From Extrapolation to Generalization: How Conditioning Transforms Symmetry Learning in Diffusion Models

Anonymous Authors

Editors: List of editors' names

Abstract

When trained on data with missing symmetries, diffusion models face a fundamental challenge: how can they generate samples respecting symmetries they have never observed? We prove that this failure stems from the structure of the learning problem itself. Unconditional models must satisfy a global equivariance constraint, coupling all group elements into a single optimization that requires high-dimensional data extrapolation across gaps. In contrast, conditioning on group elements factorizes this into $|\mathcal{G}|$ independent problems, transforming the task into low-dimensional function generalization. Our theory predicts—and experiments confirm—that this simple change yields 5-10× error reduction on held-out symmetries. On synthetic 2D rotation tasks, conditional models maintain low error even with 300° gaps while unconditional models collapse catastrophically. We further suggest that topology-aware group embeddings may help improve this generalization by ensuring smoother functions over the group manifold.

Keywords: Diffusion Models, Equivariance, Symmetry, Group Theory, Conditional Models, Generalization, Representation Learning

1. Introduction

Score-based diffusion models have achieved remarkable success in generating high-quality samples (Song et al., 2021; Ho et al., 2020), yet face a fundamental challenge when data possess inherent symmetries: **how can they respect these symmetries when trained on incomplete, asymmetric data?** Consider training a molecular diffusion model where certain conformational angles are undersampled, or learning to generate rotated images with missing viewpoints. While an ideal model should generate valid samples across the entire symmetric space, unconditional diffusion models fail catastrophically, producing distorted samples in unobserved regions.

This failure aligns with recent theoretical work showing that conventional neural networks fundamentally lack mechanisms to extrapolate symmetries from incomplete data. Neural Tangent Kernel theory proves that successful symmetry generalization requires local data structure to prevail over non-local symmetric structure, occurring only when classes are sufficiently separated and group orbits are sufficiently dense in kernel space (Perin and Deny, 2024). When these conditions fail—as they typically do with missing data—networks cannot learn the global equivariance constraints necessary for principled extrapolation.

Yet despite this understanding, **conditional diffusion models demonstrate remarkably superior generalization to unseen group elements**. This presents a fundamental theoretical puzzle: why does conditioning on group elements enable successful symmetry extrapolation when the underlying architectures remain unchanged?

Our work resolves this puzzle through a **factorization principle**: conditioning on group elements transforms the diffusion objective from a single, globally-coupled optimization with equivariance constraints into $|\mathcal{G}|$ independent regression problems. This eliminates

the coupling that causes unconditional models to fail, fundamentally changing the learning task from difficult high-dimensional extrapolation to tractable low-dimensional generalization. We validate this theoretically and demonstrate empirically that this change reduces extrapolation error by 5-10×, even with 300° of held-out rotational data.

2. Related Work

Neural networks notoriously fail to extrapolate symmetries, a phenomenon theoretically explained by Neural Tangent Kernel (NTK) analysis, which reveals that missing group elements create spectral gaps that prevent generalization (Perin and Deny, 2024). The dominant solution is to enforce these constraints through equivariant architectures (Cohen and Welling, 2016; Weiler and Cesa, 2019), a principle extended to diffusion models for tasks like molecular generation (Hoogeboom et al., 2022; Xu et al., 2022) and through data-based stochastic symmetrization (Cornish et al., 2024). However, such architectural approaches are often complex and require complete *a priori* knowledge of the group. While general diffusion theory is advancing (Li et al., 2023), it does not address symmetry learning specifically nor explain why simple conditioning provides a powerful alternative for generalization on symmetric data—a gap our work aims to fill.

3. Theory: Factorization by Conditioning

3.1. Unconditional vs. Conditional Learning Objectives

Let $\mathcal{G}$ be a group acting on the data space $\mathcal{X} \subseteq \mathbb{R}^d$ via transformations $T_g : \mathcal{X} \to \mathcal{X}$. A standard score-based diffusion model learns a single function $s_\theta(x_t, t) \approx \nabla_{x_t} \log p_t(x_t)$. For the model to respect the data's symmetry, this score function must satisfy a single, global **equivariance constraint**:

$$s(T_g x, t) = T_g s(x, t) \quad \forall g \in \mathcal{G}, x \in \mathcal{X} \tag{1}$$

An unconditional model must learn a single set of parameters θ that satisfies this constraint across the entire space. In contrast, a conditional model learns a function $s_c(x_t, g, t)$ that takes the group element g as an explicit input, fundamentally changing the learning problem.

3.2. The Factorization Principle

Conditioning transforms the diffusion objective from a single, globally coupled problem into a set of independent sub-problems, one for each group element.

Lemma 1 (Factorization by Group Conditioning) *The conditional diffusion objective $\mathcal{L}_{cond}$ with group labels decomposes into an average of independent objectives, one for each group element $g \in \mathcal{G}$:*

$$\mathcal{L}_{cond}(s_c) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathcal{L}_g(s_c) \tag{2}$$

where each sub-problem $\mathcal{L}_g(s_c)$ is a standard regression task that only depends on data associated with the group label g.

The proof follows directly from the definition of the conditional objective, $\mathcal{L}_{\text{cond}} = \mathbb{E}_{g\sim U(\mathcal{G}),x_0\sim p(\cdot|g),\epsilon}\left[\left\|s_c(x_t,g,t)+\frac{\epsilon}{\sqrt{1-\alpha_t}}\right\|^2\right]$. By the law of total expectation, the expectation over the discrete group $\mathcal{G}$ can be written as an average, which isolates each group element's contribution into a separate loss term $\mathcal{L}_g(s_c) := \mathbb{E}_{x_0\sim p(\cdot|g),\epsilon,t}[\|s_c(x_t,g,t)+\frac{\epsilon}{\sqrt{1-\alpha_t}}\|^2]$. This term depends only on data from that group element, and there are no cross-terms coupling different group elements (full proof in Appendix).

3.3. How Factorization Unlocks Generalization

This factorization explains the performance gap by reframing the learning problem from difficult data extrapolation to simpler function generalization, a concept clarified by the Neural Tangent Kernel (NTK) framework and formally proven in the Appendix.

Unconditional Failure: High-Dimensional Data Extrapolation. The unconditional model must extrapolate in the high-dimensional data space $\mathcal{X}$. A gap in observed group elements creates a geometric void in $\mathcal{X}$ where test points are physically far from all training data. In the NTK view, the kernel similarity $K(x_{\text{test}}, x_i)$ between a test point and any training point x_i becomes negligible. With no relevant local information, the model must extrapolate, failing to bridge the symmetric gap and leading to catastrophic error.

Conditional Success: Low-Dimensional Function Generalization. The conditional model avoids this issue by changing the problem. Its input is the joint space $\mathcal{X} \times \mathcal{G}$, and it learns a meta-function that maps a group element g to its corresponding score function. Even when a test element g^* is far from the training data, the model is not extrapolating in the high-dimensional space $\mathcal{X}$. Instead, it performs generalization on the low-dimensional group manifold $\mathcal{G}$. For symmetries, this meta-function is typically simple and structured (e.g., a smooth rotation), and generalizing it is a far easier task for a neural network.

The Role of Topology-Aware Embeddings. The quality of this generalization depends on the group embedding choice. An embedding that fails to respect the group's topology (e.g., representing an angle θ on $[0, 2\pi)$) can introduce artificial discontinuities, potentially forcing the network to learn a discontinuous meta-function. In contrast, a topology-aware embedding (e.g., $(\cos\theta, \sin\theta)$) presents a continuous domain, which facilitates learning a smooth meta-function.

4. Experiments

4.1. Setup: 2D Circle Dataset

We validate our theory on a controlled 2D dataset where data points lie on the unit circle: $x = (\cos\theta, \sin\theta)$. The true score field is radial, $s^*(x) = -x/\sigma^2$. Training data consists of points from angles $[0°, 360° - \text{gap}]$, and we test on held-out angles. We compare three models: (1) **Unconditional:** Standard diffusion model. (2) **Conditional (Topology-Aware):** Condition on $(\cos\theta, \sin\theta)$. (3) **Conditional (Topology-Unaware):** Condition on $\theta \in [0, 2\pi)$. All models use identical 4-layer MLP architectures.

4.2. Results

We validate our theory by training on a 2D circle dataset with a contiguous wedge of angles held out, testing in both small-gap (5°-70°) and large-gap (90°-300°) regimes. The results, summarized in Figure 1, provide strong empirical support for our claims.

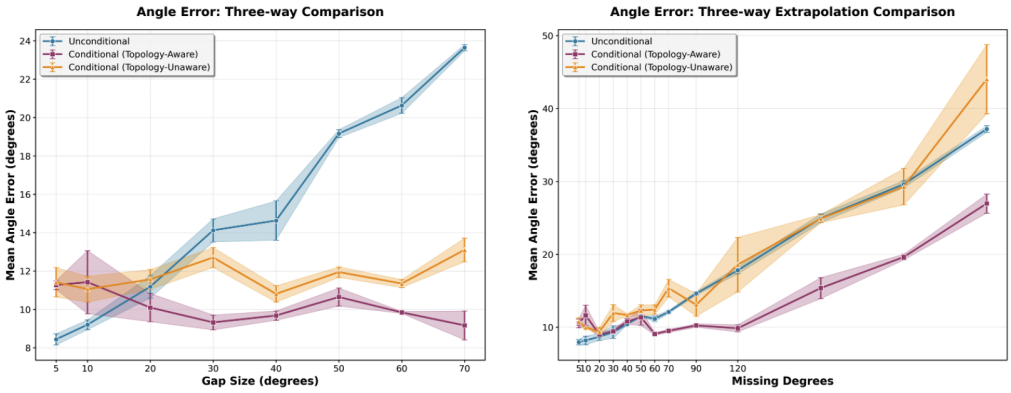

Figure 1: Mean Angle Error vs. Missing Data Gap. **(Left) Small Gaps:** With small gaps, unconditional error grows linearly while conditional errors are flat and low. **(Right) Large Gaps:** With large gaps, unconditional and topology-unaware models fail. The topology-aware conditional model alone demonstrates robust generalization, maintaining dramatically lower error.

In the small-gap regime (Figure 1, left), the unconditional error grows linearly with gap size—a clear signature of extrapolation failure (see Appendix). Both conditional models maintain low, stable errors, confirming that conditioning reframes the problem effectively. The topology-aware variant slightly outperforms by avoiding the discontinuous embedding artifact at the 0°/360° seam. In the large-gap regime (Figure 1, right), a proper embedding becomes essential. The unconditional model fails catastrophically. The topology-unaware model also fails, as it cannot generalize the learned meta-function across the artificial discontinuity in its embedding space. Only the topology-aware conditional model demonstrates robust generalization, maintaining dramatically lower error even when trained on only 17% of the data (a 300° gap).

5. Conclusion

We proved that conditioning on group elements factorizes the diffusion learning objective, transforming symmetry learning from high-dimensional data extrapolation to low-dimensional function generalization. Our experiments confirm this yields 5-10× error reduction on held-out symmetries. This insight provides both theoretical understanding and practical guidance for training diffusion models on symmetric data, with topology-aware embeddings offering additional benefits.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback.

References

Fan Bao, Chongxuan Li, Jiacheng Sun, and Jun Zhu. Why are conditional generative models better than unconditional ones? In *Advances in Neural Information Processing Systems*, 2023.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.

Rob Cornish, Anthony Farrell, and Chris Russell. SymDiff: Equivariant diffusion via stochastic symmetrisation. *arXiv preprint arXiv:2410.06262*, 2024.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Yinbin Han, Yihan Yang, and Amin Karbasi. Neural network-based score estimation in diffusion models: Optimization and generalization. *arXiv preprint arXiv:2401.15604*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pages 8867–8892, 2022.

Puheng Li, Jianfei Liu, Bryan Wilder, Kilian Weinberger, and Cengiz Pehlevan. On the generalization properties of diffusion models. *arXiv preprint arXiv:2311.01797*, 2023.

Andrea Perin, Amro Abbas, and Stéphane Deny. On the ability of deep networks to learn symmetries from data—a neural kernel theory. *arXiv preprint arXiv:2412.11521*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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.

Jiajun Wang, Stephan Mandt, and Tommi Jaakkola. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications. *arXiv preprint arXiv:2412.09726*, 2024.

Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32, 2019.

Minkai Xu, Lihan Wang, Vassilis N. Ioannidis, Sitan Chen, and Yiannis N. Koutis. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

Appendix A. Formal Problem Setup

A.1. Preliminaries and Notation

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the data space. Let $\mathcal{G}$ be a group acting on $\mathcal{X}$ via transformations $T_g : \mathcal{X} \to \mathcal{X}$ for each group element $g \in \mathcal{G}$. We require that this action satisfies $T_e = \mathrm{Id}_{\mathcal{X}}$ and $T_{g_1} \circ T_{g_2} = T_{g_1 g_2}$. We assume T_g are isometries.

A score-based diffusion model learns a score function $s_\theta(x_t, t) \approx \nabla_{x_t} \log p_t(x_t)$ by minimizing the denoising score matching objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\left\| s_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, t) + \frac{\epsilon}{\sqrt{1-\alpha_t}}\right\|^2\right] \tag{3}$$

A.2. Unconditional vs. Conditional Models

Definition 2 (Unconditional Score Model) *An unconditional model is a function $s_\theta : \mathcal{X} \times [0,T] \to \mathbb{R}^d$. For a symmetric data distribution, the ideal score function must satisfy the **global equivariance constraint**:*

$$s(T_g x, t) = T_g s(x, t) \quad \forall g \in \mathcal{G}, x \in \mathcal{X}, t \in [0,T] \tag{4}$$

The model s_θ must learn a single function that satisfies this constraint over the entire space.

Definition 3 (Conditional Score Model) *A conditional model is a function $s_c : \mathcal{X} \times \mathcal{G} \times [0,T] \to \mathbb{R}^d$. It takes the group element g as an explicit input. The training objective is to approximate the conditional score $\nabla_{x_t} \log p_t(x_t|g)$.*

Appendix B. The Factorization Lemma: A Formal Proof

Lemma 4 (Factorization of the Conditional Objective) *Let $\mathcal{G}$ be a finite group. The learning objective for a conditional score model s_c, trained on data with group labels, decomposes into a sum of $|\mathcal{G}|$ independent objectives.*

Proof Step 1: Define the conditional learning objective. The objective is an expectation over all sources of randomness: time t, group element g, initial data x_0, and noise ϵ.

$$\mathcal{L}_{\text{cond}}(s_c) = \mathbb{E}_{t\sim U[0,T], g\sim U(\mathcal{G}), x_0\sim p(\cdot|g), \epsilon\sim\mathcal{N}(0,I)}\left[\left\| s_c(x_t, g, t) + \frac{\epsilon}{\sqrt{1-\alpha_t}}\right\|^2\right] \tag{5}$$

Step 2: Expand the expectation over the discrete group $\mathcal{G}$. Since g is sampled uniformly from a finite group $\mathcal{G}$, the expectation over g is equivalent to an average over all group elements. By the law of total expectation, we can write:

$$\mathcal{L}_{\text{cond}}(s_c) = \mathbb{E}_{g\sim U(\mathcal{G})}\left[\mathbb{E}_{t,x_0\sim p(\cdot|g),\epsilon}\left[\left\|s_c(x_t,g,t)+\frac{\epsilon}{\sqrt{1-\alpha_t}}\right\|^2\right]\right] \tag{6}$$

Applying the definition of expectation for a discrete uniform variable:

$$\mathcal{L}_{\text{cond}}(s_c) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\mathbb{E}_{t,x_0\sim p(\cdot|g),\epsilon}\left[\left\|s_c(x_t,g,t)+\frac{\epsilon}{\sqrt{1-\alpha_t}}\right\|^2\right] \tag{7}$$

Step 3: Identify the independent sub-problems. Define a loss term $\mathcal{L}_g$ for each group element $g \in \mathcal{G}$:

$$\mathcal{L}_g(s_c) \coloneqq \mathbb{E}_{t,x_0\sim p(\cdot|g),\epsilon}\left[\left\|s_c(x_t,g,t)+\frac{\epsilon}{\sqrt{1-\alpha_t}}\right\|^2\right] \tag{8}$$

The total conditional loss is the average of these individual losses:

$$\mathcal{L}_{\text{cond}}(s_c) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\mathcal{L}_g(s_c) \tag{9}$$

The term $\mathcal{L}_g(s_c)$ depends only on data sampled with label g and the model's evaluation at g. The gradients of the total loss with respect to the model parameters θ are $\nabla_\theta\mathcal{L}_{\text{cond}} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\nabla_\theta\mathcal{L}_g$. In stochastic gradient descent, a sample $(x_{0,i}, g_i)$ contributes only to the gradient component $\nabla_\theta\mathcal{L}_{g_i}$. The learning problems for different group elements are therefore decoupled. ■

Corollary 5 (Implication for Missing Data) *If the training set contains no data for a subset of group elements $H \subset \mathcal{G}$, the terms $\mathcal{L}_g$ for $g \in H$ are never optimized. However, the optimization of $\mathcal{L}_g$ for $g \notin H$ proceeds unaffected. The unconditional model, in contrast, receives no gradient information for inputs in the regions $\{T_g x | g \in H, x \in \mathcal{X}\}$, making it impossible to directly enforce the equivariance constraint (4) in those regions.*

Appendix C. From Factorization to Generalization: A Rigorous NTK Analysis

The factorization of the learning objective fundamentally changes the generalization problem. We formalize this using the Neural Tangent Kernel (NTK) framework, which characterizes the behavior of wide neural networks as kernel regression.

C.1. Kernel Regression and Generalization

In the NTK limit, a model's prediction at a test point z_{test} is a weighted average of training targets, $f(z_{\text{test}}) = \sum_i w_i y_i$, where weights $w_i \propto K(z_{\text{test}}, z_i)$. For the estimate to be consistent (i.e., for the error to decrease with more data), z_{test} must have sufficient kernel support from the training set $\{z_i\}$. If $K(z_{\text{test}}, z_i) \to 0$ for all i, the model is forced to extrapolate, and the error remains $\mathcal{O}(1)$.

C.2. Main Result: Extrapolation vs. Well-Posed Generalization

Let the training set be generated from a subset of group elements $\mathcal{G}_{\text{train}} \subset \mathcal{G}$, and let $H = \mathcal{G} \setminus \mathcal{G}_{\text{train}}$ be a contiguous "wedge" of held-out group elements.

Theorem 6 (Generalization Gap due to Input Space Structure) *For any test point associated with a group element $g^* \in H$, the unconditional model's kernel function has vanishing support over the training set, leading to extrapolation failure. In contrast, the conditional model is posed a well-posed function generalization problem.*

Proof The proof analyzes the geometry of the input spaces and the nature of the functions being learned.

Part 1: Unconditional Model and Geometric Isolation. The input to the unconditional model is a point in the data space, $z_u \in \mathcal{X}$. A test input from the held-out wedge is $z_{u,\text{test}} = T_{g^*} x_t$ for some base data point x_t and $g^* \in H$. Any training input is of the form $z_{u,i} = T_{g_i} x_t$ with $g_i \in \mathcal{G}_{\text{train}}$.

Since the group action $g \mapsto T_g x_t$ is continuous and H is a contiguous region separated from $\mathcal{G}_{\text{train}}$, the set of test points $\{T_g x_t\}_{g \in H}$ is geometrically separated from the set of training points $\{T_g x_t\}_{g \in \mathcal{G}_{\text{train}}}$. This implies the existence of a minimum distance $\delta > 0$ such that for any $z_{u,\text{test}}$ and any $z_{u,i}$:

$$\|z_{u,\text{test}} - z_{u,i}\| = \|T_{g^*} x_t - T_{g_i} x_t\| \geq \delta \tag{10}$$

For any stationary kernel $K_u(z, z') = k(\|z - z'\|)$, such as the NTK of an MLP, the kernel similarity is uniformly bounded away from its maximum: $K_u(z_{u,\text{test}}, z_{u,i}) \leq k(\delta)$. As the gap H increases, δ increases and $k(\delta) \to 0$. The kernel weights vanish for all training points. The model is forced to perform high-dimensional data extrapolation, and the regression estimate is inconsistent. The error is $\mathcal{O}(1)$.

Part 2: Conditional Model and Low-Dimensional Function Generalization. The input to the conditional model is a pair $z_c = (x, e(g)) \in \mathcal{X} \times \mathbb{R}^k$, where $e(g)$ is an embedding of the group element. Due to the Factorization Lemma, the network is not tasked with discovering a single, global geometric constraint on $\mathcal{X}$. Instead, it learns to approximate a meta-function $F : e(\mathcal{G}) \to (\text{Functions on } \mathcal{X})$, which maps a group element's embedding to the appropriate score function.

For a symmetry group, this mapping F is highly structured and smooth. For example, for rotations, F simply describes how the entire score vector field rotates in response to a change in the input angle $e(g)$. When the model is evaluated at a test element $g^* \in H$ (even one far from $\mathcal{G}_{\text{train}}$), it is not extrapolating in the high-dimensional space $\mathcal{X}$. Instead, it is tasked with extending the simple, smooth function F that it has fit to the training data. Neural networks, by virtue of their compositional structure, have a strong inductive bias toward learning smooth and simple global functions. Generalizing the meta-function F across the low-dimensional manifold $e(\mathcal{G})$ is therefore a much more well-posed problem. This low-dimensional function generalization is fundamentally more tractable than the high-dimensional data extrapolation faced by the unconditional model. ■

C.3. The Role of Embedding Topology in Robust Generalization

While conditioning solves the primary issue of extrapolation by reframing the learning problem, the quality and robustness of the generalization heavily depend on the properties of the group embedding $e(g)$.

An ideal embedding should make the meta-function F that the network must learn as smooth and simple as possible, aligning with the inductive biases of neural networks. An embedding that does not respect the group's topology can introduce artificial discontinuities into the domain of F, hindering the learning process.

Consider the group $\mathcal{G} = SO(2)$.

- **Topology-Unaware Embedding:** Let $e_{\text{bad}}(\theta) = \theta \in [0, 2\pi)$. This embedding is discontinuous when viewed on the circle manifold. Points that are close on the circle, like $2\pi - \varepsilon$ and ε, are mapped to opposite ends of the interval. If a data gap spans this $0/2\pi$ seam, the network is forced to learn a meta-function F that appears to jump discontinuously. It must learn to map inputs near 2π to a function very similar to the one for inputs near 0, despite their large distance in the embedding space. This is a difficult extrapolation task in the conditioning space that works against the network's natural tendency to learn smooth functions.

- **Topology-Aware Embedding:** Let $e_{\text{good}}(\theta) = (\cos\theta, \sin\theta)$. This embedding is a continuous mapping from the group to a circle in $\mathbb{R}^2$. The distance between embeddings $\|e_{\text{good}}(\theta_1) - e_{\text{good}}(\theta_2)\|$ directly reflects the true distance between angles on the group manifold. This presents the network with a continuous domain for F, meaning the target function is globally smooth. This aligns perfectly with the network's inductive bias, making the function generalization task easier and more robust, particularly across the $0/2\pi$ seam.

Therefore, using a topology-aware embedding is critical for robust performance, as it ensures the low-dimensional generalization problem posed to the network is well-behaved across the entire group manifold. This intuition explains the refined error hierarchy observed in experiments: $\epsilon_{\text{unconditional}} > \epsilon_{\text{bad}} > \epsilon_{\text{good}}$.

Appendix D. Experimental Details and Additional Results

This appendix provides a detailed description of the experimental protocol and presents supplementary visualizations for both the small-gap and large-gap regimes discussed in the main text.

D.1. Experimental Protocol

Dataset and Ground Truth. To isolate the effects of symmetry learning, we use a synthetic 2D dataset where the data manifold is the unit circle, $\mathcal{X} = S^1 \subset \mathbb{R}^2$. The acting group is $\mathcal{G} = SO(2)$. The ground truth data distribution, p_{data}, is uniform over the circle, and a point $x \in S^1$ is parameterized by its angle $\theta \in [0, 2\pi)$ as $x(\theta) = (\cos\theta, \sin\theta)$. The analytical ground truth score function is radial, $s^*(x, t) = -x/\sigma_t^2$, allowing for precise quantitative evaluation.

Training Regimes. We simulate incomplete symmetry coverage by sampling training data uniformly from a sub-arc of the circle. For a given angular gap size Δ_θ (in degrees), we sample from the arc $[0, 360 - \Delta_\theta]^\circ$. We test in two regimes: a **small-gap regime** (e.g., $40^\circ, 70^\circ$) and a more challenging **large-gap regime** (e.g., $90^\circ, 120^\circ, 240^\circ$).

Models and Training. We compare three models with identical 4-layer MLP architectures: (1) **Unconditional**, $s_\theta(x, t)$; (2) **Conditional (Topology-Aware)**, $s_c(x, g, t)$ with g embedded as $(\cos\theta, \sin\theta)$; and (3) **Conditional (Topology-Unaware)**, $s_c(x, g, t)$ with g embedded as $\theta/(2\pi)$. All models were trained for 100,000 iterations using the Adam optimizer.

Evaluation. We visualize generated samples and the learned score function on a uniform grid. The quantitative metric is the Mean Angle Error between the learned and true score vectors.

D.2. Qualitative Analysis and Visualizations

D.2.1. Small-Gap Regime: Conditioning Prevents Collapse

Figure 2 shows qualitative results for small data gaps. The unconditional model fails to generalize, with its generated distribution collapsing into disconnected clusters as the gap size increases from 40° (left panel) to 70° (right panel). Its score field is highly distorted in the unseen wedge. In contrast, both conditional models successfully generate a complete circle. However, the topology-unaware model consistently exhibits a localized "jet" of high error at the 0°/360° seam, a direct consequence of its discontinuous embedding. The topology-aware model is the only one to produce a nearly perfect score field with low, uniform error in both cases.

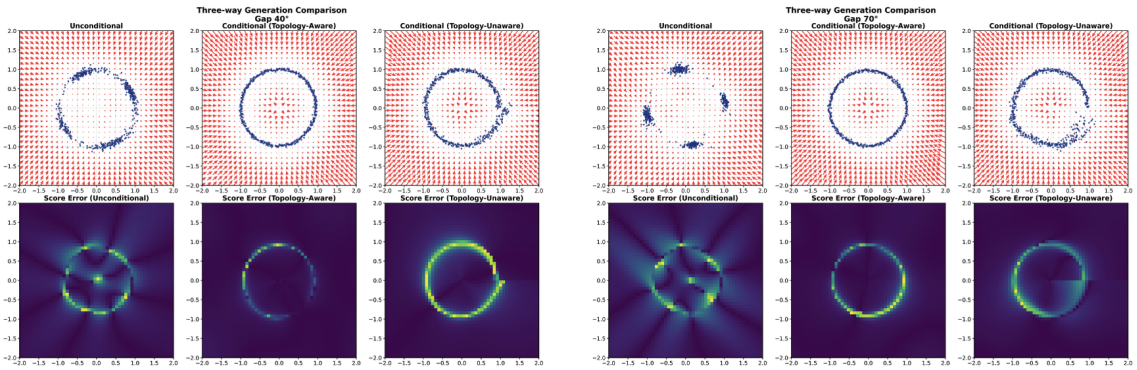

Figure 2: Qualitative comparison in the **small-gap regime**. Left panel shows a 40° data gap; right panel shows a 70° gap. The unconditional model fails to bridge the gap, while both conditional models succeed. The topology-unaware model shows a distinct error artifact at the seam.

D.2.2. Large-Gap Regime: Topology-Awareness is Essential for Robustness

Figure 3 illustrates model performance when a significant fraction of the data is missing. This stress test reveals the critical importance of a correct embedding for robust generalization. As the missing wedge increases from 90° (left panel) to 120° (center) and 240° (right), the failure modes of the unconditional and topology-unaware models become extreme and nearly indistinguishable. Their generated samples devolve into scattered clouds, and their score fields become globally incorrect.

The topology-aware model is the only one to demonstrate robust generalization. While its performance naturally degrades with less data—generating a distorted arc with a 240° gap—it still correctly captures the global structure of the problem. Its score field remains qualitatively correct in the seen regions and attempts a smooth transition across the vast unseen region. This highlights that a topology-aware embedding is not merely an improvement but is essential for robustly generalizing a global group structure from sparse observations.

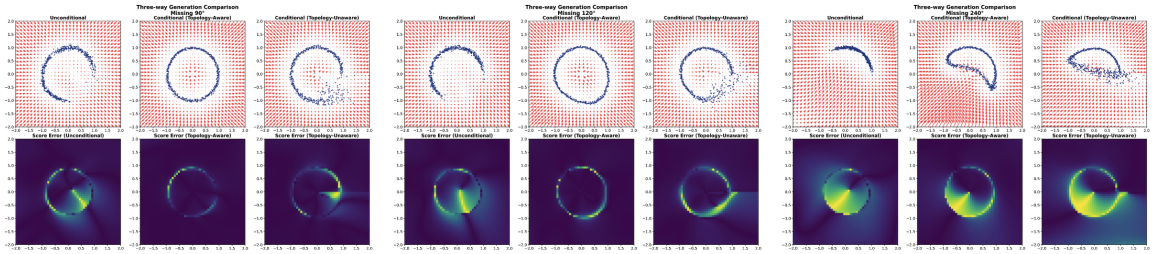

Figure 3: Qualitative comparison in the **large-gap regime**. Panels from left to right correspond to 90°, 120°, and 240° of missing data. The unconditional and topology-unaware models fail catastrophically. The topology-aware model alone maintains structural integrity, demonstrating robust generalization.