODG fields.

Each rule is implemented as a small, self-contained behavior that can be independently enabled or disabled at run time (Chapter 8). In the user study (Chapter 8), the full ODG condition enables all seven; the baseline condition uses a standard screen reader and the recipe’s original text.

The following subsections elaborate on each rule in detail, grounding them in the findings of the prior chapters.

7.3.3 R1: Object Status and Step Narration

What the rule does. R1 requires that every time the assistant refers to an ingredient, tool, or cooking action, it uses a grounded [verb]+[noun] pair that reflects the current state of that object. For example, instead of saying “add the eggs to the bowl,” the assistant says “add your *cracked* eggs to your *beaten* egg yolks.” Instead of “cook until done,” the assistant says “cook your *seared* beef until it *sounds* quiet and the surface *smells* caramelized.” This verb-noun grounding carries forward through the recipe: by the time a cook reaches the state *sear* + *beef fillets*, the ODG has already established that these are *your* beef fillets, already *seasoned*, already *brought to room temperature*. The assistant never refers to the ingredient in the abstract.

This representation originates directly from OSCAR (Chapter 6), where I showed that vision-language models perform substantially better at tracking recipe progress when they reason about object *status transitions* rather than about raw ingredient names. The ODG extends this idea from perception to narration: a cook following a recipe through a screen reader hears “Add the eggs” at every step that involves eggs, regardless of whether the eggs have already been cracked, whisked, or folded into batter. R1 eliminates this ambiguity by attaching the verb that describes the current object status to every reference.

The contextual inquiry study (Chapter 5) further validated that blind cooks need to know not just *what* an object is, but *what state it is in*. Participants in that study described repeatedly checking ingredients and tools because the recipe text never clarified whether, for instance, the tomatoes had already been chopped or were still whole. R1 operationalizes this need by encoding the object state in the ODG state label itself, so that every verbal reference from the assistant carries the current status of every ingredient in context.

Beyond grounding ingredient references, R1 also governs how step instructions are phrased. Consider a traditional recipe step for Jamaican Jerk Rice & Peas: “*Heat olive oil in a saucepan over medium heat. Add onion and salt and sauté until onion starts to soften, 4 to 5 minutes. Clear space in the middle of the pan, add tomato paste, and sauté for 2 to 3 minutes. Season with paprika, curry powder, cumin, thyme, pepper, and turmeric; continue cooking for 1 to 2 minutes until everything is well combined.*” This one step bundles seven distinct actions (heating, adding, sautéing, clearing space, adding again, seasoning with six spices, and waiting for combination), requiring a blind cook to hold all of it in working memory and track which sub-action they are on mid-step. R1 breaks this into micro-steps: each distinct action becomes its own ODG state, with its own sensory cues, time fields, and safety annotations. The cook hears one atomic action at a time, confirms completion, and then receives the next. This means a blind cook never has to hold a compound instruction in working memory and parse it mid-action; each atomic action is delivered and confirmed before the next begins.

In practice, R1 produces a running commentary that sounds like this: “*Your oven is preheated. Next: layer your brushed phyllo sheets with your melted butter. Use a pastry brush to coat each sheet, working quickly so the phyllo doesn’t dry out.*” The italicized status words are the ODG’s verb+noun groundings carrying forward from prior states.

7.3.4 R2: Sensory Cues

What the rule does. R2 requires the assistant to deliver a non-visual sensory cue (auditory, olfactory, tactile, thermal, or timing-based) every time the recipe reaches a state where the cook

would otherwise need to visually inspect the food. The cue is drawn from the state's nonVisualCues list, which is populated from the contextual inquiry study (Chapter 5) and the non-visual cooking practices study (Chapter 3).

Consider a traditional recipe step: *"Toast the nuts in the oven until browned."* The phrase "browned" describes a purely visual state and is entirely inaccessible to a blind cook. R2 replaces this visual cue with multi-channel non-visual alternatives: for the state `toast + nuts`, the ODG encodes auditory: "nuts shift from a soft thud to a dry rattle when stirred" (high reliability), olfactory: "warm, toasted aroma" (medium reliability), and timing: "3--5 minutes in a preheated oven at 350°F" (high reliability). When the state is entered, the assistant delivers: *"Toast the nuts in the oven for 3 to 5 minutes at 350 degrees. You'll know they're ready when you hear them shift from a soft thud to a dry rattle when you stir them, and when you smell a warm, toasted aroma."*

The non-visual cooking study (Chapter 3) catalogued twelve different cooking activities and found that blind cooks consistently rely on sensory modalities other than vision to accomplish them. For instance, cooks in that study described hearing the shift from a soft thud to a dry rattle when stirred nuts are ready, smelling the transition from raw to caramelized onion, feeling the texture change of dough coming together, and timing the sizzle of oil to judge temperature. The contextual inquiry study (Chapter 5) further identified five sensory channels (tactile, auditory, olfactory, thermal, and timing) as the primary non-visual means by which blind cooks assess food status.

R2 encodes these findings directly into the ODG. Each cue is tagged with a reliability level (high, medium, low) based on how consistently participants in the prior studies reported it. High-reliability cues are delivered proactively when a state begins; medium- and low-reliability cues are delivered on request or as supplementary information. If one channel is unavailable (e.g., the kitchen is too noisy for an auditory cue, or the cook has a cold that dulls smell), the multi-channel fallback ensures another channel can serve as confirmation.

7.3.5 R3: Safety Alerts

What the rule does. R3 requires the assistant to deliver a safety alert immediately before any state transition that involves a hazard such as hot oil, open flame, sharp tools, extreme temperatures, or heavy objects. The alert is drawn from the edge's safety field in the ODG, which encodes the specific hazard, the risk it poses, and the precaution the cook should take.

The non-visual cooking study (Chapter 3) identified safety and consequence as one of its eight challenge themes, reporting that all twelve interview participants expressed strong concerns about burns, fires, and accidental injuries in the kitchen. Participants described hazards including reaching into hot ovens, handling boiling liquids, working near open flame, and accidentally knocking objects over onto hot surfaces. The recipe access study (Chapter 4) reinforced these concerns: cooking instructors at the vision rehabilitation center emphasized that blind cooks must be warned proactively about hazards before encountering them, because the cook cannot visually detect dangerous conditions in time to react.

R3 operationalizes these findings by encoding safety information on the directed edges of the ODG. When the ODG engine determines that a transition $s_i \rightarrow s_j$ involves a hazard, it interrupts the normal flow to deliver the safety alert. Consider the transition from `dredge + chicken` to `lower + chicken` into oil. The ODG encodes an edge with type: thermal

and safety: "Oil at 350F can cause severe burns. Make sure your hands are completely dry before lowering the chicken. Lower the basket slowly and step back immediately after release." When the cook attempts to mark dredge + chicken as complete, R3 triggers the alert: *"Before you lower the chicken into the oil, safety alert: the oil is at 350 degrees Fahrenheit. Severe burns can occur. Please make sure your hands are completely dry. Lower the basket slowly, and step back as soon as you release it. Say 'done' when you've lowered the chicken and stepped back."*

The safety field is also populated proactively during passive states. When the cook enters the state bake + baklava in oven, the ODG engine recognizes that the incoming edge carries a thermal hazard (the oven is hot). R3 delivers: *"Reminder: the oven is now at 350 degrees. Use a thick oven mitt on your dominant hand when opening the door. When you pull out the rack, place it on the counter, never on the stovetop, which may still be hot from earlier cooking."*

R3 is designed to be unobtrusive during safe transitions. For states with no safety annotation, the rule produces no output, avoiding the fatigue that would come from safety messages on every step. The threshold for what counts as a hazard is calibrated against the practices and incident reports from the non-visual cooking study: edges are annotated with safety content whenever the prior research documented a participant injury, a near-miss, or an expressed safety concern.

7.3.6 R4: Parallel Tasks

What the rule does. R4 requires the assistant to proactively suggest independent cooking tasks during any passive wait period (while the oven preheats, while sauce simmers, while dough rests) so that the cook can use time efficiently rather than standing idle. The suggestions are drawn from the ODG's `parallelGroups` and `state.time.passive` fields.

The recipe access study (Chapter 4) found that blind cooks highly value knowing when they can perform multiple tasks simultaneously. Interview participants described the frustration of cooking pasta and realizing only after the fact that they could have prepared the sauce in parallel. One participant in the recipe access study specifically noted that recipes which told her what she could do *during* a wait period ("while the pasta cooks for 11 minutes, you can make the sauce") were substantially more helpful than recipes that listed passive time as dead time. The contextual inquiry study (Chapter 5) further showed that blind cooks develop elaborate organizational strategies to maximize parallel work, such as arranging ingredients in the order they will be used and staging tools at multiple workstations.

R4 encodes these findings in two ways. First, the ODG's `parallelGroups` explicitly mark which states are order-independent. For the Greek Baklava recipe, for instance, the states preheat + oven, chop + nuts, butter + baking pan, and unroll + phyllo are all members of the same parallel group, meaning any subset can be completed in any order before the next prerequisite-dependent state (layer + phyllo + butter) begins. Second, each state carries a `time.passive` duration that represents the wait time during which independent work can be inserted.

When the cook enters a passive state (for example, preheat + oven with a passive time of 10 minutes), R4 triggers: *"Your oven is preheating and will be ready in about 10 minutes. While you wait, here are tasks you can do in parallel: chop the nuts, butter the baking pan, or unroll the phyllo and keep it covered with a damp towel. These are all independent of the oven preheat and can be done in any order. Which would you like to start with?"*

R4 also handles nested parallelism: when the cook is already in a passive state and completes a parallel task, R4 re-evaluates the ODG and updates the suggestion list. If the cook finishes chop + nuts while the oven is still preheating, R4 might say: *“Great, the nuts are chopped. Your oven still has about 6 minutes to go. You can now butter the pan or unroll the phyllo. Would you like guidance on any of these?”*

The rule never forces parallel work on a cook who prefers to focus on one thing at a time. The suggestions are offered as options, not instructions, and the cook can decline or defer without penalty. The ODG tracks which parallel suggestions have been offered so that the assistant does not repeatedly recommend the same already-declined task.

7.3.7 R5: Precise Measurements and Technique

What the rule does. R5 requires the assistant to surface exact quantities, tool specifications, time ranges, and technique descriptions on demand, drawing from the state’s quantity field and the prep track’s items list. Rather than embedding measurement information in the prose of the step instruction, the ODG stores it as structured data that can be retrieved conversationally at the moment the cook needs it.

The recipe access study (Chapter 4) found that blind cooks frequently struggled with imprecise or missing measurements in recipes. Participants described frustration with phrases like “add a pinch of salt,” “cook until golden brown,” or “use a large skillet,” all of which require visual estimation or prior knowledge that blind cooks may not have. Instructors at the vision rehabilitation center emphasized the importance of giving blind cooks specific measurements rather than visual approximations: “two sticks of butter” rather than “enough butter to coat the pan,” and “a 10-inch cast iron skillet” rather than “a large skillet.” The contextual inquiry study (Chapter 5) reinforced this, showing that blind cooks use tactile and spatial references (e.g., the width of two fingers, the depth of a container) to anchor measurements, and that providing explicit numeric or physical anchors substantially reduces cognitive load.

R5 encodes these findings as structured quantity data on every ODG state. For the state brush + phyllo + butter, the quantity field encodes: amount: “2 sticks (1 cup total)”, tool: “pastry brush, medium stiffness”, method: “brush every sheet thoroughly, especially at the corners”, temperature: “butter should be melted and slightly cooled, not hot”. When the cook asks *“How much butter do I need for this step?”*, R5 triggers: *“You’ll need 2 sticks of butter, melted; that’s 1 cup total. Use a medium-stiffness pastry brush. Melt the butter first, let it cool slightly so it’s warm but not hot, then brush each phyllo sheet thoroughly, paying extra attention to the corners so the layers stick together.”*

The rule also covers technique. Rather than assuming the cook already knows *how* to perform an action, R5 delivers the technique information as a follow-up: *“Would you like guidance on how to brush butter onto phyllo?”* If the cook accepts, R5 provides the technique description from the ODG’s description field: *“Hold the phyllo sheet flat with one hand, and use your other hand to brush from the center outward. Work quickly; phyllo dries out fast. If you need to pause, cover the unbrushed sheets with a damp towel.”*

R5 is queryable at any time. A cook who remembers the ingredient but not the quantity can ask *“How much butter for the phyllo?”* and receive the structured answer regardless of their current position in the recipe. This directly addresses the finding from the recipe access study

that blind cooks often had to rewash their hands and scroll back through the recipe to find a measurement they had missed.

7.3.8 R6: Progress Tracking and Recovery

What the rule does. R6 requires the assistant to maintain a running model of which states have been completed, which are in progress, which are blocked by unmet prerequisites, and which have been skipped. The cook can query this model at any time to get an orientation summary, to find out what they skipped, and to re-orient after an interruption.

The OSCAR chapter (Chapter 6) demonstrated that real-time progress tracking substantially improves recipe step-prediction accuracy for blind cooks, but it also revealed that tracking alone is insufficient: cooks need to be able to *ask* about their progress, not just receive updates. The recipe access study (Chapter 4) confirmed this: participants described losing their place in the recipe after interruptions, having to re-read from the beginning after a screen reader accidentally jumped to a different section, and needing to verbally confirm with a collaborator where they were in the process. One participant described the challenge of remembering which step she was on after her dog barked and broke her concentration.

R6 operationalizes these needs through four engine methods. First, `markCompleted(s)` marks a state as complete and updates the ODG’s completed set. Second, `detectSkips()` compares the completed set against the prerequisite graph to identify any state whose prerequisites are all satisfied but which does not appear in completed, flagging it as potentially skipped. Third, `queryWhereAmI()` returns a structured re-orientation summary that includes: the number of states completed and the percentage of the recipe finished; the current cooking phase (prep, cook, rest, assembly, or finish); any states that may have been skipped; and the set of currently available next states. Fourth, `queryPrerequisites(s)` returns what must be done before a given state can begin.

Consider a cook who is following Greek Baklava and has completed the first four steps but then paused to answer a phone call. When they return and ask “*Where am I?*”, R6 triggers `queryWhereAmI()`: “*You’ve completed 4 of 18 steps, which is 22 percent. You’re currently in the preparation phase. One thing to note: it looks like you may have skipped ‘spice the nuts.’ If you skipped it, the nuts won’t have the cinnamon and sugar coating. Would you like to go back and add that now? Your next available steps are ‘layer phyllo and butter’ and ‘brush phyllo with butter.’*”

The skip detection is particularly important for blind cooks, who cannot visually scan ahead and notice that a step was missed. Without R6, a skipped step means a silently degraded outcome: underproofed dough, a missing ingredient, an unsliced vegetable. With R6, the assistant flags the gap and gives the cook the option to recover before the error compounds.

R6 also handles the case where the cook accidentally marks the wrong state complete. When `markCompleted(s)` is called, the engine checks whether the state’s prerequisites are all satisfied. If the cook attempts to mark a blocked state as complete (for example, trying to mark `assemble + beef + Wellington` complete without having first completed `sear + beef`), R6 blocks the action and says: “*Hold on, you can’t assemble the Wellington yet. First, you need to sear the beef. Once you’ve seared the beef, come back and I’ll mark assembly as complete.*”

7.3.9 R7: Automatic Timers

What the rule does. R7 requires the assistant to automatically set and track timers for every ODG state that carries a time duration: both `time.active` (time the cook is actively working) and `time.passive` (time the cook is waiting, e.g., while something bakes or rests). The timers are managed by the ODG engine and announced proactively when they expire, rather than requiring the cook to remember to set them.

The recipe access study (Chapter 4) found that timing was one of the most frequently cited pain points for blind cooks. Participants described the frustration of losing track of time during a recipe, forgetting to set a timer after a distraction, and having to rely on sighted assistance to know when a passive phase was complete. One participant described burning a dish because she lost track of the 20-minute baking time while managing other components. The contextual inquiry study (Chapter 5) showed that blind cooks are acutely aware of the relationship between time and cooking outcomes (overcooking is common when visual timers are unavailable), and that explicit time announcements are strongly preferred over vague phrases like “cook until done.”

R7 encodes these findings as `time.active` and `time.passive` fields on every ODG state. For the state `bake + baklava`, the ODG encodes `time.active`: `null` (the cook is not actively working during baking) and `time.passive`: `"35--45 minutes"` with a nominal value of 40 minutes. When the cook marks `bake + baklava` as in-progress, R7 starts the passive timer and delivers: *“Timer set: 40 minutes for the baklava to bake. I’ll let you know when it’s ready. While you wait, your oven is at 350 degrees, so use your oven mitt when opening the door. Would you like to start on the honey syrup?”*

When the timer expires, R7 delivers: *“Bake timer complete. Your baklava should be golden and fragrant. Open the oven and use your oven mitt; the rack and dish will be very hot. Say ‘done’ when you’d like to check if it’s ready for the final step.”*

R7 also handles the relationship between active and passive time. When the cook is performing an active step that has both active and passive components (for example, `sear + beef` with `time.active`: `"3 minutes per side"` and `time.passive`: `"5 minutes rest after searing"`), R7 sets both timers in sequence: first the active sear timer, then the passive rest timer, announcing each transition. This means the cook never has to remember to set a rest timer after the searing step; the ODG engine handles the full temporal structure of the state automatically.

For states where timing is approximate (e.g., “bake until golden brown, 35–45 minutes”), R7 uses the upper bound as the nominal timer value but encodes the range so the assistant can say: *“Timer set for 45 minutes. Use your senses to check before the timer goes off: the baklava should smell very fragrant and the top should sound slightly crisp when you tap it gently. I’ll check in with you before the timer expires.”*

Each rule is implemented as a small, self-contained function within the ODG engine that can be independently enabled or disabled at run time, allowing the user study to evaluate the contribution of each rule in isolation or in combination.

7.3.10 The Object Dependency Graph: A Common Substrate for the Seven Rules

The seven rules described in the previous subsections are not seven independent systems. They are seven *consumers* of the same underlying representation: a single, machine-readable structure that encodes everything a non-visual cooking assistant needs to know about a recipe. This subsection makes that representation explicit. It defines the Object Dependency Graph (ODG) formally, lists the fields each of the seven rules reads from, and walks through a worked example on a real recipe used in the study.

An Object Dependency Graph $G = (S, E, P, T)$ for a recipe consists of:

- **States S .** Each state $s \in S$ corresponds to a cooking object-status transition in the sense introduced by OSCAR [113, 114], represented as a grounded [verb]+[noun] pair (e.g., sear + beef fillets, beat + egg whites). A state carries a human-readable label and description; a group (e.g., beef_prep, sauce_cook, assembly) and mode (prep, cook, rest, assembly, finish); a time field with active and passive durations; a quantity dictionary of precise measurements; a list of contextCategories (drawn from the eight categories identified in Chapter 5); and a list of nonVisualCues, each cue tagged with its sensory channel (tactile, auditory, olfactory, thermal, timing) and reliability level (high, medium, low). *States are the data R1, R2, R5, and R7 read from.*
- **Edges E .** A directed edge $s_i \rightarrow s_j$ indicates that s_i is a prerequisite of s_j . Edges carry an optional type (thermal, mechanical, chemical, sequence) and an optional safety string that is read aloud before s_j begins, encoding (for example) that lowering a basket into 350°F oil requires dry hands and a controlled approach. The graph is required to be acyclic. *Edges are the data R3 and R6 read from.*
- **Parallel groups P .** A set of state subsets within which any two states are order-independent and can be interleaved without affecting correctness (e.g., pre-heating the oven while chopping nuts). The system uses P to answer “what can I do while I wait?”. *Parallel groups are the data R4 reads from.*
- **Prep track T .** An ordered list of preparation bundles that must be completed before the main graph begins. Each bundle groups ingredients and tools by workstation (cutting zone, stove zone, pastry zone), and carries tactile and olfactory notes drawn from Chapter 5 for every item. This track operationalizes the “mise en place before you turn anything on” strategy recommended by the rehabilitation instructors in Chapter 4. *The prep track is shared input for R1, R2, and R5.*

An ODG satisfies three structural properties: *acyclicity* (the prerequisite graph is a DAG), *completeness* (every ingredient in the recipe’s ingredient list is consumed by some state), and *groundedness* (every state has at least one nonVisualCue or a timed completion condition, so a blind cook is never asked to rely on a purely visual check).

Figure 7.1 makes the four components concrete on a simplified Chicken Pot Pie, one of the recipes used in the study (Chapter 8) and explicitly chosen for its significant parallelism. The prep track T groups items by workstation; the states S are [verb]+[noun] nodes colored by mode; the prerequisite edges E are directed arrows, with one carrying a safety annotation (read aloud before the dependent state begins); and the dashed envelopes mark two distinct parallel groups in which states can be interleaved without affecting correctness.

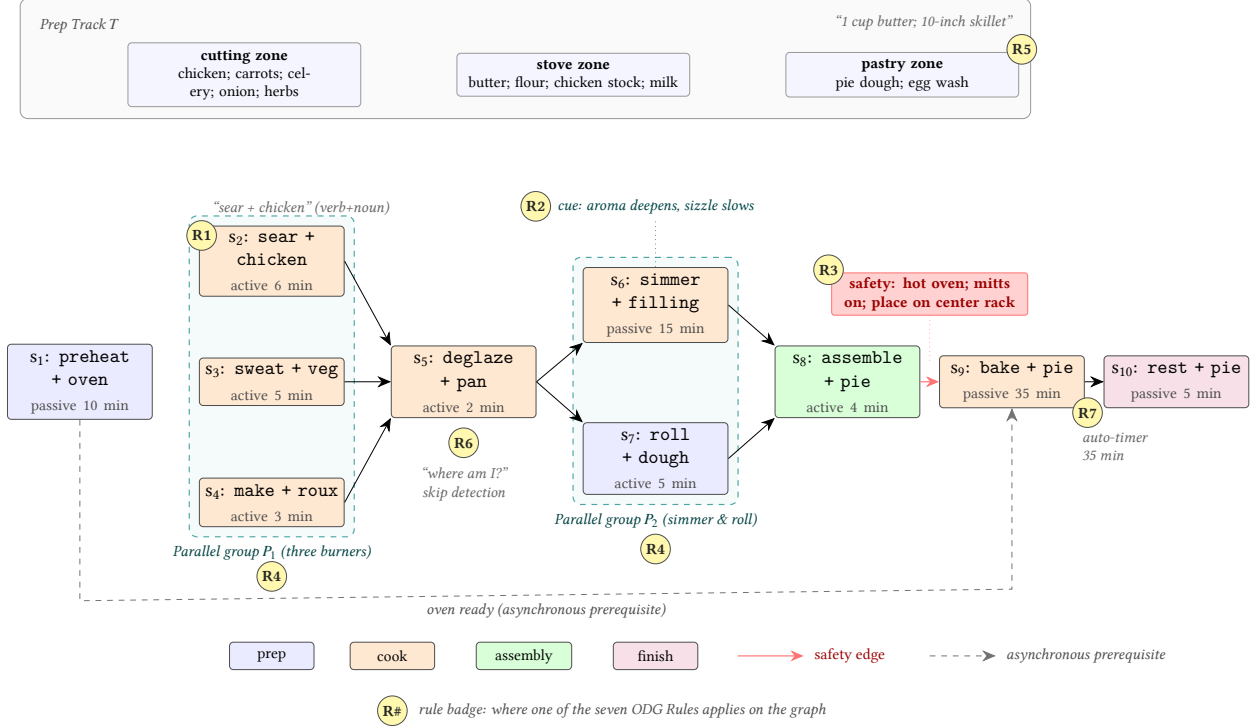


Figure 7.1: An Object Dependency Graph for a simplified Chicken Pot Pie, one of the high-parallelism recipes used in the study (Chapter 8). The prep track T (top, gray) groups ingredients and tools across three workstations; the underlying JSON attaches tactile and olfactory notes to every item. The states S are [verb]+[noun] nodes colored by mode (prep / cook / assembly / finish). The directed edges E encode prerequisites: the three stovetop states s_2, s_3, s_4 converge into a deglaze step s_5 , which forks again into a simmer/dough pair s_6, s_7 , which reconverge into the assembly $s_8 \rightarrow s_9 \rightarrow s_{10}$, where s_{10} is the final rest + pie finishing step before serving. The dashed teal envelopes mark two parallel groups: $P_1 = \{s_2, s_3, s_4\}$ can run across three burners, and $P_2 = \{s_6, s_7\}$ lets the cook roll dough while the filling simmers. The oven preheat s_1 is an *asynchronous prerequisite* of s_9 : it runs from the start of the recipe and must complete before s_9 , but is independent of every other state. The red edge $s_8 \rightarrow s_9$ carries a safety string the assistant reads aloud before sliding the pie into the hot oven. One non-visual cue (under s_6) is rendered inline as an exemplar; every other state carries similar cues in the underlying JSON. The graph is acyclic, every ingredient is consumed by some state (completeness), and every state has at least one non-visual cue or a timed completion condition (groundedness). **The R1–R7 badges (all drawn in a single color for clarity)** mark concrete points on the graph where each of the seven ODG Rules applies: R1 on the verb+noun label of s_2 ; R2 on the inline sensory cue under s_6 ; R3 on the safety tag of the $s_8 \rightarrow s_9$ edge; R4 on each of the two parallel envelopes P_1 and P_2 ; R5 on the prep track (which holds quantities and tool specifications); R6 on the convergence at s_5 (the DAG topology that supports “where am I?” and skip-detection queries); and R7 on the timer annotation under s_9 .

How the Graph Serves the Seven Rules

Having seen each rule in detail, Figure 7.1 can now be read as the literal surface every one of those rules consumes. **R1 (Object Status & Step Narration)** reads the [verb]+[noun] labels inside every state node (“sear + chicken”, “deglaze + pan”, “bake + pie”) and turns them into the assistant’s running commentary. **R2 (Sensory Cues)** reads the per-state `nonVisualCues` field; the inline annotation “cue: aroma deepens, sizzle slows” under s_6 is one example, omitted on the other nodes only for legibility. **R3 (Safety Alerts)** reads the red safety string on $s_8 \rightarrow s_9$: before the assistant says “slide the pie into the oven,” it first says “hot oven; mitts on; place on center rack.” **R4 (Parallel Tasks)** reads the dashed teal envelopes P_1 and P_2 ; when the cook asks “what can I do while the filling simmers?” the assistant returns s_7 (roll + dough) because P_2 declares them concurrent. **R5 (Precise Measurements and Technique)** reads the quantity fields embedded in each state and the items listed in the prep track at top. **R6 (Progress Tracking and Recovery)** reads the DAG topology itself: the three-way convergence into s_5 and the diamond $s_5 \rightarrow \{s_6, s_7\} \rightarrow s_8$ are exactly what the assistant traverses to answer “where am I right now?” or to detect that the cook has skipped a state. **R7 (Automatic Timers)** reads the active/passive time annotations under every state; the long passive 15 min under s_6 is precisely the slot in which R4 then proposes s_7 as concurrent work. The figure is therefore the single object from which seven different consumers (the seven rules introduced earlier in this section) each derive their own behavior. This is the sense in which the ODG is a *substrate*: removing any one rule does not change the graph, and adding new rules in future work would simply mean adding new consumers of the same fields.

Recipes used in the user study (Chapter 8) were annotated in this representation on demand, based on dishes participants identified before their session. Across the ten participants, the resulting recipe set covered a range of complexity levels: Doro Wat (Ethiopian Chicken), Beef Rouladen, and Baked Alaska represent more complex, multi-stage dishes; Jamaican Jollof Rice, Tomato Basil Soup, Shrimp Scampi, and Spaghetti Carbonara represent medium-complexity recipes with multiple but mostly linear steps; Chicken Pot Pie and General Tso’s Chicken cover dishes with significant parallelism and fried components. Two participants (P1 and P4) used Doro Wat, allowing one within-recipe pairing across the cohort.

7.4 Summary

This chapter first introduced seven **ODG Rules** (R1–R7), derived from the four prior chapters of this thesis, as the empirical statement of what a non-visual cooking assistant must do at run time: ground references in [verb]+[noun] object status (R1), replace visual doneness checks with multi-channel sensory cues (R2), alert proactively before hazardous transitions (R3), surface independent parallel work during passive waits (R4), give precise measurements and techniques on demand (R5), maintain a queryable progress model with skip detection and re-orientation (R6), and manage active and passive timers automatically (R7). The chapter then introduced the **Object Dependency Graph** $G = (S, E, P, T)$ as the common substrate from which all seven rules read: a single machine-readable structure whose state, edge, parallel-group, and prep-track fields each correspond to one or more rule behaviors, making the rules independently toggleable without changing the underlying recipe representation.

The next chapter (Chapter 8) describes the implementation of the ODG as a voice-first cooking assistant with a paired researcher dashboard, and reports the results of the within-subjects user study with 10 blind and low-vision participants comparing the ODG-based assistant against a screen-reader baseline.

Chapter 8

ODG System Implementation and User Evaluation

8.1 Overview

This chapter describes the implementation of the ODG as a voice-first cooking assistant with a paired researcher dashboard, and reports the results of the within-subjects user study that evaluates the ODG-based assistant against a screen-reader baseline with 10 blind and low-vision participants.

The system implementation section describes the reasoning engine, voice layer, researcher dashboard, and screen-reader baseline. The user study section details the participants, study design with condition assignments, procedure, task battery, measures, data sources, and analysis plan. The findings section reports the quantitative and qualitative results, including custom Likert items mapped to specific ODG rules and a deductive thematic analysis of the open-ended responses and the closing semi-structured interview; the findings are reported per ODG Rule (R1–R7) and per adoption cell rather than per research question, so that the quantitative Likert evidence and the qualitative interview evidence for each rule and each adoption question appear together.

The discussion then steps above that per-rule reporting and is organized around three higher-level observations across the rules: the gains concentrate where mainstream recipes fail blind cooks, the assistant’s contribution is best understood as coverage and recovery rather than throughput, and personalization is a first-class design dimension that lets the same rule set serve novices, daily home cooks, and experts.

The chapter is organized around three research questions:

- **RQ1 (Experience):** *How* do blind cooks experience an ODG-aware voice-first cooking assistant while following a recipe end-to-end?
- **RQ2 (Rule mechanism):** *How* does each of the seven ODG Rules (R1–R7) shape what the assistant says and when it speaks (atomic [verb] + [noun] decomposition, non-visual sensory cues, edge-level safety, parallel-task suggestions, precise measurements, dependency-aware re-orientation, automatic passive-wait timers), and *why* does each rule matter (or fail to matter) for non-visual cooking practice?

- **RQ3 (Adoption in practice):** *How* would blind cooks fit an ODG-aware assistant into their everyday cooking and *why* do those adoption preferences differ across skill levels, dish complexity, and cooking contexts?

8.2 System Implementation

The ODG is realized as a voice-first cooking assistant with a paired researcher dashboard, deployed as a single Node.js web service that the participant accesses from any modern browser on their laptop, phone, or tablet. This section first describes the implementation architecture, then walks through the two interaction surfaces (the participant-facing assistant and the researcher dashboard), and finally documents the screen-reader baseline used for Condition A.

8.2.1 Implementation Architecture

The system is built around an *input-agnostic* reasoning engine (ODGEngine). The same engine runs whether state updates arrive from a researcher’s Wizard-of-Oz click in the dashboard, from the participant’s own voice (“I’m done searing the chicken”), or, in principle, from an OSCAR-style computer-vision pipeline observing the kitchen. This separation directly reflects a lesson from Chapter 6: the reasoning about *what to say* is decoupled from the sensing of *what has happened*.

Stack and topology. The server (`server.js`, ~2,100 LOC) is a Node.js / Express process that exposes a static client (vanilla JavaScript, no frontend framework), a Socket.IO event bus, two REST endpoints, and a relay to the OpenAI Realtime API. The browser does not hold the OpenAI key and does not connect to OpenAI directly: every audio frame the participant speaks and every audio frame the assistant returns is brokered by the server over a single per-session WebSocket. This routing is what makes the dashboard’s live transcript and rule-toggle mirroring possible: the server is the one place every event passes through, so it can copy each event to both the participant’s socket and the matching researcher socket, and write it to the session log on disk.

Reasoning engine. The ODGEngine JavaScript class (`public/js/odg-engine.js`, 421 LOC) takes a recipe’s ODG and exposes a small set of methods that operate as pure functions of two sets, `completed` and `inProgress`, and the underlying graph: `attemptState`, `markCompleted`, `getAvailableNextStates`, `getInitialStates`, `getPrerequisites`, `getUnmetPrerequisites`, `checkPrerequisitesMet`, `detectSkips`, `queryPrerequisites`, `queryCues`, `queryParallel`, `queryNext`, `getProgressSummary`, and `queryWhereAmI`. No recipe text is read at run time; every spoken reply is grounded in the engine’s structured return value, which in turn is read from the ODG fields described in Chapter 7. The engine is loaded both in the server (where it drives the Realtime API’s tool calls) and in the browser (where the dashboard uses it to render the live state diagram), guaranteeing that participant device, server, and researcher dashboard always agree on what is completed, in progress, and available.

Voice layer. Condition B uses OpenAI’s Realtime API (`gpt-4o-realtime-preview`) for both speech-to-text and speech-to-speech, configured with the coral voice, 16-bit PCM audio in both

directions, Whisper-1 for input transcription, and server-side voice activity detection (`server_vad`, 500 ms threshold, 1,000 ms silence padding). A speech-speed multiplier (default 1.15, adjustable by the moderator between 0.5 and 2.0 mid-session) is forwarded into the session config so the assistant’s delivery matches the participant’s pace preference. Every model turn that asserts a fact about the recipe (which step is current, which are pending, what the quantity is, which parallel branches are open) is routed through one of fourteen function-call *tools* that the server exposes to the model; each tool reads from the engine rather than from the recipe prose. The tools the model can call are:

- Recipe and prep: `get_recipe_overview`, `get_prep_item` (delivers preparation items one at a time so the assistant cannot dump the entire mise-en-place at once).
- Step lifecycle: `start_step`, `complete_step`, `get_step_details`, `check_prerequisites`.
- Structural queries: `get_current_status`, `get_next_suggestions`, `get_parallel_tasks`, `get_progress_summary`, `where_am_i`.
- Timers: `set_timer`, `cancel_timer`, `list_timers`.

Rule-aware prompt composition. Each of the seven ODG Rules (R1–R7) corresponds to (a) a positive prompt block in a `RULE_PROMPTS` table that spells out the behavior the assistant must exhibit when that rule is enabled (for R2, for example, this prescribes that sensory cues, namely what the cook should hear, smell, feel, or time, accompany every cooking step), and (b) a paired negative instruction injected when the rule is disabled (for the same R2 example, this forbids sensory descriptions and restricts the assistant to timing-based instructions). The same prompt-construction code path drives every condition: with all seven rules on, the prompt activates all seven positive rule blocks and no negative instructions; as rules are toggled off, the corresponding positive blocks are replaced with their negative counterparts. The participant’s view of the recipe (an automatically generated step list annotated with attention-demand classifications, a dependencies list, a parallel-groups list, and a `TIMER-ELIGIBLE` list of passive waits with exact durations) is appended unchanged in every condition, so that the only thing varying across rule toggles is the behavioral guidance the model receives.

Server-side timer manager. R7’s automatic timers are implemented server-side with `set-Timeout`: when the assistant calls `set_timer` the server registers the timer, mirrors a `timer-started` event to the dashboard, and, when the timer fires, injects a synthetic user message back into the OpenAI conversation (“time’s up; announce it out loud and tell the cook what to do next”) so the audible alert is delivered as the assistant’s own voice rather than as an out-of-band beep. Timers tied to a step are automatically cancelled when that step is marked complete.

Researcher dashboard. A second web page (`odg-researcher-combined.html`), connected through the same Socket.IO bus, shows the live ODG, the streaming transcript, the running timer set, and a panel of rule toggles. Switching toggles dispatches a `combined-update-rules` event; the server immediately calls `session.update` on the OpenAI Realtime connection with the freshly rebuilt prompt, so a toggle made during cooking takes effect on the very next assistant turn without restarting the session.

Screen-reader baseline (Condition A). A deliberately plain HTML page (`baseline.html`, Figure 8.1) renders the same recipe as linear text, the layout the cook would encounter on a mainstream recipe site. Participants were free to navigate the page however they preferred: either using the page’s built-in Previous/Next/Repeat keys, or using their own screen reader (VoiceOver, TalkBack, NVDA, or JAWS) to traverse the document with its native reading commands. A small platform-rate calibration normalizes the Web Speech API’s per-platform baseline so that the same rate-1.25 setting sounds comparable on iOS, Windows, Android, and macOS. The baseline emits the same logged events as Condition B (navigations, repeats, position changes), so the saved per-condition logs share a common event schema across the two conditions.

Annotated recipe corpus. The system ships with a corpus of nine recipes annotated in the ODG schema described in Chapter 7 (`public/js/recipes.js`); these are the same nine recipes used in the user study and are listed in Table 8.2. Each recipe declares its `prepTrack` (workstation-grouped mise-en-place), `states` (`[verb]+[noun]` cooking transitions with quantities, time, mode, non-visual cues, and safety notes), `edges` (prerequisites with optional safety strings), and `parallelGroups` (pre-vetted concurrent state sets).

Session logging. Every event flowing through the server (session start, condition change, rule toggle, transcript line in both directions, tool call, timer start/fire/cancel, step transition, pause/resume, baseline navigation, participant-device state change) is appended to an in-memory event log keyed by session ID, carrying both an epoch-ms timestamp and an ISO wall-clock copy. When the researcher leaves a condition or stops the session, the server writes a JSON file containing five derived rollups (per-condition summary, chronological phase timeline, pause timeline, complete rules-toggle history, and the raw event firehose) plus per-condition checkpoint files. This logging structure preserves the full session record for post-hoc inspection.

Using the same web stack for both conditions keeps the interaction surfaces comparable and holds the underlying recipe content constant; the only manipulation is whether the cook accesses the recipe through linear text (Condition A) or through the ODG-backed voice assistant (Condition B).

8.2.2 Participant-Facing Assistant (Condition B)

The participant-facing assistant is a voice-first web page (Figure 8.2) that runs on a phone or tablet beside the cook.

The page acquires a Screen Wake Lock on load so the device does not blank mid-recipe (which would otherwise stop the microphone and speech synthesis on iOS) and re-acquires it automatically whenever the page returns to the foreground. The cook speaks to the assistant; the page streams 16-bit PCM audio to the server over Socket.IO; the server forwards each frame to the OpenAI Realtime WebSocket, which performs voice activity detection, Whisper transcription, conversation management, and speech synthesis end-to-end. The participant’s microphone and the assistant’s voice run in parallel, so the cook can interrupt mid-sentence at any time.

Every model turn that asserts a recipe fact is grounded in a tool call against ODGEngine as described above; the model is system-prompted to phrase the response naturally but to read the facts (quantities, current step, available next steps, parallel suggestions, timer durations) from the

ODG Cooking Study Condition A · Screen Reader Recipe Connected Session: RS-P01-fd8v Participant: P00

Condition A: Screen Reader Recipe Read All Stop

BASELINE P00

← Previous SECTION 1 OF 23 **Overview** Next →

Repeat

Keyboard: ← / ↑ Previous, → / ↓ Next, Space Repeat, R Read all, Esc Stop.

OVERVIEW

Chicken Pot Pie

Serves 8 · 1 hr 20 mins

This recipe has **14 ingredients** and **8 steps**.

Press **Next** to start with the first ingredient. You can use the Previous and Next buttons at the top, or the arrow keys.

Ingredients

1. 1 pound Chicken breasts
2. 1 cup sliced Carrots
3. 1 cup Frozen green peas
4. 1/2 cup sliced Celery
5. 1/3 cup Butter
6. 1/3 cup chopped Onion
7. 1/3 cup All-purpose flour
8. 1/2 teaspoon Salt

Figure 8.1: The Condition A screen-reader baseline. The recipe is rendered as a linear, sectioned web page; participants navigated it either with the page’s built-in Previous / Next / Repeat controls at the top, or with their own screen reader’s native reading commands, whichever they preferred. The current section is announced via the position display (“Section 1 of 23: Overview”); the screenshot shows the Chicken Pot Pie recipe at its overview section, with the ingredient list beginning below. This layout deliberately mirrors how blind cooks currently access recipes on mainstream cooking sites.

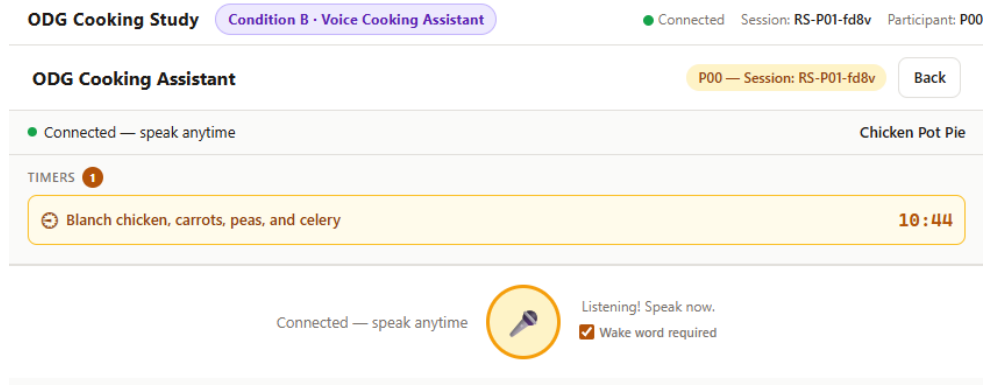


Figure 8.2: The Condition B participant interface. A deliberately minimal voice-first surface: a connection-status banner (“Connected — speak anytime / Listening! Speak now.”), a live mirror of any active assistant-managed timers (here, a “Blanch chicken, carrots, peas, and celery” timer counting down to 10:44), and a single push-to-talk microphone control. The wake-word toggle (“Hey Chef”) lets the participant call the assistant hands-free; pressing the microphone bypasses the wake word. The page also acquires a Screen Wake Lock so the device does not blank mid-recipe.

structured tool return rather than from its own training data. Consider a blind cook preparing Greek Baklava with all seven rules enabled:

- “What should I do first?” The model calls `get_next_suggestions`; the engine returns the initial states (states with no prerequisites). Under R4, the reply lists them as independent options: “You have options: *preheat the oven, prepare the nuts, butter the pan, or take the phyllo out of the fridge. Any of these can be done now.*”
- The cook says “I’ll preheat the oven.” The model calls `start_step`; the engine marks `oven_preheated_baklava` in-progress. Under R7, the model is prompted to ask for confirmation and then call `set_timer` for the state’s passive duration; under R3, the incoming edge’s safety field is surfaced as “Safety: *the oven rack should be in the middle position before you turn it on.*” Under R4, the closing line is “You can also chop the nuts and butter the pan in the meantime; they’re independent.”
- “How do I know when the nuts are ready?” Under R2, `get_step_details` returns the highest-reliability cue from `nonVisualCues`: “You’ll hear the nuts shift from a soft thud to a dry rattle when you stir them, and you’ll smell a warm, toasted aroma. That means they’re done.”
- (Participant steps away for 90 seconds.) “Where am I?” Under R6, `where_am_i` returns the structured re-orientation: “You’ve completed 4 of 18 steps (22 percent). You’re in the preparation phase. You may have skipped: *spice the nuts. Next available: layer phyllo, brush butter.*”
- “How much butter?” Under R5, `get_step_details` returns the quantity field: “Two sticks, melted; one cup total.”

The same questions, asked with rules disabled, reveal the contribution of each rule. With R2 off, the negative-instruction block prevents the model from naming sensory cues and doneness collapses to “cook for 30–40 minutes”; with R4 off, the assistant recommends only one next state and never suggests parallel work; with R6 off, “Where am I?” returns only the current step label

with no skipped-step detection. Because the toggles edit the prompt rather than the engine, the underlying ODG is unchanged across conditions: the engine is always answering correctly, and only the assistant’s license to use that answer is gated.

8.2.3 Researcher Dashboard (Wizard-of-Oz Rule Overlays)

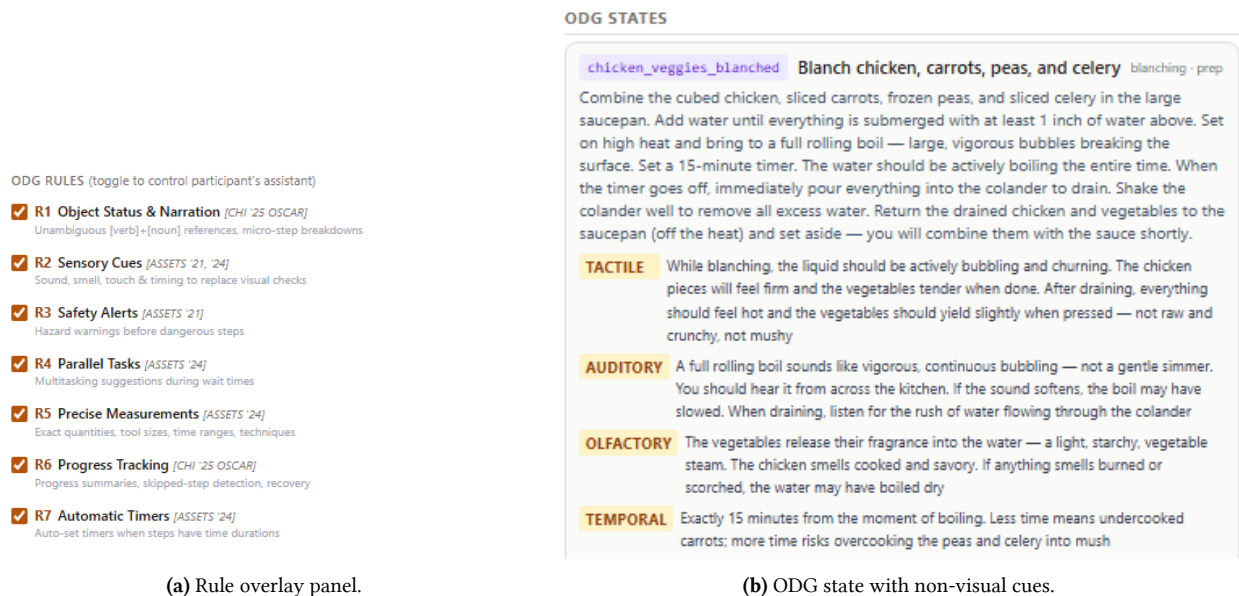
The researcher dashboard (Figure 8.3) runs on a separate device monitored by the moderator during each session. It provides: (1) a session code that the participant joins from their device, with a paired participant link generated automatically; (2) a Condition A / Condition B switcher that drives which page is loaded on the participant device; (3) a live mirror of the transcript between the participant and the assistant, plus a mirror of the active timer set; (4) a live ODG visualization (vis-network) with states colored by cooking group, prerequisite edges drawn between them, and per-node decoration for the live state (completed, in-progress, available, blocked); (5) a panel of seven rule checkboxes that toggle R1–R7 in real time on the participant’s assistant; (6) a speech-speed slider that mid-session re-tunes the Realtime API’s delivery rate; and (7) a Pause/Resume control that clears the OpenAI audio input buffer so the participant can take a break without the assistant continuing to speak. Two of these panels are shown close up in Figure 8.4: the rule-toggle overlay (Figure 8.4a) and a single expanded ODG state with its non-visual cues grouped by sensory channel (Figure 8.4b).

The dashboard serves two roles. As a *demonstration surface* it shows exactly which rule produces which behavior: a viewer can toggle R2 off, ask “how do I know when the nuts are ready?”, toggle it back on, and ask the same question to hear the rule’s contribution in isolation, with the underlying graph unchanged. As a *within-subjects instrument* it lets the moderator drive condition order, recipe selection, and rule configuration for each participant, while every event (condition changes, rule toggles, pause windows, transcript lines, tool calls, timer events, and step transitions) is logged with paired epoch and ISO timestamps so the saved JSON can be sliced by condition or by rule after the fact. The Wizard-of-Oz framing in the subsection title refers specifically to this rule overlay: the assistant’s verbal responses are generated autonomously by the model from the engine’s tool returns; the moderator is wizarding which behaviors are licensed at any moment, not the content of any individual utterance.

The screenshot displays the ODG Researcher Dashboard for a session titled "Chicken Pot Pie". The interface is divided into several panels:

- Left Panel (Recipe Preview):** Contains a list of ingredients (e.g., 1 pound Chicken breasts, 1 cup sliced Carrots) and a "REMOTE SESSION" section with a participant ID (P01) and a copyable link. It also features a "SHOW ON PARTICIPANT" section with a "Waiting" status and a "Voice + ODG" button.
- Top Panel (Session Controls):** Includes buttons for "Stop Session", "Resume Participant", "Save Session Log", "Finish & Save", and "Back".
- Right Panel (Session Active):** Displays the "Active Rules" (R1-R7) and a "PARTICIPANT DEVICE" status block showing various sensors (Socket, Visibility, Mic, Wake lock) and their states (Connected, Visible, Muted, Paused). Below this is an "ACTIVE TIMERS" section showing a timer for "Blanch chicken, carrots, peas, and celery" at 11:22.
- Bottom Panel (Live Transcript):** Shows a "LIVE TRANSCRIPT (READ-ONLY MIRROR)" with a list of rules synced and a session log. It includes a chat interface with participant messages (e.g., "Hello.", "Gathering ingredients", "Jump to cook it") and assistant responses (e.g., "Hello. We're making chicken pot pie. It serves eight and takes about one hour twenty minutes total...").

Figure 8.3: The ODG researcher dashboard (full view). The left panel contains the session controls: a recipe selector at the top, the remote session ID and copy-pasteable participant link, an A / B condition switcher (“Show on participant”), the seven ODG Rule toggles (R1–R7), and, below the toggles, the live preparation track and the list of every ODG state in the current recipe. The right panel mirrors the participant’s runtime: an Active Rules pill row, a live participant-device status block (socket, visibility, microphone, wake lock, assistant, wake word, audio output, paused), the Active Timers list, and the streaming transcript with the participant’s utterances on the right (orange) and the assistant’s replies on the left. Every visible element is driven by the same Socket.IO event stream the server writes to the session log, so what the moderator sees during the study is exactly what the saved JSON reconstructs after the fact.



(a) Rule overlay panel.

(b) ODG state with non-visual cues.

Figure 8.4: Two close-ups of the researcher dashboard’s left panel. (a) The seven ODG Rule toggles. Each rule shows its title (e.g., R2 Sensory Cues), the chapter that motivates it as an italicized citation tag (e.g., *ASSETS* '21, '24), and a one-line description of the behavior it licenses (“Sound, smell, touch & timing to replace visual checks”). Toggling a checkbox rebuilds the assistant’s system prompt server-side and pushes a `session.update` to the OpenAI Realtime connection, so the rule change takes effect on the very next assistant turn. (b) A single ODG state expanded to show the data the assistant reads from. The state header carries the grounded [verb]+[noun] label (`chicken_veggies_blanched` / “Blanch chicken, carrots, peas, and celery”), its cooking mode (`blanching · prep`), and its description. Below the description, the `nonVisualCues` list is grouped by sensory channel (TACTILE, AUDITORY, OLFACTORY, and TEMPORAL), exactly as it is consumed by R2 at run time.

8.3 User Study

To evaluate the ODG-based assistant, I conducted a within-subjects remote study with **10 blind and low-vision participants** (P1 through P10) comparing it to a screen-reader recipe baseline. The sample size of 10 participants is consistent with the sample sizes used in the prior chapters of this thesis and is sufficient to surface recurring qualitative patterns within each cell of the deductive analysis framework (the seven ODG Rules and the three adoption sub-themes, Section 8.3.4) while providing directional quantitative signal from the custom Likert measures.

8.3.1 Participants

I recruited 10 adult participants who are blind or have low vision and cook at least occasionally, through mailing lists, social media, and snowball sampling from participants of the studies in Chapters 3, 4, and 5. Participants were compensated at the rate used in prior chapters. Sessions were conducted remotely over video call: the participant used the screen reader and voice assistant already installed on their own devices, and the moderator used the researcher dashboard to drive conditions and record the session.

The 10 participants (P1–P10) spanned a wide range of ages (24–70), gender, vision levels (legally blind to totally blind), onset (congenital to recent onset), cooking frequency (once a month to daily), and prior experience with voice assistants (Table 8.1). Cooking experience also varied substantially, from P8, who began cooking in 2023 at a vision rehabilitation center, to P2, a former chef of decades’ experience, and this variation is reflected in the qualitative findings reported below.

8.3.2 Design, Conditions, and Recipes

The study used a within-subjects design over the factor *access method* with two levels:

- **Condition A (Baseline, Screen Reader).** The recipe is presented as a standard web page. Participants were free to navigate it however they preferred, either with the page’s built-in Previous/Next/Repeat keys, or with their own screen reader (VoiceOver, TalkBack, NVDA, or JAWS) using its native reading commands. This represents the most common way blind cooks access recipes today [117]: linear text, navigated by the participant’s own device and preferences.
- **Condition B (ODG Assistant, Voice First).** The same recipe is presented through the voice-first ODG assistant with all seven rules (R1–R7) enabled throughout.

Condition order was counterbalanced across participants using a systematic assignment based on participant number: odd-numbered participants (P1, P3, P5, P7, P9) completed Condition B first, then Condition A; even-numbered participants (P2, P4, P6, P8, P10) completed Condition A first, then Condition B. This ensures approximately balanced order effects across conditions. Each participant completed the same task battery under both conditions using the same recipe, allowing direct within-participant comparison of the two assistance modalities.

Table 8.2 shows the full participant assignment. Before the study, participants were asked to identify dishes they wanted to learn or improve upon; those recipes were then extracted from

PID	Age	Gender	Vision Level (Onset)	Cooks	Cook Conf.	VA Comfort	Platform
P1	34	F	Legally blind, finger-counting only (2019)	Several / week	7	7	Windows / JAWS
P2	70	M	Legally blind, congenital	Daily	7	7	macOS / VoiceOver
P3	64	M	Totally blind, no light perception (age 22)	Weekly	5	7	Windows / JAWS
P4	41	M	Totally blind (2014)	Daily	7	6	macOS / VoiceOver
P5	44	F	Legally blind, light perception (age 17)	Several / week	6	7	iOS / VoiceOver
P6	51	F	Legally blind, light perception (congenital, premature)	Several / week	6	6	Windows / JAWS
P7	40	F	Legally blind, light perception (congenital)	Daily	6	7	iOS / VoiceOver
P8	24	F	Legally blind, minimal light perception (congenital)	Once / month	5	3	Windows / JAWS
P9	42	F	Totally blind (2010); legally blind since birth	Several / week	6	7	macOS / VoiceOver
P10	53	F	Totally blind, no light perception (~10 years ago)	Weekly	5	7	iOS / VoiceOver

Table 8.1: Demographic and background information for the 10 study participants. *Cook Conf.* is self-rated cooking confidence on a 1–7 scale (“How confident are you cooking a new recipe independently?”); *VA Comfort* is self-rated comfort with voice assistants on a 1–7 scale.

Participant	Condition Order	Recipe Used
P1	B then A	Doro Wat (Ethiopian Chicken) [†]
P2	A then B	Jamaican Jollof Rice [†]
P3	B then A	Baked Alaska [†]
P4	A then B	Doro Wat (Ethiopian Chicken) [†]
P5	B then A	Tomato Basil Soup [†]
P6	A then B	Beef Rouladen [†]
P7	B then A	Shrimp Scampi [†]
P8	A then B	Spaghetti Carbonara [†]
P9	B then A	Chicken Pot Pie [†]
P10	A then B	General Tso's Chicken [†]

Table 8.2: Participant assignment to condition order and recipe. Odd-numbered participants (P1, P3, P5, P7, P9) completed Condition B (ODG assistant) first; even-numbered participants (P2, P4, P6, P8, P10) completed Condition A (screen reader) first. Each participant used the same recipe under both conditions. Recipe names marked with [†] are hyperlinks to the public source recipe that was extracted and adapted into the ODG structure.

publicly available sources (Table 8.2) and adapted into the ODG structure. Each participant used the same recipe under both conditions to allow direct within-participant comparison of the two assistance modalities.

8.3.3 Procedure

Each session took approximately 60 minutes and proceeded in four phases:

1. **Welcome, Setup, and Pre-Study Questionnaire [5 min].** The moderator welcomed the participant, explained the study purpose, obtained verbal consent, helped the participant test their audio and screen reader setup, and administered the demographics, cooking background, and technology comfort items (Section 8.3.3.1).
2. **First Condition + Post-Task Questionnaires [20 min].** The participant completed the task battery under their first assigned condition, followed by the post-task questionnaire for that condition (Section 8.3.3.3).
3. **Second Condition + Post-Task Questionnaires [20 min].** The participant completed the task battery under the second condition, followed by the post-task questionnaire for that condition. The two conditions ran back-to-back with no scheduled break.
4. **Closing Semi-Structured Interview [15 min].** The moderator conducted the closing interview (Section 8.3.3.4).

8.3.3.1 Pre-Study Questionnaire

Administered once at the start of the session (Demographics + Cooking Background):

- What is your age?
- What is your gender?
- Can you please describe your vision level? When was the onset?
- How often do you cook? (daily / several times a week / weekly / rarely / never)
- How do you currently access recipes?
- When did you start cooking?
- On a scale of 1–7, how confident are you cooking a new recipe independently?
- On a scale of 1–7, how comfortable are you using voice assistants?
- What device do you primarily use to access recipes? (iPhone / Android / iPad / smart speaker / laptop / other)

8.3.3.2 Task Battery (Identical Under Condition A and Condition B)

To isolate the effect of the access method from the effect of the recipe, the participant was given the same six tasks under both conditions. The moderator’s prompt to the participant was: *“Now you will be testing two ways of accessing recipes; I want you to try both and let us know how you like it. You can assume that you are cooking, but for some tasks you can pretend you finish instantly.”*

1. **Task 1 (Orientation; probes R1, R6):** Please go through the system and briefly describe what information is available to you.
2. **Task 2 (Measurement; probes R5):** Find ingredient #3 and tell me the amount.
3. **Task 3 (Forward navigation; probes R6):** Go to step 4 and read what it says.
4. **Task 4 (Backward navigation; probes R6):** Go back to step 2.
5. **Task 5 (Prerequisite reasoning; probes R1 dependencies):** Now let’s navigate to the second-to-last cooking step. What needs to be ready or finished before you can start that step?
6. **Task 6 (Parallel tasks and timing; probes R4 and R7):** Imagine you’ve just started boiling water (which takes a long time and runs by itself). What else could you safely do during that time? Please set a timer.

The tasks were designed to probe specific ODG rules. Task 1 is an open orientation probe that surfaces how each condition conveys recipe structure (R1, R6). Task 2 probes R5 (precise measurements). Tasks 3 and 4 probe R6 (forward and backward progress tracking). Task 5 probes the participant’s ability to reason about edges in the dependency graph (R1) by asking what must be complete before a specific later step. Task 6 jointly probes R4 (parallel tasks during passive waits) and R7 (the automatic-timer mechanism).

8.3.3.3 Post-Task Questionnaires (Administered After Each Condition)

After each condition, the participant completed the following instruments verbally.

Custom Likert Scales (1–7, Strongly Disagree to Strongly Agree). Each item is mapped in Table 8.3 to the rule it is designed to probe.

- *Understanding & Orientation.*
 - I understood what I needed to do at each step.
 - I knew which step I was on and how much was left.
 - I could easily find my way back if I lost track.
- *Safety & Confidence.*
 - I would feel safe performing the cooking steps while using the system.
 - I was aware of potential hazards before encountering them.
- *Ingredient & Measurement Clarity.*
 - I always knew exactly which ingredient was being referred to.
 - The quantities and measurements were clear enough to follow.
- *Task Management.*
 - I understood when I could do multiple things at the same time.
- *Overall.*
 - I would use this system for cooking.
 - I would recommend this system to a blind friend who cooks.

Open-Ended Questions (per condition).

- What was the most helpful thing about this system?
- What was the most frustrating or confusing moment?
- Was there anything you wished the system had told you but didn't?

8.3.3.4 Closing Semi-Structured Interview

After both conditions and their post-task questionnaires, the moderator conducted a semi-structured interview organized around the seven ODG Rules and the system's overall use, with the rule-specific themes mapped one-to-one to the rule each is designed to interrogate:

- *Object Status (probes R1).*
 - How would you describe the amount of information in each recipe step? Was it about right, too much, or too little?
- *Sensory Cues (probes R2).*

Custom Likert Item	Probes Rule(s)
I understood what I needed to do at each step.	R1, R2
I knew which step I was on and how much was left.	R6
I could easily find my way back if I lost track.	R6
I would feel safe performing the cooking steps while using the system.	R2, R3
I was aware of potential hazards before encountering them.	R2, R3
I always knew exactly which ingredient was being referred to.	R1, R5
The quantities and measurements were clear enough to follow.	R5
I understood when I could do multiple things at the same time.	R4, R7
I would use this system for cooking.	Overall
I would recommend this system to a blind friend who cooks.	Overall

Table 8.3: Mapping from each custom Likert item to the ODG Rule(s) it is designed to probe. Five items map to a pair of rules because a single Likert dimension is supported by more than one rule in the assistant: *understanding what to do* is jointly driven by R1 (step instructions) and R2 (the sensory cues that tell a blind cook how to execute and verify each step); the two *safety* items are driven by R3 (explicit safety annotations on edges) together with R2 (sensory hazard signals such as the smell of burning or the sound of oil spattering); *ingredient identification* is driven by R1 (the atomic [verb] + [noun] reference to the current object) together with R5 (the precise quantity attached to that object); and *multi-tasking* is driven by R4 (parallel-task suggestions) together with R7 (the passive-wait durations that make the R4 suggestions actionable). This mapping is used to attribute differences between conditions to specific rules. R2 is also probed directly by the closing interview’s Sensory Cues theme and indirectly by Task 6 (the boiling-water parallel-and-timer probe).

- What was your experience with the non-visual cues (sounds, smells, textures, timing) that the assistant mentioned, if any?
- *Safety (probes R3).*
 - How did each system handle safety-related information during the cooking task?
- *Parallel Tasks (probes R4).*
 - How did you experience the suggestions for what to do while waiting on a passive step?
- *Precise Measurement (probes R5).*
 - Tell me about the information each system gave you on tools, time ranges, and ingredient quantities.
- *Error Recovery (probes R6).*
 - Have you ever made a mistake or skipped a step while cooking? If so, can you describe what happened?
 - How did each system handle the moments when you had to recover or re-orient?
- *Timer (probes R7).*
 - How do you normally handle timers when you cook?
 - What was your experience with the assistant’s automatic timers?
- *Future Use (overall).*
 - For what kinds of recipes, if any, can you see yourself using each system?
 - Is there anything else you’d want a cooking assistant to do that neither system did?

8.3.4 Data Sources and Analysis

Each session produced four streams of data: (1) pre-study questionnaire responses (demographics, cooking background, technology comfort); (2) the ten custom Likert ratings per condition per participant ($n = 10 \times 2 \times 10 = 200$ ratings in total); (3) open-ended responses after each condition plus the closing semi-structured interview transcript; and (4) the per-session JSON event log written by the server (every Socket.IO event flowing through the session, plus the five derived rollups described in Section 8.2.1: per-condition summary, phase timeline, pause timeline, complete rule-toggle history, and the raw event firehose). Interview audio was recorded with participant consent and transcribed verbatim. The analysis triangulates these four streams across the three research questions.

Quantitative analysis (Likert-scale ratings). Given the within-subjects design with $n = 10$, descriptive statistics (mean and standard deviation) were computed for each Likert item under each condition. Rather than inferential test statistics, which would not be meaningful at this sample size, I report these descriptive statistics together with the within-participant gap between conditions for each item; any differences are read as directional rather than conclusive.

Qualitative analysis (open-ended responses and interview). The open-ended responses and closing semi-structured interview transcripts were analyzed with a *deductive* thematic analysis [30], because the interview instrument was already structured around the seven ODG Rules (the closing-interview script in Section 8.3.3.4 has one question block per rule: Object Status, Sensory Cues, Safety, Parallel Tasks, Precise Measurement, Error Recovery, Timer; plus a Future Use block addressing RQ3) and the analytical categories were therefore fixed before the data was read. The pre-specified framework had ten cells: one per ODG Rule (R1–R7) plus three for RQ3 (*when the assistant is the strongest primary tool*, *complementary uses of the screen reader*, and *adoption preconditions*). Coding proceeded in two passes: first, every relevant utterance from each participant was charted into the framework cell it belonged to (an utterance could be charted into more than one cell when it bridged rules, e.g., a sensory hazard remark would be charted into both the R2 and the R3 cells); second, the charted data was read across participants per cell to identify the recurring patterns within each cell. A pattern was reported when it recurred across at least four of the ten participants within a framework cell or across closely related cells. The findings are then reported per ODG Rule (Section 8.4.3) and per adoption cell (Section 8.4.4), so that the quantitative and qualitative evidence for each rule and each adoption question appears together. Quotations are reproduced verbatim from the transcripts with minimal editing for readability (filler words collapsed; bracketed clarifications added only where the referent would otherwise be ambiguous). All coding was performed by the author and the chapter does not claim inter-rater reliability; the deductive framework is what constrains coder drift here, since the seven rules and the three RQ3 sub-themes are fixed externally to the data, so the analytical surface is the framework itself rather than an emergent codebook.

8.4 Findings

The user study yields four kinds of evidence, reported in the four subsections that follow:

- *Recipe corpus and rule coverage* (Section 8.4.1). The naturalistic spread of participant-chosen recipes across procedural complexity and ODG-rule exposure, which determines how the rule-level analyses below should be read.
- *Aggregate Likert comparison* (Section 8.4.2). Whether the ODG-based assistant (Condition B) improves the participant experience relative to the screen-reader baseline (Condition A) on the custom Likert items, with each item mapped via Table 8.3 to the rule (or pair of rules) it probes.
- *Preferences for the ODG rules* (Section 8.4.3). For each of R1–R7 in turn, the per-rule Likert gap reported together with the qualitative interview evidence charted into the corresponding cell of the deductive thematic-analysis framework, so quantitative and qualitative evidence for each rule appear together.
- *Preferences for adopting the assistant in cooking scenarios* (Section 8.4.4). The cooking contexts in which participants would adopt the ODG-based assistant as their primary tool, where they would retain the screen reader as a complementary fallback, and the onboarding, verbosity, persona, and per-rule customization controls they requested.

8.4.1 Recipe Characteristics and Rule Coverage

Because the recipes were chosen by the participants rather than by the researcher (Table 8.2), they offered a naturalistic spread of cooking styles. The shape of that spread, and the rule coverage it produced across the ten sessions, is itself one of the first findings from the study, because it determines how the per-rule analyses below should be read. I therefore characterize the nine annotated recipes (P1 and P4 shared Doro Wat) along two dimensions: *procedural complexity* and *which ODG rules they exercise most heavily*.

Along the complexity axis, the recipes form three clusters:

- *High-complexity, multi-stage recipes.* Doro Wat, Beef Rouladen, and Baked Alaska each involve a long preparation phase that runs in parallel with a long passive cook or chill phase. They stress **R4 (parallel tasks)** because there are genuine independent branches a sighted cook would naturally interleave, and they stress **R6 (progress tracking)** because the participant has to remember where they are after a long passive wait.
- *Medium-complexity, more linear recipes.* Jamaican Jollof Rice, Tomato Basil Soup, Shrimp Scampi, and Spaghetti Carbonara are dominated by short, timing-sensitive steps. They stress **R7 (automatic timers)** and **R1 (object status)** because doneness is determined by transitions that a sighted cook would judge visually (rice grains separate, shrimp curl and turn opaque, carbonara eggs thicken without scrambling), and they stress **R2 (non-visual sensory cues)** because text recipes typically describe these moments with color words.
- *High-parallelism, hazard-rich recipes.* Chicken Pot Pie and General Tso's Chicken combine a filling/sauce prep with a high-temperature finish (oven for pot pie, deep frying for General Tso). They stress **R3 (safety alerts)** most directly, since hot oil and a hot oven are both common cooking hazards highlighted in Chapter 3, and they also exercise **R4** and **R5 (precise measurements and technique)** because oil temperature, dredging order, and seal-the-crust technique all have to be communicated explicitly.

Aggregated across the ten sessions, the nine recipes cover all seven ODG rules at least three times: every participant's recipe touched R1, R5, and R6; eight of ten touched R2 and R7; six touched R4; and four touched R3 directly through frying, broiling, or oven-handling steps. This coverage is what allows the per-rule Likert analysis below to be read as evidence about the rules themselves, rather than as evidence about a single recipe, and it explains why the qualitative patterns identified under the deductive framework (e.g., parallel-task awareness, recovery after a long passive wait, oil-temperature safety) recur in roughly the proportions in which the recipes load on each rule.

8.4.2 Likert Ratings by Condition and Rule

This subsection reports the Likert ratings under the analysis protocol described in Section 8.3.4: descriptive statistics per condition (mean and standard deviation across the ten participants) together with the within-participant gap between conditions for each item, read as a directional indicator rather than as an inferential test given the sample size. The within-subjects design means each participant contributes a paired observation per item, so Condition A and Condition B are compared on the same person rather than on two independent groups; with $n = 10$,

statistical inferences are framed as directional rather than conclusive, and the rule mapping in Table 8.3 is used to read each item as evidence about the rule (or pair of rules) it probes rather than as a standalone outcome. Items that map to the same rule are summarized together to assess the rule-level effect, and the one item that trends against Condition B is reported alongside the items that favor it rather than averaged into a single composite, so that the direction and magnitude of every per-item gap remains visible.

Figure 8.5 summarizes mean Likert scores (1 = Strongly Disagree, 7 = Strongly Agree) for each condition, with error bars showing standard deviation. Across all 10 participants, Condition B (ODG assistant) scored higher than Condition A (screen reader) on safety-related items (felt safe: $M=6.9$ vs. $M=5.8$; warned about hazards: $M=6.5$ vs. $M=3.4$) and on task management (parallel tasks: $M=6.6$ vs. $M=5.5$). The two conditions are roughly comparable on understanding and orientation, with one R6 item (knowing which step one is on) trending against Condition B ($M=5.7$ vs. $M=6.2$), which I attribute to the screen-reader baseline still providing the cook with random-access awareness of the full recipe length.

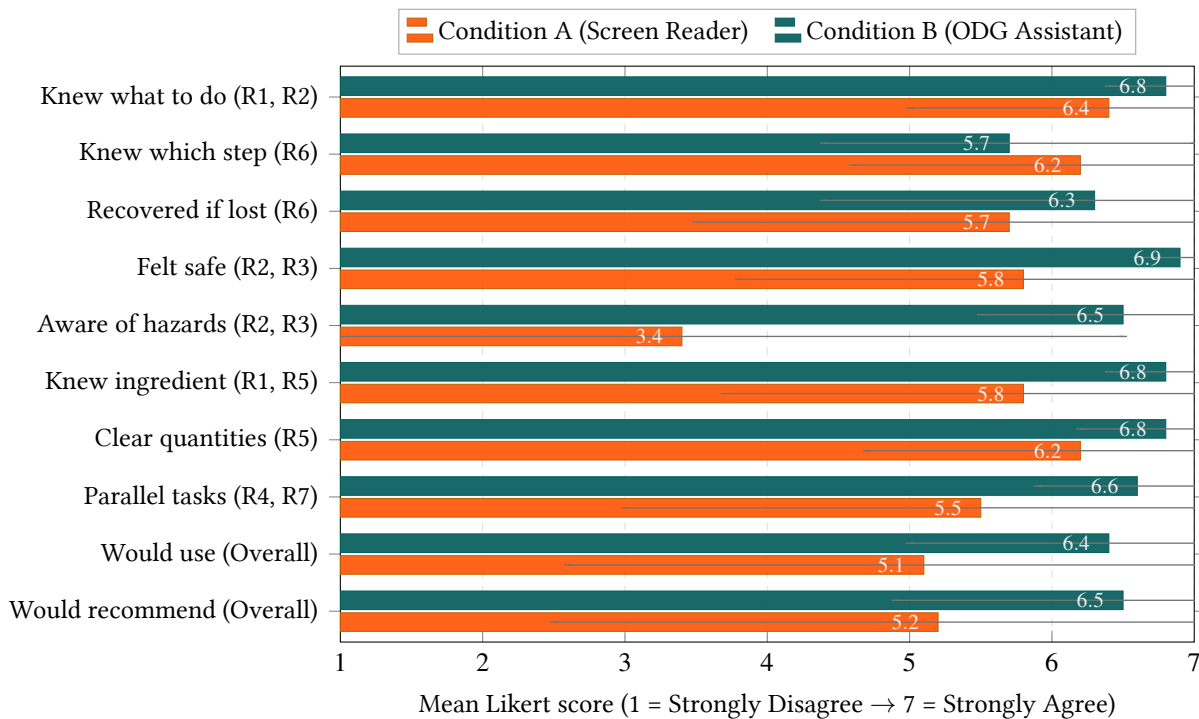


Figure 8.5: Custom Likert scores (1–7, mean $\pm$ SD) by condition and rule mapping, $n = 10$ participants. Each item pairs Condition A (screen reader, top bar) with Condition B (ODG assistant, bottom bar); the probed rule appears in parentheses. The bar length is the group mean (printed inside the bar) and the horizontal whisker is one standard deviation. The largest gap appears on the safety items (*aware of hazards*: 3.4 vs. 6.5; *felt safe*: 5.8 vs. 6.9), with smaller but consistent advantages for Condition B on task management (*multiple things at once*: 5.5 vs. 6.6) and overall adoption (*would use*: 5.1 vs. 6.4). One R6 item (*knew which step I was on*) trends in the opposite direction (6.2 vs. 5.7), reflecting that the screen-reader baseline still affords stronger document-level situational awareness than the voice assistant. The systematically shorter whiskers under Condition B on most items also indicate that responses to the ODG assistant were less variable across participants than responses to the screen reader.

8.4.3 Preferences for the ODG Rules

This subsection reports the per-rule findings: for each of R1–R7 in turn, the quantitative gap (from the Likert items mapped to the rule in Table 8.3) is reported together with the qualitative evidence charted into the corresponding cell of the deductive thematic-analysis framework (Section 8.3.4). Rules are presented in order of the strength and breadth of the evidence rather than in numerical order.

8.4.3.1 R3 (Safety) Is the Rule with the Largest Demonstrated Benefit

R3 produces the single largest condition gap in the study on both safety items (*aware of potential hazards*: 3.4 → 6.5; *felt safe*: 5.8 → 6.9), and is the rule participants most consistently described as making them feel competent and protected in the kitchen. R3’s proactive hazard annotations on recipe edges — delivered *before* the cook reaches the dangerous step rather than as an after-the-fact warning — are not present in mainstream recipes at all, which is exactly why the screen-reader baseline could not match them. The interview data is unambiguous about how valuable participants found this: P3 (who scored Condition A 1/7 on *hazards*) said “safety is my number one concern” and called the ODG’s safety cues “little precise steps that can be very precious”; P5 said “safety cannot be reiterated or not” (i.e., cannot be over-reiterated); P7 liked that safety features were “built in” for cooks earlier in their journey; P8 said the safety cues helped her “feel more comfortable” with cooking methods she was unfamiliar with. In-session capture showed R3 also working as confirmation for cooks who already practice good habits: when the assistant reminded P5 to keep a pan handle turned inward, she replied “oh, that’s perfect that you said for me to keep the handle inward. I always do that.” R3 is therefore both a learning aid for newer cooks and a reassurance layer for experienced ones — the rule with the broadest reach across the population.

8.4.3.2 R4 (Parallel Tasks) Recovers Passive-Wait Windows for Productive Work

R4 produces a clear advantage on the *multiple things at once* item (5.5 → 6.6) by turning what would otherwise be dead waiting time into structured opportunities to clean utensils, prep the next ingredient, or stage the next station. This matters more for blind cooks than for sighted ones, because a sighted cook can scan the kitchen at a glance to spot the next useful task, while a blind cook depends on the recipe itself to surface that opportunity. Participants embraced exactly this framing: P3 said “it is a great idea, I am the one like to best use my time, so I can clean the utensils while waiting”; P1 and P5 likewise named concrete parallel work the assistant unlocked, and described the rule as enabling independence rather than just productivity. The right design implication is that R4’s [isOptional] flag should be honored at the surface level so the suggestions remain inviting rather than directive — a customization concern (taken up in Section 8.4.4) rather than a problem with the rule itself.

8.4.3.3 R1 (Object Status) Makes Recipes Followable While Cooking by Atomizing Long Steps

R1 produced a clear advantage on both items it maps to: *knew what to do* (6.4 → 6.8) and *knew which ingredient/utensil I needed* (5.8 → 6.8), and it is the rule that underwrites the conversa-

tional, step-at-a-time experience participants praised. Mainstream recipe steps are written for sighted readers who can re-scan a long sentence at will; they routinely bundle several physical actions and multiple ingredients into one paragraph. A screen reader plays that bundle back as a single uninterrupted utterance, which leaves a blind cook reconstructing the action sequence from memory in the middle of an active kitchen. R1’s decomposition into atomic [verb] + [noun] units — “pick up the small saucepan,” “add the room-temperature butter,” “stir until combined” — removes that memory burden entirely: each utterance is one physical action on one named object that the cook can execute and confirm before the next one arrives. Participants asked for this directly: P3 wanted ingredients “separated step-by-step” rather than bundled and called the ODG assistant “a cooking teacher” walking him through one move at a time; P2 said the voice “is more fluent, I can move back and forth” through the atomized steps; P10’s “if you give me too much info at once, I will stop listening” is a statement of exactly the constraint R1 is designed to satisfy. The larger gap on *knew which ingredient/utensil* (5.8 vs. 6.8) is consistent with this reading: with atomized steps the relevant ingredient is always the most recent noun the assistant has spoken, never a name the cook must hold from earlier in a paragraph.

8.4.3.4 R7 (Timer) Externalizes Elapsed Time So the Cook Can Plan Around It

R7 was useful as a participant-controlled *tracker* that externalizes “I have started this” and “approximately this long has elapsed,” freeing the cook from holding multiple parallel durations in working memory while cooking. Participants used R7 exactly this way: P10 articulated the value directly — “if I know something will be cooking for 10 minutes, I can easily figure out by myself” — meaning that once the assistant tells the cook the duration, the cook can confidently plan parallel work around it (and R7 thereby makes R4’s parallel-task suggestions actionable). Cooks naturally combined this tracker affordance with their own non-visual doneness judgment, reading state from sound, smell, and touch and dismissing the timer once those cues had advanced, which is the right division of labor for the kitchen: the system holds the clock, the cook holds the judgment.

8.4.3.5 R6 (Progress and Recovery) Is the Rule That Recovers Blind Cooks from Real, Common Errors

R6’s most valuable contribution is its *recovery* function: when a cook loses their place, skips an ingredient, or realizes mid-step that something has gone wrong, R6’s dependency-aware re-orientation tells them exactly where they are and what to do next. The corresponding Likert item (*recovered if lost*) registers a clear advantage for Condition B (5.7 → 6.3), and the interview evidence behind that gap is striking: every participant who was asked about errors reported routine skips and mistakes that the screen reader could not help with — P3 (“happened a lot, went too fast”), P6 (“yep, we all have done that”), P7 (burned garlic and compensated by adding water), P8 (“plenty of times,” including a forgotten cinnamon discovered only by feeling the jar still on the counter), P9, P10. Participants treated R6’s re-orientation as the assistant earning its keep precisely when things went wrong, with P3 describing the experience as having “a cooking teacher” on hand. The companion item (*knew which step*: 6.2 → 5.7) trends in the opposite direction not because R6 fails at recovery but because the prototype did not yet expose multi-step backward

navigation cleanly (P10: “I am on step 5, to say go back to step 3”); that gap is an interface affordance to be added, not a content limitation of the rule.

8.4.3.6 R2 (Sensory Cues) Is the Rule Participants Endorsed Most Enthusiastically in Their Own Words

R2 collected the strongest unprompted endorsements in the entire interview corpus, and is what made cooks feel that the assistant was genuinely designed *for non-visual cooking* rather than retrofitted from a sighted reading experience. P3 called the tactile cue for egg-white stiffness “priceless” and said thermal and “golden brown” cues let him know when the meringue was ready without sight; P5 said “non-visual cues is the way to go. I always want to recommend to non-visual participants,” and emphasized cues that are “independent of vision”; P6 said the cues gave her “more details of how to do the steps”; P8 (cooking only since 2023) said they were “incredibly helpful every time” for a newer cook. The unprompted P5 affirmation captured the rule’s distinctive value with unusual precision: after the assistant described a doneness state as something the cook “should feel soft,” she responded mid-task, “wow, that’s a good way of explaining it that it should feel soft.” Beyond doneness signaling, R2 also functions as hazard signaling in its own right (smell of burning, sound of oil spattering, heat radiation from a hot pan), which is one reason the R3 safety items show such a large gap in Condition B.

8.4.3.7 R5 (Measurements) Provides On-Demand Precision the Cook Can Pull at the Moment of Action

R5 scored highly on both conditions (*clear quantities*: $6.2 \rightarrow 6.8$), and that already-high baseline is itself the point: mainstream recipes do reliably include numeric quantities, so R5’s value in Condition B is not to introduce measurements but to make them *retrievable on demand* at the moment the cook actually needs them. Asking “how much butter?” mid-step and getting an immediate answer is materially easier than scrolling back through the ingredient list with a screen reader while the pan is heating, and the gap on the Likert item reflects this affordance even though the underlying quantities are the same. R5 thus closes a small but consistently useful gap: it converts a piece of information that was technically available into a piece of information that is operationally at hand.

8.4.4 Preferences for Adopting the ODG-Aware Assistant in Cooking Scenarios

This subsection reports the findings from the three RQ3 cells of the deductive thematic-analysis framework: when participants said the ODG-aware assistant was the strongest primary tool, when they retained a complementary role for the screen reader, and the personalization and customization needs that emerged across the interviews.

8.4.4.1 Broad Adoption Willingness Across Participants

Participants endorsed the ODG-aware assistant strongly and consistently. Eight of the ten participants rated *would use again* at 6/7 or higher for Condition B (mean 6.4), and the same pattern

held for *would recommend to a blind friend* (mean 6.5); P5 gave the maximum score of 7/7 on both adoption items. The interview data carried the same message: P2 said “I would definitely use” the assistant and “would recommend it to a blind friend who cooks”; P3 went furthest with “voice can fully substitute the screen reader one” for any recipe; P8 said the assistant gave her the confidence to attempt cooking methods she had previously avoided. Adoption willingness was not contingent on novelty or curiosity — it was grounded in concrete benefits participants had just experienced under the seven ODG rules in their own kitchens.

8.4.4.2 Primary-Tool Conditions: Unfamiliar, Complex, and Hands-Occupied Cooking

Participants converged on the use cases where the ODG-aware assistant most obviously outperforms a screen-reader workflow: unfamiliar recipes (P2, P3, P8), complex or baking recipes that require precision and dependency tracking (P3, P6, P8), and any moment when the cook’s hands are occupied with the food rather than free to hold a phone or scrub through text (P1, P4, P5). These are exactly the conditions under which the seven ODG rules contribute the most — R3 safety annotations on edges, R2 sensory cues for unfamiliar doneness states, R1 atomized steps for procedures the cook has not internalized, R6 recovery for the slips that complex recipes invite, R4 and R7 for the multi-stream parallelism that baking and multi-pot cooking require. In other words, the assistant’s primary-tool reach grows exactly with the cooking complexity that mainstream recipe formats handle worst for blind cooks.

8.4.4.3 Complementary Use of the Screen Reader

For recipes the cook has already mastered — simple boils, routine bakes, dishes cooked many times before (P6, P7, P8: “boil methods and baking”) — the screen reader remains a serviceable lightweight fallback, mostly because the cook no longer needs the rules to fire: the safety habits, doneness signals, step decomposition, and recovery strategies are already internalized. The picture this gives is not a competition the assistant partially loses, but a deployment in which the ODG-aware assistant is the primary tool that grows with the cook’s repertoire while the screen reader fills a narrowing complementary slot for the dishes the cook has fully mastered. Several participants explicitly described their imagined long-term use in exactly these terms, naming the assistant as the tool they would reach for whenever the cooking went beyond routine.

8.4.4.4 Onboarding and Verbosity Controls

Three independent participants asked, unprompted, for a discoverability layer at session start (P8, P9, P10), and several more requested controllable speed, persona, and verbosity (P1, P2, P4, P5, P7). The right reading of these requests is not that the assistant is hard to use — participants gave it high marks anyway — but that a structured onboarding tour and a global verbosity/persona control would let an even broader audience reach the same benefits on first contact. These are growth affordances for the system, not corrective patches: they widen the population for whom the ODG-aware assistant is approachable from the very first session without changing the rule set that already produces the gains documented in Section 8.4.3.

8.4.4.5 Personalization Across Expertise Levels

The two participants who gave Condition B lower adoption scores than Condition A (P4, an expert daily cook who started at age 5; P10, a linear-paced reader who reported ADHD) endorsed the underlying ODG rule content but asked for control over how that content is delivered. P4’s request was for restraint in tone and density — expert cooks “have our personal ways of understanding what was going on,” so rule output that was well-calibrated for less experienced participants registered as patronizing (“feels like a mother standing over you”). P6 (cooking since age 6–7) made the constructive version of the same point: safety hints were “telling you things that you should already know,” but she still preferred Condition B for the rule-driven step detail and explicitly proposed exposing tips and safety as “a separate section” that an experienced cook can opt into. P10’s preference for linear pacing (“if you give me too much info at once, I will stop listening”) and the navigation request “I am on step 5, to say go back to step 3” similarly point to controls the surface should expose — not problems with R1 or R6. A global verbosity dial, per-rule toggles (especially for R2 density and R3 hazard prompts), persona and pacing selection, and structured multi-step backward navigation would let the same seven rules serve novices, daily home cooks, and expert cooks from the same engine.

8.5 Discussion

The findings in Section 8.4 report what the ODG-aware assistant changed for each rule and for adoption. This discussion steps above that per-rule reporting to interpret the pattern across the rules and the implications it has for how voice-first ODG-aware assistants should be designed and deployed. Three observations organize the discussion: the gains concentrate exactly where mainstream recipes fail blind cooks, suggesting that the ODG framework is closing a coverage gap rather than competing with the screen-reader baseline on its own terms (Section 8.5.1); the assistant’s contribution is best understood as *what* is communicated and *how recovery is supported*, not as *how fast* the cooking goes (Section 8.5.2); and personalization is a first-class design dimension that lets the same rule set serve novices, daily home cooks, and experts (Section 8.5.3).

8.5.1 The Gains Concentrate Where Mainstream Recipes Fail Blind Cooks

Looking across the per-rule findings, the rules that produced the largest Condition B advantages are exactly the rules that encode information mainstream recipes either do not include at all or do not surface in a form a blind cook can act on while cooking. R2 (non-visual sensory cues) and R3 (proactive safety annotations on edges) describe content that simply is not in a sighted-reader recipe; R1 (atomic [verb]+[noun] decomposition) and R6 (dependency-aware re-orientation) describe structural transformations of content that nominally is in the recipe but is delivered in a form that assumes random-access visual scanning; R4 (parallel-task suggestions) and R7 (passive-wait durations) describe relational information — which tasks can interleave with which waits — that a sighted cook reconstructs at a glance from the recipe’s layout but that a screen reader cannot deliver. The aggregate pattern is therefore not that voice is intrinsically a better modality than text plus a screen reader, but that the ODG framework is exposing a class of information

that the mainstream recipe format never made accessible to blind cooks in the first place. The Condition B advantage is closing a coverage gap, not winning a head-to-head on the same content.

This reading also explains why R5 (precise measurements) registers a smaller, though still positive, gap than R2 or R3: numeric quantities are the one slice of recipe content that mainstream formats already make explicit, so R5’s contribution is the convenience of on-demand retrieval rather than the introduction of new information. The size of each rule’s gap in Section 8.4.3 can be read as a proxy for how much of that rule’s content was previously inaccessible to blind cooks under the screen-reader workflow.

8.5.2 Coverage and Recovery, Not Throughput

This chapter deliberately did not report a wall-clock comparison between conditions and does not claim that the ODG-aware assistant accelerates cooking. Two reasons motivate that choice. First, completion time across a voice-first conversational modality and a screen-reader-paced text page is not a meaningful index of usefulness: the two conditions differ in what they deliver, not just in how fast they deliver the same thing, so any time difference would conflate modality cost (voice is intrinsically slower than skim-reading text) with task structure (Condition B asked participants to engage with sensory cues and safety annotations that Condition A did not). Second, the participants themselves did not frame the assistant’s value in terms of speed. They framed it as a different kind of deliverable — “a person doing it with me together” (P2), “a cooking teacher” (P3) — and as a system that earned its keep precisely when things went wrong (the recovery evidence in Section 8.4.3’s R6 paragraph: P3, P6, P7, P8, P9, P10 all reported routine errors the screen reader could not help with). The right outcome construct for an ODG-aware assistant is therefore content coverage and recovery support, not throughput; future evaluations that adopt this work should resist the temptation to add wall-clock timing as a headline metric.

8.5.3 Personalization Is a First-Class Design Dimension for the Same Rule Set

The single most actionable interpretive observation from this study is that adoption willingness varies across participants not because the underlying ODG rules vary in value but because the same rule, delivered at the same density and tone, lands differently on a novice and on an expert. The cases of P4 (expert daily cook), P6 (cooking since age 6–7), and P10 (linear-paced reader) make this concrete: all three endorsed R-level content but asked for control over how it was surfaced — restraint in tone (P4), tips and safety as “a separate section” (P6), linear pacing and multi-step backward navigation (P10). The implication is that a production deployment of an ODG-aware assistant should expose its rule set through a small, finite set of surface controls: a global verbosity dial, per-rule toggles (especially for R2 density and R3 hazard prompts for experienced cooks), persona and pacing selection, and structured multi-step backward navigation. With those controls present, the same seven rules can serve a first-time cook attempting an unfamiliar baking recipe and an expert reaching for a routine weekday dish — the rule set does not have to fork to accommodate the population spread, only the surface does.

There is a broader design principle in this observation. Accessibility systems that try to serve a heterogeneous population are commonly built either as a single fixed deliverable that targets the median user, or as a constellation of mode-specific deliverables (one per user segment). The

pattern in this study suggests a third option for rule-based accessibility systems: a single rule set with a tunable surface, where personalization adjusts *how* the rules fire rather than *whether* the rules apply. This is cheaper to build and maintain than mode-specific deliverables and serves a wider population than a fixed median target.

8.6 Summary

This chapter described the implementation of the ODG as a voice-first cooking assistant with a paired researcher dashboard, including the input-agnostic reasoning engine, the voice layer, the participant-facing interface, and the Wizard-of-Oz researcher dashboard. It reported the design and results of a within-subjects user study with 10 blind and low-vision participants (P1 through P10), comparing the ODG-based assistant (Condition B) against a screen-reader baseline (Condition A) under a counterbalanced condition order. Findings were reported per ODG Rule (R1–R7) and per adoption cell (broad willingness, primary-tool conditions, complementary screen-reader use, onboarding/verbosity, personalization), with quantitative Likert evidence and qualitative interview evidence presented together for each rule and each adoption question. The discussion stepped above the per-rule reporting to interpret the pattern across rules: the Condition B gains concentrate exactly where mainstream recipes fail blind cooks (R2 and R3 add content that is simply not present in sighted-reader recipes; R1 and R6 restructure content the recipe nominally contains but cannot deliver in an actionable form; R4 and R7 surface relational information a sighted cook reconstructs visually), so the ODG framework is closing a coverage gap rather than competing with the screen reader on the same content; the assistant’s contribution is best understood as *what* is communicated and *how recovery is supported* rather than as *how fast* the cooking goes; and personalization is a first-class design dimension that lets the same seven rules serve novices, daily home cooks, and experts through a small finite set of surface controls (verbosity dial, per-rule toggles, persona and pacing selection, multi-step backward navigation).

Chapter 9

Contributions, Discussion, and Future Work

9.1 Overview

This chapter is the final chapter of the thesis. It draws together the contributions made across the preceding eight chapters, situates them as a single coherent body of work, distills concrete design guidelines from the work, states the limitations of what was built plainly, charts directions in which the broader research community can take the work forward, and closes with a reflection on the research process.

The chapter is organized as follows. Section 9.2 draws the threads of the thesis together, showing how the four empirical studies (Chapters 3–6) led to the ODG representation (Chapter 7) and its evaluation (Chapter 8), and what the combined body of work means for the design of AI-enabled assistive technologies. Section 9.3 enumerates the concrete contributions of the thesis (empirical, conceptual, technical, and dataset) in a single consolidated list. Section 9.4 distills the combined body of work into seven concrete design guidelines for AI-enabled assistive technology in physical activities. Section 9.5 states the limitations of the work plainly, including several surfaced in the thesis defense. Section 9.6 outlines five high-level directions in which the broader research community can take the work forward, including concrete evidence (AROMA [140]) that the substrate is already being extended by independent research groups. Section 9.7 closes with a reflection on the research process and the broader meaning of the work.

9.2 Synthesis: From Understanding to Delivery

This thesis explored the entire pipeline of generating a physical information space for non-visual cooking and using it to develop AI-enabled assistive technologies. The work proceeded through four empirical stages before arriving at the ODG representation and its evaluation.

Stage 1: Characterizing non-visual cooking practices and challenges (Chapter 3). The initial study, a content analysis of 122 YouTube videos and 12 semi-structured interviews, revealed that blind cooks employ a rich repertoire of sensory strategies to compensate for the absence of vision, ranging from auditory cues for oil temperature to tactile techniques for ingredient doneness. It also surfaced eight distinct challenge themes organized around a core tension: blind

cooks must constantly switch between *doing* and *monitoring*, because the information that tells them whether they are on track is locked inside the recipe’s prose or hidden in the physical environment. This chapter established that the challenge of non-visual cooking is not merely an accessibility problem; it is a fundamental information delivery problem.

Stage 2: Improving recipe access (Chapter 4). The second study, with 20 blind cooks and 4 rehabilitation instructors, revealed that the recipe itself is a significant barrier. Linear text, visual descriptions, and inaccessible web layouts conspire to force blind cooks into constant re-reading, hand-washing, and mental integration. The chapter produced a design framework for accessible recipe content and structure (including non-visual alternatives to visual descriptors, modular step organization, and non-linear navigation) that informed the ODG’s state decomposition and the prep track.

Stage 3: Understanding contextual object information needs (Chapter 5). The third study, a contextual inquiry with 12 blind cooks in their own kitchens, moved from the recipe to the physical environment. It produced a taxonomy of five primary contextual information categories (position, orientation, proximity and grouping, similarity and duplicates, and internal state) and two secondary categories (safety-related and health-related information), plus one application-specific category (plating and serving), along with a detailed account of how blind cooks use touch, sound, and smell to gather this information. This chapter directly populated the ODG’s `contextCategories` field and informed the sensory channel taxonomy for `nonVisualCues`.

Stage 4: Building OSCAR (Chapter 6). The fourth stage built the perception layer. OSCAR demonstrated that object status tracking (reasoning about what has been done to an ingredient, not just that it exists) substantially improves recipe step prediction for blind cooks. More importantly, OSCAR showed that the gap between *perceiving* object status and *delivering* it to a blind cook at the right moment was still unbridged. The time-causal model and the prediction log from OSCAR directly informed the ODG’s state representation and the progress tracking methods in R6.

Stage 5: The ODG (Chapter 7). The ODG represents the synthesis of all four prior stages into a single unified representation. It does not replace any of the prior chapters; rather, it provides the common substrate on which all of their findings can operate together. The seven ODG Rules (R1–R7) are each traceable to a specific prior chapter, and each is grounded in the empirical findings of that chapter. R1 uses grounded object-status narration from OSCAR and the contextual inquiry. R2 delivers multi-channel sensory cues from the non-visual cooking and contextual inquiry studies. R3 issues safety alerts from the safety concerns documented in the non-visual cooking study. R4 suggests parallel tasks from the multitasking findings in the recipe study. R5 surfaces precise measurements and techniques from the recipe and contextual inquiry studies. R6 maintains progress tracking from OSCAR and the interruption-recovery findings in the recipe study. R7 manages timers from the timing challenges in the recipe and contextual inquiry studies.

Stage 6: Implementation and evaluation (Chapter 8). The system implementation comprised an input-agnostic reasoning engine, a voice-first participant assistant, a paired researcher dashboard for Wizard-of-Oz rule toggling, and a screen-reader baseline. A within-subjects remote study with 10 blind and low-vision participants (P1–P10), organized around three research questions (aggregate effect, rule-level attribution, and adoption), evaluated the ODG-aware assistant against the baseline using ten custom Likert items mapped to specific ODG rules and a closing semi-structured interview analyzed under a deductive thematic-analysis framework structured

around R1–R7 and three RQ3 adoption sub-themes. The evaluation showed that the assistant materially changes the non-visual cooking experience on safety, parallel-task awareness, measurement clarity, and recovery, and that participants expressed broad adoption willingness, with the screen reader retained as a complementary lightweight fallback for cooking the participant has already fully mastered and with a finite set of surface controls (onboarding, verbosity, per-rule toggles, persona/pacing, multi-step backward navigation) identified to extend the assistant’s reach further.

Taken together, these six stages form a complete arc: *understand* blind cooks’ practices, *design* accessible recipe systems, *inquire* into contextual information needs, *build* a perception system (OSCAR), *represent* the unified information space (ODG), and *deliver* it through a voice-first interface. This arc demonstrates how a physical information space can be assembled and operationalized end-to-end for non-visual cooking assistance.

9.3 Contributions of This Thesis

Pulling across the six stages, this thesis makes the following concrete contributions:

9.3.1 Empirical Contributions

- **A characterization of non-visual cooking practices and challenges** (Chapter 3), grounded in a content analysis of 122 YouTube videos and 12 semi-structured interviews, that identifies twelve cooking-activity practices and eight cross-cutting challenge themes for blind cooks.
- **A design framework for accessible recipe content, structure, and access** (Chapter 4), derived from interviews with 20 blind cooks and 4 rehabilitation instructors, that articulates the requirements for non-visual recipe content (non-visual alternatives to visual descriptors, time ranges, substitutions), structure (discrete steps, logical chunks, consistent terminology), and supporting technology (level-of-detail adaptation, non-linear navigation, hands-free interaction).
- **A taxonomy of contextual object information for non-visual cooking** (Chapter 5), constructed from a contextual inquiry with 12 blind cooks in their own kitchens, that organizes the information blind cooks gather about kitchen objects into five primary categories (position, orientation, proximity/grouping, similarity/duplicates, internal state), two secondary categories (safety, health), and one application-specific category (plating).
- **Empirical evidence that an ODG-aware voice assistant materially changes the non-visual cooking experience** (Chapter 8), through a within-subjects user study with 10 blind and low-vision participants comparing the ODG assistant against a screen-reader baseline on matched task probes and custom Likert measures.

9.3.2 Conceptual Contributions

- **The notion of a physical information space for non-visual cooking** (Chapter 1), a structured account of what an AI assistant needs to track, why each piece matters, and

how it should be communicated, framed as a precondition for AI assistance in any physical activity.

- **The Object Dependency Graph (ODG)** (Chapter 7), a unifying representation that integrates object status, prerequisites, safety constraints, non-visual cues, precise measurements, progress tracking, and parallel structure into a single machine-readable substrate.
- **Seven ODG Rules (R1–R7)** (Chapter 7), each grounded explicitly in one of the prior empirical chapters and each implemented as an independently toggleable behavior of a voice-first assistant, that operationalize what an ODG-aware system actually does at run time.

9.3.3 Technical and Artifact Contributions

- **OSCAR** (Chapter 6), a vision-language pipeline that extracts object-status transitions from a recipe, aligns video frames to recipe steps using a time-causal model, and improves step-prediction accuracy by more than 20% over text-only baselines across two evaluation datasets.
- **A runnable end-to-end ODG voice-first cooking assistant** (Chapter 8), comprising an input-agnostic reasoning engine, a participant-facing voice interface, and a paired researcher dashboard with rule-toggle Wizard-of-Oz controls.

9.3.4 Dataset Contributions

- **The non-visual cooking dataset** (Chapter 6), the first dataset of unprocessed egocentric cooking videos captured by blind cooks in their own kitchens, annotated with timestamps and step labels and released to the community for evaluating vision and language models on real-world non-visual cooking.

9.4 Design Guidelines for AI-Enabled Assistive Technology in Physical Activities

The four empirical studies, the ODG representation, and the within-subjects evaluation converge on seven concrete design guidelines that travel beyond the specifics of this thesis. Each guideline is stated as an imperative, anchored to the chapters that support it, and accompanied by a short rationale that names the evidence behind it. Together they form a checklist that practitioners and researchers can apply to any AI-enabled assistive technology for a structured physical activity, not just to cooking.

G1. Architect the physical information space as an explicit, structured object. Recipe prose, linear documents, and vision-language model embeddings are insufficient containers for what a blind user needs to know while acting in the physical world. A purpose-built representation with named fields for state, prerequisite, sensory cue, safety, measurement, and parallelism is what lets an AI assistant deliver the right information at the right moment. The single largest finding across this thesis is that this representation needs to exist as a first-class object; once it

does, individual rules and behaviors fall out of it. [Chapter 3 *challenges*; Chapter 5 *taxonomy*; Chapter 7 *ODG*.]

G2. Track object status, not just step position. “Which numbered step am I on” is the wrong abstraction for non-visual physical activities. The right abstraction is “what has happened to which object,” because that is what supports re-entry after interruption, parallel-task reasoning, doneness inference, and safety checks. The 20+% step-prediction improvement OSCAR achieved over text-only baselines on two evaluation datasets is the quantitative form of this lesson. [Chapter 6; Chapter 7 *R1*.]

G3. Encode non-visual sensory information by channel, not as paraphrases of visual cues. Auditory, tactile, olfactory, thermal, and temporal cues are first-class evidence about doneness, safety, and progress, and they cannot be obtained by rewording a visual description. The interview data in this thesis surfaced unprompted endorsements (“priceless” from P3, “the way to go” from P5, “wow, that’s a good way of explaining it that it should feel soft” from P5 mid-task) when sensory cues were authored by channel rather than paraphrased from sight. [Chapter 3 *practices*; Chapter 5 *internal state*; Chapter 7 *R2*; Chapter 8 *R2 per-rule finding*.]

G4. Surface safety, parallelism, and timing proactively, and let the user turn each off. Safety information was the single largest gap between the screen-reader baseline and the ODG assistant in the summative evaluation (3.4 vs. 6.5 on hazard awareness), and parallel-task awareness showed a clear secondary gain (5.5 vs. 6.6). Proactive delivery is what makes these gains possible, but the same evaluation also showed that an expert user (P4) experienced unsolicited delivery as patronizing. The discipline the guideline encodes is therefore not “always proactive,” but “proactive by default, with the user holding the off switch.” [Chapter 7 *R3, R4, R7*; Chapter 8 *R3/R4/R7 per-rule findings; personalization*.]

G5. Design for interruption and re-entry, not for linear completion. Blind users will be interrupted: by hand-washing, by phone calls, by family members, by accidental skips. A linear text rendered through a screen reader does not help them re-orient; what they need is a structured re-entry answer to “where am I?” that names what is completed, what is in progress, what is available next, and what may have been skipped. This is the rule that the evaluation’s R6 evidence most consistently endorses (*recovered if lost*: 5.7 vs. 6.3) even where overall-progress orientation regressed. [Chapter 4 *interruption recovery*; Chapter 6 *progress tracking*; Chapter 7 *R6*.]

G6. Make verbosity, speech rate, and persona user-controlled, not designer-fixed. The same content lands differently for a first-time cook learning from scratch (P8, who started cooking in 2023) and an expert who lost their vision later in life (P4, who described unsolicited cues as “a mother standing over you”). The thesis’s findings repeatedly surface the same point: the designer’s default verbosity, default speech rate, and default persona are wrong for somebody, and the right architectural response is to make those parameters first-class and adjustable rather than to tune the default harder. [Chapter 4 *level-of-detail adaptation*; Chapter 8 *onboarding/verbosity and personalization findings*.]

G7. Separate perception from delivery and evaluate each before closing the loop. Conflating “can the system see what is happening” with “does the system say the right thing” makes both untestable. Building the perception layer (OSCAR-style) and the delivery layer (ODG-style) as independent components with a defined interface let this thesis attribute each finding to the right layer (OSCAR’s gains to perception, the user study’s gains to delivery) and let AROMA later extend the perception side without re-litigating the delivery side. The discipline is to keep the layers separately evaluable, not to keep them permanently separate. [Chapter 6; Chapter 7; Chapter 8 *Wizard-of-Oz researcher dashboard*.]

9.5 Limitations

This section discusses the main limitations of the work presented in this thesis. Each limitation is described together with the opportunity it opens for future research, and is tied to the chapter it bears on.

Need to personalize ODG systems. The ODG-aware assistant currently treats the cook as a largely generic blind user. It exposes per-rule toggles (Chapter 8) that let a single session be tuned toward a novice or an expert, but it carries no persistent profile of an individual cook—their skill level, repertoire of already-mastered dishes, available tools and appliances, dietary needs, or cuisine preferences—and it does not carry such information across sessions. The evaluation makes the value of closing this gap concrete: the same content, at the same density and tone, served an expert daily cook (P4) and a first-time cook (P8) very differently. Because the rule layer is already parameterized and toggleable, a persistent personalization layer is a natural addition on top of the existing system rather than a redesign of it (Section 9.6), and it is among the most promising next steps the work opens up.

Need to deeply consider the diversity within the blind and low-vision population. The thesis models a prototypical blind cook and delivers the same non-visual content regardless of a cook’s residual or functional vision, spatial and proprioceptive skill, or onset of vision loss—even though participants ranged from congenitally to recently blind and from totally blind to legally blind with light perception (Chapters 4–8). A cook with usable peripheral vision may need less sensory-cue narration than the rules supply, and a cook with strong tactile expertise may find some proactive guidance redundant. This points to a rich design space: the same per-rule machinery that enables personalization could be driven by a cook’s vision and skill profile, adapting how much is narrated, and how, to whom. Mapping that space is a clear opportunity for follow-on work.

Need to close the perception–delivery loop. The thesis separated perception (OSCAR, Chapter 6) from delivery (the ODG assistant, Chapter 8) and did not yet join them end to end. OSCAR improves step-prediction accuracy by more than 20% over text-only baselines, but on real-world non-visual footage the best configuration reaches roughly two-thirds accuracy—short of the reliability a cook would need to trust an unsupervised system, since a single misread step can compromise a dish. Rather than a flaw in either component, this is precisely the interface the modular architecture was built to support: improving perception (as AROMA [140] has already

begun to do) and establishing the accuracy threshold required for safe unsupervised delivery are well-defined, independently tractable problems, and gains in perception flow into the delivery layer without redesigning it.

Need to support sensing beyond a single camera. OSCAR and the current perception story rest on a single camera stream. The `nonVisualCues` field of the ODG already names auditory, tactile, olfactory, thermal, and temporal cues, and the engine is architected so that additional sensors could feed it, but no non-visual sensing channel—audio event detection, thermal or weight sensing, smell, or even simple elapsed-time signals—is yet implemented or evaluated; those cues are presently authored into the recipe rather than sensed from the kitchen. Because the representation is sensor-agnostic by design, adding these channels is an extension the substrate explicitly anticipates, and one of the most direct ways to let the assistant sense the very cues blind cooks rely on most.

Need to support cooking familiar dishes, not only new ones. The studies and the assistant are framed largely around following a recipe for the first time, yet much everyday cooking is of dishes a cook already knows. At present the system accommodates familiarity only negatively, by letting an experienced cook switch rules off. Modeling what a cook has already mastered—and supporting the transition from learning a dish to having internalized it—is an appealing direction that builds directly on the personalization layer above, and would let a single system grow with a cook over time.

Need to consider cost and everyday practicality. The prototype is expensive to run: a single end-to-end session on a frontier model costs on the order of several dollars in model tokens, and the thesis does not include a formal cost analysis. This is a real barrier to the everyday, repeated use the work envisions—and also a tractable engineering frontier. The most generalizable component, atomized and memory-light step delivery, is exactly the part that could run on far cheaper models or reach a much wider population through low-cost smart speakers under a universal-design framing; pursuing that is how the work moves from a research prototype toward something deployable at scale.

Need to support cooking as a social, interdependent activity. Following a long tradition in assistive technology, the thesis frames cooking around individual independence. Cooking is also a deeply social and interdependent activity, and the study itself surfaced this—participants wanted the assistant to take on the persona of a family member, or to support cooking alongside others. Treating these relational and collaborative dimensions as first-class, rather than out of scope, is a promising broadening of the work: the assistant could be designed to support co-located and shared cooking, not only solo independence.

Taken together, these are limitations of what the thesis built, but several of them also point to where the work can be taken next. The architectural choices that bound the work—separating perception from delivery, keeping the ODG declarative and sensor-agnostic, and making the rule layer parameterizable—should help make personalization, richer sensing, a closed perception-delivery loop, and collaborative cooking reachable extensions rather than rewrites. The next section develops the most important of these directions in detail.

9.6 Directions for the Research Community

This thesis is intended as a foundation that other researchers can build on, not as a finished product. The artifacts it contributes to the field (the ODG representation, the seven rules R1–R7, the OSCAR perception pipeline, the non-visual cooking dataset, the nine annotated recipes, the contextual-information taxonomy, the recipe-access design framework, and the empirical findings from four formative studies and one summative evaluation) are released so that the accessibility, HCI, AI, and rehabilitation research communities can extend, challenge, and re-purpose them. The subsections below are not a roadmap of what I personally plan to do next; they are five high-level invitations to the broader research community, each framed around a question that no single dissertation can settle on its own and each leveraging multiple chapters of this thesis as its starting substrate. Narrower technical and methodological follow-ups (factorial rule ablations, voice-prosody studies, automatic ODG extraction, skill-level personalization, community co-design of authoring tools) are situated as concrete entry points *within* the broader directions they actually serve.

9.6.1 Generalizing the Physical Information Space Beyond Cooking

The deepest question this thesis raises is whether the move it makes, of treating the physical information space of an activity as an explicit, structured object that an AI system reasons over and delivers from, generalizes beyond the kitchen. The ODG’s representational pattern (states, prerequisites, parallel groups, typed sensory cues, safety-annotated edges) and the seven rules that ride on top of it are not, by construction, cooking-specific; nor are the contextual taxonomy from Chapter 5 (position, orientation, proximity, similarity, internal state) and the recipe-access design framework from Chapter 4 (non-visual alternatives to visual descriptors, modular structure, non-linear navigation). Researchers working on other structured physical activities where blind users currently lack real-time guidance (flat-pack furniture assembly, medication management, equipment maintenance, physical-therapy routines, lab procedures, gardening, woodworking) are invited to take this thesis’s full stack as a starting substrate and instantiate it for their own domain. The open community-scale questions are which elements of the substrate transfer unchanged, which require domain-specific extensions (a furniture-assembly graph has no passive-wait phase; a medication graph has no sensory doneness; both have failure modes the ODG was not designed for), and whether voice-first delivery remains the right modality outside the kitchen at all. Cross-domain replications would tell the field whether the contribution of this thesis is a general pattern for non-visual access to physical activity or a cooking-specific construction.

9.6.2 Scaling the Substrate: From AI-Drafted ODGs to Community-Stewarded Authoring

The nine recipes used in this thesis’s evaluation were drafted by a large language model from publicly sourced recipe text using the ODG schema as the target structure, and then verified by the author before being used in the user study. AI authoring is therefore not the open problem. The open problem is everything that comes after the AI draft: how to verify it efficiently, who has the authority to do so, and how the corpus scales beyond what any single researcher can personally

review. The natural-language-processing and applied-AI communities can sharpen the extraction side, ingesting arbitrary recipe text and, eventually, recipe videos to produce richer candidate ODGs whose sensory cues, safety annotations, and parallel-group structure are reliable enough to ship without per-recipe researcher review, with the nine ODGs released with this thesis serving as evaluation targets and the non-visual cooking video dataset released alongside OSCAR (Chapter 6) providing a benchmark for the harder case of grounding extraction in real-world, unedited blind-cook footage rather than in the curated YouCook2-style datasets vision-language models for cooking have historically been benchmarked against. The deeper move, however, is on the human side: for the blind and visually impaired community, together with rehabilitation centers and organizations such as the American Council of the Blind and the National Federation of the Blind, to take co-design leadership of the verification and stewardship process itself, choosing which AI-proposed sensory cues are actually useful in their kitchens, which safety annotations are patronizing versus necessary, which terminology the community adopts, and which recipes the corpus should prioritize first. The most consequential research questions live at the intersection: what does an accessible authoring tool look like that lets a blind cook (or a rehabilitation instructor) audit, correct, and contribute AI-drafted ODGs at scale, with the model providing the first pass and the community providing the editorial judgment that decides what enters the public corpus? Answering that question is a multi-team, multi-year effort that this thesis can seed but cannot itself carry out.

9.6.3 From Perception-Augmented Assistance to Embodied Physical Intelligence

This thesis deliberately separated the perception layer (OSCAR, Chapter 6) from the delivery layer (the ODG assistant, Chapter 8) and evaluated them in isolation, a separation that made each contribution tractable but left the most consequential research questions open. The next horizon is to ask what becomes possible when this separation closes: when perception, the structured physical information space (the ODG and the contextual taxonomy from Chapter 5), and conversational delivery operate as one loop that watches a physical activity unfold and intervenes only when the cook’s mental model of the kitchen diverges from what is actually happening. The hard questions in that regime are normative rather than architectural: when should a perception-grounded assistant speak unprompted, when should it stay silent, how should it surface its own uncertainty about a hot pan or a partially mixed batter, and how should it negotiate authority with a blind cook who has decades of tactile expertise? Factorial or partial-factorial follow-ups that vary subsets of R1–R7 are a natural concrete entry point here (the seven rules, the matched task probes, and the custom Likert measures are documented in sufficient detail for other groups to replicate and extend), and would tell the field which rules become more critical, less critical, or redundant once perception is in the loop rather than wizarded.

That this is a tractable research agenda rather than a speculative one is supported by work that has already begun. AROMA [140], published independently at UIST 2025 by a different research team, extends OSCAR’s object-status representation to ground a wearable camera feed against a *video* recipe rather than the text recipe OSCAR aligned to, and adopts a mixed-initiative interaction style whose proactive components align directly with the delivery philosophy of the seven rules (R2 sensory cues, R3 safety alerts, R4 parallel-task suggestions, R7 automatic timers).

That AROMA was able to extend the perception side without re-litigating the delivery side is itself evidence for guideline G7 (the perception/delivery separation), and converging evidence that the empirical findings of this thesis (safety alerts, doneness cues, parallel-task awareness, re-orientation after interruption) are durable design requirements rather than artifacts of a single study. The thesis contributes the substrate (the ODG schema, R1–R7 as a delivery contract, OSCAR and its dataset, and the empirical accounts of where blind cooks need versus resist proactive intervention) on which the AI, HCI, and accessibility communities can study the open normative questions above.

9.6.4 From Snapshots to Long-Term, Personalized Assistive Partnerships

Every study in this thesis (the YouTube content analysis in Chapter 3, the recipe-access interviews in Chapter 4, the in-home contextual inquiry in Chapter 5, and the within-subjects evaluation in Chapter 8) captured snapshots of blind cooks’ practices at a single moment in time, and presented the same ODG-aware assistant to every participant. The question the field still has to answer is what happens *across* time and *across* people: how an AI-enabled assistant for non-visual cooking becomes (or fails to become) a durable partner in someone’s kitchen, how its role evolves from instructor to collaborator to ambient backstop as the cook’s expertise grows, how the same ODG should be *rendered differently* for a first-time cook learning from scratch versus a former chef who lost their vision later in life (the two ends of the skill spectrum surfaced in Chapter 4 and re-encountered in Chapter 8’s P4 and P8), and how the assistant negotiates the moments where the cook’s hands, attention, kitchen noise, and emotional state make the default verbosity wrong. Personalization, longitudinal adaptation, and the question of whether the assistant supports skill development or fosters dependence are facets of the same broader question: what an assistive technology grounded in a physical information space looks like when it lives in someone’s kitchen for a year rather than for a single 60-minute session. The artifacts of this thesis (the design framework, the contextual taxonomy, the ODG and its seven rules, the runnable assistant, the recipe corpus, the dataset, and the participant base assembled across four studies) together comprise the foundation that multi-year deployment programs, run in partnership with the rehabilitation and blind-community organizations that already steward long-term independence training, can build upon.

9.6.5 The Relational Dimensions of AI Assistance for Blind Users

The interviews and qualitative observations across four chapters of this thesis surfaced a vocabulary that the chapter-level measures alone could not. Blind cooks described an AI cooking assistant variously as “a cooking teacher” (P3, Chapter 8), as “a person doing it with me together” (P2), and, at the other end, as “a mother standing over you” (P4). The non-visual cooking practices study (Chapter 3) and the recipe-access study (Chapter 4) similarly surfaced strong preferences about who is in charge of pacing, what counts as helpful versus patronizing, and how visible the assistant should make itself in the middle of a cooking task. These observations point at a question this thesis was not designed to answer: what does it mean for a blind person to live alongside an AI assistant that, by design, knows more about the ongoing state of their kitchen

than any prior voice interface has? Researchers in HCI, conversational AI, ethics, and disability studies are invited to take the seven rules, the design framework, and the empirical material released with this thesis as starting points for studying the relational, identity, and autonomy dimensions of AI assistance for blind users: when an assistant should defer to a cook's expertise versus challenge them, how it should represent itself (or refuse to be represented) as a persona, and what kinds of long-term self-conceptions of independence and competence are reinforced or undermined by living with one. The narrower technical questions (voice persona, prosody, alert tone, interruption handling) are one concrete entry point into this larger question; the community is invited to take on the whole.

9.7 Reflection on the Research Process

This thesis pursued a question that emerged from close listening to blind cooks: what would it take to give a blind cook the same confidence in their kitchen that a sighted cook takes for granted? The answer, the work suggests, is not a single technology but a pipeline: from understanding the practices and needs of blind cooks, to designing accessible representations and systems, to building perception systems that can track the physical world, to delivering the right information at the right moment through the right channel.

The ODG is one concrete instantiation of this pipeline for non-visual cooking. The broader contribution is the demonstration that a physical information space for assistive technology can be assembled from empirical studies, encoded in a structured representation, and delivered through a voice-first interface that respects blind cooks' autonomy, expertise, and sensory capabilities. The seven ODG Rules, grounded in the experiences of dozens of blind cooks across four empirical studies and stress-tested in one summative evaluation, offer a principled foundation for the next generation of AI-enabled cooking assistance, and, the thesis argues, a template for how an AI assistant for any complex physical activity could be built around an explicit information space rather than around recipe prose or model embeddings alone. This thesis does not claim that the ODG or OSCAR are the only viable foundations for non-visual cooking assistance. What it claims, and what work like AROMA [140] helps to support, is that the underlying *move* of making the physical information space an explicit, structured object that an AI assistant reasons over, rather than leaving it implicit in recipe prose or in vision-language model embeddings, is durable, useful, and extensible to new interaction styles, new perception modalities, and new collaborators.

What I take from the research process itself is that the most useful contributions in this space are the ones that resist being collapsed into a single artifact. The ODG matters because OSCAR's object-status representation, the contextual taxonomy, the recipe-access design framework, and the practices catalogue underwrite it; the ODG-aware assistant matters because the seven rules give it something specific to do; and the user study matters because the limit cases (P4's expert agency and P10's reliance on linear pacing) are at least as informative as the headline gains. None of these pieces is sufficient on its own; their value is in the connections between them. That, more than any individual finding, is what I hope future work on AI-enabled assistive technology will carry forward.

This thesis is therefore not an ending but a foundation: a set of empirical findings, design principles, technical artifacts, datasets, and open questions that the research community and the blind and visually impaired community can build upon in the years ahead.

Bibliography

- [1] Ali Abdolrahmani, Ravi Kuber, and Stacy M Branham. 2018. "Siri Talks at You" An Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 249–258.
- [2] Ali Abdolrahmani, Kevin M Storer, Antony Rishin Mukkath Roy, Ravi Kuber, and Stacy M Branham. 2020. Blind leading the sighted: drawing design insights from blind users towards more productivity-oriented voice interfaces. *ACM Transactions on Accessible Computing (TACCESS)* 12, 4 (2020), 1–35.
- [3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [4] Dustin Adams and Sri Kurniawan. 2014. A blind-friendly photography application for smartphones. *ACM SIGACCESS Accessibility and Computing* 108 (2014), 12–15.
- [5] Niran Adetoro. 2016. *INFORMATION ACCESS FOR THE VISUALLY IMPAIRED IN THE DIGITAL AGE*.
- [6] Ken Albala and Rosanna Nafziger. 2010. *The Lost Art of Real Cooking*. Perigee Trade.
- [7] Allrecipe. 2023. Allrecipes | Recipes, How-Tos, Videos and More. <https://www.allrecipes.com/>. (2023). (Accessed on 04/18/2023).
- [8] Amazon. 2023. Amazon Alexa. <https://alexa.amazon.com/>. (2023). (Accessed on 05/11/2023).
- [9] Lisa Anthony, YooJin Kim, and Leah Findlater. 2013. Analyzing user-generated youtube videos to understand touchscreen use by people with motor impairments. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1223–1232.
- [10] Riku Arakawa, Hiromu Yakura, Vimal Mollyn, Suzanne Nie, Emma Russell, Dustin P. DeMeo, Haarika A. Reddy, Alexander K. Maytin, Bryan T. Carroll, Jill Fain Lehman, and Mayank Goel. 2023. PrISM-Tracker: A Framework for Multimodal Procedure Tracking Using Wearable Sensors and State Transition Information with User-Driven Handling of Errors and Uncertainty. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 156 (jan 2023), 27 pages. DOI:<http://dx.doi.org/10.1145/3569504>
- [11] Andrea Begley and Danielle Gallegos. 2010. What's cooking for dietetics? A review of the literature. *Nutrition & Dietetics* 67, 2 (2010), 74–79.
- [12] Warren Belasco. 2008. *Food: The key concepts*. Berg.
- [13] Genevieve Bell, Mark Blythe, and Phoebe Sengers. 2005. Making by Making Strange: Defamiliarization and the Design of Domestic Technologies. *ACM Transactions on Computing-Human Interaction* 12, 2 (2005), 149–173.
- [14] Karissa Bell. 2023. Microsoft's 'talking camera' app for the blind isn't as magical as it sounds | Mashable. <https://mashable.com/article/microsoft-seeing-ai-app-for-blind>. (2023). (Accessed on 09/10/2023).
- [15] BeMyEyes. 2023. Introducing: Be My AI. <https://www.bemyeyes.com/blog/introducing-be-my-ai>. (2023). (Accessed on 11/22/2023).
- [16] David Benyon. 2001. The new HCI? Navigation of information space. *Knowledge-Based Systems* 14, 8 (2001), 425–430.

-
- [17] Hugh Beyer and Karen Holtzblatt. 1999. Contextual design. *interactions* 6, 1 (1999), 32–42.
 - [18] Alexy Bhowmick and Shyamanta M Hazarika. 2017. An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends. *Journal on Multimodal User Interfaces* 11, 2 (2017), 149–172.
 - [19] Jeffrey P Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010. VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 65–72.
 - [20] Marie Claire Bilyk, Jessica M Sontrop, Gwen E Chapman, Susan I Barr, and Linda Mamer. 2009. Food experiences and eating patterns of visually impaired and blind people. *Canadian Journal of Dietetic practice and research* 70, 1 (2009), 13–18.
 - [21] Rubén Blasco, Álvaro Marco, Roberto Casas, Diego Cirujano, and Richard Picking. 2014. A smart kitchen for ambient assisted living. *Sensors* 14, 1 (2014), 1629–1653.
 - [22] Paul Blenkhorn. 1995. Producing a text-to-speech synthesizer for use by blind people. *Extra-ordinary Human-Computer Interaction: Interfaces for Users with Disabilities*, ADN Edwards (ed.), Cambridge University Press, New York (1995), 307–314.
 - [23] Andrea Borghini. 2015. What is a Recipe? *Journal of Agricultural and Environmental Ethics* 28 (2015), 719–738.
 - [24] Jullien Bouchet and Laurence Nigay. 2004. ICARE: a component-based approach for the design and development of multimodal interfaces. In *CHI'04 extended abstracts on Human factors in computing systems*. 1325–1328.
 - [25] Anne L. Bower. 1997. Our Sisters' Recipes: Exploring "Community" in a Community Cookbook. *The Journal of Popular Culture* 31, 3 (1997), 137–151.
 - [26] Ian Bradbury, James Shell, and Ian Knowles. 2003. eyeCOOK: A gaze and voice-controlled recipe system. In *Proceedings of the 2003 International Conference on Multimedia and Expo*. 996–999.
 - [27] Gary Bradski. 2000. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer* 25, 11 (2000), 120–123.
 - [28] Stacy M Branham and Shaun K Kane. 2015. Collaborative accessibility: How blind and sighted companions co-create accessible home spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2373–2382.
 - [29] Stacy M Branham and Antony Rishin Mukkath Roy. 2019. Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 446–458.
 - [30] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
 - [31] Da Cao, Zhiwang Yu, Hanling Zhang, Jiansheng Fang, Liqiang Nie, and Qi Tian. 2019. Video-based cross-modal recipe retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1685–1693.
 - [32] Michele Capella-McDonnall. 2007. The need for health promotion for adults who are visually impaired. *Journal of Visual Impairment & Blindness* 101, 3 (2007), 133–145.
 - [33] Martin Caraher, Philippa Dixon, Tim Lang, and Roy Carr-Hill. 1999. The state of cooking in England: the relationship of cooking skills to food choice. *British Food Journal* 101, 8 (1999), 590–609.
 - [34] Luis Cavazos Quero, Jorge Iranzo Bartolomé, Seonggu Lee, En Han, Sunhee Kim, and Jundong Cho. 2018. An interactive multimodal guide to improve art accessibility for blind people. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 346–348.
 - [35] Minsuk Chang, Léonore V Guillain, Hyeunghsik Jung, Vivian M Hare, Juho Kim, and Maneesh Agrawala. 2018. Recipescape: An interactive tool for analyzing cooking instructions at scale. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

- [36] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 381–389.
- [37] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.
- [38] Jen-Hao Chen, Peggy Pei-Yu Chi, Hao-Hua Chu, Cheryl Chia-Hui Chen, and Polly Huang. 2010. A smart kitchen for nutrition-aware cooking. *IEEE Pervasive Computing* 9, 4 (2010), 58–65.
- [39] Quentin Chibaudel, Wafa Johal, Bernard Oriola, Marc JM Macé, Pierre Dillenbourg, Valérie Tartas, and Christophe Jouffrais. 2020. "If you've gone straight, now, you must turn left"-Exploring the use of a tangible interface in a collaborative treasure hunt for people with visual impairments. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–10.
- [40] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. "Nobody Speaks that Fast!" An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [41] Jaz Hee-jeong Choi, Marcus Foth, and Greg Hearn. 2014. *Eat, cook, grow: Mixing human-computer interactions with human-food interactions*. MIT Press.
- [42] Jenny Cooper. 2015. Cooking trends among millennials: Welcome to the digital kitchen. *Think with Google* (2015), 18.
- [43] Aira Tech Corp. 2021. Home - Aira : Aira. <https://aira.io/>. (April 2021). (Accessed on 04/08/2021).
- [44] Hilary Davis, Bjorn Nansen, Frank Vetere, Toni Robertson, Margot Brereton, Jeannette Durick, and Kate Vaisutis. 2014. Homemade cookbooks: a recipe for sharing. In *Proceedings of the 2014 conference on Designing interactive systems*. 73–82.
- [45] Antonella De Angeli, Uday Athavankar, Anirudha Joshi, Lynne Coventry, and Graham I Johnson. 2004. Introducing ATMs in India: a contextual inquiry. *Interacting with Computers* 16, 1 (2004), 29–44.
- [46] Nem Khan Dim, Chaklam Silpasuwanchai, Sayan Sarcar, and Xiangshi Ren. 2016. Designing mid-air TV gestures for blind people using user-and choice-based elicitation approaches. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. 204–214.
- [47] Lucia Donatelli, Theresa Schmidt, Debanjali Biswas, Arne Köhn, Fangzhou Zhai, and Alexander Koller. 2021. Aligning actions across recipe graphs. In *Proceedings of the 2021 conference on empirical methods in natural language processing*. 6930–6942.
- [48] Bryan Dosono, Jordan Hayes, and Yang Wang. 2015. "I'm Stuck!": A Contextual Inquiry of People with Visual Impairments in Authentication. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*. 151–168.
- [49] Bengisu Dost. 2021. How can we improve Microsoft's "Seeing AI" application? | by Bengisu Dost | Medium. <https://bengisudost.medium.com/how-can-we-improve-microsofts-seeing-ai-application-4cb4b074e0d6>. (Jan 2021). (Accessed on 09/10/2023).
- [50] Paul Dourish. 2001. *Where the action is: The foundations of embodied interaction*. Vol. 210. MIT Press.
- [51] Paul Dourish and Matthew Chalmers. 1994. Running out of space: Models of information navigation. In *Short paper presented at HCI'94 (People and Computers IX)*.
- [52] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, and others. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378* (2023).
- [53] Rachel Engler-Stringer. 2010. Food, cooking skills, and health: A literature review. *Canadian Journal of Dietetic Practice and Research* 71, 3 (2010), 141–145.
- [54] Thomas Erickson and David W McDonald. 2007. *HCI Remixed: Reflections on Works That Have Influenced the HCI Community*. Mit Press.

-
- [55] Maurizio Ficocelli and Goldie Nejat. 2012. The design of an interactive assistive kitchen system. *Assistive Technology* 24, 4 (2012), 246–258.
 - [56] Leah Findlater and Lotus Zhang. 2020. Input accessibility: A large dataset and summary analysis of age, motor ability and input performance. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–6.
 - [57] Amy Rae Fox, Philip Guo, Clemens Nylandsted Klokmose, Peter Dalsgaard, Arvind Satyanarayan, Haijun Xia, and James D Hollan. 2020. Towards a dynamic multiscale personal information space: beyond application and document centered views of information. In *Companion Proceedings of the 4th International Conference on Art, Science, and Engineering of Programming*. 136–143.
 - [58] Giovanni Fusco, Ender Tekin, Richard E Ladner, and James M Coughlan. 2014. Using computer vision to access appliance displays. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 281–282.
 - [59] Franck Ganier. 2004. Factors affecting the processing of procedural instructions: Implications for document design. *IEEE Transactions on Professional Communication* 47, 1 (2004), 15–26.
 - [60] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2023. Physically Grounded Vision-Language Models for Robotic Manipulation. *arXiv preprint arXiv:2309.02561* (2023).
 - [61] David Gonçalves, André Rodrigues, and Tiago Guerreiro. 2020. Playing With Others: Depicting Multiplayer Gaming Experiences of People With Visual Impairments. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.
 - [62] Tao Gong, Kai Chen, Xinjiang Wang, Qi Chu, Feng Zhu, Dahua Lin, Nenghai Yu, and Huamin Feng. 2021. Temporal ROI align for video object recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1442–1450.
 - [63] Google. 2022. Google Forms: Online Form Creator | Google Workspace. <https://www.google.com/forms/about/>. (2022). (Accessed on 08/24/2022).
 - [64] Google. 2024. Lookout - Assisted vision - Apps on Google Play. https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en_US&gl=US. (2024). (Accessed on 03/30/2024).
 - [65] GoPro. 2023. GoPro HERO11 Black Action Camera (Waterproof + Stabilization). <https://gopro.com/en/us/shop/cameras/hero11-black/CHDHX-111-master.html>. (Sept 2023). (Accessed on 09/04/2023).
 - [66] Rúben Gouveia and Evangelos Karapanos. 2013. Footprint tracker: supporting diary studies with lifelogging. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2921–2930.
 - [67] Catherine Grevet, Anthony Tang, and Elizabeth Mynatt. 2012. Eating alone, together: new forms of communality. In *Proceedings of the 2012 ACM International Conference on Supporting Group Work*. 103–106.
 - [68] Andrea Grimes and Richard Harper. 2008. Celebratory technology: New directions for food research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 467–476.
 - [69] William Grussenmeyer and Eelke Folmer. 2017. Accessible touchscreen technology for people with visual impairments: a survey. *ACM Transactions on Accessible Computing (TACCESS)* 9, 2 (2017), 1–31.
 - [70] João Guerreiro and Daniel Gonçalves. 2014. Text-to-speeches: evaluating the perception of concurrent speech by blind people. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 169–176.
 - [71] Anhong Guo, Xiang’Anthony’ Chen, Haoran Qi, Samuel White, Suman Ghosh, Chieko Asakawa, and Jeffrey P Bigham. 2016. Vizlens: A robust and interactive screen reader for interfaces in the real world. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 651–664.

- [72] Anhong Guo, Jeeun Kim, Xiang'Anthony' Chen, Tom Yeh, Scott E Hudson, Jennifer Mankoff, and Jeffrey P Bigham. 2017. Facade: Auto-generating tactile interfaces to appliances. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5826–5838.
- [73] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.
- [74] Christine Ha. 2020. The Blind Cook – Christine Ha's Adventures. <http://www.theblindcook.com/>. (April 2020). (Accessed on 03/14/2021).
- [75] Reiko Hamada, Jun Okabe, Ichiro Ide, Shin'ichi Satoh, Shuichi Sakai, and Hidehiko Tanaka. 2005a. Cooking navi: assistant for daily cooking in kitchen. In *Proceedings of the 13th annual ACM international conference on Multimedia*. 371–374.
- [76] Ryusuke Hamada, Itiro Siio, and Owen Frank. 2005b. HappyCooking: Recipe task organization system for managing multiple parallel cooking tasks. In *Proceedings of the International Conference on Ubiquitous Computing (UbiComp 2005)*. 132–135.
- [77] Rex Hartson and Pardha S Pyla. 2012. *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier.
- [78] Ed Henkler. 2020. How Does a Blind Person Cook? - The Blind Guide. <https://theblindguide.com/how-does-blind-person-cook/>. (April 2020). (Accessed on 03/14/2021).
- [79] Steve Hodges, Emma Berry, and Ken Wood. 2011. SenseCam: A wearable camera that stimulates and rehabilitates autobiographical memory. *Memory* 19, 7 (2011), 685–696.
- [80] James Hollan, Edwin Hutchins, and David Kirsh. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7, 2 (2000), 174–196.
- [81] Karen Holtzblatt and Hugh Beyer. 1995. Contextual Design: using customer work models to drive system design. In *Conference companion on human factors in computing systems*. 391–392.
- [82] Juan Pablo Hourcade, Sarah L Mascher, David Wu, and Luiza Pantoja. 2015. Look, my baby is using an iPad! An analysis of YouTube videos of infants and toddlers using tablets. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1915–1924.
- [83] Mina Huh, Zihui Xue, Ujjaini Das, Kumar Ashutosh, Kristen Grauman, and Amy Pavel. 2025. Vid2Coach: Transforming How-To Videos into Task Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- [84] Zoom Video Communications Inc. 2012. Video Conferencing, Cloud Phone, Webinars, Chat, Virtual Events | Zoom. <https://zoom.us/>. (September 2012). (Accessed on 07/20/2021).
- [85] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. 2011. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. 203–210.
- [86] Nabila Jones, Hannah Elizabeth Bartlett, and Richard Cooke. 2019. An analysis of the impact of visual impairment on activities of daily living and vision-related quality of life in a visually impaired adult population. *British Journal of Visual Impairment* 37, 1 (2019), 50–63.
- [87] Wendy Ju, Roger Hurwitz, Tim Judd, and Brian Lee. 2001. CounterActive: an interactive cookbook for video and audio instruction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 269–270.
- [88] Holtzblatt Karen and Jones Sandra. 2017. Contextual inquiry: A participatory technique for system design. In *Participatory design*. CRC Press, 177–210.
- [89] Shuaib Karim, Amin Andjomshoaa, and A Min Tjoa. 2006. Exploiting sensecam for helping the blind in business negotiations. In *Computers Helping People with Special Needs: 10th International Conference, ICCHP 2006, Linz, Austria, July 11-13, 2006. Proceedings 10*. Springer, 1147–1154.

-
- [90] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and others. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.
 - [91] Fumihiko Kato and Shoichi Hasegawa. 2013. Interactive Cooking Simulator: Showing food ingredients appearance changes in frying pan cooking. In *Proceedings of the 5th international workshop on Multimedia for cooking & eating activities*. 33–38.
 - [92] Seita Kayukawa, Keita Higuchi, João Guerreiro, Shigeo Morishima, Yoichi Sato, Kris Kitani, and Chieko Asakawa. 2019. BBeep: A sonic collision avoidance system for blind travellers and nearby pedestrians. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
 - [93] Andrey Khusid and Oleg Shardin. 2011. An Online Visual Collaboration Platform for Teamwork | Miro. <https://miro.com/>. (2011). (Accessed on 06/02/2021).
 - [94] Mohammad Kianpisheh, Franklin Mingzhe Li, and Khai N Truong. 2019. Face recognition assistant for people with visual impairments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.
 - [95] Minyung Kim, Sooyoung Hwang, Kyoungmin Choi, Youkeun Oh, and Dokshin Lim. 2022. Vision-based cooking assistance system for visually impaired people. In *International Conference on Human-Computer Interaction*. Springer, 540–547.
 - [96] Yonghee Kim, Youngju Park, and Jeongil Choi. 2017. A study on the adoption of IoT smart home service: using Value-based Adoption Model. *Total Quality Management & Business Excellence* 28, 9-10 (2017), 1149–1165.
 - [97] Hubert Knoblauch, Bernt Schnettler, Jürgen Raab, and Hans-Georg Soeffner. 2012. *Video analysis: Methodology and methods*. Peter Lang Frankfurt am Main.
 - [98] Aida Komkaite, Liga Lavrinovica, Maria Vranka, and Mikael B Skov. 2019. Underneath the Skin: An Analysis of YouTube Videos to Understand Insertable Device Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
 - [99] A König and K Thongpull. 2015. Lab-on-Spoon—a 3-D integrated hand-held multi-sensor system for low-cost food quality, safety, and processing monitoring in assisted-living systems. *Journal of Sensors and Sensor Systems* 4, 1 (2015), 63–75.
 - [100] Eliza Kostyra, Sylwia Żakowska-Biemans, Katarzyna Śniegocka, and Anna Piotrowska. 2017. Food shopping, sensory determinants of food choice and meal preparation by visually impaired people. Obstacles and expectations in daily food experiences. *Appetite* 113 (2017), 14–22.
 - [101] Sven Krome, Steffen P Walz, and Stefan Greuter. 2016. Contextual inquiry of future commuting in autonomous cars. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 3122–3128.
 - [102] Sanna Kuoppamäki, Sylvaine Tuncer, Sara Eriksson, and Donald McMillan. 2021. Designing Kitchen Technologies for Ageing in Place: A Video Study of Older Adults’ Cooking at Home. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 2 (2021), 1–19.
 - [103] Bineeth Kuriakose, Raju Shrestha, and Frode Eika Sandnes. 2020. Multimodal navigation systems for users with visual impairments—a review and analysis. *Multimodal Technologies and Interaction* 4, 4 (2020), 73.
 - [104] Kazuma Kusu, Nozomi Makino, Takamitsu Shioi, and Kenji Hatano. 2017. Calculating Cooking Recipe’s Difficulty based on Cooking Activities. In *Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with The 2017 International Joint Conference on Artificial Intelligence*. 19–24.
 - [105] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 213–224.

- [106] Jaewook Lee, Andrew D Tjahjadi, Jiho Kim, Junpu Yu, Minji Park, Jiawen Zhang, Jon E Froehlich, Yapeng Tian, and Yuhang Zhao. 2024. CookAR: Affordance Augmentations in Wearable AR to Support Kitchen Tool Interactions for People with Low Vision. *arXiv preprint arXiv:2407.13515* (2024).
- [107] Jinna Lei, Xiaofeng Ren, and Dieter Fox. 2012. Fine-grained kitchen activity recognition using rgb-d. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 208–211.
- [108] Wentao Lei, Mingming Fan, and Juliann Thang. 2022. “I Shake The Package To Check If It’s Mine” A Study of Package Fetching Practices and Challenges of Blind and Low Vision People in China. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [109] Franklin Mingzhe Li, Di Laura Chen, Mingming Fan, and Khai N Truong. 2019. FMT: A Wearable Camera-Based Object Tracking Memory Aid for Older Adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–25.
- [110] Franklin Mingzhe Li, Jamie Dorst, Peter Cederberg, and Patrick Carrington. 2021. Non-visual cooking: exploring practices and challenges of meal preparation by people with visual impairments. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–11.
- [111] Franklin Mingzhe Li, Michael Xieyang Liu, Shaun K Kane, and Patrick Carrington. 2024. A Contextual Inquiry of People with Vision Impairments in Cooking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
- [112] Franklin Mingzhe Li, Michael Xieyang Liu, Yang Zhang, and Patrick Carrington. 2022. Freedom to Choose: Understanding Input Modality Preferences of People with Upper-body Motor Impairments for Activities of Daily Living. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16.
- [113] Franklin Mingzhe Li, Kaitlyn Ng, Bin Zhu, and Patrick Carrington. 2025a. Exploring Object Status Recognition for Recipe Progress Tracking in Non-Visual Cooking. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 1–15.
- [114] Franklin Mingzhe Li, Kaitlyn Ng, Bin Zhu, and Patrick Carrington. 2025b. OSCAR: Object Status and Contextual Awareness for Recipes to Support Non-Visual Cooking. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–6.
- [115] Franklin Mingzhe Li, Akihiko Oharazawa, Chloe Qingyu Zhu, Misty Fan, Daisuke Sato, Chieko Asakawa, and Patrick Carrington. 2025c. More than One Step at a Time: Designing Procedural Feedback for Non-visual Makeup Routines. *arXiv preprint arXiv:2507.03942* (2025).
- [116] Franklin Mingzhe Li, Franchesca Spektor, Meng Xia, Mina Huh, Peter Cederberg, Yuqi Gong, Kristen Shinohara, and Patrick Carrington. 2022. “It Feels Like Taking a Gamble”: Exploring Perceptions, Practices, and Challenges of Using Makeup and Cosmetics for People with Visual Impairments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [117] Franklin Mingzhe Li, Ashley Wang, Patrick Carrington, and Shaun K Kane. 2024. A Recipe for Success? Exploring Strategies for Improving Non-Visual Access to Cooking Instructions. *arXiv preprint arXiv:2407.19065* (2024).
- [118] Franklin Mingzhe Li, Lotus Zhang, Maryam Bandukda, Abigale Stangl, Kristen Shinohara, Leah Findlater, and Patrick Carrington. 2023. Understanding Visual Arts Experiences of Blind People. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [119] Angela S Lin, Sudha Rao, Asli Celikyilmaz, Elnaz Nouri, Chris Brockett, Debadeepta Dey, and Bill Dolan. 2020. A recipe for creating multimodal aligned datasets for sequential tasks. *arXiv preprint arXiv:2005.09606* (2020).
- [120] Qing Lin and Youngjoon Han. 2014. A context-aware-based audio guidance system for blind people using a multimodal profile model. *Sensors* 14, 10 (2014), 18670–18700.
- [121] Michael Xieyang Liu. 2023. *Tool Support for Knowledge Foraging, Structuring, and Transfer During Online Sense-making*. Ph.D. Dissertation. Ph. D. Dissertation. Carnegie Mellon University. <http://reports-archive.adm>

- [122] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A Myers. 2019. Unakite: Scaffolding developers’ decision-making using the web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 67–80.
- [123] Michael Xieyang Liu, Aniket Kittur, and Brad A Myers. 2021. To reuse or not to reuse? A framework and system for evaluating summarized knowledge. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–35.
- [124] Michael Xieyang Liu, Andrew Kuznetsov, Yongsung Kim, Joseph Chee Chang, Aniket Kittur, and Brad A Myers. 2022. Wigglyte: Low-cost Information Collection and Triage. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [125] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.
- [126] Xingyu Liu, Patrick Carrington, Xiang’Anthony’ Chen, and Amy Pavel. 2021. What makes videos accessible to blind and visually impaired people?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [127] Lucy M Long. 2001. Nourishing the academic imagination: The use of food in teaching concepts of culture. *Food and Foodways* 9, 3-4 (2001), 235–262.
- [128] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860* (2021).
- [129] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people’s experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5988–5999.
- [130] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2021), 187–203.
- [131] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. 2021. Orbit: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10818–10828.
- [132] Matthew Louis Mauriello, Brenna McNally, and Jon E Froehlich. 2019. Thermportal: An easy-to-deploy temporal thermographic sensor system to support residential energy audits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [133] Matthew Louis Mauriello, Manaswi Saha, Erica Brown Brown, and Jon E Froehlich. 2017. Exploring novice approaches to smartphone-based thermographic energy auditing: A field study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1768–1780.
- [134] Sarah Mennicken, Tanja Karrer, Paul Russell, and Jan Borchers. 2010. PersonalChef: Mobile Projection for Cooking Support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3403–3408.
- [135] Microsoft. 2023. Seeing AI App from Microsoft. <https://www.microsoft.com/en-us/ai/seeing-ai>. (Sept 2023). (Accessed on 09/03/2023).
- [136] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A survey on food computing. *ACM Computing Surveys (CSUR)* 52, 5 (2019), 1–36.
- [137] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P Bigham, and Shaun K Kane. 2016. ” With most of it being pictures now, I rarely use it” Understanding Twitter’s Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5506–5516.
- [138] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking Vision Language Models for Cultural Understanding. *arXiv preprint arXiv:2407.10920* (2024).

- [139] Food Network. 2023. Recipes, Dinners and Easy Meal Ideas | Food Network. <https://www.foodnetwork.com/recipes>. (2023). (Accessed on 04/18/2023).
- [140] Zheng Ning, Leyang Li, Daniel Killough, JooYoung Seo, Patrick Carrington, Yapeng Tian, Yuhang Zhao, Franklin Mingzhe Li, and Toby Jia-Jun Li. 2025. AROMA: Mixed-Initiative AI Assistance for Non-Visual Cooking by Grounding Multimodal Information Between Reality and Videos. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. ACM.
- [141] Patrick Olivier, Guangyou Xu, Andrew Monk, and Jesse Hoey. 2009. Ambient kitchen: designing situated services using a high fidelity prototyping environment. In *Proceedings of the 2nd international conference on pervasive technologies related to assistive environments*. 1–7.
- [142] Shotaro Omori and Ikuko Eguchi Yairi. 2013. Collaborative music application for visually impaired people with tangible objects on table. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–2.
- [143] Sharon Oviatt, Björn Schuller, Philip Cohen, Daniel Sonntag, and Gerasimos Potamianos. 2017. *The handbook of multimodal-multisensor interfaces, volume 1: Foundations, user modeling, and common modality combinations*. Morgan & Claypool.
- [144] Rémi Parlouar, Florian Dramas, Marc MJ Macé, and Christophe Jouffrais. 2009. Assistive device for the blind based on object recognition: an application to identify currency bills. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. 227–228.
- [145] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.
- [146] Mahika Phutane, Julie Wright, Brenda Veronica Castro, Lei Shi, Simone R Stern, Holly M Lawson, and Shiri Azenkot. 2022. Tactile materials in practice: Understanding the experiences of teachers of the visually impaired. *ACM Transactions on Accessible Computing (TACCESS)* 15, 3 (2022), 1–34.
- [147] Azzurra Pini, Jer Hayes, Connor Upton, and Medb Corcoran. 2019. AI inspired recipes: Designing computationally creative food combos. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.
- [148] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological Review* 106, 4 (1999), 643–675.
- [149] Michael Pollan. 2013. *Cooked: A Natural History of Transformation*. Penguin.
- [150] Karthir Prabhakar, Sangmin Oh, Ping Wang, Gregory D Abowd, and James M Rehg. 2010. Temporal causality for the analysis of visual events. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1967–1974.
- [151] Jennifer Preece, Yvonne Rogers, and Helen Sharp. 2015. *Interaction Design: Beyond Human-Computer Interaction* (4 ed.). Wiley, Hoboken, NJ.
- [152] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [153] Mary Elizabeth Raven and Alicia Flanders. 1996. Using contextual inquiry to learn about your audiences. *ACM SIGDOC Asterisk Journal of Computer Documentation* 20, 1 (1996), 1–13.
- [154] Leah M Reeves, Jennifer Lai, James A Larson, Sharon Oviatt, TS Balaji, Stéphanie Buisine, Penny Collings, Phil Cohen, Ben Kraal, Jean-Claude Martin, and others. 2004. Guidelines for multimodal user interface design. *Commun. ACM* 47, 1 (2004), 57–59.
- [155] Gisele Yumi Arabori Ribeiro, Maria Lilian de Araújo Barbosa, Maria Lúcia Leite Ribeiro Okimoto, and Rafael Lima Vieira. 2019. Information for Tactile Reading: A Study of Tactile Ergonomics of Packaging for Blind People. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VII: Ergonomics in Design, Design for All, Activity Theories for Work Analysis and Design, Affective Design 20*. Springer, 1682–1688.

- [156] André Rodrigues, Kyle Montague, Hugo Nicolau, and Tiago Guerreiro. 2015. Getting smartphones to talkback: Understanding the smartphone adoption process of blind users. In *Proceedings of the 17th international acm sigaccess conference on computers & accessibility*. 23–32.
- [157] Mohammad Farid Saaid, AM Mohammad, and MSA Megat Ali. 2016. Smart cane with range notification for blind people. In *2016 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. IEEE, 225–229.
- [158] Jaime Sánchez, Ticianne Darin, and Rossana Andrade. 2015. Multimodal videogames for the cognition of people who are blind: trends and issues. In *Universal Access in Human-Computer Interaction. Access to Learning, Health and Well-Being: 9th International Conference, UAHCI 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part III 9*. Springer, 535–546.
- [159] Ayaka Sato, Keita Watanabe, and Jun Rekimoto. 2014. MimiCook: a cooking assistant system with situated guidance. In *Proceedings of the 8th international conference on tangible, embedded and embodied interaction*. 121–124.
- [160] Daisuke Sato, Uran Oh, Kakuya Naito, Hironobu Takagi, Kris Kitani, and Chieko Asakawa. 2017. Navcog3: An evaluation of a smartphone-based blind indoor navigation assistant with semantic features in a large-scale environment. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 270–279.
- [161] Baifeng Shi, Qi Dai, Judy Hoffman, Kate Saenko, Trevor Darrell, and Huijuan Xu. 2021. Temporal action detection with multi-level supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8022–8032.
- [162] Lei Shi, Yuhang Zhao, and Shiri Azenkot. 2017. Designing interactions for 3D printed models with blind people. In *Proceedings of the 19th international acm sigaccess conference on computers and accessibility*. 200–209.
- [163] Yayoi Shimomura, Ebba Thora Hvannberg, and Hjalmtyr Hafsteinsson. 2010. Accessibility of audio and tactile interfaces for young blind people performing everyday tasks. *Universal Access in the Information Society* 9 (2010), 297–310.
- [164] Itiro Siio, Ryusuke Hamada, and Yasuaki Mima. 2007. Kitchen of the Future: A context-aware kitchen with multiple displays to enhance meal preparation. In *Proceedings of the International Conference on Ubiquitous Computing (UbiComp 2007)*. 471–474.
- [165] Gina Simmons and Diana Chapman. 2012. A recipe for success: Improving cooking skills to enhance well-being. *Health Promotion Journal of Australia* 23, 4 (2012), 225–230.
- [166] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
- [167] Martine Stead, Laura McDermott, Anne MacKintosh, and Ashley Adamson. 2004. Why healthy eating is bad for you: Perceptions of food among working-class mothers. *Sociology of Health & Illness* 26, 4 (2004), 351–372.
- [168] Michad Steehouder, Joyce Karreman, and Nicole Ummelen. 2000. Making sense of step-by-step procedures. In *18th Annual Conference on Computer Documentation. ipcc sigdoc 2000. Technology and Teamwork. Proceedings. IEEE Professional Communication Society International Professional Communication Conference an*. IEEE, 463–475.
- [169] Sebastian Stein and Stephen J McKenna. 2013. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 729–738.
- [170] Simon Stephan, Ralf Mayrhofer, and Michael R Waldmann. 2020. Time and singular causation—a computational model. *Cognitive Science* 44, 7 (2020), e12871.
- [171] Sean Stitt. 1996. An international perspective on food and cooking skills in education. *British Food Journal* 98, 10 (1996), 27–34.

- [172] Kevin M Storer, Tejinder K Judge, and Stacy M Branham. 2020. "All in the Same Boat": Tradeoffs of Voice Assistant Ownership for Mixed-Visual-Ability Families. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [173] Ryo Suzuki, Abigale Stangl, Mark D Gross, and Tom Yeh. 2017. Fluxmarker: Enhancing tactile graphics with dynamic tactile markers. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 190–199.
- [174] TapTapSee. 2023. TapTapSee - Blind and Visually Impaired Assistive Technology - powered by CloudSight.ai Image Recognition API. <https://taptapseeapp.com/>. (Sept 2023). (Accessed on 09/10/2023).
- [175] Dan Tasse and Noah A Smith. 2008. SOUR CREAM: Toward semantic processing of recipes. *Carnegie Mellon University, Pittsburgh, Tech. Rep. CMU-LTI-08-005* (2008).
- [176] Kimberly Tee, Karyn Moffatt, Leah Findlater, Eve MacGregor, Joanna McGrenere, Barbara Purves, and Sidney S Fels. 2005. A visual recipe book for persons with language impairments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 501–510.
- [177] Ender Tekin, James M Coughlan, and Huiying Shen. 2011. Real-time detection and reading of LED/LCD displays for visually impaired persons. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 491–496.
- [178] ChunYuen Teng, Yu-Ru Lin, and Lada A. Adamic. 2011. Recipe recommendation using ingredient networks. *CoRR* abs/1111.3919 (2011). <http://arxiv.org/abs/1111.3919>
- [179] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-first dataset creation: Lessons from constructing a dataset for teachable object recognition with blind and low vision data collectors. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.
- [180] YingLi Tian, Chucai Yi, and Aries Arditi. 2010. Improving computer vision-based indoor wayfinding for blind persons with context information. In *Computers Helping People with Special Needs: 12th International Conference, ICCHP 2010, Vienna, Austria, July 14–16, 2010, Proceedings, Part II 12*. Springer, 255–262.
- [181] Graham Tomlinson. 1986. Thought for food: A study of written instructions. *Symbolic interaction* 9, 2 (1986), 201–216.
- [182] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.
- [183] Quan Tran, Giang Calcaterra, and Elizabeth Mynatt. 2005. Cook’s Collage: Bulletin board for the kitchen. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 5 (2005), 15–18.
- [184] Daisuke Uriu, Kazutoshi Kobayashi, and Kris Kitani. 2012. Panavi: Supporting Recipe Execution with Sensor-integrated Frying Pan and Projector-based Augmented Reality. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 129–138.
- [185] Marieke Van Erp, Christian Reynolds, Diana Maynard, Alain Starke, Rebeca Ibáñez Martín, Frederic Andres, Maria CA Leite, Damien Alvarez de Toledo, Ximena Schmidt Rivera, Christoph Trattner, and others. 2021. Using natural language processing and artificial intelligence to explore the nutrition and sustainability of recipes and food. *Frontiers in Artificial Intelligence* 3 (2021), 621577.
- [186] Maarten Van Someren, Yvonne F Barnard, and J Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress* 11 (1994), 29–41.
- [187] VisionAware. 2021. Safe Cooking Techniques for Cooks Who Are Blind or Have Low Vision - VisionAware. <https://visionaware.org/everyday-living/essential-skills/cooking/safe-cooking-techniques/>. (2021). (Accessed on 04/05/2021).
- [188] Trieu Minh Vu and Riva Khanna. 2018. Application of artificial intelligence in smart kitchen. *International Journal of Innovative Technology and Interdisciplinary Sciences* 1, 1 (2018), 1–8.

- [189] Sagar M Waghmare, Kimberly Wilber, Dave Hawkey, Xuan Yang, Matthew Wilson, Stephanie Debats, Catalyya Nuengsigkapien, Astuti Sharma, Lars Pandikow, Huisheng Wang, and others. 2025. Sanpo: A scene understanding, accessibility and human navigation dataset. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 7866–7875.
- [190] Ru Wang, Nihan Zhou, Tam Nguyen, Sanbrita Mondal, Bilge Mutlu, and Yuhang Zhao. 2023. Practices and Barriers of Cooking Training for Blind and Low Vision People. (2023).
- [191] WHO. 2021. Blindness and vision impairment. <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>. (Feb 2021). (Accessed on 06/14/2021).
- [192] Hans Jørgen Wiberg. 2021. Be My Eyes - See the world together. <https://www.bemyeyes.com/>. (April 2021). (Accessed on 04/10/2021).
- [193] Wikipedia. 2021. Christine Ha - Wikipedia. https://en.wikipedia.org/wiki/Christine_Ha. (March 2021). (Accessed on 03/14/2021).
- [194] Edward K Wong and Minya Chen. 2003. A new robust algorithm for video text extraction. *Pattern Recognition* 36, 6 (2003), 1397–1406.
- [195] Zihui Xue, Kumar Ashutosh, and Kristen Grauman. 2024. Learning object state changes in videos: An open-world perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18493–18503.
- [196] Hanlu Ye, Meethu Malu, Uran Oh, and Leah Findlater. 2014. Current and future mobile and wearable device use by people with visual impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3123–3132.
- [197] YouTube. 2023. (51) YouTube. <https://www.youtube.com/>. (2023). (Accessed on 05/11/2023).
- [198] Nikola Žarić, Milutin Radonjić, Nikola Pavličević, and Sanja Paunović Žarić. 2021. Design of a kitchen-monitoring and decision-making system to support AAL applications. *Sensors* 21, 13 (2021), 4449.
- [199] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11975–11986.
- [200] Yuhang Zhao, Sarit Szpiro, Jonathan Knighten, and Shiri Azenkot. 2016. CueSee: exploring visual cues for people with low vision to facilitate a visual search task. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 73–84.
- [201] Jianlong Zhou, Heimo Müller, Andreas Holzinger, and Fang Chen. 2023. Ethical ChatGPT: Concerns, challenges, and commandments. *arXiv preprint arXiv:2305.10646* (2023).
- [202] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16816–16825.
- [203] Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI Conference on Artificial Intelligence*. 7590–7598. <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344>
- [204] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11477–11486.