

# **From Causal Structure to Conditional Generation**

Yuan Guan

CMU-CS-26-120

August 2026

Computer Science Department  
School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15213

**Thesis Committee:**

Pradeep Ravikumar, Chair  
Bryan Wilder

*Submitted in partial fulfillment of the requirements  
for the degree of Masters of Science in Computer Science.*

**Keywords:** Conditional Generation, Causal Representation Learning, Compositional Generation, Score-Based Diffusion, Trajectory Forecasting, Information Fusion

## Abstract

Conditional generation sometimes relies on structural knowledge that is not fully represented in the training data. This thesis studies two ways such knowledge can enter a generative model: as an inference-time rule for constructing an unseen target distribution and as a training-time inductive bias.

The first project studies inference-time composition of learned diffusion scores. Under independent causal mechanisms, a compound intervention can be identified at the clean-data level as a geometric composition of single-intervention distributions. However, this identification result does not imply that the same score composition is exact at intermediate diffusion times. We distinguish approximation error, caused by the noncommutativity of noising and composition, from estimation error in the learned source scores. In the controlled analytical and empirical settings studied here, estimation error can dominate when the composed target lies in low-density regions of the source distributions. In these settings, Feynman–Kac Correctors may reduce sampling approximation error while worsening final sample quality by more faithfully following an inaccurate learned composition.

The second project considers multimodal vessel trajectory prediction under a domain-informed directional prior: environmental context influences vessel dynamics over the forecasting horizon, whereas vessel motion does not meaningfully determine the surrounding mesoscale environment. The main contribution is the Conditional Informer, which encodes this asymmetry through directional cross-attention while preserving vessel state as the primary prediction stream. We use standard modality masking during training to evaluate the same architecture under different patterns of auxiliary-input availability. In the dataset and deployment conditions considered, the results suggest that the architectural prior provides a useful organization of contextual information, while exposure to missing-modality cases is still needed for robust fallback when auxiliary inputs are absent.

Together, the two projects examine distinct ways of incorporating structural knowledge into conditional generation: through inference-time composition in diffusion models and through architectural design in multimodal trajectory forecasting.



## Acknowledgments

I am deeply grateful to my advisor, Professor Pradeep Ravikumar, for his guidance, patience, and intellectual support throughout this work. His feedback sharpened both the technical arguments in this thesis and the broader questions connecting its two projects. I am also grateful to Chandler Squires for his mentorship, collaboration, and generosity throughout my research at CMU.

I would like to thank Professors Krzysztof Mierzewski and Anil Ada for their patient guidance during my undergraduate years and for supporting my path to graduate study. Professor Ada's course, *Great Ideas in Theoretical Computer Science*, sparked my interest in computer science and helped motivate my change of major. Professor Mierzewski, who taught me while I was studying philosophy, gave me an interdisciplinary foundation and ways of thinking that have continued to shape my work.

I am thankful to my friends for their sincere conversations and companionship, and to my parents for making my education possible. Above all, I thank my girlfriend, Ronan, for putting up with my madness. Her encouragement, understanding, and support made the difficult and uncertain parts of this process much easier to carry.



# Contents

|          |                                                                      |           |
|----------|----------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                                  | <b>1</b>  |
| 1.1      | Limits of inference-time compositional sampling . . . . .            | 1         |
| 1.2      | Encoding known structure at training time . . . . .                  | 2         |
| 1.3      | A shared perspective . . . . .                                       | 2         |
| 1.4      | Outline . . . . .                                                    | 3         |
| <b>2</b> | <b>Background</b>                                                    | <b>5</b>  |
| 2.1      | Overview . . . . .                                                   | 5         |
| 2.2      | Causal Representation Learning . . . . .                             | 6         |
| 2.3      | Independent Causal Mechanisms and Geometric Compositions . . . . .   | 6         |
| 2.4      | Conditional Diffusion Models . . . . .                               | 7         |
| 2.5      | Structured Multimodal Trajectory Prediction . . . . .                | 9         |
| <b>3</b> | <b>Limits of Inference-Time Diffusion Composition</b>                | <b>11</b> |
| 3.1      | The Compositional Generation Task . . . . .                          | 12        |
| 3.1.1    | Setup and Notation . . . . .                                         | 12        |
| 3.1.2    | Two Sources of Error . . . . .                                       | 12        |
| 3.2      | Inference-Time Approximation Error . . . . .                         | 13        |
| 3.2.1    | The Naive Composition Heuristic . . . . .                            | 13        |
| 3.2.2    | Gaussian Analysis of Approximation Error . . . . .                   | 14        |
| 3.2.3    | Factorized Conditionals: When Approximation Error Vanishes . . . . . | 14        |
| 3.3      | Experiments . . . . .                                                | 15        |
| 3.3.1    | Two-dimensional Gaussian experiments . . . . .                       | 16        |
| 3.3.2    | Gaussian mixture experiments . . . . .                               | 17        |
| 3.3.3    | Semi-synthetic data: Objects in a room . . . . .                     | 18        |
| 3.3.4    | Discussion and conclusion . . . . .                                  | 19        |
| <b>4</b> | <b>Structured Multimodal Vessel Trajectory Prediction</b>            | <b>23</b> |
| 4.1      | Problem Formulation . . . . .                                        | 23        |
| 4.1.1    | States, Contexts, and Metadata . . . . .                             | 24        |
| 4.1.2    | Directional Structural Prior . . . . .                               | 24        |
| 4.1.3    | Prediction Task and Training Regimes . . . . .                       | 24        |
| 4.2      | Design Motivation . . . . .                                          | 25        |
| 4.3      | The Conditional Informer . . . . .                                   | 26        |

|          |                                             |           |
|----------|---------------------------------------------|-----------|
| 4.3.1    | Modality-Specific Embeddings . . . . .      | 26        |
| 4.3.2    | Conditional Attention . . . . .             | 27        |
| 4.3.3    | Conditional Encoder and Decoder . . . . .   | 27        |
| 4.3.4    | Training under Missing Modalities . . . . . | 28        |
| 4.4      | Experiments . . . . .                       | 28        |
| 4.4.1    | Data and Preprocessing . . . . .            | 28        |
| 4.4.2    | Models and Baselines . . . . .              | 29        |
| 4.4.3    | Training and Evaluation Protocol . . . . .  | 29        |
| 4.4.4    | Results . . . . .                           | 29        |
| 4.4.5    | Discussion and conclusion . . . . .         | 30        |
| <b>5</b> | <b>Conclusion</b>                           | <b>31</b> |
| 5.1      | Summary of Findings . . . . .               | 31        |
| 5.2      | Limitations . . . . .                       | 32        |
| 5.3      | Future Work . . . . .                       | 33        |
|          | <b>Bibliography</b>                         | <b>35</b> |

# List of Figures

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |    |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1 | <b>Out-of-distribution Gaussian composition with analytic and learned scores.</b> The sources are $P^0 = \mathcal{N}(0, I)$ , $P^{a_1} = \mathcal{N}(0, \text{diag}(10, 1))$ , and $P^{a_2} = \mathcal{N}(0, \text{diag}(1, 10))$ , giving the base-composed target $P^w(x) \propto P^{a_1}(x)P^{a_2}(x)/P^0(x) = \mathcal{N}(0, 10I)$ . Analytic scores recover the intended composition, whereas learned scores are inaccurate in the target-relevant low-density region; increasing FKC correction follows this inaccurate learned composition more strongly. . . . .                                                                                                                               | 20 |
| 3.2 | <b>Cross-shaped Gaussian-mixture composition in the out-of-distribution setting.</b> The first condition has component means $\{(-6, 0), (6, 0)\}$ and the second has means $\{(0, -6), (0, 6)\}$ , so the base composition emphasizes regions weakly supported by either learned source model. Analytic-score sampling remains comparatively stable as the FKC particle count increases, while learned-score sampling deteriorates because stronger correction follows errors in the estimated composition. . . . .                                                                                                                                                                                   | 20 |
| 3.3 | <b>Compositional room generation across factorized and OOD settings.</b> Naive denoising ( $K = 1$ ) is compared with FKC ( $K = 16$ ) in four settings. Panels (a) and (c) compose couches with paintings, an approximately factorized pairing because the objects typically occupy disjoint spatial regions; panels (b) and (d) compose couches with coffee tables, a non-factorized pairing because both objects compete for floor space. The top row is in-distribution and the bottom row is out-of-distribution. FKC is most effective in the in-distribution non-factorized setting, whereas the OOD samples remain incomplete or visibly distorted as learned-score error accumulates. . . . . | 21 |
| 4.1 | <b>Architecture of the Conditional Informer.</b> (a) Encoder with state self-attention, context cross-attention, and metadata bias. (b) Decoder with masked self-attention and conditional cross-attention over encoder representations. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 25 |
| 4.2 | <b>Example state and context trajectories for context-informed ship trajectory prediction.</b> In (a), LAT, LON, SOG, and COG are short for <i>latitude</i> , <i>longitude</i> , <i>speed over ground</i> , and <i>course over ground</i> , respectively. In (b), SWH and WS are short for <i>significant wave height</i> and <i>wind speed</i> , respectively. . . . .                                                                                                                                                                                                                                                                                                                                | 28 |



# List of Tables

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |    |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1 | <b>Gaussian composition results across factorization, target overlap, training-set size, and FKC particle count.</b> FKC is applied with $K$ particles to a shared conditional diffusion model trained on $N$ samples, or to analytic source scores. Each entry is the mean sliced Wasserstein-2 distance ( $SW_2$ ) between 5,000 generated samples and 5,000 target samples over 30 independently trained models; lower is better. . . . .                                                                                      | 16 |
| 3.2 | <b>Gaussian-mixture composition results for in-distribution and revised cross-shaped out-of-distribution targets.</b> Entries report mean sliced Wasserstein-2 distance ( $SW_2$ ), with lower values indicating better agreement with the reference target. The in-distribution results average 30 independently trained models; the revised OOD results average 3 runs. $K$ is the FKC particle count and $N$ is the number of training samples per source condition. . . . .                                                   | 18 |
| 4.1 | <b>Trajectory forecasting results across training and test-time modality subsets.</b> Each cell reports mean $\pm$ standard deviation over five random seeds for mean Haversine distance (km) and normalized-state MSE ( $\times 10^{-3}$ ). $S$ , $C$ , and $M$ denote vessel state, environmental context, and static metadata. Bold entries mark the best result within each test-time column. The no-masking row uses the same Conditional Informer architecture but is trained only with complete multimodal inputs. . . . . | 30 |



# Chapter 1

## Introduction

Conditional generative models are often expected to operate beyond the exact conditions represented in their training data. A model may need to combine separately observed factors, use auxiliary context that follows a known physical direction, or remain useful when some inputs are unavailable at deployment. These are not arbitrary extrapolation problems: in each case, domain knowledge provides structure that narrows the intended behavior.

This thesis asks how such structure can be used without assuming that specifying the desired relation is sufficient to realize it. The first project studies an inference-time setting: a clean target distribution is specified by composing observed source distributions, and we ask whether pretrained diffusion models can sample from it accurately. The second studies a training-time setting: a directional prior between vessel state and environmental context is encoded in the architecture of a multimodal trajectory predictor.

The two projects use different models, data, and evaluation criteria, and are not intended as a controlled comparison. They address a shared practical question: once structural knowledge specifies how conditions should combine or interact, what additional modeling, estimation, or training constraints are needed for reliable prediction and generation? The following sections summarize how each project approaches this question; the formal background is deferred to Chapter 2.

### 1.1 Limits of inference-time compositional sampling

The diffusion project asks whether inference-time procedures can sample from a geometrically composed target using pretrained conditional diffusion models. This setting is motivated by the CRL framework: under ICM and suitable intervention assumptions, a compound-intervention distribution can be identified from the corresponding single-intervention and control distributions. The clean target is therefore specified at the distributional level, but the model has not been trained directly on samples from that target.

We distinguish two sources of error. *Inference-time approximation error* arises because geometric composition of clean densities does not, in general, commute with the Gaussian noising process used by diffusion models. Consequently, even exact source scores need not yield the exact score of the noised compound target when combined naively at intermediate diffusion times.

*Score estimation error* arises because learned source scores may be inaccurate in low-density regions that become important for an out-of-distribution composed target.

We analyze approximation error in tractable Gaussian settings and identify factorized cases in which noising preserves the relevant separation and naive score composition becomes exact. In the controlled learned-score experiments considered here, score estimation error can dominate the remaining sampling approximation error when the target lies far from the concentration regions of the source distributions. We further study Feynman–Kac Correctors (FKC), a particle-based method designed to reduce inference-time approximation error. Stronger correction improves sampling in some settings with accurate source scores, but can reduce final sample quality in the out-of-distribution learned-score settings studied here by following an inaccurate composed score more faithfully. These results show that, within the evaluated models and tasks, inference-time correction does not compensate for source scores that are poorly constrained in the regions required by the target.

## 1.2 Encoding known structure at training time

The trajectory project studies multimodal vessel trajectory prediction under a domain-informed directional prior. Over the forecasting horizon considered here, environmental variables such as wind and wave conditions plausibly influence vessel motion, whereas the motion of an individual vessel is unlikely to materially alter the surrounding mesoscale weather. This partial ordering motivates an asymmetric model: vessel state is treated as the primary dynamical stream, while weather and vessel metadata act as conditioning information.

Standard multimodal architectures often concatenate modalities or encode them in parallel before symmetric fusion. Such designs are flexible, but they do not explicitly distinguish the predicted dynamical state from auxiliary context. The principal methodological contribution of this project is the Conditional Informer, which encodes the assumed directional structure through cross-attention: vessel-state representations query environmental context, while the context stream is not updated symmetrically by vessel state.

We additionally use modality masking during training to evaluate the model under missing-input conditions. This is a conventional robustness technique rather than a separate methodological contribution. Its purpose is to expose the structurally organized model to the modality-availability patterns considered at deployment. In the dataset studied here, the results suggest that directional conditioning is a useful way to organize contextual information, while exposure to missing-modality cases remains important for maintaining a usable fallback pathway. This project should therefore be understood as an ICM-inspired use of a domain prior, not as a direct application of the compound-intervention identification result or as evidence that the assumed ordering is a complete causal model.

## 1.3 A shared perspective

The two projects address different tasks and are not a controlled comparison of training-time and inference-time methods. They nevertheless motivate the same practical question: once structural

assumptions specify a target distribution or interaction pattern, what else is required for a learned model to realize it reliably?

In the diffusion setting, performance depends on the accuracy of source scores in the regions emphasized by the composed target. In the trajectory setting, directional attention organizes how context enters the predictor, while performance under missing inputs also depends on whether those conditions appear during training. The thesis therefore treats structural specification, model estimation, and training coverage as distinct parts of reliable conditional generation.

## 1.4 Outline

The remainder of this thesis is organized as follows:

- Chapter 2 introduces the causal representation learning framework, the geometric composition implied by modular interventions, and the diffusion-model notation required for the compositional-generation analysis. It also distinguishes this formal identification result from the domain-informed structural prior used in the trajectory project.
- Chapter 3 analyzes inference-time composition of diffusion scores, separating noising-induced approximation error from estimation error in learned source scores.
- Chapter 4 presents the Conditional Informer and evaluates directional multimodal fusion and Modality Masking for vessel trajectory prediction under changing modality availability.
- Chapter 5 summarizes the findings, states their limitations, and outlines directions for future work.



# Chapter 2

## Background

This chapter develops the background required for the thesis’s two research questions. We first state the questions and explain how the chapter is organized. We then introduce causal representation learning, independent causal mechanisms, and conditional diffusion models, which provide the formal basis for the diffusion composition project in Chapter 3. Finally, we review structured multimodal trajectory prediction, directional attention, and missing-modality robustness for Chapter 4. The trajectory project uses a domain-informed physical asymmetry as an architectural prior; it is not derived from the compound-intervention identification result.

### 2.1 Overview

This thesis studies two distinct uses of structural knowledge in conditional generation. The first concerns whether an unseen distribution identified through causal modularity can be sampled using pretrained diffusion models. The second concerns how a domain-informed directional prior and deployment-time modality availability should shape model architecture and training. We state these questions first so that the technical background that follows is motivated by the role it plays in the thesis.

**Question 1: Sampling from an identified compositional target.** Under the assumptions reviewed in Section 2.3, a compound-intervention distribution can be identified at the clean-data level through (2.4). This does not imply that a diffusion sampler can recover it by linearly composing source scores at every noise level. Chapter 3 asks when inference-time score composition succeeds, separates error caused by the sampling approximation from error in the learned source scores, and studies why the latter can dominate when the target occupies low-density regions of the source distributions. Sections 2.2–2.4 develop the causal and diffusion-model background for this question.

**Question 2: Encoding directional structure during training.** In many forecasting problems, domain knowledge provides a partial ordering among inputs even when a complete causal graph is unavailable. In the maritime setting studied here, environmental variables influence vessel motion over the prediction horizon, whereas an individual vessel does not meaningfully determine

the surrounding mesoscale environment. How should this asymmetry be represented in the architecture, and what training procedure is needed to preserve performance when auxiliary modalities are unavailable at deployment? Chapter 4 studies this question through directional cross-attention and Modality Masking. The resulting model should be understood as incorporating a physical structural prior, not as recovering or identifying a full causal graph. Section 2.5 reviews the trajectory-forecasting, multimodal-conditioning, and robustness background for this question.

The questions are complementary but not equivalent. Chapter 3 concerns whether an already trained model contains sufficiently accurate information to realize a newly identified target at inference time, whereas Chapter 4 concerns how a model is trained to use structured inputs robustly.

## 2.2 Causal Representation Learning

We follow the causal representation learning (CRL) framework of [24]. A **causal representation model** is a tuple  $\mathcal{M} = (\mathcal{S}, g)$ , where  $\mathcal{S}$  is a structural causal model (SCM) over a latent space  $\mathcal{Z} = \mathbb{R}^d$  and  $g: \mathcal{Z} \rightarrow \mathcal{X}$  is a diffeomorphism onto the observation space  $\mathcal{X}$ . The SCM  $\mathcal{S}$  is associated with a directed acyclic graph (DAG)  $G$  over latent variables  $z_1, \dots, z_d$ , and induces a latent observational distribution that factorizes along  $G$ :

$$Q^{\text{obs}}(z) = \prod_{i=1}^d Q^{\text{obs}}(z_i \mid \text{pa}_G(z_i)). \quad (2.1)$$

The corresponding observation-space distribution is

$$P^{\text{obs}}(x) = Q^{\text{obs}}(g^{-1}(x)) \cdot |\det J_{g^{-1}}(x)|, \quad (2.2)$$

where  $J_{g^{-1}}$  denotes the Jacobian of  $g^{-1}$ .

**Interventions.** An **intervention**  $\mathcal{I}$  on  $\mathcal{M}$  with targets  $T(\mathcal{I}) \subseteq [d]$  replaces the conditional mechanisms of the targeted variables while leaving the remaining mechanisms unchanged, yielding the interventional latent distribution

$$Q^{\mathcal{I}}(z) = \prod_{i \notin T(\mathcal{I})} Q^{\text{obs}}(z_i \mid \text{pa}_G(z_i)) \cdot \prod_{i \in T(\mathcal{I})} Q^{\mathcal{I}}(z_i \mid \text{pa}_G(z_i)). \quad (2.3)$$

The observation-space interventional distribution is then  $P^{\mathcal{I}}(x) = Q^{\mathcal{I}}(g^{-1}(x)) \cdot |\det J_{g^{-1}}(x)|$ . For notational convenience, we write  $P^a$  for the distribution associated with intervention  $\mathcal{I}_a$ , and  $P^0 := P^{\text{obs}}$  for the observational distribution.

## 2.3 Independent Causal Mechanisms and Geometric Compositions

A central principle in causal representation learning is the **independence of causal mechanisms** (ICM) [19]: the conditional mechanisms associated with different variables are treated as autonomous modules. For the intervention model considered here, this modularity has a useful

consequence. Suppose interventions  $\mathcal{I}_a$  and  $\mathcal{I}_{a'}$  replace disjoint conditional mechanisms,

$$T(\mathcal{I}_a) \cap T(\mathcal{I}_{a'}) = \emptyset,$$

and the compound intervention  $\mathcal{I}_{a,a'}$  applies both replacements while leaving all remaining mechanisms unchanged. Under the modular intervention framework of Varici et al. [24], the corresponding observation-space densities satisfy

$$\frac{P^{a,a'}(x)}{P^0(x)} = \frac{P^a(x)}{P^0(x)} \frac{P^{a'}(x)}{P^0(x)}, \quad (2.4)$$

where the ratios are defined on their common support. Equivalently,

$$P^{a,a'}(x) = \frac{P^a(x)P^{a'}(x)}{P^0(x)}. \quad (2.5)$$

In this modular-intervention setting the right-hand side is already normalized. We nevertheless introduce a normalized geometric-composition notation below because the diffusion analysis also considers more general weights for which a normalizing constant is required.

**Definition 2.1** (Geometric composition [28]). *Fix observed conditions  $\mathcal{A}_o$  with associated densities  $\mathbf{P} = (P^a)_{a \in \mathcal{A}_o}$ . Let  $w = (w_a)_{a \in \mathcal{A}_o} \in \mathbb{R}^{|\mathcal{A}_o|}$  be such that  $\prod_{a \in \mathcal{A}_o} P^a(x)^{w_a}$  is integrable and has a finite, positive normalizing constant. The **geometric composition** is*

$$P^w(x) := \frac{1}{Z_w} \prod_{a \in \mathcal{A}_o} P^a(x)^{w_a}, \quad Z_w := \int_{\mathcal{X}} \prod_{a \in \mathcal{A}_o} P^a(x)^{w_a} dx. \quad (2.6)$$

The two-intervention identity (2.5) is the **base composition** with weights  $w_a = w_{a'} = 1$  and  $w_0 = -1$ . More generally, combining  $k$  disjoint interventions gives unit weight to each intervention distribution and weight  $1 - k$  to the base distribution. This connection motivates Chapter 3: given models for the observed source distributions, can an inference-time procedure realize the identified compound target?

**Clean-data score identity.** On points where all relevant densities are positive and differentiable, the logarithm of (2.6) converts the geometric composition into a linear score identity. Defining  $S^a(x) := \nabla_x \log P^a(x)$ , we obtain

$$S^w(x) := \nabla_x \log P^w(x) = \sum_{a \in \mathcal{A}_o} w_a S^a(x). \quad (2.7)$$

The derivative of the constant  $\log Z_w$  vanishes, so the identity is exact at the clean-data level. It does not imply that the same linear combination remains the score of the correctly noised target at intermediate diffusion times. That distinction between clean-target identification and diffusion-time realization is central to Chapter 3.

## 2.4 Conditional Diffusion Models

We review the score-based diffusion notation needed for Chapter 3 [9, 21].

**Forward noising process.** Given a data distribution  $P^a$  over  $\mathcal{X}$ , consider the linear forward SDE

$$dx_t = \mu_t x_t dt + \sigma_t dW_t, \quad (2.8)$$

where  $W_t$  is a standard Wiener process,  $\mu_t$  is the drift schedule, and  $\sigma_t$  is the diffusion-coefficient schedule. We take  $t = 0$  to denote clean data and  $t = 1$  to denote the terminal noise level. Conditional on  $x_0$ , the transition kernel is Gaussian:

$$q_t(x_t | x_0) = \mathcal{N}(x_t; \alpha_t x_0, \gamma_t^2 I), \quad \alpha_t := \exp\left(\int_0^t \mu_s ds\right), \quad \gamma_t := \alpha_t \sqrt{\int_0^t \frac{\sigma_s^2}{\alpha_s^2} ds}. \quad (2.9)$$

The noised marginal is therefore the convolution and scaling induced by this kernel,

$$P_t^a(x_t) = \int q_t(x_t | x_0) P^a(x_0) dx_0, \quad (2.10)$$

which is not generally Gaussian unless  $P^a$  is Gaussian. We write  $S_t^a(x) := \nabla_x \log P_t^a(x)$  for its score.

**Reverse-time generation.** With the exact score, the reverse-time SDE has drift

$$v_t^a(x) := \mu_t x - \sigma_t^2 S_t^a(x), \quad (2.11)$$

and, when integrated from the terminal distribution at  $t = 1$  back to  $t = 0$ , recovers samples from  $P^a$  [21]. In practice, a neural network estimates the time-dependent conditional scores from noisy training examples. Under an  $\epsilon$ -prediction parameterization for the kernel in (2.9),

$$S_t^a(x_t) \approx -\frac{\epsilon_\theta(t, x_t, a)}{\gamma_t}. \quad (2.12)$$

A single conditional model may share parameters across all observed conditions  $a \in \mathcal{A}_o$  while receiving  $a$  as an input.

**Naive score composition.** For a clean target  $P^w$  defined by (2.6), let

$$P_t^w := \text{NOISE}_t(P^w), \quad S_t^w(x) := \nabla_x \log P_t^w(x) \quad (2.13)$$

denote the correctly noised target and its score. A common inference-time heuristic [8, 13] instead uses

$$S_t^{w, \text{nv}}(x) := \sum_{a \in \mathcal{A}_o} w_a S_t^a(x). \quad (2.14)$$

At  $t = 0$ , this equals the clean-target score by (2.7). At intermediate times, however,

$$\text{NOISE}_t(\text{COMP}_w(\mathbf{P})) \neq \text{COMP}_w((\text{NOISE}_t(P^a))_{a \in \mathcal{A}_o}) \quad (2.15)$$

in general. Thus  $S_t^{w, \text{nv}}$  need not equal  $S_t^w$ . This noncommutativity creates inference-time approximation error even when the source scores are known exactly. Chapter 3 separates this error from estimation error in learned source scores and studies how the two interact with Feynman–Kac correction.

## 2.5 Structured Multimodal Trajectory Prediction

The second research question concerns how directional structure and expected modality availability should shape a conditional predictor during training. This section reviews the sequence-modeling and robustness concepts needed for Chapter 4.

**Vessel trajectory forecasting.** AIS-based vessel forecasting has been studied with recurrent, convolutional, and attention-based sequence models. Capobianco et al. [3] survey and compare deep-learning approaches for predicting vessel motion from historical AIS records. Informer-TP [30] adapts the Informer architecture [32] to long-horizon ship trajectory prediction and provides the state-only forecasting backbone adapted in Chapter 4. Other recent work incorporates multiple vessel attributes or environmental measurements through joint multimodal representations [4, 31]. These approaches establish that auxiliary information can improve predictive accuracy, but they do not by themselves determine how performance should degrade when those inputs are absent at deployment.

**Multimodal conditioning and side information.** A standard strategy for incorporating auxiliary variables is early fusion, in which all features are concatenated before sequence modeling. More structured alternatives encode modalities separately and combine their representations later. The broader literature on learning with privileged or side information studies how information available during training can improve a predictor even when it may not be available at test time [11, 14, 22, 23]. The setting of Chapter 4 differs in that auxiliary modalities may be available for some deployment examples but absent for others. The goal is therefore a single predictor that benefits from context when present while retaining a useful state-only pathway.

**Directional attention as a structural prior.** Transformers provide a modular mechanism for restricting which representation supplies queries, keys, and values [27]. Causal and direction-aware Transformer variants have used attention structure to encode assumptions about temporal or causal relationships [16, 17]. In this thesis, directional cross-attention is not treated as a causal-effect estimator or as a method for recovering a causal graph. It is an architectural constraint expressing the intended role of environmental context and vessel metadata as conditioning information for the primary state stream.

**Shortcut dependence and missing modalities.** Deep models often exploit highly predictive cues that perform well under the training distribution but fail when those cues are removed or shifted [6]. In a multimodal forecaster, this can appear as excessive dependence on an auxiliary stream. Evaluating the same trained model under different modality-availability patterns provides an operational test of this behavior. Modality Masking functions as structured input dropout by exposing the model during training to the fallback conditions that will be evaluated at deployment. This distinction is central to the second research question: architectural structure determines which interactions are permitted, while the training distribution determines whether the model actually learns to operate when particular conditioning variables are absent.



# Chapter 3

## Limits of Inference-Time Diffusion Composition

**Reproduction and collaboration note.** This chapter is adapted from the collaborative manuscript *Catastrophic Compositional Generation: Why Vanilla Diffusion Models Fail to Extrapolate*, joint work with Duncan Soiffer, Chandler Squires, Jason Hartford, and Pradeep Ravikumar.

Chapter 2 showed that, under the independent causal mechanisms assumption and non-overlapping intervention targets, an unseen compound-intervention distribution can be identified at the clean-data level as a geometric composition of observed source distributions. This identification result reduces a causal question to a distributional one, but it does not provide an exact sampler when the available source models are diffusion models.

This chapter studies inference-time composition of learned conditional diffusion scores. We separate two failure modes. First, even exact source scores generally do not compose correctly at intermediate noise levels because Gaussian noising and geometric composition do not commute. Second, learned source scores may be inaccurate in the low-density regions emphasized by an out-of-distribution composition. These are respectively an inference-time approximation problem and a model-estimation problem.

The distinction matters when evaluating Feynman–Kac Correctors (FKC) [20]. FKC is designed to reduce approximation error by correcting the reverse-time particle system toward the composition implied by the supplied score fields. It does not improve those score fields themselves. Consequently, when the learned composition is inaccurate, a stronger correction can reduce approximation error while increasing discrepancy from the true target.

Throughout the chapter, we restrict attention to *vanilla conditional diffusion models*: shared-parameter networks trained with standard denoising objectives and combined only at inference time, without architectural or training objectives tailored to compositional extrapolation. The results characterize this setting and should not be interpreted as an impossibility result for compositional generation more broadly.

## 3.1 The Compositional Generation Task

### 3.1.1 Setup and Notation

We work within the geometric composition framework introduced in Chapter 2. Let  $\mathcal{A}_o \subseteq \mathcal{A}$  denote the set of observed conditions, with ground-truth source distributions  $\mathbf{P}_\star = (P_\star^a)_{a \in \mathcal{A}_o}$ . For a valid weighting function  $w: \mathcal{A}_o \rightarrow \mathbb{R}$ , the desired clean-data target is

$$P_\star^w := \text{COMP}_w(\mathbf{P}_\star). \quad (3.1)$$

We focus on the **base-composed distribution**, which arises from the ICM identity for non-overlapping intervention targets:

$$P_\star^w(x) \propto P_\star^0(x)^{w_0} \prod_{a \in \mathcal{A}_o^-} P_\star^a(x), \quad \mathcal{A}_o^- := \mathcal{A}_o \setminus \{0\}, \quad (3.2)$$

with  $w_a = 1$  for  $a \in \mathcal{A}_o^-$  and  $w_0 = 1 - |\mathcal{A}_o^-|$ . Here  $P_\star^0$  denotes the observational or base distribution.

A conditional diffusion model provides learned source scores  $\hat{S}_t^a$  and therefore implicitly defines learned source distributions  $\hat{P}^a$ . Their clean-data geometric composition is

$$\hat{P}^w := \text{COMP}_w((\hat{P}^a)_{a \in \mathcal{A}_o}). \quad (3.3)$$

This distribution is the target implied by the learned source models. An inference-time sampler does not generally sample from  $\hat{P}^w$  exactly; instead, it produces an output distribution  $\tilde{P}^w$  by integrating a reverse-time process based on naively composed or corrected learned scores.

The compositional generation task is therefore to construct  $\tilde{P}^w$  that is close to the true target  $P_\star^w$  using only the pretrained conditional model and an efficient inference procedure.

### 3.1.2 Two Sources of Error

For any metric  $d$  satisfying the triangle inequality,

$$d(\tilde{P}^w, P_\star^w) \leq \underbrace{d(\tilde{P}^w, \hat{P}^w)}_{\text{inference-time approximation error}} + \underbrace{d(\hat{P}^w, P_\star^w)}_{\text{model or score estimation error}}. \quad (3.4)$$

Equation (3.4) separates three distributions that are easily conflated: the true composed target  $P_\star^w$ , the composition  $\hat{P}^w$  implied by the learned source models, and the actual sampler output  $\tilde{P}^w$ .

**Inference-time approximation error.** Even if the learned source scores were exact for their learned distributions, naively composing their noisy-time scores would not generally recover the noisy marginals of  $\hat{P}^w$ . The resulting reverse process therefore targets a proxy distribution rather than  $\hat{P}^w$ . This error is caused by the inference procedure and can, in principle, be reduced by a more accurate sampler or correction method.

**Model or score estimation error.** The learned source scores may differ from the ground-truth source scores, especially in low-density regions that receive little constraint from training data. Geometric composition can place substantial target mass precisely in such regions. Consequently, even an exact sampler for  $\hat{P}^w$  may remain far from  $P_\star^w$ . We use the term *score estimation error* because the diffusion models are learned through their score fields, while recognizing that the second term in (3.4) is ultimately a discrepancy between composed distributions.

The two errors require different remedies. Approximation error is an inference problem; estimation error is a learning problem. An inference-time corrector can reduce the first term without reducing the second. Indeed, if  $\hat{P}^w$  is inaccurate, following it more faithfully can worsen distance to the true target. This distinction organizes the theoretical and empirical results that follow.

## 3.2 Inference-Time Approximation Error

We first isolate the error that remains even when the source score functions are known exactly. The key issue is that geometric composition is exact at the clean-data level, whereas diffusion sampling requires scores at every intermediate noise level.

### 3.2.1 The Naive Composition Heuristic

At  $t = 0$ , the clean composed score satisfies

$$S^w(x) = \sum_{a \in \mathcal{A}_o} w_a S^a(x). \quad (3.5)$$

A tempting extension is to use the same linear combination after independently noising each source distribution:

$$S_t^{w,\text{nv}}(x) := \sum_{a \in \mathcal{A}_o} w_a S_t^a(x). \quad (3.6)$$

However, the score required to sample from the desired target is

$$S_t^w(x) := \nabla_x \log[\text{NOISE}_t(P^w)](x), \quad (3.7)$$

whereas (3.6) is the score of a geometric composition of the separately noised source marginals. In general,

$$\text{NOISE}_t(\text{COMP}_w(\mathbf{P})) \neq \text{COMP}_w((\text{NOISE}_t(P^a))_{a \in \mathcal{A}_o}). \quad (3.8)$$

Thus, ICM identifies the clean target but does not imply exact score composition along the diffusion path.

The naive reverse-time SDE substitutes  $S_t^{w,\text{nv}}$  for  $S_t^w$ :

$$dx_t^{w,\text{nvs}} = v_t^{w,\text{nv}}(x_t^{w,\text{nvs}}) dt + \sigma_t dW_t, \quad v_t^{w,\text{nv}}(x) := \mu_t x - \sigma_t^2 S_t^{w,\text{nv}}(x). \quad (3.9)$$

The corresponding probability-flow ODE is

$$dx_t^{w,\text{nvo}} = \tilde{v}_t^{w,\text{nv}}(x_t^{w,\text{nvo}}) dt, \quad \tilde{v}_t^{w,\text{nv}}(x) := \mu_t x - \frac{1}{2} \sigma_t^2 S_t^{w,\text{nv}}(x). \quad (3.10)$$

The discrepancy induced by (3.8) is the approximation error defined in Section 3.1.2.

### 3.2.2 Gaussian Analysis of Approximation Error

To characterize this error analytically, we consider weighted compositions of zero-mean multivariate Gaussians.

**Theorem 3.1** (Gaussian composition). *Let  $P^a = \mathcal{N}(0, \Sigma_a)$  for  $a \in \mathcal{A}_o$ . Then  $w$  is valid if and only if  $\sum_{a \in \mathcal{A}_o} w_a \Sigma_a^{-1} \succ 0$ , in which case*

$$P^w = \mathcal{N}\left(0, \left(\sum_{a \in \mathcal{A}_o} w_a \Sigma_a^{-1}\right)^{-1}\right). \quad (3.11)$$

The clean composition combines covariance matrices through a weighted sum of precision matrices. The naive probability-flow ODE produces a different covariance.

**Theorem 3.2** (Naive denoising ODE under Gaussians). *Let  $P^a = \mathcal{N}(0, \Sigma_a)$  for  $a \in \mathcal{A}_o$ , assume that the matrices  $\{\Sigma_a\}_{a \in \mathcal{A}_o}$  commute, and assume  $\sum_{a \in \mathcal{A}_o} w_a = 1$ . Then the naive denoising ODE (3.10) yields  $\tilde{P}^{w, \text{nvo}} = \mathcal{N}(0, \Sigma^{w, \text{nvo}})$  with*

$$\Sigma^{w, \text{nvo}} = \prod_{a \in \mathcal{A}_o} \Sigma_a^{w_a}. \quad (3.12)$$

Theorem 3.2 shows that the desired precision-weighted composition is replaced by a matrix geometric mean. The two can differ arbitrarily. For example, let  $|\mathcal{A}_o| = 2$ ,  $w_1 = w_2 = 1/2$ ,  $\Sigma_1 = \varepsilon I$ , and  $\Sigma_2 = \varepsilon^{-1} I$ . Then  $\Sigma^{w, \text{nvo}} = I$ , whereas

$$\Sigma^w = \frac{2\varepsilon}{1 + \varepsilon^2} I. \quad (3.13)$$

Consequently,  $\|\Sigma^{w, \text{nvo}}\|_{\text{op}} / \|\Sigma^w\|_{\text{op}} = (1 + \varepsilon^2) / (2\varepsilon) \rightarrow \infty$  as  $\varepsilon \rightarrow 0^+$ . This example establishes that approximation error can be large even with exact source scores.

### 3.2.3 Factorized Conditionals: When Approximation Error Vanishes

The noncommutativity in (3.8) is not unavoidable. It disappears when the intervention effects act on independent coordinate blocks that remain separated under the Gaussian noising kernel.

**Definition 3.3** (Factorized Conditionals, adapted from [1]). *A collection  $(P^0, P^{a_1}, \dots, P^{a_k})$  over  $\mathbb{R}^n$  consists of **Factorized Conditionals** if there exists a partition  $M_0, M_1, \dots, M_k$  of  $[n]$  such that*

$$P^{a_i}(x) = P^{a_i}(x_{M_i}) P^0(x_{M_i^c}), \quad P^0(x) = P^0(x_{M_0}) \prod_{i \in [k]} P^0(x_{M_i}). \quad (3.14)$$

This definition requires genuine separability across disjoint coordinate blocks; it is stronger than the ordinary DAG factorization satisfied by any distribution Markov to a causal graph. Each condition modifies one block while all other blocks retain their base mechanisms. Because independent Gaussian noising acts coordinate-wise, this product structure is preserved at every noise level.

**Theorem 3.4** (FKC under Factorized Conditionals). *Assume  $(P^0, P^{a_1}, \dots, P^{a_k})$  are Factorized Conditionals. Then, for the base composition (3.2), the naively composed score equals the true composed noisy score at every  $t \in [0, 1]$ . Equivalently, the Feynman–Kac Corrector log-weight drift satisfies  $g_t(x) = 0$  for all  $x$  and  $t$ , and the FKC process coincides with the naive reverse-time SDE (3.9).*

The theorem isolates a regime in which approximation error is exactly zero. It also provides a useful experimental control: with analytic scores, neither naive composition nor FKC should incur an approximation penalty in this setting. Any degradation observed after replacing analytic scores with learned scores must therefore arise from estimation and numerical effects rather than from the clean noncommutativity analyzed above.

*Remark 3.5.* Theorem 3.4 does not by itself imply that FKC amplifies learned-score errors. It only shows that, under exact factorized scores, the correct correction signal is zero. With learned scores, the estimated FKC weights may become nonzero because the score identities no longer cancel exactly. Whether resampling then improves or worsens final sample quality is an empirical question examined in Section 3.3.

### 3.3 Experiments

The experiments are designed to separate, as far as possible, the two terms in the decomposition of Section 3.1.2. We vary two properties of the problem. First, we compare *in-distribution* settings, in which the composed target concentrates in regions well represented by the source distributions, with *out-of-distribution* settings, in which the target concentrates in low-density regions of those sources. Second, we compare factorized and non-factorized compositions. Factorization controls whether naive score composition is exact under the idealized assumptions of Section 3.2.3; target overlap controls how accurately learned source scores are expected to be constrained in the region relevant to the composition.

For each synthetic task, we specify a base distribution  $P^0$  and two condition distributions  $P^{a_1}$  and  $P^{a_2}$ . The target is their base composition,

$$P^w(x) \propto \frac{P^{a_1}(x)P^{a_2}(x)}{P^0(x)}. \quad (3.15)$$

A shared conditional diffusion model is trained on samples from the three source distributions. We then apply naive score composition or a Feynman–Kac Corrector with  $K$  particles. We vary the number of training samples  $N$  and also report an oracle condition in which the analytic source scores are used. The oracle condition removes learned score-estimation error and therefore isolates the behavior of the inference procedure more cleanly.

We evaluate 5,000 generated samples against 5,000 reference target samples using sliced Wasserstein-2 distance ( $SW_2$ ). Unless otherwise noted, the tables report means over 30 independently trained models; the revised OOD Gaussian-mixture experiment reports means over 3 runs. In all result tables, rows index the FKC particle count  $K$ , columns index the training-set size  $N$ , and the analytic column uses exact source scores. Lower  $SW_2$  is better. Differences across  $K$  should be interpreted as trends in final sample quality, not as direct numerical estimates of the two error terms themselves.

Table 3.1: **Gaussian composition results across factorization, target overlap, training-set size, and FKC particle count.** FKC is applied with  $K$  particles to a shared conditional diffusion model trained on  $N$  samples, or to analytic source scores. Each entry is the mean sliced Wasserstein-2 distance ( $\text{SW}_2$ ) between 5,000 generated samples and 5,000 target samples over 30 independently trained models; lower is better.

| (a) Factorized, in-distribution |        |        |        |          | (b) Non-factorized, in-distribution |        |        |        |          |
|---------------------------------|--------|--------|--------|----------|-------------------------------------|--------|--------|--------|----------|
| $K \backslash N$                | 100    | 1000   | 10000  | analytic | $K \backslash N$                    | 100    | 1000   | 10000  | analytic |
| 1                               | 0.3552 | 0.0918 | 0.0727 | 0.0366   | 1                                   | 0.4403 | 0.1052 | 0.0861 | 0.0630   |
| 4                               | 0.4118 | 0.0909 | 0.0698 | 0.0349   | 4                                   | 0.5112 | 0.0831 | 0.0716 | 0.0403   |
| 16                              | 0.4443 | 0.0871 | 0.0697 | 0.0369   | 16                                  | 0.5442 | 0.0792 | 0.0671 | 0.0345   |
| 64                              | 0.4625 | 0.0915 | 0.0720 | 0.0367   | 64                                  | 0.5575 | 0.0841 | 0.0666 | 0.0339   |
| 256                             | 0.4642 | 0.0902 | 0.0716 | 0.0344   | 256                                 | 0.5610 | 0.0824 | 0.0690 | 0.0345   |

| (c) Factorized, out-of-distribution |        |        |        |          | (d) Non-factorized, out-of-distribution |        |        |        |          |
|-------------------------------------|--------|--------|--------|----------|-----------------------------------------|--------|--------|--------|----------|
| $K \backslash N$                    | 100    | 1000   | 10000  | analytic | $K \backslash N$                        | 100    | 1000   | 10000  | analytic |
| 1                                   | 0.9685 | 0.5932 | 0.3540 | 0.1149   | 1                                       | 1.0977 | 0.9998 | 0.7877 | 0.6434   |
| 4                                   | 1.9100 | 1.7264 | 1.1655 | 0.1092   | 4                                       | 2.0135 | 1.9324 | 1.2279 | 0.1760   |
| 16                                  | 2.5163 | 2.4914 | 1.8375 | 0.1173   | 16                                      | 2.6265 | 2.7259 | 1.7346 | 0.1012   |
| 64                                  | 2.7828 | 2.8479 | 2.2443 | 0.1165   | 64                                      | 2.9078 | 3.1444 | 2.0733 | 0.0908   |
| 256                                 | 2.9036 | 2.9945 | 2.4543 | 0.1055   | 256                                     | 3.0345 | 3.3345 | 2.2483 | 0.0846   |

### 3.3.1 Two-dimensional Gaussian experiments

Figure 3.1 visualizes the factorized OOD setting with  $P^0 = \mathcal{N}(0, I)$ ,  $P^{a_1} = \mathcal{N}(0, \text{diag}(10, 1))$ , and  $P^{a_2} = \mathcal{N}(0, \text{diag}(1, 10))$ . The analytic composition is recovered accurately, whereas the learned composition deteriorates under stronger correction because the target lies outside the well-constrained regions of the source models.

The Gaussian experiments provide a controlled  $2 \times 2$  design. Throughout, we set

$$P^{a_1} = \mathcal{N}(0, \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}), \quad P^{a_2} = \mathcal{N}(0, \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}), \quad (3.16)$$

and vary  $P^0$ .

**Factorized, in-distribution.** We take  $P^0 = \mathcal{N}(0, 10I)$ , yielding the target  $P^w = \mathcal{N}(0, I)$ . Under the factorized construction, naive composition is exact with analytic scores, so increasing the particle count should not provide a systematic benefit. Table 3.1 matches this prediction: the analytic column remains essentially unchanged across  $K$ . With learned scores, additional particles also provide no consistent improvement and can slightly worsen performance, especially when only  $N = 100$  samples are available. This setting shows that FKC is unnecessary when approximation error is absent.

**Factorized, out-of-distribution.** We instead take  $P^0 = \mathcal{N}(0, I)$ , yielding  $P^w = \mathcal{N}(0, 10I)$ . The factorized structure again removes approximation error in the oracle setting, and the analytic results remain stable as  $K$  increases. The learned models behave very differently: for every training-set size in the table,  $\text{SW}_2$  rises sharply with particle count. Because this degradation disappears when analytic scores are used, the result is consistent with FKC magnifying the consequences of learned-score error rather than correcting a genuine noising–composition mismatch.

**Non-factorized, in-distribution.** We set  $P^0 = \mathcal{N}(0, 20I)$ , for which  $P^w = \mathcal{N}(0, \frac{20}{21}I)$ . Here naive score composition has nonzero approximation error. In the analytic column, increasing  $K$  reduces  $\text{SW}_2$  from 0.0630 at  $K = 1$  to approximately 0.034 at larger particle counts. The same improvement appears for the better-trained learned models, although it is smaller and eventually saturates. For  $N = 100$ , however, increasing  $K$  worsens performance. Thus FKC helps only when the correction signal is more reliable than the score-estimation noise on which it is computed.

**Non-factorized, out-of-distribution.** Finally, we take  $P^0 = \mathcal{N}(0, 1.1I)$ , yielding  $P^w = \mathcal{N}(0, \frac{110}{21}I)$ . With analytic scores, FKC substantially improves the samples, confirming that approximation error is present and correctable. With learned scores, the trend reverses: performance worsens monotonically with  $K$  for all three values of  $N$ . This is the clearest controlled example of the distinction developed in this chapter. A more accurate inference procedure improves the oracle model but harms the learned model because the latter does not provide sufficiently accurate scores in the target-relevant region.

Taken together, the four Gaussian settings show that neither factorization nor target overlap alone determines the outcome. Factorization controls whether an inference correction is needed; target overlap and data quantity affect whether the learned scores are accurate enough for that correction to be useful.

### 3.3.2 Gaussian mixture experiments

We next consider non-factorized mixtures of Gaussians in  $\mathbb{R}^2$ . Every mixture component has isotropic covariance  $\sigma I$ . Because the target is a ratio of mixtures, it is generally not itself a Gaussian mixture and its normalizing constant is not available in closed form. We therefore construct reference samples using rejection sampling in the in-distribution task and importance sampling in the out-of-distribution task.

**In-distribution mixture.** The component means lie on a  $2 \times 3$  grid,  $\{-3, 0, 3\} \times \{-2.5, 2.5\}$ , with  $\sigma = 0.6$ . The base distribution assigns uniform mass to all six components, while the two conditions retain overlapping subsets. The resulting target remains concentrated near modes represented by both condition distributions. Table 3.2 shows a large benefit from FKC when scores are analytic or trained with  $N \geq 1000$  samples. For example, at  $N = 10,000$ ,  $\text{SW}_2$  falls from 1.2105 at  $K = 1$  to 0.1748 at  $K = 16$ . By contrast, the  $N = 100$  model deteriorates as  $K$  increases. This reproduces the Gaussian conclusion in a multimodal density: correction is

Table 3.2: **Gaussian-mixture composition results for in-distribution and revised cross-shaped out-of-distribution targets.** Entries report mean sliced Wasserstein-2 distance ( $SW_2$ ), with lower values indicating better agreement with the reference target. The in-distribution results average 30 independently trained models; the revised OOD results average 3 runs.  $K$  is the FKC particle count and  $N$  is the number of training samples per source condition.

| (a) In-distribution |        |        |        |          | (b) Out-of-distribution |        |        |        |          |
|---------------------|--------|--------|--------|----------|-------------------------|--------|--------|--------|----------|
| $K \backslash N$    | 100    | 1000   | 10000  | analytic | $K \backslash N$        | 100    | 1000   | 10000  | analytic |
| 1                   | 2.8470 | 1.6443 | 1.2105 | 1.2388   | 1                       | 0.8399 | 0.6451 | 0.3779 | 0.3575   |
| 4                   | 2.8598 | 0.4442 | 0.2120 | 0.2873   | 4                       | 1.2986 | 1.3118 | 0.7824 | 0.3682   |
| 16                  | 3.0575 | 0.3394 | 0.1748 | 0.1252   | 16                      | 1.8565 | 2.6129 | 1.6719 | 0.4037   |
| 64                  | 3.1566 | 0.3401 | 0.1877 | 0.1064   | 64                      | 2.8347 | 3.2287 | 2.3752 | 0.3858   |
| 256                 | 3.2169 | 0.3059 | 0.2097 | 0.1197   | 256                     | 3.3122 | 3.3269 | 3.0585 | 0.2860   |

effective when estimation quality is adequate, but particle count cannot substitute for accurate learned scores.

**Out-of-distribution mixture.** For the revised OOD task, the two condition distributions form the cross-shaped arrangement shown in Figure 3.2:  $P^{a_1}$  has component means  $\{(-6, 0), (6, 0)\}$ , while  $P^{a_2}$  has component means  $\{(0, -6), (0, 6)\}$ . This construction places the composition in regions that are weakly constrained by the individual source distributions. Table 3.2 shows a sharp separation between analytic and learned scores. With analytic scores, performance remains broadly stable across particle counts and reaches its best reported value at  $K = 256$ . With learned scores, however,  $SW_2$  increases strongly as  $K$  grows for every training-set size. At  $N = 10,000$ , for example, the error rises from 0.3779 at  $K = 1$  to 3.0585 at  $K = 256$ . The same qualitative pattern appears at  $N = 100$  and  $N = 1000$ . Thus stronger correction consistently drives the sampler toward inaccuracies in the learned composition rather than toward the true target.

The mixture experiments provide a clearer counterpart to the Gaussian analysis. In distribution, FKC can substantially reduce approximation error when the learned scores are sufficiently accurate. In the revised OOD setting, the analytic oracle remains stable while learned-score performance deteriorates sharply with particle count. This does not yield an exact numerical decomposition of approximation and estimation error, but it provides stronger evidence that score-estimation error can dominate the behavior of an otherwise more accurate inference procedure.

### 3.3.3 Semi-synthetic data: Objects in a room

To evaluate these dynamics on more realistic, higher-dimensional visual data, we create a semi-synthetic dataset consisting of realistic room images generated via a text-to-image model. The dataset is partitioned into discrete underlying classes: empty rooms, rooms containing a single object, and rooms containing exactly two objects.

Each base and conditional distribution is a mixture of images from four underlying classes: “empty,” “couch,” a second object “ $X$ ,” and “couch +  $X$ .” We alter object  $X$  to either “framed

painting” or “coffee table” to control the factorization of the distributions. When  $X$  is a framed painting, the conditionals are approximately factorized, as couches and paintings typically occupy disjoint spatial regions in a room (floor versus wall). Conversely, when  $X$  is a coffee table, factorization breaks, as both objects compete for overlapping spatial support on the floor.

We define the ID and OOD settings by adjusting the probability weights across the four underlying classes, represented as the vector  $(P_{\text{empty}}, P_{\text{couch}}, P_X, P_{\text{couch}+X})$ . In the in-distribution setting,  $P^0$  is a uniform mixture  $(0.25, 0.25, 0.25, 0.25)$ , and the perturbation distributions are heavily biased toward the presence of their respective objects:  $P^{a_1}$  is set to  $(\alpha, 0.5 - \alpha, \alpha, 0.5 - \alpha)$  and  $P^{a_2}$  is set to  $(\alpha, \alpha, 0.5 - \alpha, 0.5 - \alpha)$ , where we use a small smoothing factor  $\alpha = 0.01$  to ensure the composition remains well-defined. Under this construction, the composed target distribution places nearly all of its probability mass on the “couch +  $X$ ” class, which is well-supported within each condition’s training data. In the OOD setting, the base distribution is defined as  $(1.0 - 3\beta, \beta, \beta, \beta)$  with  $\beta = 0.05/3$ , heavily skewing it toward the “empty” class.  $P^{a_1}$  and  $P^{a_2}$  are identically skewed toward the isolated “couch” and “ $X$ ” classes, respectively. As a result, the composed target still concentrates almost entirely on the “couch +  $X$ ” class, which is *not* well-supported within each condition’s training data.

### 3.3.4 Discussion and conclusion

The central conclusion is that the quality of inference-time composition depends on both the approximation induced by the sampler and the accuracy of the learned source scores in the region emphasized by the target. FKC can be useful when noncommutativity creates appreciable approximation error and the supplied scores are reliable. It provides benefit when naive composition is already exact, but it can also worsen final sample quality when the learned composition is inaccurate out of distribution.

First, analytic scores recover the expected distinction between factorized and non-factorized settings: FKC provides little benefit when naive composition is already exact, but can substantially reduce error when noncommutativity creates approximation error. Second, increasing the number of training samples generally improves the learned composition, yet even the largest learned models can remain qualitatively different from the analytic-score oracle in OOD settings. Third, increasing particle count strengthens the correction toward the distribution implied by the supplied scores; it does not make those scores more accurate. Consequently, larger set of particles for FKC can improve final quality when approximation error dominates and worsen it when learned-score error dominates.

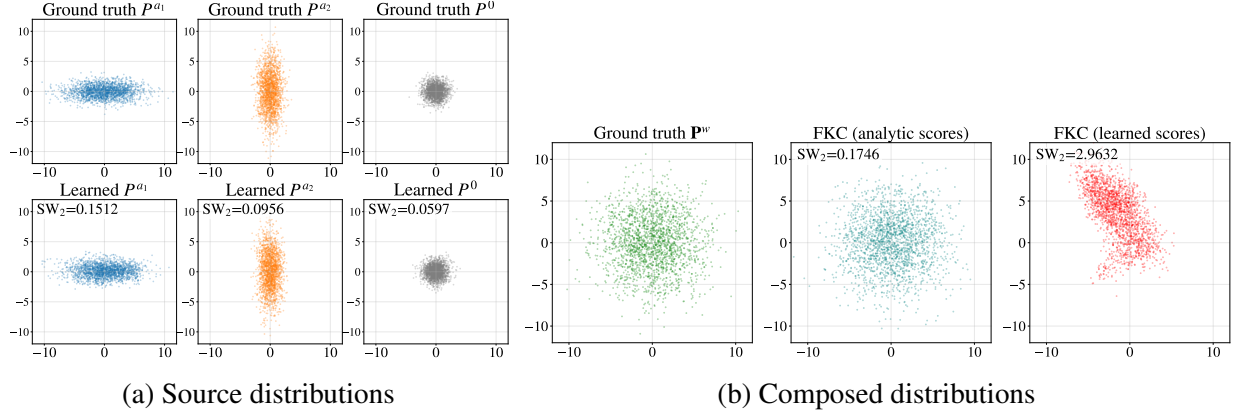


Figure 3.1: **Out-of-distribution Gaussian composition with analytic and learned scores.** The sources are  $P^0 = \mathcal{N}(0, I)$ ,  $P^{a_1} = \mathcal{N}(0, \text{diag}(10, 1))$ , and  $P^{a_2} = \mathcal{N}(0, \text{diag}(1, 10))$ , giving the base-composed target  $P^w(x) \propto P^{a_1}(x)P^{a_2}(x)/P^0(x) = \mathcal{N}(0, 10I)$ . Analytic scores recover the intended composition, whereas learned scores are inaccurate in the target-relevant low-density region; increasing FKC correction follows this inaccurate learned composition more strongly.

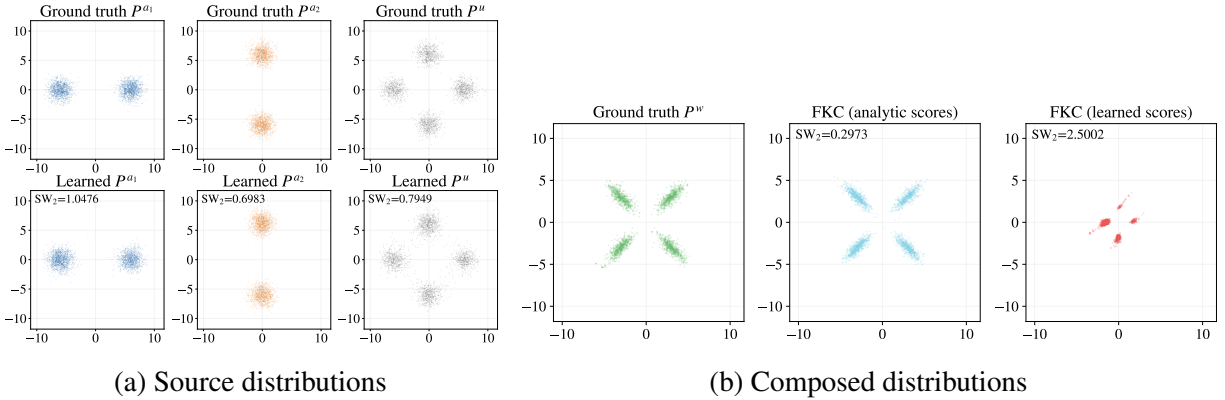


Figure 3.2: **Cross-shaped Gaussian-mixture composition in the out-of-distribution setting.** The first condition has component means  $\{(-6, 0), (6, 0)\}$  and the second has means  $\{(0, -6), (0, 6)\}$ , so the base composition emphasizes regions weakly supported by either learned source model. Analytic-score sampling remains comparatively stable as the FKC particle count increases, while learned-score sampling deteriorates because stronger correction follows errors in the estimated composition.



Figure 3.3: **Compositional room generation across factorized and OOD settings.** Naive denoising ( $K = 1$ ) is compared with FKC ( $K = 16$ ) in four settings. Panels (a) and (c) compose couches with paintings, an approximately factorized pairing because the objects typically occupy disjoint spatial regions; panels (b) and (d) compose couches with coffee tables, a non-factorized pairing because both objects compete for floor space. The top row is in-distribution and the bottom row is out-of-distribution. FKC is most effective in the in-distribution non-factorized setting, whereas the OOD samples remain incomplete or visibly distorted as learned-score error accumulates.



# Chapter 4

## Structured Multimodal Vessel Trajectory Prediction

**Reproduction and collaboration note.** This chapter is adapted from *Context-Informed Ship Trajectory Prediction via Conditional Attention*, joint work with Chandler Squires, Timothy Hu, and Pradeep Ravikumar.

The preceding chapter studied structural knowledge applied after model training. This chapter considers the complementary design choice of encoding a domain-informed directional prior directly in the model architecture.

We study *context-informed trajectory prediction*: forecasting a vessel’s future kinematic state from its past motion and, when available, contemporaneous environmental conditions and static vessel attributes. Over the forecasting horizon, wind and wave conditions influence vessel dynamics, whereas an individual vessel does not materially determine the surrounding mesoscale environment. We use this physical asymmetry as an architectural prior, without claiming a complete causal model of maritime navigation.

The central contribution is the Conditional Informer, an encoder–decoder architecture in which vessel-state representations query environmental context through directional cross-attention. Separate state, context, and metadata pathways preserve the distinction between the predicted dynamical state and auxiliary information. A simple modality-masking procedure is used only as a supporting training device for experiments in which auxiliary inputs may be absent at deployment.

### 4.1 Problem Formulation

We formalize the trajectory prediction task and the directional structural prior used by the proposed model.

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

### 4.1.1 States, Contexts, and Metadata

Let  $k$  index the vessels whose trajectories we wish to model. Each vessel  $k$  is associated with a **state trajectory**  $X^k = \{x_t^k\}_{t=1}^{T_k}$ , where  $x_t^k \in \mathcal{X} = \mathbb{R}^{d_x}$  encodes the vessel’s kinematic state at time  $t$ . In our application,  $d_x = 5$  and  $x_t^k$  comprises elapsed time since trajectory start, latitude, longitude, speed over ground (SOG), and course over ground (COG).

Each vessel may additionally be associated with a **context trajectory**  $W^k = \{w_t^k\}_{t=1}^{T_k}$ , where  $w_t^k \in \mathcal{W} = \mathbb{R}^{d_w}$  encodes environmental conditions at the vessel’s position and time  $t$ . Here  $d_w = 3$ , comprising significant wave height, wind speed, and a missingness indicator for wave height. Finally, each vessel may be associated with a static **metadata vector**  $a^k \in \mathcal{A} = \mathbb{R}^{d_a}$  encoding vessel hull characteristics and cargo type ( $d_a = 9$ , including binary missingness indicators).

### 4.1.2 Directional Structural Prior

Over the forecasting horizon, wind and wave conditions exert forces on a vessel and thereby influence its speed and heading. Conversely, the motion of an individual vessel does not materially determine the surrounding mesoscale weather field. We therefore adopt the partial ordering

$$W \longrightarrow X, \quad X \not\longrightarrow W, \quad (4.1)$$

for the variables represented by the model.

This ordering is a domain-informed simplification, not a complete causal graph of maritime navigation. Operator decisions, route constraints, currents, measurement procedures, and unobserved environmental variables may affect both the observed context and vessel motion. Accordingly, our aim is not to identify causal effects or recover a structural causal model. We use the ordering only as an architectural prior: vessel-state representations may query environmental representations, while the context representation is constructed independently of the state pathway.

This distinction is important. The ICM composition result in Chapter 2 formally identifies a compound-intervention distribution under specific assumptions. No analogous identification claim is required here. The trajectory project instead asks whether a physically motivated asymmetry is useful for organizing multimodal prediction and whether that architectural prior remains reliable when modalities are missing at deployment.

### 4.1.3 Prediction Task and Training Regimes

Fix a history length  $L$  and a prediction horizon  $P$ . At inference time, we are given the historical state trajectory  $X_{\text{past}}^{k*} = \{x_t^{k*}\}_{t=1}^L$ , and optionally the historical context trajectory  $W_{\text{past}}^{k*}$  and metadata vector  $a^{k*}$ . The true future trajectory and its prediction are

$$Y^{k*} = \{x_t^{k*}\}_{t=L+1}^{L+P}, \quad \hat{Y}^{k*} = \{\hat{x}_t^{k*}\}_{t=L+1}^{L+P}. \quad (4.2)$$

We train by minimizing mean squared error (MSE) over the full predicted state vector, and evaluate using the average Haversine distance between predicted and true geographic coordinates:

$$m(Y^{k*}, \hat{Y}^{k*}) = \frac{1}{P} \sum_{t=L+1}^{L+P} \text{hav}(x_t^{k*}, \hat{x}_t^{k*}), \quad (4.3)$$

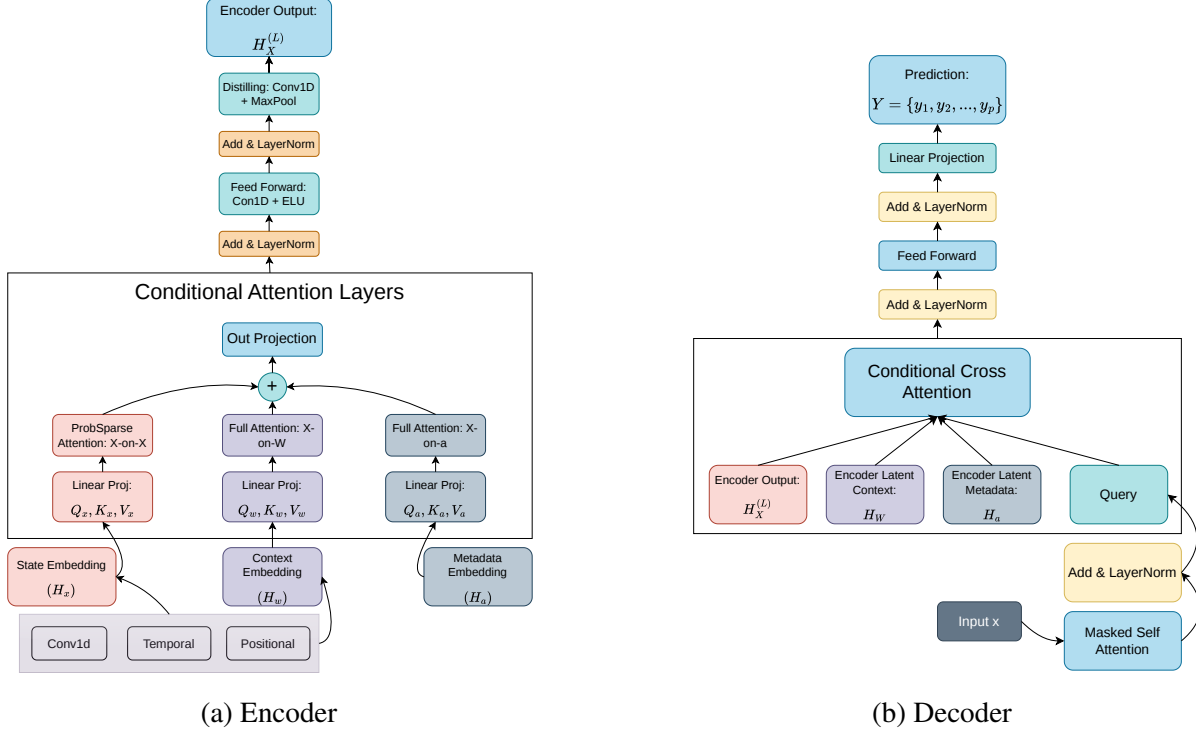


Figure 4.1: **Architecture of the Conditional Informer.** (a) Encoder with state self-attention, context cross-attention, and metadata bias. (b) Decoder with masked self-attention and conditional cross-attention over encoder representations.

where  $\text{hav}(x, x')$  denotes the great-circle distance computed from the latitude and longitude coordinates of  $x$  and  $x'$ .

A key design consideration is that the modalities available at training time may differ from those available at inference time. We refer to this as the deployment modality mismatch (DMM) problem, and consider the following training regimes:

- **State-only (S):** training data consists of state trajectories  $X^k$  only.
- **State + Context (S+C):** training data consists of pairs  $(X^k, W^k)$ .
- **State + Metadata (S+M):** training data consists of pairs  $(X^k, a^k)$ .
- **State + Context + Metadata (S+C+M):** training data consists of triples  $(X^k, W^k, a^k)$ .

Each training regime may be evaluated under any subset of modalities at inference time, yielding a matrix of performance outcomes. The DMM problem is most acute for models trained in the S+C+M regime but evaluated with context or metadata absent.

## 4.2 Design Motivation

Figure 4.1 summarizes the Conditional Informer. Its design follows the directional prior introduced in Section 4.1.2: vessel state is the primary dynamical stream, while environmental context and static metadata provide conditioning information.

Standard early fusion concatenates heterogeneous variables before sequence modeling. Although flexible, this representation does not distinguish the quantity being predicted from variables that influence it. The Conditional Informer instead preserves separate embedding pathways and uses state-derived queries with context-derived keys and values. Information therefore flows from environmental representations into the state prediction without a reciprocal operation in which the context stream is updated by vessel state. This asymmetry is an architectural inductive bias motivated by the physical relation between weather and vessel motion.

## 4.3 The Conditional Informer

The Conditional Informer is an encoder–decoder forecasting model built on Informer [32] and its vessel-trajectory adaptation, Informer-TP [30]. Informer provides a long-horizon forecasting backbone with modular attention blocks, making it convenient to introduce separate state, context, and metadata pathways. The model contains modality-specific embeddings, a conditional encoder, and a conditional decoder. We use  $d_{\text{model}}$  for the shared hidden dimension and  $d_k$  for the per-head query and key dimension.

### 4.3.1 Modality-Specific Embeddings

**State embedding.** The normalized state vector  $x_t \in \mathbb{R}^{d_x}$  is projected to the shared hidden dimension,

$$u_t = x_t W^{\text{val}}, \quad W^{\text{val}} \in \mathbb{R}^{d_x \times d_{\text{model}}}. \quad (4.4)$$

A fixed sinusoidal positional encoding  $\text{PE}_t$  and a learned calendar-time embedding  $\text{TE}_t$  are added to the value projection:

$$H_X^{(t)} = u_t + \text{PE}_t + \text{TE}_t. \quad (4.5)$$

The calendar embedding is computed from minute, hour, weekday, month, and year features. It is distinct from elapsed time within a trajectory: elapsed time represents local progression through a segment, whereas the calendar features represent absolute temporal context.

**Context embedding.** The environmental vector  $w_t \in \mathbb{R}^{d_w}$  is normalized and embedded through its own value and temporal projections, yielding  $H_W^{(t)} \in \mathbb{R}^{d_{\text{model}}}$ . The context stream uses the same timestamps as the state stream but does not share its learned projection parameters.

**Metadata embedding.** The static metadata vector  $a^k \in \mathbb{R}^{d_a}$  is normalized and projected as

$$H_a = a^k W^{\text{meta}}, \quad W^{\text{meta}} \in \mathbb{R}^{d_a \times d_{\text{model}}}. \quad (4.6)$$

Because  $H_a$  is constant over the trajectory window, it is broadcast across the temporal dimension when fused with the state representation.

### 4.3.2 Conditional Attention

Let  $H_X, H_W \in \mathbb{R}^{B \times L \times d_{\text{model}}}$  denote state and context embeddings, and let  $H_a \in \mathbb{R}^{B \times d_{\text{model}}}$  denote metadata embeddings. Conditional Attention combines three branches.

**State self-attention.** State queries, keys, and values are obtained from independent projections of  $H_X$ :

$$Q_X = H_X W_X^Q, \quad K_X = H_X W_X^K, \quad V_X = H_X W_X^V. \quad (4.7)$$

The branch uses Informer’s ProbSparse attention,

$$O_X = \text{ProbSparseAttn}(Q_X, K_X, V_X), \quad (4.8)$$

which approximates full self-attention for long sequences.

**Context cross-attention.** The state stream queries the environmental stream:

$$Q_{X \rightarrow W} = H_X W_{X \rightarrow W}^Q, \quad K_W = H_W W_W^K, \quad V_W = H_W W_W^V, \quad (4.9)$$

$$O_W = \text{softmax}\left(\frac{Q_{X \rightarrow W} K_W^\top}{\sqrt{d_k}}\right) V_W. \quad (4.10)$$

There is no reciprocal branch in which context tokens query vessel-state tokens. The attention structure therefore encodes the intended direction of conditioning without implying that attention weights recover causal effects.

**Metadata conditioning.** Since the metadata representation is a single static vector, attention over it would reduce to a projected value with unit softmax weight. We therefore write the operation directly as

$$O_a = \mathbf{1}_L \otimes (H_a W_a^V), \quad (4.11)$$

where  $\mathbf{1}_L \otimes$  denotes broadcasting over the sequence length.

**Fusion.** The three branches are added and passed through a shared output projection:

$$O_{\text{final}} = (O_X + O_W + O_a) W^O. \quad (4.12)$$

This additive form preserves a distinct state computation and permits an unavailable auxiliary branch to contribute zero without changing tensor dimensions.

### 4.3.3 Conditional Encoder and Decoder

**Encoder.** Each encoder layer applies Conditional Attention to the evolving state representation, followed by the residual, normalization, and feed-forward operations used by the Informer backbone. Informer’s distilling layers downsample the sequence between encoder blocks. The context and metadata representations remain auxiliary sources rather than being updated by the state pathway.

**Decoder.** The decoder predicts the future sequence using masked self-attention over decoder inputs and cross-attention to the encoded history. The decoder query attends to the encoded state history and, when available, the environmental and metadata representations. The causal mask prevents a decoder position from using later target positions during training.

#### 4.3.4 Training under Missing Modalities

For experiments involving unavailable auxiliary inputs, we apply standard modality dropout: environmental context is removed with probability  $p_C = 0.4$  and metadata with probability  $p_M = 0.3$ . A removed branch contributes zero in (4.12), while the state sequence is never masked.

### 4.4 Experiments

We evaluate whether environmental context improves trajectory forecasting and whether the directional Conditional Informer is more effective than state-only and early-concatenation baselines. We also report missing-modality results using the standard modality-dropout procedure described in Section 4.3.4.

#### 4.4.1 Data and Preprocessing

| $t$      | LAT ( $^{\circ}$ N) | LON ( $^{\circ}$ W) | SOG (kn) | COG ( $^{\circ}$ ) | $t$      | SWH ( $m$ ) | WS ( $ms^{-1}$ ) | $t$      | DateTime (UTC)      |
|----------|---------------------|---------------------|----------|--------------------|----------|-------------|------------------|----------|---------------------|
| 1        | 23.8                | -83.5               | 14.2     | 072                | 1        | 0.95        | 8.59             | 1        | 2023-06-30 14:31:00 |
| 2        | 24.9                | -81.6               | 14.8     | 075                | 2        | 0.99        | 7.52             | 2        | 2023-06-30 14:36:00 |
| 3        | 26.3                | -79.8               | 15.1     | 078                | 3        | 0.99        | 6.49             | 3        | 2023-06-30 14:41:00 |
| $\vdots$ | $\vdots$            | $\vdots$            | $\vdots$ | $\vdots$           | $\vdots$ | $\vdots$    | $\vdots$         | $\vdots$ | $\vdots$            |

(a) State Trajectory  $\mathbf{X}^k$ 
(b) Context Trajectory  $\mathbf{W}^k$ 
(c) Timestamps

Figure 4.2: **Example state and context trajectories for context-informed ship trajectory prediction.** In (a), LAT, LON, SOG, and COG are short for *latitude*, *longitude*, *speed over ground*, and *course over ground*, respectively. In (b), SWH and WS are short for *significant wave height* and *wind speed*, respectively.

**Vessel trajectories.** Raw Automatic Identification System (AIS) data were obtained from the Marine Cadastre vessel traffic archive [2]. We use records from June through September 2023 in the Gulf of Mexico region bounded by 23–27 $^{\circ}$ N latitude and 84–77 $^{\circ}$ W longitude. Following Informer-TP [30], we restrict the analysis to cargo vessels (IMO type 70). The preprocessing pipeline consists of DBSCAN-based outlier removal, segmentation at temporal gaps longer than 60 minutes, interpolation or filling of short gaps, and resampling to a uniform five-minute

interval. Segments with unrecoverable missing values or fewer than 48 resampled observations are removed.

**Environmental context.** Environmental variables are drawn from ERA5 reanalysis [5, 7]. We use ten-meter wind speed and significant wave height. Each AIS observation is paired with the nearest ERA5 grid point and the preceding full-hour estimate. Missing wave-height values are zero-imputed after normalization, with a corresponding missingness indicator.

**Static metadata.** The metadata vector contains vessel draft, length, width, cargo type, and binary missingness indicators. Continuous attributes are normalized using training-set statistics.

Figure 4.2 illustrates the aligned state and context sequences. The resulting 1,571 trajectory segments are split at the trajectory level into 1,256 training, 157 validation, and 158 test segments. A sliding window with history length  $L = 24$  and prediction horizon  $P = 12$  yields two hours of observed motion and one hour of prediction at five-minute resolution. The final dataset contains 98,944 training, 12,785 validation, and 12,042 test windows.

## 4.4.2 Models and Baselines

We compare state-only Informer-TP [30], an early-concatenation Informer that treats all modalities as input features, and the Conditional Informer with separate state, context, and metadata pathways. We also include a Conditional Informer trained without modality dropout to isolate the effect of exposure to missing-input cases. This comparison evaluates the proposed directional fusion within one forecasting backbone rather than across all modern sequence models.

## 4.4.3 Training and Evaluation Protocol

All models use model dimension  $d_{\text{model}} = 512$ , eight attention heads, a two-layer encoder, and a one-layer decoder. We train with Adam using learning rate  $5 \times 10^{-6}$  and betas (0.9, 0.98), for at most 100 epochs with early stopping after 20 epochs without improvement in validation Haversine distance. Each result is the mean and standard deviation over five random seeds. Models are evaluated using mean Haversine distance over the forecast horizon and MSE over the normalized predicted state vector.

## 4.4.4 Results

Table 4.1 reports the complete comparison. With context available, the Conditional Informer achieves  $5.10 \pm 0.33$  km in the state-plus-context regime, compared with  $6.03 \pm 0.22$  km for state-only Informer-TP. With all modalities, the Conditional Informer obtains  $5.10 \pm 0.36$  km, while early concatenation obtains  $6.12 \pm 0.33$  km. Metadata provides little additional improvement in this relatively homogeneous cargo-vessel dataset.

A model trained only with complete inputs degrades sharply when context is removed. Standard modality dropout substantially reduces this failure, although state-only performance remains worse than that of a model trained exclusively on state. This ablation concerns robustness to input availability and is secondary to the architectural comparison.

Table 4.1: **Trajectory forecasting results across training and test-time modality subsets.** Each cell reports mean  $\pm$  standard deviation over five random seeds for mean Haversine distance (km) and normalized-state MSE ( $\times 10^{-3}$ ).  $S$ ,  $C$ , and  $M$  denote vessel state, environmental context, and static metadata. Bold entries mark the best result within each test-time column. The no-masking row uses the same Conditional Informer architecture but is trained only with complete multimodal inputs.

| Model<br>(Training Time Data) | Haversine (km) [Test Time Data]   |                                   |                                   |                                   | MSE ( $\times 10^{-3}$ ) [Test Time Data] |                                   |                                   |                                   |
|-------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-------------------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
|                               | <i>State Only</i>                 | <i>S+C</i>                        | <i>S+M</i>                        | <i>S+C+M</i>                      | <i>State Only</i>                         | <i>S+C</i>                        | <i>S+M</i>                        | <i>S+C+M</i>                      |
| InformerTP-(S)                | 6.03 $\pm$ 0.22                   | N/A                               | N/A                               | N/A                               | 1.46 $\pm$ 0.21                           | N/A                               | N/A                               | N/A                               |
| Concat-(S+C+M)                | 9.28 $\pm$ 1.26                   | 9.20 $\pm$ 1.19                   | 6.47 $\pm$ 0.56                   | 6.12 $\pm$ 0.33                   | 1.93 $\pm$ 0.20                           | 1.81 $\pm$ 0.16                   | 1.55 $\pm$ 0.14                   | 1.50 $\pm$ 0.17                   |
| Conditional-(S)               | <b>5.97 <math>\pm</math> 0.44</b> | N/A                               | N/A                               | N/A                               | <b>1.41 <math>\pm</math> 0.17</b>         | N/A                               | N/A                               | N/A                               |
| Conditional-(S+M)             | 6.04 $\pm$ 0.52                   | N/A                               | <b>6.02 <math>\pm</math> 0.58</b> | N/A                               | 1.45 $\pm$ 0.19                           | N/A                               | <b>1.45 <math>\pm</math> 0.19</b> | N/A                               |
| Conditional-(S+C)             | 8.89 $\pm$ 1.16                   | <b>5.10 <math>\pm</math> 0.33</b> | N/A                               | N/A                               | 1.52 $\pm$ 0.22                           | <b>1.40 <math>\pm</math> 0.18</b> | N/A                               | N/A                               |
| Conditional-(S+C+M)           | 8.76 $\pm$ 1.64                   | 5.13 $\pm$ 0.31                   | 8.75 $\pm$ 1.67                   | <b>5.10 <math>\pm</math> 0.36</b> | 1.50 $\pm$ 0.22                           | 1.41 $\pm$ 0.17                   | 1.53 $\pm$ 0.22                   | <b>1.40 <math>\pm</math> 0.16</b> |
| No masking-(S+C+M)            | 67.05 $\pm$ 12.84                 | 7.04 $\pm$ 0.44                   | 66.19 $\pm$ 13.36                 | 6.44 $\pm$ 0.49                   | 8.40 $\pm$ 2.06                           | 1.50 $\pm$ 0.21                   | 8.49 $\pm$ 2.09                   | 1.51 $\pm$ 0.21                   |

#### 4.4.5 Discussion and conclusion

The proposed architecture encodes a physical asymmetry by allowing state representations to query environmental context while keeping the context pathway separate. Within the evaluated setting, incorporating environmental context improves trajectory prediction, and the proposed directional fusion outperforms both the state-only and early-concatenation baselines.

**Data availability.** The AIS data are publicly available from the Marine Cadastre vessel traffic archive [2]. ERA5 reanalysis data are available from the Copernicus Climate Data Store [5] under the applicable Copernicus data terms.

# Chapter 5

## Conclusion

This thesis examined two questions related to the use of structural knowledge in conditional generative models: under what conditions inference-time composition procedures can approximate causally identified target distributions, and how a domain-informed directional prior can be encoded as a training-time inductive bias in a multimodal predictor. The two studies use different tasks and methods, so their relationship is interpretive rather than a controlled comparison. We conclude by summarizing the findings within their experimental scope, discussing limitations, and outlining directions for future work.

### 5.1 Summary of Findings

**Limits of inference-time compositional sampling.** Chapter 3 distinguishes approximation error, caused by the failure of noising and geometric composition to commute, from estimation error in learned source scores. Exact clean-density composition does not generally imply exact score composition at intermediate diffusion times, although the approximation error vanishes under the factorization conditions considered. In the evaluated settings, learned-score error can dominate when the composed target lies in low-density regions of the source distributions. Feynman–Kac Correctors may then improve fidelity to the learned composition without improving fidelity to the true target.

**Structured multimodal trajectory prediction.** Chapter 4 studies vessel trajectory prediction under a domain-informed directional prior. The Conditional Informer represents vessel state as the primary prediction stream and introduces environmental context through directional cross-attention. On the dataset considered, it performs better than the evaluated early-concatenation baseline when context is available. Modality masking during training improves performance when auxiliary inputs are absent at inference time.

**Bridging the two projects.** The two projects address different uses of structural knowledge. In the diffusion project, causal assumptions specify a composed target distribution, while the main question is whether available score models and inference procedures can realize that target. In the trajectory project, a domain-informed directional relation motivates the architecture itself,

while the experiments evaluate that design within a specific forecasting setting. Taken together, the two projects indicate that structural knowledge is not self-executing. It may identify a target distribution or motivate an architectural constraint, but the resulting behavior still has to be supported by the learned model. The shared lesson is that structural assumptions should be evaluated not only by whether they are conceptually appropriate, but also by whether the training data, model architecture, and estimation procedure sufficiently constrain behavior in the regimes where those assumptions must be realized.

## 5.2 Limitations

Several limitations constrain the scope of the conclusions drawn in this thesis.

The trajectory prediction study uses a relatively homogeneous population of cargo vessels from a single geographic region and season. The resulting findings may therefore not transfer directly to more heterogeneous fleets, waterways, weather regimes, or forecasting horizons. The limited variation in vessel metadata also makes it difficult to determine whether the weak contribution of metadata reflects the architecture or simply the dataset. In addition, the experiments vary modality availability through a fixed masking scheme rather than a broad family of deployment shifts. Modality masking improves robustness to the missing-input patterns studied here, but it does not establish robustness to arbitrary corruption, delayed measurements, sensor bias, or changes in the statistical relationship between context and vessel motion.

The directional structure used by the Conditional Informer should also be interpreted as a domain-informed partial ordering rather than a fully specified or identified causal graph. Over the forecasting horizon, mesoscale environmental variables plausibly influence vessel dynamics without being materially changed by an individual vessel. Nevertheless, unobserved variables, measurement procedures, operator decisions, and finer-scale feedback may complicate this picture. The experiments suggest that the proposed directional organization is useful relative to the evaluated baselines; they do not establish that the assumed ordering is uniquely correct or that the learned attention weights recover causal effects.

The compositional generation analysis is restricted to conditional diffusion models trained with standard score-matching objectives and composed only at inference time. The results therefore do not characterize architectures or objectives explicitly designed for compositional extrapolation. The analytical results are obtained in Gaussian settings, while the empirical studies provide controlled evidence rather than a complete theory for high-dimensional learned scores. In particular, separating approximation and estimation error exactly is possible only in settings where reference scores or target samples are available.

Finally, the two projects are not a controlled comparison of training-time and inference-time uses of causal structure. They involve different tasks, data, models, and evaluation criteria. Their shared interpretation should therefore be read as a recurring pattern in the settings studied here, not as a general theorem about all uses of structural knowledge in generative modeling.

## 5.3 Future Work

The findings of this thesis motivate several possible directions for future work.

**Training for compositional extrapolation.** One direction from the diffusion project is to modify the model rather than only the sampler. Candidate approaches include architectures that represent intervention-specific mechanisms more explicitly, objectives that expose the model to controlled compositions during training, and post-training adaptation using a small amount of target-relevant data. Object-centric and causal generative models [10, 12, 29] may provide useful building blocks because they encourage more modular representations of the factors being composed. A central empirical question is whether such structure improves score accuracy specifically in the low-density regions traversed by an unseen compound target.

**Diagnosing estimation error in high dimensions.** Future work on compositional diffusion could develop diagnostics for identifying where a composed sampler evaluates its source scores, how strongly those regions are supported by training data, and whether improvements in sampler accuracy move samples toward the intended target or merely toward the distribution implied by the learned scores.

**Robustness to modality availability.** Future work on multi-modal trajectory forecasting could adapt the modality-masking curriculum to expected deployment conditions or to the model’s measured dependence on each auxiliary input. It could also train and evaluate under delayed, corrupted, biased, or distribution-shifted context, allowing a broader assessment of whether the state pathway remains usable under realistic sensor failures.

**Relaxing structural assumptions.** The two projects rely on different forms of structural knowledge: a domain-informed partial ordering in the trajectory setting and non-overlapping interventions under ICM in the diffusion setting. Future work could study partial or uncertain orderings, overlapping intervention targets, and violations of mechanism independence. Recent work on partial identifiability in causal representation learning [25, 26] may provide tools for characterizing what remains identifiable under weaker assumptions.

**Broader scientific applications.** Perturbational biology provides a relevant application of geometric composition, since the effect of an unseen drug combination is often predicted from single-drug experiments [15, 18]. Applying the error decomposition developed here could help assess whether failures arise primarily from the composition procedure or from inaccurate source models in the regions occupied by combination treatments. Similar questions may arise in other scientific settings where structured interventions are combinatorial but direct observations are sparse.



# Bibliography

- [1] Arwen Bradley, Preetum Nakkiran, David Berthelot, James Thornton, and Joshua M Susskind. Mechanisms of projective composition of diffusion models. *arXiv preprint arXiv:2502.04549*, 2025. 3.3
- [2] Bureau of Ocean Energy Management (BOEM) and National Oceanic and Atmospheric Administration (NOAA). MarineCadastre.gov AIS data. <https://marinecadastre.gov/ais/>, 2024. 4.4.1, 4.4.5
- [3] Simone Capobianco, Roberto Millefiori, Nicola Forti, Paolo Braca, and Peter Willett. Deep learning methods for vessel trajectory prediction using AIS data. *IEEE Transactions on Signal Processing*, 69:3194–3209, 2021. 2.5
- [4] Xiaogang Chen, Zikun Liu, et al. Vessel trajectory prediction based on AIS data: Dual-path spatial-temporal attention network with multi-attribute information. *Journal of Marine Science and Engineering*, 11(12), 2023. 2.5
- [5] Copernicus Climate Change Service, Climate Data Store. ERA5 hourly data on single levels from 1940 to present. <https://cds.climate.copernicus.eu/>, 2026. DOI: 10.24381/cds.adbb2d47 (Accessed on 2026-03-01). 4.4.1, 4.4.5
- [6] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2.5
- [7] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Rosal, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. 4.4.1
- [8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2.4
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf). 2.4
- [10] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. *arXiv preprint arXiv:2303.10834*, 2023. 5.3

- [11] Rico Jonschkowski, Sebastian Hofer, and Oliver Brock. Patterns for learning with side information. *arXiv: Learning*, 2015. URL <https://api.semanticscholar.org/CorpusID:18010010>. 2.5
- [12] Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling. *arXiv preprint arXiv:2310.11011*, 2023. 5.3
- [13] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pages 423–439. Springer, 2022. 2.4
- [14] D. Lopez-Paz, B. Schölkopf, L. Bottou, and V. Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*, November 2016. 2.5
- [15] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Erez Lieberman Aiden, Jay Shendure, José L. McFaline-Figueroa, Pierre Boyeau, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517. 5.3
- [16] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pages 15293–15327. PMLR, 2022. 2.5
- [17] Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36:31450–31465, 2023. 2.5
- [18] Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6. 5.3
- [19] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 2.3
- [20] Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alán Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. *arXiv preprint arXiv:2503.02819*, 2025. 3
- [21] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=PxtIG12RRHS>. 2.4, 2.4
- [22] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009. 2.5

- [23] Vladimir Vapnik, Rauf Izmailov, et al. Learning using privileged information: Similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015. 2.5
- [24] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2024. 2.2, 2.3
- [25] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. *Advances in Neural Information Processing Systems*, 2024. 5.3
- [26] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025. 5.3
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 2.5
- [28] Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. *Advances in Neural Information Processing Systems*, 2023. 2.1
- [29] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models. *Advances in Neural Information Processing Systems*, 36:50932–50958, 2023. 5.3
- [30] Caiquan Xiong, Hao Shi, Jiaming Li, et al. Informer-based model for long-term ship trajectory prediction. *Journal of Marine Science and Engineering*, 12(8):1269, 2024. 2.5, 4.3, 4.4.1, 4.4.2
- [31] Zhe Zhang, Nianyuan Wang, Yu Zhang, et al. An adaptive multimodal data vessel trajectory prediction model based on a satellite automatic identification system and environmental data. *Journal of Marine Science and Engineering*, 12(3):513, 2024. 2.5
- [32] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021. 2.5, 4.3