ORTHOGONALITY OF GENERALIZATION AND MEMORIZATION TRAJECTORIES IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training over-parameterized diffusion models with limited data typically exhibits a two-phase phenomenon in which the model generalizes early in training before progressively memorizing the training data. Although this qualitative behavior has been observed and studied in many recent works, a precise quantitative characterization of the geometry and dynamics underlying this transition is still lacking. In this paper, we characterize this transition through the geometry of the denoiser outputs. We show that the training trajectory can be decomposed into two directions associated with the population and empirical distributions, which are *orthogonal* in expectation. Across synthetic and real-world datasets, this decomposition reveals a quantitative phase-transition phenomenon with the denoising trajectory: the trajectory first moves predominantly toward the population Bayes-optimal denoiser, then it makes a sharp, nearly orthogonal turn toward the empirical Bayes-optimal denoiser as memorization emerges. We further study this behavior through the spectrum of the denoiser output-space training kernel, observing that the population direction is learned through substantially faster kernel modes than the empirical direction. Motivated by this separation, we propose a Direction-Aware Orthogonal (DAO) regularization, which selectively penalizes denoising motion along an estimated empirical direction. Experiments show that DAO is highly effective for mitigating memorization while further improving generation quality.

1 INTRODUCTION

Diffusion models have emerged as a powerful class of generative models, achieving strong performance across a wide range of tasks, including image and video generation (Ho et al., 2020; 2022), time-series modeling (Tashiro et al., 2021; Rasul et al., 2021), and molecular design (Hoogeboom et al., 2022; Guan et al., 2023). A central question in deep generative models is when these models learn the underlying data distribution well enough to generate high-quality novel samples and when they memorize individual training samples. Recent studies have revealed a striking progression: in over-parameterized settings with limited training data, diffusion models tend to generalize during the early stages of training before transitioning toward memorization as training continues (Favero et al., 2025; Bonnaire et al., 2025; Li et al., 2024; Wang et al., 2026).

Existing studies provide complementary perspectives on this progression. Analyses of learned denoisers link generalization beyond the training set to inductive biases toward geometry-adaptive representations and approximately Gaussian structure, which favor the capture of shared statistical regularities (Kadkhodaie et al., 2024; Li et al., 2024). Analytic models show how locality and equivariance enable novel images to emerge through locally consistent recombinations of training-image patches (Kamb & Ganguli, 2025). Moreover, studies of training dynamics identify distinct timescales for learning population structure and memorizing training samples (Favero et al., 2025; Bonnaire et al., 2025; Wang et al., 2026), while complementary spectral analyses characterize the order and rates at which different data modes are learned (Wang & Pehlevan, 2025). Related theoretical work examines how empirical score estimation promotes memorization and how regularization can mitigate it (Baptista et al., 2025; Ye et al., 2026). We refer the reader to Appendix A for a more comprehensive review of related work. Building on these insights, we seek a finer-grained characterization that quantifies how the denoiser evolves throughout training. This motivates our main question: *Can we precisely characterize the transition from generalization to memorization by disentangling population learning from empirical fitting along the training trajectory?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

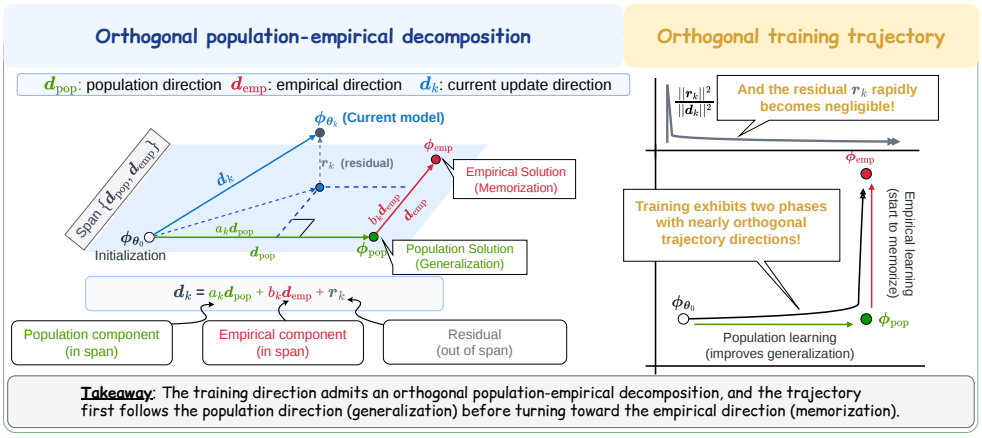


Figure 1: **Overview of two orthogonal phases of the denoising training trajectory.** We decompose the displacement of denoiser outputs from their values at initialization into components along the population and empirical directions, together with a residual orthogonal to their span. Real diffusion model training follows a trajectory that first follows the population direction before turning toward the empirical direction as memorization begins, while the residual rapidly becomes negligible. The population and empirical components are nearly orthogonal.

Summary of contributions. To address this question, we study diffusion model training through the geometry of denoiser outputs and use the resulting characterization to develop a targeted strategy for mitigating memorization. Our main contributions are as follows:

- **A geometric characterization of training dynamics.** We introduce a population-empirical decomposition of the displacement in denoiser output relative to initialization and show that the two reference directions are orthogonal in expectation (see Figure 1). Across synthetic and real-world datasets, we observe that the displacement increasingly concentrates in their span. Training initially progresses primarily along the population direction, with the best observed generalization attained near the saturation of this component, before making a sharp, nearly orthogonal turn toward the empirical Bayes-optimal denoiser (see Figure 2).
- **A spectral analysis of the two-stage trajectory.** We derive the denoiser output-space kernel dynamics and show that progress along the population and empirical directions is governed by distinct spectral rates. Empirically, the population direction exhibits a substantially larger rate throughout training. Building on this separation, we derive a bound establishing a population-first window in which empirical progress remains suppressed while the model rapidly advances along the population direction, providing insight into the observed nearly orthogonal trajectory.
- **Direction-aware orthogonal regularization for better generalization.** We show that this newly revealed geometry can be effectively exploited to improve the generalization of diffusion models in the limited-data regime. We propose **D**irection-**A**ware **O**rthogonal (DAO) regularization, which selectively penalizes denoiser output displacement along an estimated empirical direction to mitigate memorization while largely preserving population-level learning. Using an early checkpoint trained on the same limited dataset as a surrogate population reference, DAO substantially reduces memorization under prolonged training while consistently improving generation quality.

Notation. Given a vector $\boldsymbol{a}$, we use $\|\boldsymbol{a}\|$ to denote its Euclidean norm and a_i to denote its i-th entry. We denote by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma} \succeq \mathbf{0}$. Given a Gaussian random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succ \mathbf{0}$, we denote its probability density function by $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

2 An Empirical Phenomenon

In this section, we first review the preliminaries on diffusion models in Section 2.1, then introduce a population–empirical decomposition in Section 2.2, and finally present our empirical findings on two-phase training trajectories characterized by a sharp, nearly orthogonal turn in Section 2.3.

2.1 PRELIMINARIES ON DIFFUSION MODELS

We consider a training dataset $\mathcal{D}_N = \{\boldsymbol{x}^{(i)}\}_{i=1}^N \subseteq \mathbb{R}^d$, where the data points are independently and identically distributed (i.i.d.) samples from some underlying data distribution $p_{\text{data}}(\boldsymbol{x})$, i.e., $\boldsymbol{x}^{(i)} \overset{\text{i.i.d.}}{\sim} p_{\text{data}}$ for all $i = 1, \ldots, N$.

Forward and reverse processes. Diffusion models generate samples by learning to reverse a noise-injection process indexed by time $t \in [0, 1]$ (Song et al., 2021; Ho et al., 2020). Specifically, the forward process is governed by the following stochastic differential equation:

$$\mathrm{d}\boldsymbol{x}_t = f(t)\boldsymbol{x}_t \,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{w}_t, \quad \boldsymbol{x}_0 \sim p_{\text{data}},$$

where $(\boldsymbol{w}_t)_{t\geq 0}$ is a standard Wiener process. Let p_t denote the density of $\boldsymbol{x}_t$ with $p_0 = p_{\text{data}}$. The transition kernel is[1]

$$p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}\big(\boldsymbol{x}_t; s_t\boldsymbol{x}_0, s_t^2\sigma_t^2\boldsymbol{I}_n\big)\,, \;\; s_t = \exp\left(\int_0^t f(\xi)\,\mathrm{d}\xi\right),\; \sigma_t = \sqrt{\int_0^t (g^2(\xi)/s^2(\xi))\,\mathrm{d}\xi}.$$

The corresponding reverse process can be implemented by drawing $\boldsymbol{x}_1 \sim p_1$ and integrating the probability-flow ordinary or stochastic differential equation backward from $t = 1$ to $t = 0$.

Training objective. However, the score function $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)$ is generally unavailable because the underlying data distribution p_{data} is unknown. Tweedie's formula (Efron, 2011) relates the score $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)$ to the posterior mean $\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$: $s_t\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] = \boldsymbol{x}_t + s_t^2\sigma_t^2\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)$ for all $t \in (0, 1]$. This identity allows us to estimate the score by learning the posterior mean. To this end, we train a time-dependent denoising autoencoder (DAE) $\boldsymbol{\phi}_{\boldsymbol{\theta}} : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ to learn $\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$ by minimizing the empirical denoising loss on $\mathcal{D}_N$:

$$\min_{\boldsymbol{\theta}} \; \ell(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}_0\sim p_{\text{emp}}, t\sim\mathcal{U}[0,1], \boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)}\left[\left\|\boldsymbol{\phi}_{\boldsymbol{\theta}}(s_t\boldsymbol{x}_0 + \gamma_t\boldsymbol{\epsilon}, t) - \boldsymbol{x}_0\right\|^2\right], \tag{1}$$

where $\gamma_t := s_t\sigma_t$ and $p_{\text{emp}} := \sum_{i=1}^N \delta_{\boldsymbol{x}^{(i)}}/N$ is the empirical distribution of $\mathcal{D}_N$ with $\delta_{\boldsymbol{x}^{(i)}}$ being the Dirac measure concentrated at $\boldsymbol{x}^{(i)}$.

Population and empirical Bayes-optimal denoiser. The population counterpart of the empirical objective in Problem (1) is obtained by replacing p_{emp} with the data distribution p_{data}. Assuming that $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\cdot, t)$ is sufficiently expressive, the empirical and population objectives are minimized by their respective posterior means (Zhang et al., 2024; Karras et al., 2022):

$$\boldsymbol{\phi}_{\text{emp}}(\boldsymbol{x}, t) := \frac{\sum_{i=1}^N \boldsymbol{x}^{(i)} p_t(\boldsymbol{x} \mid \boldsymbol{x}^{(i)})}{\sum_{i=1}^N p_t(\boldsymbol{x} \mid \boldsymbol{x}^{(i)})}, \quad \boldsymbol{\phi}_{\text{pop}}(\boldsymbol{x}, t) := \frac{\int \boldsymbol{x}_0 p_{\text{data}}(\boldsymbol{x}_0) p_t(\boldsymbol{x} \mid \boldsymbol{x}_0)\,\mathrm{d}\boldsymbol{x}_0}{\int p_{\text{data}}(\boldsymbol{x}_0) p_t(\boldsymbol{x} \mid \boldsymbol{x}_0)\,\mathrm{d}\boldsymbol{x}_0}. \tag{2}$$

Notably, $\boldsymbol{\phi}_{\text{emp}}$ and $\boldsymbol{\phi}_{\text{pop}}$ are the *empirical* and *population* Bayes-optimal denoisers under squared-error loss for the distributions p_{emp} and p_{data}, respectively.

2.2 POPULATION–EMPIRICAL DECOMPOSITION

Population and empirical directions. Note that $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\cdot, t)$ denotes the DAE at diffusion time t. To distinguish the training iteration from the diffusion time t, we use k to denote the training iterations and $\boldsymbol{\theta}_k$ to denote the network parameters at iteration k. For each noisy input $(\boldsymbol{x}_t, t)$, we define two direction vectors to characterize the training. The first, called *the population direction*, points from the output of the DAE at initialization to that of the population Bayes-optimal denoiser:

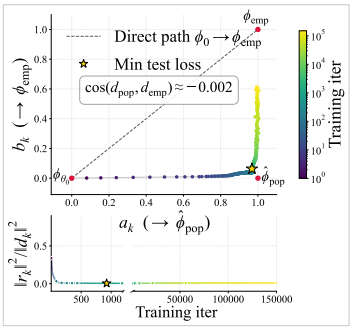
$$\boldsymbol{d}_{\text{pop}}(\boldsymbol{x}_t, t) := \boldsymbol{\phi}_{\text{pop}}(\boldsymbol{x}_t, t) - \boldsymbol{\phi}_{\boldsymbol{\theta}_0}(\boldsymbol{x}_t, t). \tag{3}$$

The second, called *the empirical direction*, points from the output of the population Bayes-optimal denoiser to that of the empirical Bayes-optimal denoiser:

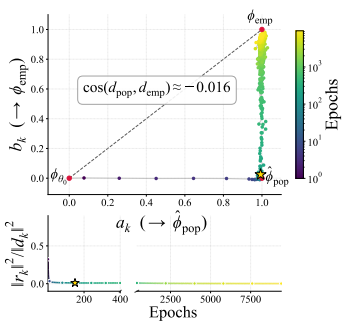
$$\boldsymbol{d}_{\text{emp}}(\boldsymbol{x}_t, t) := \boldsymbol{\phi}_{\text{emp}}(\boldsymbol{x}_t, t) - \boldsymbol{\phi}_{\text{pop}}(\boldsymbol{x}_t, t). \tag{4}$$

Under the randomness of the training dataset and the noise corruption, we show that these two directions are orthogonal to each other in expectation, whose proof is deferred to Appendix B.1.

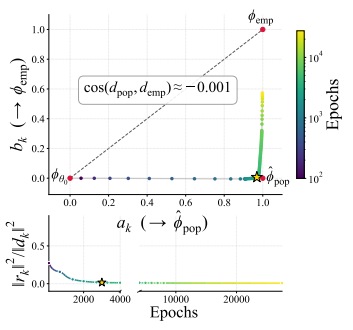
[1] With a slight abuse of notation, we denote $s_t := s(t)$ and $\sigma_t := \sigma(t)$.



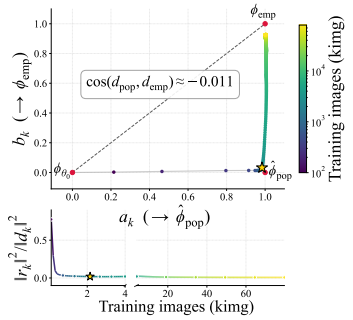
(a) MLP on MoG ($N = 256$)



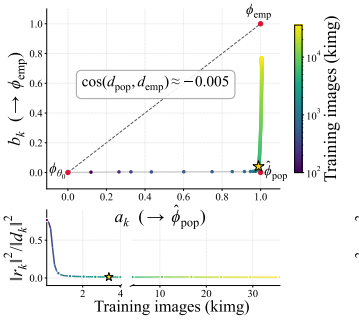
(b) DDPM on CelebA ($N = 4096$)



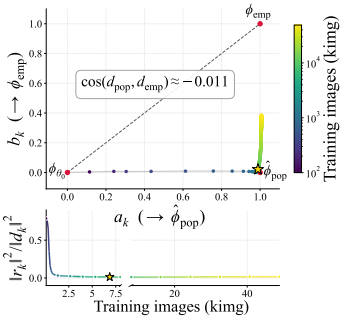
(c) JiT on ImageNet ($N = 5000$)



(d) EDM on CIFAR-10 ($N = 2048$)



(e) EDM on FFHQ ($N = 2048$)



(f) EDM on AFHQ ($N = 2048$)

Figure 2: **Prevalence of the orthogonal training trajectory. (a)** An experiment with a synthetic dataset generated from a mixture of Gaussians illustrates the characteristic population-to-empirical geometry. **(b)–(f)** Across diverse real-image datasets, training-set sizes, and model architectures, diffusion training exhibits the same qualitative behavior: the trajectory first makes rapid progress along the population direction while the relative residual energy $\|\boldsymbol{r}_k\|^2/\|\boldsymbol{d}_k\|^2$ quickly diminishes, and then makes a nearly orthogonal turn toward the empirical direction.

Lemma 1. *Suppose that $\boldsymbol{\theta}_0$ is independent of $\mathcal{D}_N$, $\boldsymbol{x}_0$, and $\boldsymbol{\epsilon}$, and that all the random vectors involved are square-integrable. Then, almost surely with respect to the random initialization $\boldsymbol{\theta}_0$,*

$$\mathbb{E}[\langle \boldsymbol{d}_{\text{pop}}(\boldsymbol{x}_t, t), \boldsymbol{d}_{\text{emp}}(\boldsymbol{x}_t, t)\rangle \,|\, \boldsymbol{\theta}_0] = 0. \tag{5}$$

Our experiments with diffusion models trained on synthetic and real-world datasets provide empirical support for this orthogonality, as shown in Figure 2. Across these experiments, the cosine similarity between $\boldsymbol{d}_{\text{pop}}$ and $\boldsymbol{d}_{\text{emp}}$ is always close to zero.

Decomposition of the training trajectory. Moreover, for each $k \geq 0$ and time $t > 0$, we define the cumulative training displacement in the denoiser output space, from the DAE at initialization to the DAE at iteration k, as

$$\boldsymbol{d}_k(\boldsymbol{x}_t, t) := \boldsymbol{\phi}_{\boldsymbol{\theta}_k}(\boldsymbol{x}_t, t) - \boldsymbol{\phi}_{\boldsymbol{\theta}_0}(\boldsymbol{x}_t, t). \tag{6}$$

For a fixed diffusion time t and suppressing the dependence on $(\boldsymbol{x}_t, t)$ for notational simplicity, we decompose $\boldsymbol{d}_k$ at iteration k into components along the population and empirical directions, together with a residual:

$$\boldsymbol{d}_k = a_k \boldsymbol{d}_{\text{pop}} + b_k \boldsymbol{d}_{\text{emp}} + \boldsymbol{r}_k, \tag{7}$$

where a_k and b_k are respectively scalar coefficients associated with the population and empirical directions and $\boldsymbol{r}_k \perp \text{span}(\boldsymbol{d}_{\text{pop}}, \boldsymbol{d}_{\text{emp}})$. Notably, the coefficient pair (a_k, b_k) thus provides a two-dimensional representation of the training trajectory projected onto the span of these directions in output space (see Figure 1).

2.3 The Sharp Orthogonal Training Trajectory

Based on the setup in Sections 2.1 and 2.2, we now conduct numerical experiments to visualize the training trajectory of diffusion models for solving Problem (1). According to (7), the initial DAE

output and the population and empirical Bayes-optimal outputs correspond to $(a, b) = (0, 0)$, $(1, 0)$, and $(1, 1)$, respectively, with zero residual in each case. In our experiments, we consider an over-parameterized regime in which the number of network parameters far exceeds the number of training samples and observe an initial generalization phase followed by memorization (Favero et al., 2025; Li et al., 2024; Bonnaire et al., 2025; Wang et al., 2026). Throughout these experiments, we plot the trajectories at a fixed intermediate noise level, with results for other noise levels provided in Appendix C (see Figure 7).

Experiments on synthetic datasets. We study a synthetic setting in which the data distribution is an equally weighted mixture of K isotropic Gaussians (MoG), $p_{\mathrm{data}}(\boldsymbol{x}) = \frac{1}{K}\sum_{r=1}^{K}\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_r, \boldsymbol{I}_d)$. In our experiments, we set $K = 4$, $d = 16$ and draw $N = 256$ training samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ from this distribution. This setup allows us to compute both the empirical Bayes-optimal denoiser and population Bayes-optimal denoiser in (2), respectively. We then train a standard multilayer perceptron (MLP) with SiLU activations using full-batch gradient descent (GD) over the N training samples to minimize the objective in Problem (1). The resulting training trajectory is shown in Figure 2(a).

Experiments on real-world datasets. We conduct experiments on CelebA (Liu et al., 2015), CIFAR-10 (Krizhevsky et al., 2009), FFHQ (Karras et al., 2021), AFHQv2 (Choi et al., 2020), and ImageNet (Deng et al., 2009). Across these datasets, we consider DDPM (Ho et al., 2020), EDM (Karras et al., 2022), and JiT (Li & He, 2026) configurations, training each model on a limited subset to optimize the objective in Problem (1). Further experimental details are provided in Appendix D.2. We compute the empirical Bayes-optimal denoiser using (2). Since the underlying data distribution p_{data} is unknown, the true population Bayes-optimal denoiser $\boldsymbol{\phi}_{\mathrm{pop}}$ is unavailable. We therefore approximate it with a model trained on the full available dataset when analyzing the training geometry of these limited-data runs. The resulting trajectories are shown in Figure 2(b)–(f). In all panels, a_k and b_k are plotted on the horizontal and vertical axes, respectively, color indicates training progress, and the relative residual energy $\|\boldsymbol{r}_k\|^2/\|\boldsymbol{d}_k\|^2$ is shown separately. Across these experiments, we observe that the transition from generalization to memorization consistently manifests as a sharp, nearly right-angle turn in the training trajectory:

- **Early generalization phase.** Training initially makes little progress along the empirical direction, with b_k remaining close to zero. Meanwhile, a_k rises rapidly and the relative residual energy $\|\boldsymbol{r}_k\|^2/\|\boldsymbol{d}_k\|^2$ decreases, indicating that the training displacement increasingly aligns with the population direction. The model attains its best observed generalization performance as the population component approaches saturation and the relative residual energy reaches its minimum.

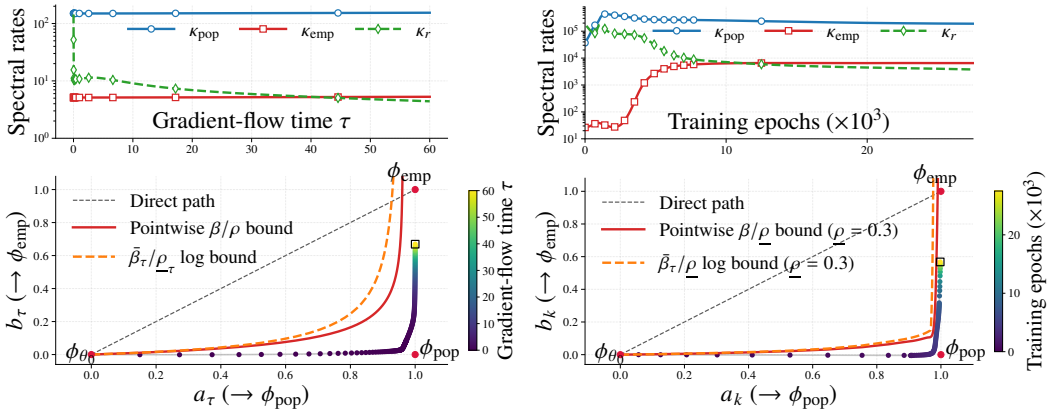
- **Late memorization phase.** Once the population component saturates, b_k begins to increase, while a_k remains near its plateau and the relative residual energy stays negligible. The trajectory thus advances predominantly along the empirical direction, producing a sharp, nearly right-angle turn in the population–empirical plane. This change in direction marks the transition from population learning to empirical fitting.

Together, these observations reveal a characteristic geometric transition from early generalization to late memorization: *diffusion models first concentrate their learning along the population direction and, as this component saturates, takes a sharp, nearly orthogonal turn toward the empirical Bayes-optimal denoiser*. We refer to this geometric pattern as the *nearly orthogonal training trajectory*.

3 Theoretical Justification

In this section, we take a first step to gain insights into the observed two-stage training trajectory by analyzing the denoiser's time-dependent output-space kernel. Throughout this section, we condition on a fixed training set $\mathcal{D}_N = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ and omit the explicit dependence on $\mathcal{D}_N$ from the notation. Let $I \sim \mathrm{Unif}(\{1, \dots, N\})$, $t \sim \mathrm{Unif}([0, 1])$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ be mutually independent. Let $\mathcal{H}$ denote the space of square-integrable $\mathbb{R}^d$-valued functions under the conditional distribution of $(\boldsymbol{x}_t, t)$. For $\boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathcal{H}$, define

$$\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle_{\mathcal{H}} := \mathbb{E}[\langle \boldsymbol{h}_1(\boldsymbol{x}_t, t), \boldsymbol{h}_2(\boldsymbol{x}_t, t) \rangle], \quad \|\boldsymbol{h}\|_{\mathcal{H}}^2 := \langle \boldsymbol{h}, \boldsymbol{h} \rangle_{\mathcal{H}}.$$



(a) Synthetic Setting (b) JiT on ImageNet ($N = 5000$)

Figure 3: **Spectral-rate comparison and evaluation of Theorem 1.** Each subfigure shows the directional Rayleigh quotients (top) and the training trajectory with the calculated bounds (Theorem 1) on b (bottom). In both settings, κ_{pop} substantially exceeds κ_{emp} throughout training and exceeds κ_r shortly after training begins. **(a)** The synthetic experiment evaluates these bounds under gradient flow of the expected empirical objective. **(b)** The practical JiT run uses the same expressions with the assumed uniform lower bound $\rho(\tau) \geq 0.3$ over the plotted interval.

3.1 Time-Dependent Kernel Dynamics in the Output Space

The empirical objective in (1) uses clean training samples $\boldsymbol{x}_0$ as regression targets. The following lemma shows that replacing these targets with the empirical Bayes denoiser leaves the expected parameter gradient unchanged. Thus, ϕ_{emp} is the effective regression target of the objective. For $\boldsymbol{x}_t = s_t \boldsymbol{x}_0 + \gamma_t \boldsymbol{\epsilon}$, define the output Jacobian $\boldsymbol{J}_{\boldsymbol{\theta},t} := \partial \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)/\partial \boldsymbol{\theta}$.

Lemma 2. *Suppose that $\phi_{\boldsymbol{\theta}}$ is differentiable in $\boldsymbol{\theta}$ and that differentiation can be interchanged with expectation. Then, we have*

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = 2\,\mathbb{E}\big[\boldsymbol{J}_{\boldsymbol{\theta},t}^{\top}\big(\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \phi_{\text{emp}}(\boldsymbol{x}_t, t)\big)\big]\,. \tag{8}$$

The proof is deferred to Section B.2. This lemma allows us to analyze the mean training dynamics using ϕ_{emp} as the regression target, since the clean-target residual contributes zero to the expected parameter gradient.

Gradient flow of the denoiser. To express the training dynamics in the denoiser's output space, we assume that $\boldsymbol{\theta} \mapsto \phi_{\boldsymbol{\theta}}$ is differentiable as a map from $\mathbb{R}^p$ to $\mathcal{H}$. Denote its derivative by $\mathcal{J}_{\boldsymbol{\theta}} : \mathbb{R}^p \to \mathcal{H}$, so that $(\mathcal{J}_{\boldsymbol{\theta}} \boldsymbol{v})(\boldsymbol{x}_t, t) = \boldsymbol{J}_{\boldsymbol{\theta},t} \boldsymbol{v}$. Its adjoint with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is $\mathcal{J}_{\boldsymbol{\theta}}^* \boldsymbol{h} = \mathbb{E}[\boldsymbol{J}_{\boldsymbol{\theta},t}^{\top} \boldsymbol{h}(\boldsymbol{x}_t, t)]$. Let τ denote gradient flow training time. By (8), the empirical objective has the gradient-flow dynamics

$$\frac{\mathrm{d}\boldsymbol{\theta}(\tau)}{\mathrm{d}\tau} = -\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}(\tau)) = -2\,\mathcal{J}_{\boldsymbol{\theta}(\tau)}^*\big(\phi_{\boldsymbol{\theta}(\tau)} - \phi_{\text{emp}}\big).$$

Applying the chain rule yields the output-space dynamics

$$\dot{\phi}_{\boldsymbol{\theta}(\tau)} = \frac{\mathrm{d}\phi_{\boldsymbol{\theta}(\tau)}}{\mathrm{d}\tau} = \boldsymbol{K}_\tau\big(\phi_{\text{emp}} - \phi_{\boldsymbol{\theta}(\tau)}\big), \quad \boldsymbol{K}_\tau := 2\,\mathcal{J}_{\boldsymbol{\theta}(\tau)} \mathcal{J}_{\boldsymbol{\theta}(\tau)}^*. \tag{9}$$

The operator $\boldsymbol{K}_\tau : \mathcal{H} \to \mathcal{H}$ is self-adjoint and positive semidefinite with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Importantly, equation 9 holds exactly along the parameter gradient flow. Unlike the frozen-kernel approximation commonly used in neural tangent kernel analyses (Jacot et al., 2018; Lee et al., 2019), we evaluate $\boldsymbol{K}_\tau$ at the current parameters throughout training, without linearizing the network.

3.2 Main Assumptions and Results

The output-space dynamics governed by the kernel $\boldsymbol{K}_\tau$ allow us to relate population and empirical learning to the kernel's response along their respective directions. To examine this connection, we express the population–empirical decomposition in (7) in continuous training time:

$$\phi_{\boldsymbol{\theta}(\tau)} - \phi_{\boldsymbol{\theta}_0} = a_\tau \boldsymbol{d}_{\text{pop}} + b_\tau \boldsymbol{d}_{\text{emp}} + \boldsymbol{r}_\tau, \quad \text{where } \boldsymbol{r}_\tau \perp_{\mathcal{H}} \operatorname{span}\{\boldsymbol{d}_{\text{pop}}, \boldsymbol{d}_{\text{emp}}\}. \tag{10}$$

The reference models ϕ_{θ_0}, ϕ_{pop}, and ϕ_{emp} correspond to $(a, b) = (0, 0), (1, 0), (1, 1)$, respectively. We define the directional spectral rates through the Rayleigh quotients

$$\kappa_{\text{pop}}(\tau) := \frac{\langle \boldsymbol{d}_{\text{pop}}, \boldsymbol{K}_\tau \boldsymbol{d}_{\text{pop}} \rangle_{\mathcal{H}}}{\|\boldsymbol{d}_{\text{pop}}\|_{\mathcal{H}}^2}, \quad \kappa_{\text{emp}}(\tau) := \frac{\langle \boldsymbol{d}_{\text{emp}}, \boldsymbol{K}_\tau \boldsymbol{d}_{\text{emp}} \rangle_{\mathcal{H}}}{\|\boldsymbol{d}_{\text{emp}}\|_{\mathcal{H}}^2}, \quad \kappa_r(\tau) := \frac{\langle \boldsymbol{r}_\tau, \boldsymbol{K}_\tau \boldsymbol{r}_\tau \rangle_{\mathcal{H}}}{\|\boldsymbol{r}_\tau\|_{\mathcal{H}}^2}.$$

We set $\kappa_r(\tau) = 0$ when $\boldsymbol{r}_\tau = \boldsymbol{0}$. These quantities measure the kernel response along the corresponding directions during training. The trajectories of these quantities can already provide some insight into the observed two-phase trajectory: as shown in the top row of Figure 3, κ_{pop} remains substantially larger than κ_{emp}, which indicates that the dynamics are stronger along the population direction than along the empirical direction. In the early stage, κ_r is also large, consistent with the rapid decay of the residual component. This directional characterization can be related to spectral-bias analyses of diffusion model training (Wang & Pehlevan, 2025), although it takes a different perspective.

However, to rigorously establish the connection between the directional spectral rates and the observed two-stage trajectory, we also need to account for coupling between the population and empirical directions, as well as residual leakage. To quantify these effects, we introduce auxiliary quantities that will allow us to bound empirical progress relative to population progress. Let $a_0 = b_0 = 0$ and define

$$\chi(\tau) := \langle \boldsymbol{d}_{\text{pop}}, \boldsymbol{K}_\tau \boldsymbol{d}_{\text{emp}} \rangle_{\mathcal{H}}, \quad \varepsilon_{\text{emp}}(\tau) := \frac{\langle \boldsymbol{d}_{\text{emp}}, \boldsymbol{K}_\tau \boldsymbol{r}_\tau \rangle_{\mathcal{H}}}{\|\boldsymbol{d}_{\text{emp}}\|_{\mathcal{H}}^2},$$

$$\rho(\tau) := \frac{\dot{a}_\tau}{(1 - a_\tau)\kappa_{\text{pop}}(\tau)}, \quad \beta(\tau) := \frac{\kappa_{\text{emp}}(\tau) + |\chi(\tau)|/\|\boldsymbol{d}_{\text{emp}}\|_{\mathcal{H}}^2 + |\varepsilon_{\text{emp}}(\tau)|}{\kappa_{\text{pop}}(\tau)}.$$

For $\tau \geq 0$, we write $\bar{\beta}_\tau := \sup_{0 \leq s \leq \tau} \beta(s)$ and $\underline{\rho}_\tau := \inf_{0 \leq s \leq \tau} \rho(s)$. The following assumption specifies the setting and the conditions on population and empirical progress used in our analysis.

Assumption 1. *Consider parameter gradient flow for the expected empirical objective, with fixed, nonzero reference directions satisfying $\langle \boldsymbol{d}_{\text{pop}}, \boldsymbol{d}_{\text{emp}} \rangle_{\mathcal{H}} = 0$. Suppose that there exist $0 < T_1 < T_2$ such that the following conditions hold.*

(i) ***Population stage.*** *On $[0, T_1]$, $0 \leq a_\tau < 1$, $0 \leq b_\tau \leq 1$, $\kappa_{\text{pop}}(\tau) > 0$, $\bar{\beta}_{T_1} < \infty$, and $\underline{\rho}_{T_1} > 0$.*

(ii) ***Empirical stage.*** *On $[T_1, T_2]$, for some $\epsilon \in (0, 1)$,*

$$|1 - a_\tau| \leq \epsilon, \quad (1 - b_\tau)\kappa_{\text{emp}}(\tau) + (1 - a_\tau)\frac{\chi(\tau)}{\|\boldsymbol{d}_{\text{emp}}\|_{\mathcal{H}}^2} > |\varepsilon_{\text{emp}}(\tau)|. \tag{11}$$

We defer the empirical verification of Assumption 1 to Appendix C (see Figure 5). Under the above assumption, we can now state the main result.

Theorem 1. *Suppose that Assumption 1 holds. Then the following statements hold:*

(i) ***Population stage.*** *The population coordinate a_τ is strictly increasing on $[0, T_1]$, and for every $\tau \in [0, T_1]$,*

$$b_\tau \leq \int_0^{a_\tau} \frac{\beta(s(\alpha))}{\rho(s(\alpha))(1 - \alpha)} \, \mathrm{d}\alpha \leq \frac{\bar{\beta}_\tau}{\underline{\rho}_\tau} \log \frac{1}{1 - a_\tau}, \tag{12}$$

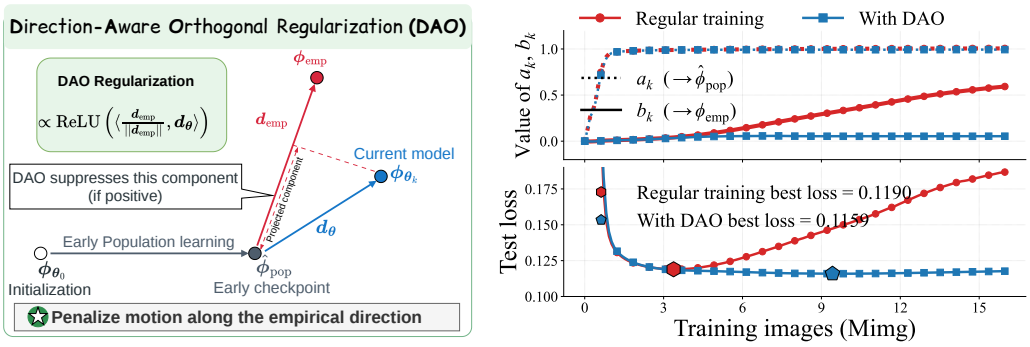
where $s(\alpha)$ is the unique time in $[0, \tau]$ such that $a_{s(\alpha)} = \alpha$.

(ii) ***Empirical stage.*** *For every $\tau \in [T_1, T_2]$,*

$$\dot{b}_\tau > 0, \quad |1 - a_\tau| \leq \epsilon. \tag{13}$$

We defer the proof to Appendix B.3. The theorem characterizes the two stages of training:

- **Why population learning comes first.** The ratio $\beta(\tau)/[\rho(\tau)(1-a_\tau)]$ controls empirical progress per unit of population progress. Integrating this ratio along the trajectory thus bounds the accumulated empirical displacement b_τ. As verified in Figure 3, these bounds allow a_τ to grow substantially while b_τ remains small.
- **Why the trajectory turns near the population corner.** Once a_τ stays within ϵ of one, little population displacement remains. In this regime, if the empirical driving term, together with the coupling term, dominates the opposing residual leakage (Assumption 1(ii)), then $\dot{b}_\tau > 0$. In other words, once the population coordinate has nearly saturated, the empirical coordinate b_τ continues to increase while a_τ remains close to one.



(a) Illustration of DAO

(b) DAO on FFHQ ($N = 2048$)

Figure 4: **Illustration and empirical effect of DAO. (a)** DAO selectively suppresses motion toward the empirical solution while largely preserving orthogonal, population-level learning. **(b)** Representative training dynamics on FFHQ ($N = 2048$). DAO slows the growth of b_k, substantially mitigates overfitting under prolonged training, and improves the best test performance, while leaving the progression of a_k nearly unchanged.

4 Mitigating Memorization through Orthogonal Regularization

The orthogonal training trajectory suggests a direct practical implication: memorization may be controlled by suppressing the late motion toward the empirical solution. In this section, we develop this idea for limited-data training and investigate two questions: (i) Can we mitigate overfitting and memorization under prolonged training? (ii) Beyond merely delaying overfitting, can we improve the best achievable generation quality over standard training?

4.1 DAO: Direction-Aware Orthogonal Regularization

A key advantage of the population-empirical training trajectory decomposition is that the two directions are orthogonal in expectation. This creates the possibility of suppressing empirical fitting without substantially interfering with population-level learning. In contrast to regularizers that act more globally on the learning process (Baptista et al., 2025), we therefore target only the component of the model's motion aligned with the empirical direction. We call this approach **DAO**: **D**irection-**A**ware **O**rthogonal Regularization.

Direction-aware regularization. Suppose an approximate population reference $\widehat{\phi}_{\text{pop}}$ is available. For each sample i in a mini-batch, define

$$\boldsymbol{d}_{\text{emp}}^{(i)} := \phi_{\text{emp}}\big(\boldsymbol{x}_t^{(i)}, \sigma_i\big) - \widehat{\phi}_{\text{pop}}\big(\boldsymbol{x}_t^{(i)}, \sigma_i\big), \quad \boldsymbol{d}_{\boldsymbol{\theta}}^{(i)} := \phi_{\boldsymbol{\theta}}\big(\boldsymbol{x}_t^{(i)}, \sigma_i\big) - \widehat{\phi}_{\text{pop}}\big(\boldsymbol{x}_t^{(i)}, \sigma_i\big),$$

which respectively represent the estimated empirical direction and the current model displacement from the population reference. We measure the model's progress along the empirical direction by

$$s_i := \left\langle \frac{\boldsymbol{d}_{\text{emp}}^{(i)}}{\|\boldsymbol{d}_{\text{emp}}^{(i)}\|_2}, \boldsymbol{d}_{\boldsymbol{\theta}}^{(i)} \right\rangle, \quad \eta_i := \frac{s_i}{\|\boldsymbol{d}_{\text{emp}}^{(i)}\|_2}.$$

Here, s_i is the signed displacement along the empirical direction, while η_i measures the fraction of the population-to-empirical distance traversed along that direction. We add the following direction-aware regularizer to the training objective in Problem (1):

$$\mathcal{L}_{\text{DAO}}(\boldsymbol{\theta}) = \sum_i \lambda_i \,\text{ReLU}(s_i) \left(1 + \frac{\text{ReLU}(\eta_i)}{\eta_0}\right), \tag{14}$$

where $\lambda_i \geq 0$ controls the regularization strength and $\eta_0 > 0$ sets the scale at which the penalty is amplified. The $\text{ReLU}(s_i)$ term makes the regularization one-sided, penalizing only motion toward the empirical solution, while the factor $1 + \text{ReLU}(\eta_i)/\eta_0$ increases the penalty as the normalized empirical progress η_i grows. We use $\eta_0 = 0.1$ throughout. Because the empirical direction is nearly orthogonal to the population direction, DAO selectively suppresses empirical fitting while largely preserving population-level learning.

Approximating the population direction. As discussed above, the population denoiser ϕ_{pop} is unavailable for real-world datasets. The larger-data surrogate used in our earlier analysis is also impractical here, since it requires additional data beyond the limited-data regime of interest. Fortunately, consistent with the early-learning implicit bias identified in prior work (Li et al., 2024; Bonnaire et al., 2025), our observation that early training predominantly follows the population direction suggests a practical alternative. Specifically, we first run standard training on the same limited dataset and select an early checkpoint from the population-learning stage as the surrogate population denoiser $\widehat{\phi}_{\text{pop}}$. We then train a new model from scratch with DAO using this frozen surrogate. This choice also connects DAO to early-learning regularization (Liu et al., 2020; 2022), which leverages early-stage predictions to improve robustness to label noise. Due to limited space, we summarize the training procedure as Algorithm 1 (in Appendix D), defer further implementation details to Appendix D.1, and present the empirical results in the next subsection.

4.2 Results

We evaluate DAO within the EDM training pipeline (Karras et al., 2022), using standard training as the baseline. We consider CIFAR-10 with $N \in \{2048, 4096\}$ and FFHQ with $N \in \{1024, 2048\}$. All runs are trained for 30M training images, except CIFAR-10 with $N = 4096$, which is trained for 40M images. We report the best FID achieved throughout training together with the memorization ratio at the final checkpoint in Table 1.

Across all four settings, DAO improves the best FID while substantially suppressing memorization under prolonged training. The final memorization ratio is reduced by more than 90% relative to standard training in every setting, reaching at most 4.11% and falling below 1% in three of the four cases. Meanwhile, DAO consistently reduces the best FID across all four settings, showing that its improved resistance to memorization is accompanied by better generation quality. Moreover, as illustrated in Figure 4, DAO does not delay the early population-learning dynamics: the rise of a_k and the decrease in test loss closely track those of standard training, while the subsequent growth of b_k and test loss is strongly suppressed, allowing the test loss to decrease further than under standard training.

Table 1: **DAO improves generation quality and reduces memorization.** Results on FFHQ and CIFAR-10 in limited-data settings. We report the best FID achieved during training and the final memorization ratio. Lower values are better for both metrics. All results are averaged over three random seeds.

Dataset	N	Method	Best FID ↓	Final memorization ratio (%) ↓
FFHQ	1024	Baseline	17.36 ± 0.29	81.64 ± 0.25
		DAO	$\mathbf{12.71} \pm 1.02$	$\mathbf{0.56} \pm 0.09$
	2048	Baseline	11.30 ± 0.22	18.40 ± 0.92
		DAO	$\mathbf{9.59} \pm 0.26$	$\mathbf{0.04} \pm 0.01$
CIFAR-10	2048	Baseline	13.44 ± 0.19	55.70 ± 0.21
		DAO	$\mathbf{11.20} \pm 0.40$	$\mathbf{4.11} \pm 0.34$
	4096	Baseline	7.00 ± 0.08	16.47 ± 0.43
		DAO	$\mathbf{6.02} \pm 0.08$	$\mathbf{0.84} \pm 0.07$

5 Conclusion

In this work, we presented a precise geometric characterization of the transition from early generalization to late memorization in diffusion models. We established that the population and empirical directions are orthogonal in expectation. Across synthetic and real-world datasets, we observed that training trajectories increasingly concentrate in the subspace spanned by these directions, first following the population direction and then making a sharp, nearly right-angle turn toward the empirical denoiser. Guided by this geometry, we proposed DAO, which improves generalization and mitigates memorization in limited-data settings. Future work could analyze the learning dynamics for a given data distribution and network architecture to verify or generalize Assumption 1.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the writing of the manuscript, including improving phrasing and transitions in both the main text and the proofs. All AI-assisted revisions were carefully reviewed and edited by the authors to ensure their correctness and consistency with the intended content. We also used generative AI tools to assist with writing code for experimental verification and plotting, and all resulting code was checked by the authors for correctness. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, clean and reformat dataset, support qualitative and thematic data analysis, and generate synthetic data sets is not applicable to this work. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings and implementation details in Appendix D to facilitate reproducibility. In addition, we include code for reproducing the training trajectory results and DAO training in the EDM setting as supplementary materials.

REFERENCES

Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 2024.

Ricardo Baptista, Agnimitra Dasgupta, Nikola B Kovachki, Assad Oberai, and Andrew M Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 2024.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Sam Buchanan, Druv Pai, Yi Ma, and Valentin De Bortoli. On the edge of memorization in diffusion models. *Advances in Neural Information Processing Systems*, 38:96113–96157, 2026.

Ruchika Chavhan, Ondrej Bohdal, Yongshuo Zong, Da Li, and Timothy Hospedales. Memorized images in diffusion models share a subspace that can be located and deleted. *arXiv preprint arXiv:2406.18566*, 2024.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Tyler Farghly, Peter Potaptchik, Samuel Howard, George Deligiannidis, and Jakiw Pidstrigach. Diffusion models and the manifold hypothesis: Log-domain smoothing is geometry adaptive. In *Advances in Neural Information Processing Systems*, volume 38, pp. 122795–122840, 2026.

Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models. *arXiv preprint arXiv:2505.16959*, 2025.

Jérôme Garnier-Brun, Luca Biggio, Davide Beltrame, Marc Mézard, and Luca Saglietti. Biased generalization in diffusion models. In *International Conference on Machine Learning*, 2026.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3D equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations*, 2023.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding NeMo: Localizing neurons responsible for memorization in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning*, volume 162, pp. 8867–8887, 2022.

Henry Hunt, Mason Kamb, and Surya Ganguli. An exact information theory of generalization phase transitions in bayesian diffusion models. *arXiv preprint arXiv:2607.08041*, 2026.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stephane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *International Conference on Machine Learning*, volume 267, pp. 28795–28831, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(12): 4217–4228, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Xiang Li, Yixiang Dai, and Qing Qu. Understanding generalizability of diffusion models requires rethinking the hidden Gaussian structure. In *Advances in Neural Information Processing Systems*, volume 37, pp. 57499–57538, 2024.

Xiao Li, Zekai Zhang, Xiang Li, Siyi Chen, Zhihui Zhu, Peng Wang, and Qing Qu. Understanding representation dynamics of diffusion models via low-dimensional modeling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 107365–107404, 2025.

Xiao Li, Yixuan Jia, Zekai Zhang, Xiang Li, Lianghe Shi, Jinxin Zhou, Zhihui Zhu, Liyue Shen, and Qing Qu. Evaluating the representation space of diffusion models via self-supervised principles. In *International Conference on Machine Learning*, 2026.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20331–20342, 2020.

Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2606–2616, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: Emergence of diffusion models from associative memory. *arXiv preprint arXiv:2505.21777*, 2025.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14532–14542, 2022.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, volume 139, pp. 8857–8868, 2021.

Kulin Shah, Alkis Kalavasis, Adam Klivans, and Giannis Daras. Does generation require memorization? Creative diffusion models using ambient diffusion. In *International Conference on Machine Learning*, pp. 54143–54166, 2025.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47783–47803, 2023.

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.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816, 2021.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 95865–95963, 2025.

Binxu Wang, Emma Lucia Byrnes Finn, and Bingbin Liu. The two clocks and the innovation window: When and how generative models learn rules. *arXiv preprint arXiv:2605.10019*, 2026.

Zeqi Ye, Qijie Zhu, Molei Tao, and Minshuo Chen. Provable separations between memorization and generalization in diffusion models. In *International Conference on Learning Representations*, 2026.

Huijie Zhang, Jinfan Zhou, Yifu Lu, Minzhe Guo, Peng Wang, Liyue Shen, and Qing Qu. The emergence of reproducibility and consistency in diffusion models. In *International Conference on Machine Learning*, 2024.

Zekai Zhang, Xiao Li, Xiang Li, Lianghe Shi, Meng Wu, Molei Tao, and Qing Qu. Generalization of diffusion models arises with a balanced representation space. In *International Conference on Learning Representations*, 2026.

Zhaoyu Zhang, Yang Hua, Guanxiong Sun, Hui Wang, and Seán McLoone. Training diffusion-based generative models with limited data. In *International Conference on Machine Learning*, 2025.

Bohao Zou. Denoising diffusion probabilistic model (ddpm). https://github.com/zoubohao/DenoisingDiffusionProbabilityModel-ddpm-. GitHub repository.

Appendix

The appendix is organized as follows: in Appendix A, we discuss related works. In Appendix B. we provide proof details for Section 2 and Section 3. In Appendix C, we provide additional complementary experiments. Lastly, in Appendix D, we discuss the implementation details of DAO and also present the detailed experimental setups for all the empirical results in the paper.

A Related Work

In this section, we discuss the relationship between our results and closely related works on diffusion models.

Generalization and memorization in diffusion models. Understanding how diffusion models generalize from finite datasets requires distinguishing the learning of population structure from the reproduction of individual training samples. Complementing studies of the transition from early generalization to late memorization during training, empirical work has identified a transition from memorization to generalization as the training set grows (Zhang et al., 2024; Kadkhodaie et al., 2024). Efforts have also been made to identifying the crossover between generalization and memorization as a function of model capacity (Buchanan et al., 2026). To explain generalization, Kadkhodaie et al. (2024) characterize the learned denoisers through geometry-adaptive harmonic representations, while Li et al. (2024) identify an inductive bias toward Gaussian structure determined by the training data's mean and covariance. Farghly et al. (2026) provide a complementary manifold-based explanation, showing that smoothing the empirical score is geometry-adaptive and preferentially acts along the underlying data manifold. Moreover, Kamb & Ganguli (2025) develop an analytical account of how locality and translation equivariance enable convolutional diffusion models to generate novel images by recombining local training patches. From a complementary theoretical perspective, Baptista et al. (2025) analyze how the exact empirical score leads to memorization, highlighting the distinction between minimizing the empirical objective and recovering the underlying distribution. Related work has also interpreted diffusion models as associative memory systems (Ambrogioni, 2024). Building on this view, Pham et al. (2025) study the transition from isolated training-sample attractors to emergent generative states as the amount of training data increases. From a complementary dynamical perspective, Biroli et al. (2024) have also identified a transition along the backward diffusion process from recovering global data structure to collapsing toward individual training samples. Additionally, efforts have been made to connect the generalization and memorization behavior of diffusion models with their representation space, and to use representation characteristics to diagnose diffusion models (Li et al., 2025; Zhang et al., 2026; Li et al., 2026). More recently, Hunt et al. (2026) derive an information-theoretic boundary between memorization and generalization in Bayesian diffusion models with restricted observations, relating this boundary to dataset size, diffusion noise, and information restriction. Our work complements these statistical and representational perspectives by studying the geometry of the training trajectory relative to the population and empirical Bayes denoisers. We establish orthogonality in expectation between the associated reference directions and observe that the model's displacement increasingly concentrates in their span.

Training dynamics and implicit regularization. Recent work shows that training duration plays a central role in the transition from generalization to memorization. Bonnaire et al. (2025) identify distinct generalization and memorization timescales through experiments and a theoretical analysis of a random-feature model. Favero et al. (2025) likewise observe that over-parameterized diffusion models progressively generalize before memorizing, and characterize the benefits and limitations of early stopping using a hierarchical data model. Related observations in Li et al. (2024) show that Gaussian structure emerges during early training before memorization develops. More recently, Garnier-Brun et al. (2026) identify a regime of *biased generalization*, in which test denoising loss continues to decrease while generated samples become increasingly biased toward individual training examples. At the level of learning mechanisms, Wang & Pehlevan (2025) derive training dynamics for linear and convolutional denoisers, relating mode-learning timescales to the data covariance spectrum. Our work complements these accounts by examining how the separation of learning timescales manifests geometrically: the observed trajectory initially follows the population direction before making a sharp, nearly orthogonal turn toward the empirical denoiser.

Mitigating memorization in diffusion models. Existing approaches mitigate memorization by modifying the training objective, model capacity, conditioning mechanism, or learned representations. Baptista et al. (2025) investigate explicit score regularization, under-parameterization, and early stopping as mechanisms for avoiding memorization. Shah et al. (2025) show that memorization is primarily required for denoising at sufficiently low noise levels and propose training with noisy data at larger noise scales to reduce memorization. For text-to-image diffusion models, Somepalli et al. (2023) demonstrate the importance of text conditioning and propose caption diversification and conditioning perturbations to reduce copying. Zhang et al. (2025) propose using a compression model together with mixed data augmentation during training to improve generation quality under limited data. More targeted interventions identify and deactivate memorization neurons (Hintersdorf et al., 2024), prune attention heads according to the importance of the denoising objective (Ye et al., 2026), or prune weights associated with activation subspaces shared by memorized prompts (Chavhan et al., 2024). DAO introduces a training penalty defined by the geometry of denoiser displacement. This connects the observed trajectory geometry to a selective intervention designed to mitigate memorization while preserving population-level learning.

B Proofs in Section 2 and Section 3

B.1 Proof of Lemma 1

Proof of Lemma 1. The square-integrability assumption ensures that all the conditional expectations and inner products above are well-defined. Fix any $t \in (0, 1]$. By the definitions of the empirical and population Bayes-optimal denoisers, we have

$$\boldsymbol{\phi}_{\mathrm{emp}}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t, \mathcal{D}_N], \quad \boldsymbol{\phi}_{\mathrm{pop}}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]. \tag{15}$$

Define the sigma-fields $\mathcal{G}_t := \sigma(\boldsymbol{x}_t, \boldsymbol{\theta}_0)$ and $\mathcal{H}_t := \sigma(\boldsymbol{x}_t, \mathcal{D}_N, \boldsymbol{\theta}_0)$. Clearly, $\mathcal{G}_t \subseteq \mathcal{H}_t$. Moreover, since $\boldsymbol{\theta}_0$ is independent of $(\mathcal{D}_N, \boldsymbol{x}_0, \boldsymbol{\epsilon})$, (15) implies

$$\boldsymbol{\phi}_{\mathrm{emp}}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid \mathcal{H}_t], \quad \boldsymbol{\phi}_{\mathrm{pop}}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid \mathcal{G}_t].$$

It follows that $\boldsymbol{d}_{\mathrm{emp}}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid \mathcal{H}_t] - \mathbb{E}[\boldsymbol{x}_0 \mid \mathcal{G}_t]$. By the tower property of conditional expectation, we have

$$\mathbb{E}[\boldsymbol{d}_{\mathrm{emp}}(\boldsymbol{x}_t, t) \mid \mathcal{G}_t] = \mathbb{E}[\mathbb{E}[\boldsymbol{x}_0 \mid \mathcal{H}_t] \mid \mathcal{G}_t] - \mathbb{E}[\boldsymbol{x}_0 \mid \mathcal{G}_t] = \mathbb{E}[\boldsymbol{x}_0 \mid \mathcal{G}_t] - \mathbb{E}[\boldsymbol{x}_0 \mid \mathcal{G}_t] = 0. \tag{16}$$

Moreover, we obtain that

$$\boldsymbol{d}_{\mathrm{pop}}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid \mathcal{G}_t] - \boldsymbol{\phi}_{\boldsymbol{\theta}_0}(\boldsymbol{x}_t, t)$$

is $\mathcal{G}_t$-measurable. Hence, using (16) and the tower property again,

$$\begin{aligned}\mathbb{E}[\langle \boldsymbol{d}_{\mathrm{pop}}(\boldsymbol{x}_t, t), \boldsymbol{d}_{\mathrm{emp}}(\boldsymbol{x}_t, t)\rangle \mid \boldsymbol{\theta}_0] &= \mathbb{E}[\mathbb{E}[\langle \boldsymbol{d}_{\mathrm{pop}}(\boldsymbol{x}_t, t), \boldsymbol{d}_{\mathrm{emp}}(\boldsymbol{x}_t, t)\rangle \mid \mathcal{G}_t] \mid \boldsymbol{\theta}_0] \\ &= \mathbb{E}[\langle \boldsymbol{d}_{\mathrm{pop}}(\boldsymbol{x}_t, t), \mathbb{E}[\boldsymbol{d}_{\mathrm{emp}}(\boldsymbol{x}_t, t) \mid \mathcal{G}_t]\rangle \mid \boldsymbol{\theta}_0] = 0.\end{aligned}$$

This proves the result for almost every realization of $\boldsymbol{\theta}_0$. □

B.2 Proof of Lemma 2

Proof of Lemma 2. For ease of exposition, we write $\boldsymbol{x}_0 = \boldsymbol{x}^{(I)}$ and $\boldsymbol{x}_t = s_t \boldsymbol{x}_0 + \gamma_t \boldsymbol{\epsilon}$. The empirical objective in (1) is

$$\ell(\boldsymbol{\theta}) = \mathbb{E}\big[\|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\|_2^2\big]. \tag{17}$$

Recall that the empirical Bayes denoiser satisfies $\boldsymbol{\phi}_{\mathrm{emp}}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t, t]$. Define the clean-target residual as $\boldsymbol{\xi} := \boldsymbol{x}_0 - \boldsymbol{\phi}_{\mathrm{emp}}(\boldsymbol{x}_t, t)$. This implies $\mathbb{E}[\boldsymbol{\xi} \mid \boldsymbol{x}_t, t] = \mathbf{0}$. For each fixed $\boldsymbol{\theta}$, the parameter Jacobian $\boldsymbol{J}_{\boldsymbol{\theta},t}$ is measurable with respect to $(\boldsymbol{x}_t, t)$. Thus, the tower property gives

$$\mathbb{E}\big[\boldsymbol{J}_{\boldsymbol{\theta},t}^\top \boldsymbol{\xi}\big] = \mathbb{E}\big[\boldsymbol{J}_{\boldsymbol{\theta},t}^\top \mathbb{E}[\boldsymbol{\xi} \mid \boldsymbol{x}_t, t]\big] = \mathbf{0}.$$

Under the stated regularity assumptions, we may differentiate (17) under the expectation. Applying the chain rule and the identity above yields

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = 2\,\mathbb{E}\big[\boldsymbol{J}_{\boldsymbol{\theta},t}^\top \big(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\big)\big] = 2\,\mathbb{E}\big[\boldsymbol{J}_{\boldsymbol{\theta},t}^\top \big(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\phi}_{\mathrm{emp}}(\boldsymbol{x}_t, t)\big)\big],$$

which proves (8). □

The objective decomposition follows from the same conditional-mean property. Specifically,

$$\mathbb{E}[\langle \boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\phi}_{\mathrm{emp}}(\boldsymbol{x}_t, t), \boldsymbol{\xi} \rangle] = 0.$$

Expanding the squared error therefore gives

$$\ell(\boldsymbol{\theta}) = \mathbb{E}\big[\|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\phi}_{\mathrm{emp}}(\boldsymbol{x}_t, t)\|_2^2\big] + \mathbb{E}\big[\|\boldsymbol{\xi}\|_2^2\big] = \ell^\star_{\mathrm{emp}} + \|\boldsymbol{\phi}_{\boldsymbol{\theta}} - \boldsymbol{\phi}_{\mathrm{emp}}\|_{\mathcal{H}}^2,$$

where $\ell^\star_{\mathrm{emp}} := \mathbb{E}[\|\boldsymbol{\xi}\|_2^2]$ is independent of $\boldsymbol{\theta}$. The same decomposition holds for any denoiser in $\mathcal{H}$, so $\boldsymbol{\phi}_{\mathrm{emp}}$ minimizes the empirical denoising objective over $\mathcal{H}$.

B.3 Proof of Theorem 1

Lemma 3 (Exact coordinate dynamics)**.** *Define* $\varepsilon_{\mathrm{pop}}(\tau) := \frac{\langle \boldsymbol{d}_{\mathrm{pop}}, \boldsymbol{K}_\tau \boldsymbol{r}_\tau \rangle_{\mathcal{H}}}{\|\boldsymbol{d}_{\mathrm{pop}}\|_{\mathcal{H}}^2}$*. Under the output-space dynamics (9) and decomposition (10), suppose that the reference directions are fixed, nonzero, and satisfy*

$$\langle \boldsymbol{d}_{\mathrm{pop}}, \boldsymbol{d}_{\mathrm{emp}} \rangle_{\mathcal{H}} = 0. \tag{18}$$

The coordinate dynamics satisfy

$$\dot{a}_\tau = (1 - a_\tau)\, \kappa_{\mathrm{pop}}(\tau) + (1 - b_\tau)\, \frac{\chi(\tau)}{\|\boldsymbol{d}_{\mathrm{pop}}\|^2} - \varepsilon_{\mathrm{pop}}(\tau), \tag{19}$$

$$\dot{b}_\tau = (1 - a_\tau)\, \frac{\chi(\tau)}{\|\boldsymbol{d}_{\mathrm{emp}}\|^2} + (1 - b_\tau)\, \kappa_{\mathrm{emp}}(\tau) - \varepsilon_{\mathrm{emp}}(\tau), \tag{20}$$

and the residual obeys

$$\dot{\boldsymbol{r}}_\tau = (1 - a_\tau)\boldsymbol{P}^\perp \boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{pop}} + (1 - b_\tau)\boldsymbol{P}^\perp \boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{emp}} - \boldsymbol{P}^\perp \boldsymbol{K}_\tau \boldsymbol{r}_\tau,$$

where $\boldsymbol{P}^\perp$ *denotes the* $\mathcal{H}$*-orthogonal projection onto* $\mathcal{S}^{\perp_{\mathcal{H}}}$*, with* $\mathcal{S} = \mathrm{span}\{\boldsymbol{d}_{\mathrm{pop}}, \boldsymbol{d}_{\mathrm{emp}}\}$*. In particular, the residual energy obeys*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau} \|\boldsymbol{r}_\tau\|^2 = (1 - a_\tau)\, \|\boldsymbol{d}_{\mathrm{pop}}\|^2\, \varepsilon_{\mathrm{pop}}(\tau) + (1 - b_\tau)\, \|\boldsymbol{d}_{\mathrm{emp}}\|^2\, \varepsilon_{\mathrm{emp}}(\tau) - \kappa_r(\tau)\, \|\boldsymbol{r}_\tau\|^2. \tag{21}$$

Proof of Lemma 3. For simplicity, we write $\boldsymbol{\phi}_\tau := \boldsymbol{\phi}_{\boldsymbol{\theta}(\tau)}$. By (10) and the assumed orthogonality in (18), we obtain

$$\boldsymbol{\phi}_\tau - \boldsymbol{\phi}_{\boldsymbol{\theta}_0} = a_\tau \boldsymbol{d}_{\mathrm{pop}} + b_\tau \boldsymbol{d}_{\mathrm{emp}} + \boldsymbol{r}_\tau,$$

where $\langle \boldsymbol{d}_{\mathrm{pop}}, \boldsymbol{d}_{\mathrm{emp}} \rangle_{\mathcal{H}} = \langle \boldsymbol{d}_{\mathrm{pop}}, \boldsymbol{r}_\tau \rangle_{\mathcal{H}} = \langle \boldsymbol{d}_{\mathrm{emp}}, \boldsymbol{r}_\tau \rangle_{\mathcal{H}} = 0$. Since $\boldsymbol{\phi}_{\mathrm{emp}} - \boldsymbol{\phi}_{\boldsymbol{\theta}_0} = \boldsymbol{d}_{\mathrm{pop}} + \boldsymbol{d}_{\mathrm{emp}}$, the prediction error relative to the empirical denoiser is

$$\boldsymbol{\phi}_{\mathrm{emp}} - \boldsymbol{\phi}_\tau = (1 - a_\tau)\boldsymbol{d}_{\mathrm{pop}} + (1 - b_\tau)\boldsymbol{d}_{\mathrm{emp}} - \boldsymbol{r}_\tau.$$

Substituting this expression into (9) gives

$$\dot{\boldsymbol{\phi}}_\tau = (1 - a_\tau)\boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{pop}} + (1 - b_\tau)\boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{emp}} - \boldsymbol{K}_\tau \boldsymbol{r}_\tau. \tag{22}$$

(i) Because the reference directions are nonzero and orthogonal in $\mathcal{H}$, the projection coefficients satisfy

$$a_\tau = \frac{\langle \boldsymbol{d}_{\mathrm{pop}}, \boldsymbol{\phi}_\tau - \boldsymbol{\phi}_{\boldsymbol{\theta}_0} \rangle_{\mathcal{H}}}{\|\boldsymbol{d}_{\mathrm{pop}}\|_{\mathcal{H}}^2}, \quad b_\tau = \frac{\langle \boldsymbol{d}_{\mathrm{emp}}, \boldsymbol{\phi}_\tau - \boldsymbol{\phi}_{\boldsymbol{\theta}_0} \rangle_{\mathcal{H}}}{\|\boldsymbol{d}_{\mathrm{emp}}\|_{\mathcal{H}}^2}.$$

Moreover, self-adjointness of $\boldsymbol{K}_\tau$ implies $\langle \boldsymbol{d}_{\mathrm{emp}}, \boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{pop}} \rangle_{\mathcal{H}} = \langle \boldsymbol{d}_{\mathrm{pop}}, \boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{emp}} \rangle_{\mathcal{H}} = \chi_\tau$. Since the reference directions and $\boldsymbol{\phi}_{\boldsymbol{\theta}_0}$ are fixed, differentiating the coordinate formulas and using (22) yields

$$\dot{a}_\tau = \frac{\langle \boldsymbol{d}_{\mathrm{pop}}, \dot{\boldsymbol{\phi}}_\tau \rangle_{\mathcal{H}}}{\|\boldsymbol{d}_{\mathrm{pop}}\|_{\mathcal{H}}^2} = (1 - a_\tau)\kappa_{\mathrm{pop}}(\tau) + (1 - b_\tau)\frac{\chi_\tau}{\|\boldsymbol{d}_{\mathrm{pop}}\|_{\mathcal{H}}^2} - \varepsilon_{\mathrm{pop}}(\tau),$$

$$\dot{b}_\tau = \frac{\langle \boldsymbol{d}_{\mathrm{emp}}, \dot{\boldsymbol{\phi}}_\tau \rangle_{\mathcal{H}}}{\|\boldsymbol{d}_{\mathrm{emp}}\|_{\mathcal{H}}^2} = (1 - a_\tau)\frac{\chi_\tau}{\|\boldsymbol{d}_{\mathrm{emp}}\|_{\mathcal{H}}^2} + (1 - b_\tau)\kappa_{\mathrm{emp}}(\tau) - \varepsilon_{\mathrm{emp}}(\tau).$$

(ii) Note that $\boldsymbol{P}^\perp$ is the orthogonal projection onto $\mathcal{S}^{\perp_{\mathcal{H}}}$, where $\mathcal{S} = \mathrm{span}\{\boldsymbol{d}_{\mathrm{pop}}, \boldsymbol{d}_{\mathrm{emp}}\}$. The residual is

$$\boldsymbol{r}_\tau = \boldsymbol{P}^\perp\big(\boldsymbol{\phi}_\tau - \boldsymbol{\phi}_{\boldsymbol{\theta}_0}\big).$$

The reference plane is fixed, so $\boldsymbol{P}^{\perp}$ is independent of τ. Differentiating and applying $\boldsymbol{P}^{\perp}$ to (22) gives

$$\dot{\boldsymbol{r}}_\tau = (1-a_\tau)\boldsymbol{P}^{\perp}\boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{pop}} + (1-b_\tau)\boldsymbol{P}^{\perp}\boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{emp}} - \boldsymbol{P}^{\perp}\boldsymbol{K}_\tau \boldsymbol{r}_\tau,$$

To obtain the residual energy identity, note that $\boldsymbol{P}^{\perp}$ is self-adjoint and $\boldsymbol{P}^{\perp}\boldsymbol{r}_\tau = \boldsymbol{r}_\tau$. Hence, for every $\boldsymbol{h} \in \mathcal{H}$, we have $\langle \boldsymbol{r}_\tau, \boldsymbol{P}^{\perp}\boldsymbol{h}\rangle_{\mathcal{H}} = \langle \boldsymbol{r}_\tau, \boldsymbol{h}\rangle_{\mathcal{H}}$. Taking the inner product of $\dot{\boldsymbol{r}}_\tau$ with $\boldsymbol{r}_\tau$ therefore yields

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\|\boldsymbol{r}_\tau\|_{\mathcal{H}}^2 &= \langle \boldsymbol{r}_\tau, \dot{\boldsymbol{r}}_\tau\rangle_{\mathcal{H}} = (1-a_\tau)\langle \boldsymbol{r}_\tau, \boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{pop}}\rangle_{\mathcal{H}} + (1-b_\tau)\langle \boldsymbol{r}_\tau, \boldsymbol{K}_\tau \boldsymbol{d}_{\mathrm{emp}}\rangle_{\mathcal{H}} - \langle \boldsymbol{r}_\tau, \boldsymbol{K}_\tau \boldsymbol{r}_\tau\rangle_{\mathcal{H}} \\ &= (1-a_\tau)\|\boldsymbol{d}_{\mathrm{pop}}\|_{\mathcal{H}}^2 \varepsilon_{\mathrm{pop}}(\tau) + (1-b_\tau)\|\boldsymbol{d}_{\mathrm{emp}}\|_{\mathcal{H}}^2 \varepsilon_{\mathrm{emp}}(\tau) - \kappa_r(\tau)\|\boldsymbol{r}_\tau\|_{\mathcal{H}}^2,\end{aligned}$$

where the last equality uses self-adjointness of $\boldsymbol{K}_\tau$ and the definitions of the directional rates and coupling terms. With the convention $\kappa_r(\tau) = 0$ whenever $\boldsymbol{r}_\tau = \boldsymbol{0}$, the identity remains valid at such times. □

Proof of Theorem 1. We first prove part (i). By the definition of ρ and Assumption 1(i), for every $u \in [0, T_1]$,

$$\dot{a}_u = \rho(u)(1-a_u)\kappa_{\mathrm{pop}}(u) > 0.$$

Hence, a_u is strictly increasing on $[0, T_1]$. Fix any $\tau \in [0, T_1]$. For every $\alpha \in [0, a_\tau]$, there exists a unique time $s(\alpha) \in [0, \tau]$ satisfying $a_{s(\alpha)} = \alpha$.

Since $\boldsymbol{K}_u$ is positive semidefinite, $\kappa_{\mathrm{emp}}(u) \geq 0$. Using $0 \leq 1 - a_u \leq 1$ and $0 \leq 1 - b_u \leq 1$ in equation 20, we obtain, for every $u \in [0, \tau]$,

$$\begin{aligned}\dot{b}_u &= (1-a_u)\frac{\chi(u)}{\|\boldsymbol{d}_{\mathrm{emp}}\|_{\mathcal{H}}^2} + (1-b_u)\kappa_{\mathrm{emp}}(u) - \varepsilon_{\mathrm{emp}}(u) \\ &\leq \frac{|\chi(u)|}{\|\boldsymbol{d}_{\mathrm{emp}}\|_{\mathcal{H}}^2} + \kappa_{\mathrm{emp}}(u) + |\varepsilon_{\mathrm{emp}}(u)| = \beta(u)\kappa_{\mathrm{pop}}(u).\end{aligned}$$

Using $\dot{a}_u = \rho(u)(1-a_u)\kappa_{\mathrm{pop}}(u)$ gives

$$\dot{b}_u \leq \frac{\beta(u)}{\rho(u)(1-a_u)}\dot{a}_u.$$

Differentiating the identity $a_{s(\alpha)} = \alpha$ with respect to α gives

$$\dot{a}_{s(\alpha)}\frac{\mathrm{d}s(\alpha)}{\mathrm{d}\alpha} = 1, \qquad \frac{\mathrm{d}s(\alpha)}{\mathrm{d}\alpha} = \frac{1}{\dot{a}_{s(\alpha)}}.$$

The chain rule therefore yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}b_{s(\alpha)} = \dot{b}_{s(\alpha)}\frac{\mathrm{d}s(\alpha)}{\mathrm{d}\alpha} &= \frac{\dot{b}_{s(\alpha)}}{\dot{a}_{s(\alpha)}} \\ &\leq \frac{\beta(s(\alpha))}{\rho(s(\alpha))(1-a_{s(\alpha)})} = \frac{\beta(s(\alpha))}{\rho(s(\alpha))(1-\alpha)}.\end{aligned}$$

Integrating from $\alpha = a_0 = 0$ to $\alpha = a_\tau$ gives

$$b_{s(a_\tau)} - b_{s(0)} \leq \int_0^{a_\tau} \frac{\beta(s(\alpha))}{\rho(s(\alpha))(1-\alpha)}\,\mathrm{d}\alpha.$$

Since $s(a_\tau) = \tau$, $s(0) = 0$, and $b_0 = 0$, we obtain

$$b_\tau \leq \int_0^{a_\tau} \frac{\beta(s(\alpha))}{\rho(s(\alpha))(1-\alpha)}\,\mathrm{d}\alpha.$$

Moreover, $s(\alpha) \in [0, \tau]$ for every $\alpha \in [0, a_\tau]$. Hence,

$$\beta(s(\alpha)) \leq \bar{\beta}_\tau, \qquad \rho(s(\alpha)) \geq \underline{\rho}_\tau > 0.$$

Applying these bounds to the preceding integral gives

$$b_\tau \le \frac{\bar{\beta}_\tau}{\underline{\rho}_\tau} \int_0^{a_\tau} \frac{\mathrm{d}\alpha}{1-\alpha} = \frac{\bar{\beta}_\tau}{\underline{\rho}_\tau} \log \frac{1}{1-a_\tau},$$

which proves part (i).

For part (ii), let $\tau \in [T_1, T_2]$. By equation 20,

$$\dot{b}_\tau = (1-b_\tau)\kappa_{\mathrm{emp}}(\tau) + (1-a_\tau)\frac{\chi(\tau)}{\|\boldsymbol{d}_{\mathrm{emp}}\|^2_{\mathcal{H}}} - \varepsilon_{\mathrm{emp}}(\tau).$$

Assumption 1(ii) therefore implies

$$\dot{b}_\tau \ge (1-b_\tau)\kappa_{\mathrm{emp}}(\tau) + (1-a_\tau)\frac{\chi(\tau)}{\|\boldsymbol{d}_{\mathrm{emp}}\|^2_{\mathcal{H}}} - |\varepsilon_{\mathrm{emp}}(\tau)| > 0.$$

The inequality $|1-a_\tau| \le \epsilon$ is also part of Assumption 1(ii). This proves part (ii). □

C Additional Experiments

Assumption validation. We numerically evaluate the conditions in Assumption 1 using a synthetic experiment. We draw $N = 100$ clean samples from a $K = 4$ mixture of Gaussian distributions in $\mathbb{R}^{20}$ and generate $m = 2$ noisy inputs per sample with noise level $\sigma = 0.2$. With the synthetic dataset, we can get its exact population target $\boldsymbol{\phi}_{\mathrm{pop}}$ and empirical target $\boldsymbol{\phi}_{\mathrm{emp}}$. We use a two-layer MLP network of width 512 with $\tanh$ activation and directly integrate its full-batch gradient-flow ODE on the noisy inputs. Thus, the optimization dynamics follow exact gradient flow for the finite Monte Carlo objective, and the fixed batch of samples approximates the expectation in the mean gradient flow analyzed in Section 3. We also verified that the raw population and empirical directions are nearly orthogonal, with cosine similarity 0.008.

We report the assumption checks in Figure 5. Specifically, in the left panel, the plot demonstrates $\rho(\tau) > 0$ throughout the population stage where the $\kappa_{\mathrm{pop}} > 0$ condition is already checked in Figure 3(top). The middle panel shows b_τ remains below both the pointwise integral bound the log bound. Finally, in the right panel, define the empirical margin

$$m_{\mathrm{emp}}(\tau) := (1-b_\tau)\kappa_{\mathrm{emp}}(\tau) + (1-a_\tau)\frac{\chi(\tau)}{\|\boldsymbol{d}_{\mathrm{emp}}\|^2_{\mathcal{H}}} - |\varepsilon_{\mathrm{emp}}(\tau)|.$$

and taking $T_1 = 0.044$ and $T_2 = 60$, we show that $m_{\mathrm{emp}}(\tau) > 0$ throughout $[T_1, T_2]$. Over the same interval, $|1-a_\tau| \le 0.05$, so the population coordinate remains close to one. At the stage boundary T_1, $a_{T_1} = 0.9506$ and $b_{T_1} = 0.0264$, which shows that population learning is nearly complete while empirical progress remains limited.

Qualitative comparison of DAO and standard training. Table 1 shows that DAO substantially mitigates memorization quantitatively. Here, we complement these results with a qualitative comparison of generated samples. We use the same models as in Table 1, trained under the EDM (Karras et al., 2022) setting on 2048 FFHQ (Karras et al., 2021) images for a total of 30M training images, and visualize generations from the final checkpoint. For each generated sample, we also show its nearest neighbor in the 2048-image training set. We mark a sample with a red box as a potential memorized copy if the ratio between its distance to the nearest neighbor and its distance to the second-nearest neighbor is below 0.7. As shown in Figure 6, standard training produces several such near-copies, whereas DAO produces none. Both models are sampled using identical latent codes with seed $= 0$, and no cherry-picking is performed for either method.

Sensitivity to the evaluation noise level. To evaluate whether the observed nearly orthogonal trajectory depends strongly on the choice of noise level, we repeat the trajectory analysis for the CIFAR-10 EDM experiment with $N = 2048$ at several values of σ, namely $\sigma \in \{0.1, 0.3, 0.5, 1.0, 2.0, 8.0\}$. The results are shown in Figure 7. Across a broad range of noise levels, we consistently observe the same qualitative behavior: the trajectory first progresses predominantly along the population direction, the relative residual energy quickly decreases, and the dynamics later make a sharp nearly orthogonal turn toward the empirical direction. The $\sigma = 0.3$ (Panel (c)) case is used in the main manuscript.

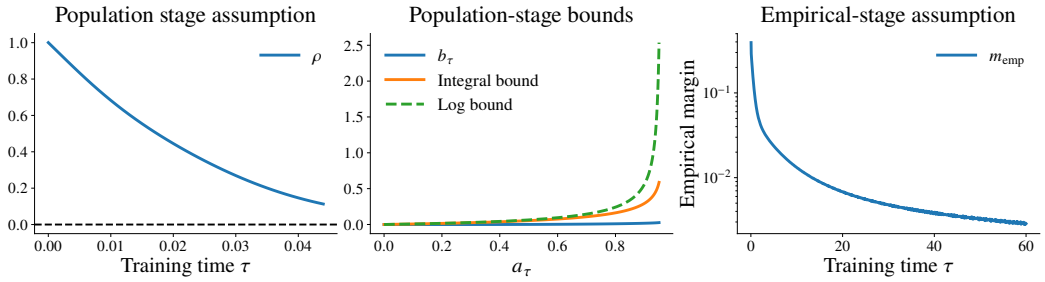
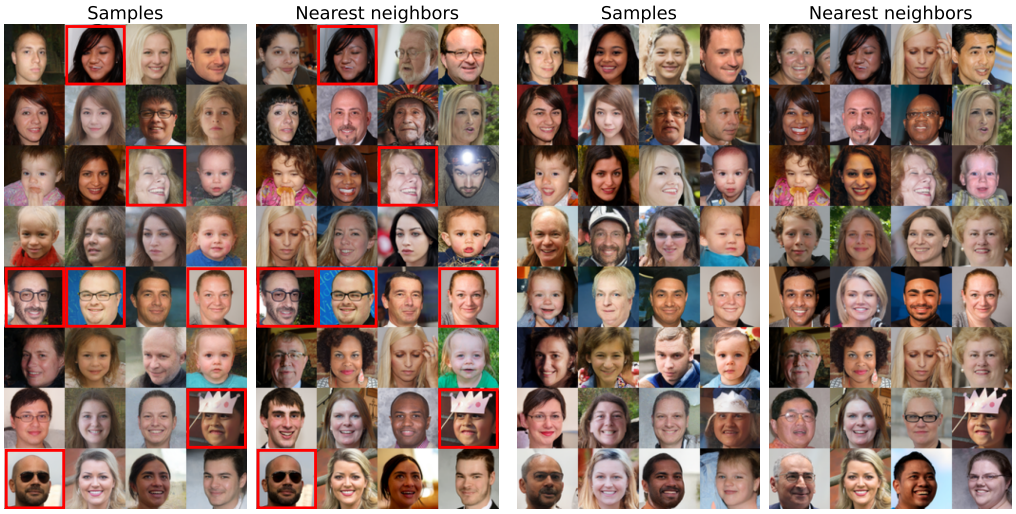


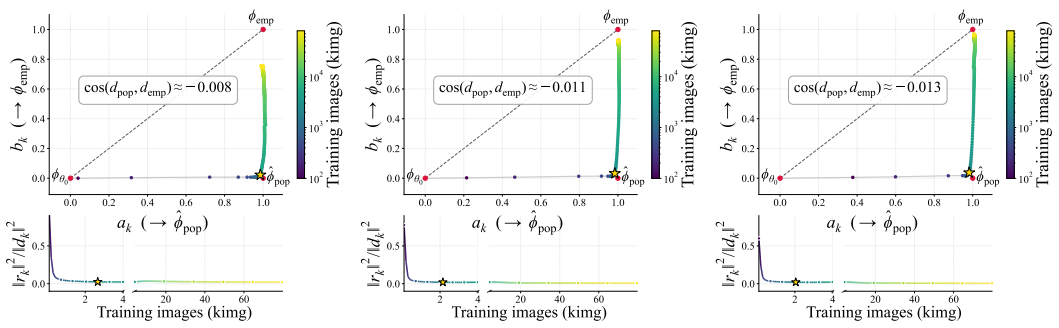
Figure 5: **Verification of Assumption 1.** We validate Assumption 1 under gradient flow on a mixture of Gaussians dataset using an MLP with $\tanh$ activations. The left panel verifies $\rho > 0$ during the population-learning stage. The middle panel verifies the two bounds in Theorem 1. The right panel verifies the positive empirical-stage margin $m_{\text{emp}}(\tau) := (1 - b_\tau)\kappa_{\text{emp}}(\tau) + (1 - a_\tau)\frac{\chi(\tau)}{\|\boldsymbol{d}_{\text{emp}}\|_{\mathcal{H}}^2} - |\varepsilon_{\text{emp}}(\tau)|$.



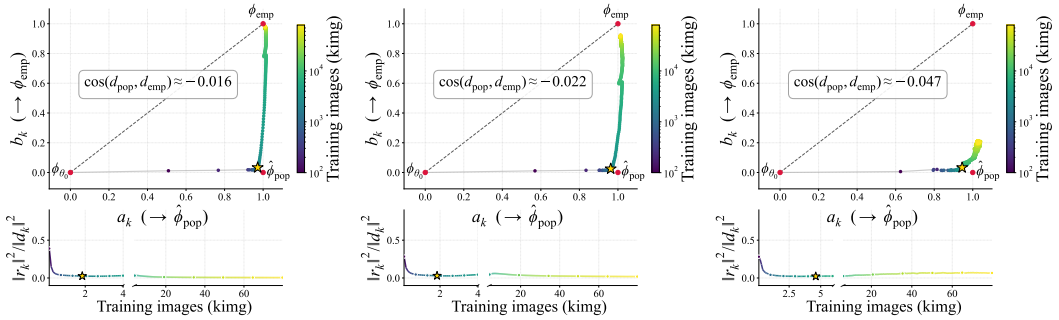
(a) Standard training (b) DAO training

Figure 6: **Qualitative comparison of DAO and standard training.** We use the final checkpoints corresponding to the FFHQ ($N = 2048$) results in Table 1. For each generated sample (left), we show its nearest neighbor in the training set (right). Samples identified as potential memorized copies are highlighted with red boxes. Standard training produces several such near-copies, whereas DAO produces none. Both models are sampled using identical latent codes with seed $= 0$, with no cherry-picking.

Training trajectory with large N. Our paper focuses on the limited-data regime, where we show that the training exhibits a two-phase trajectory with a sharp turn from the population direction toward the empirical direction. It is therefore important to verify that this turn disappears when abundant data are available, i.e., in a regime where memorization is not expected to happen. To this end, we revisit the synthetic mixture-of-Gaussians setting from the main text and compare two training-set sizes: the limited-data regime with $N = 256$ (reusing the result from Figure 2) and an abundant-data regime with $N = 50{,}000$. In both cases, the population and empirical Bayes-optimal denoisers are available in closed form, so the trajectory can be analyzed exactly in the same way as in Figure 2. As shown in Figure 8, in contrast to the limited-data case, when $N = 50000$ the empirical-turning phase essentially disappears. The trajectory remains concentrated near the population direction, and the empirical component stays negligible throughout training. This comparison supports the view that the late turn toward the empirical direction is a finite-data effect associated with memorization, and becomes largely absent when the training set is sufficiently large.



(a) $\sigma = 0.1$ (b) $\sigma = 0.3$ (c) $\sigma = 0.5$



(d) $\sigma = 1.0$ (e) $\sigma = 2.0$ (f) $\sigma = 8.0$

Figure 7: **Sensitivity of the training trajectory to the evaluation noise level.** We repeat the trajectory analysis for the EDM model trained on CIFAR-10 with $N = 2048$, evaluating the decomposition at $\sigma \in \{0.1, 0.3, 0.5, 1.0, 2.0, 8.0\}$. Across a broad range of noise levels, the same qualitative behavior persists: training first progresses predominantly along the population direction while the relative residual energy $\|\boldsymbol{r}_k\|^2/\|\boldsymbol{d}_k\|^2$ rapidly decreases, and then turns toward the empirical direction. The main manuscript uses $\sigma = 0.3$ as a representative setting.

D Experiment Details

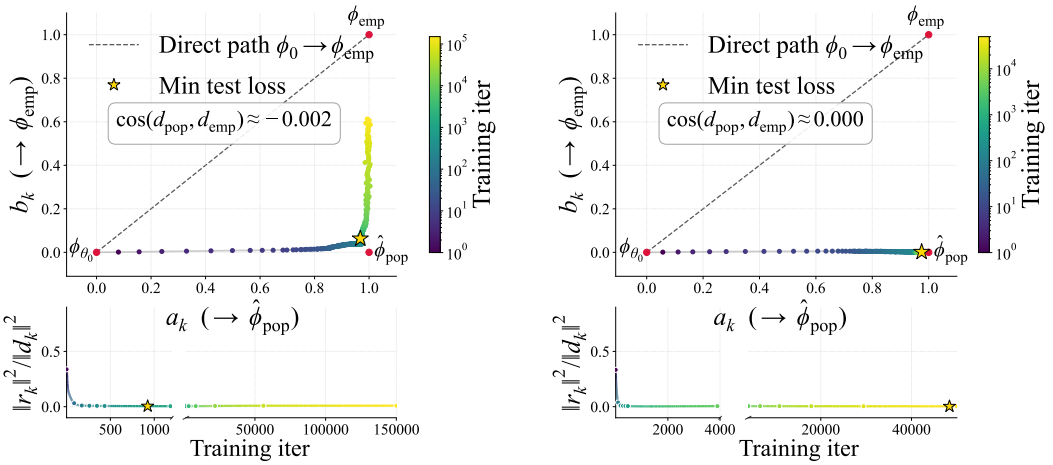
In this section, we first discuss the design choices of DAO in Appendix D.1 and then state the experiment details for reproducing each figure and table in the main manuscript in Appendix D.2.

D.1 Implementation Details for Section 4

In Section 4, we introduced the design principle of **D**irection-**A**ware **O**rthogonal Regularization (DAO). Here, we provide the full training procedure in Algorithm 1, followed by additional implementation details and discussion of the main design choices. We note that our implementation follows the algorithm and implementation described in the main paper, with both hyperparameters fixed across all experiments and without tuning. We do not exhaustively explore the alternative implementation choices discussed below or other hyperparameter settings, and some of these alternatives may further improve the performance of DAO.

Comparison with a simpler form. A simpler direction-aware regularizer would penalize only the positive empirical projection, i.e., $\mathcal{L}_{\text{simple}} = \sum_i \lambda_i \,\text{ReLU}(s_i)$. Our initial trials using this simpler form actually confirm that this simpler form alone can already improve the best generation performance. However, its regularization strength does not depend on how far the model has already progressed along the empirical direction. Consider for $s_i > 0$,

$$\frac{\partial \mathcal{L}_{\text{simple}}}{\partial s_i} = \lambda_i,$$



(a) MoG with $N = 256$ (reusing Figure 2(a))

(b) MoG with $N = 50{,}000$

Figure 8: **Training trajectories in the synthetic setting under limited and abundant data.** In the limited-data regime ($N = 256$), training first progresses along the population direction and then makes a sharp turn toward the empirical direction. In contrast, with abundant data ($N = 50000$), the trajectory stays close to the population direction and exhibits little to no late turn toward the empirical direction. This supports the view that the empirical-turning phase is a finite-data phenomenon associated with memorization.

Algorithm 1 Training with DAO

Require: training set $\mathcal{D}_N$; frozen population reference $\widehat{\phi}_{\text{pop}}$; penalty scale c_{pen}; progress scale η_0

for each training iteration **do**
  Compute the standard denoising loss $\ell(\boldsymbol{\theta})$ on a mini-batch from $\mathcal{D}_N$
  Sample a penalty mini-batch $\{\boldsymbol{x}^{(i)}\}_{i=1}^{B_p} \subseteq \mathcal{D}_N$ and noise levels $\{\sigma_i\}_{i=1}^{B_p}$
  $\boldsymbol{x}_t^{(i)} \leftarrow \boldsymbol{x}^{(i)} + \sigma_i \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  for $i = 1, \ldots, B_p$ **do**
    Compute $\boldsymbol{\phi}_{\text{emp}}^{(i)} = \boldsymbol{\phi}_{\text{emp}}(\boldsymbol{x}_t^{(i)}, \sigma_i)$ using equation 2
    $\boldsymbol{d}_{\text{emp}}^{(i)} \leftarrow \boldsymbol{\phi}_{\text{emp}}^{(i)} - \widehat{\boldsymbol{\phi}}_{\text{pop}}(\boldsymbol{x}_t^{(i)}, \sigma_i)$
    $\boldsymbol{d}_{\boldsymbol{\theta}}^{(i)} \leftarrow \boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(i)}, \sigma_i) - \widehat{\boldsymbol{\phi}}_{\text{pop}}(\boldsymbol{x}_t^{(i)}, \sigma_i)$
    $s_i \leftarrow \left\langle \frac{\boldsymbol{d}_{\text{emp}}^{(i)}}{\|\boldsymbol{d}_{\text{emp}}^{(i)}\|_2}, \boldsymbol{d}_{\boldsymbol{\theta}}^{(i)} \right\rangle$
    $\eta_i \leftarrow s_i / \|\boldsymbol{d}_{\text{emp}}^{(i)}\|_2$
    $\lambda_i \leftarrow \text{sg}[2c_{\text{pen}} \left\| w(\sigma_i)\left(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(i)}, \sigma_i) - \boldsymbol{x}^{(i)}\right)\right\|_2]$
    $\ell_{\text{DAO}}^{(i)} \leftarrow \lambda_i \,\text{ReLU}(s_i)\left(1 + \frac{\text{ReLU}(\eta_i)}{\eta_0}\right)$
  end for
  $\mathcal{L}(\boldsymbol{\theta}) \leftarrow \ell(\boldsymbol{\theta}) + \sum_{i=1}^{B_p} \ell_{\text{DAO}}^{(i)}$
  Update $\boldsymbol{\theta}$ using the total loss $\mathcal{L}(\boldsymbol{\theta})$
end for

so the resistance to the empirical motion remains constant regardless of the accumulated displacement. As a result, this simpler form can slow motion toward the empirical solution, but does not adaptively become stronger as empirical fitting progresses.

In DAO, as written in the main manuscript, we instead choose the penalty to consider the normalized empirical progress $\eta_i = s_i / \|\boldsymbol{d}_{\text{emp}}^{(i)}\|_2$. For $s_i > 0$,

$$\frac{\partial \mathcal{L}_{\text{DAO}}}{\partial s_i} = \lambda_i \left(1 + \frac{2\eta_i}{\eta_0}\right),$$

so the regularization becomes adaptively stronger as the model moves farther along the empirical direction. We can see that near the population reference, where η_i is small, DAO behaves similarly to the simpler form. But as empirical fitting accumulates, it increasingly suppresses further motion toward $\boldsymbol{\phi}_{\mathrm{emp}}$, thus has a better "mitigating" memorization effect compared with the simpler form.

Adaptive penalty strength. To account for variations in the scale of the denoising objective across noise levels and training stages, we adapt the penalty strength to the magnitude of the corresponding denoising gradient:

$$\lambda_i := \mathrm{sg}\Big[2c_{\mathrm{pen}}\, w(\sigma_i)\, \Big\|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(i)}, \sigma_i) - \boldsymbol{x}^{(i)}\Big\|_2\Big], \tag{23}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator and c_{pen} controls the overall regularization strength and $w(\sigma)$ is the weighting function in the original denoising objective. Since $2w(\sigma_i)\|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(i)}, \sigma_i) - \boldsymbol{x}^{(i)}\|_2$ is exactly the norm of the per-sample denoising gradient with respect to the model output, this scaling keeps the strength of DAO proportional to the underlying denoising force. We use $c_{\mathrm{pen}} = 1.0$ for all DAO experiments without tuning this hyperparameter.

Choice of surrogate $\widehat{\boldsymbol{\phi}}_{\mathrm{pop}}$. Motivated by the early population-learning behavior and the nearly orthogonal training trajectory observed in our analysis, we construct $\widehat{\boldsymbol{\phi}}_{\mathrm{pop}}$ using the same limited training set. Specifically, we first run standard training and select an early checkpoint from the population-learning stage that achieves the best FID as the surrogate population denoiser. We then train a new model from scratch with DAO using Algorithm 1. While DAO itself introduces little additional per-iteration overhead, obtaining $\widehat{\boldsymbol{\phi}}_{\mathrm{pop}}$ requires a short auxiliary standard-training run. This surrogate construction directly exploits the tendency of early training to remain close to the population direction. One possible alternative is to use a running model average (e.g., the EMA version of the model) to serve as the surrogate population denoiser for better efficiency, which we leave for future exploration.

Besides exploring the early learning scenario, other constructions are also possible. For example, prior work has identified a transition between generalization and memorization that depends on both the number of training samples and model capacity (Zhang et al., 2024), with smaller datasets or larger models generally making memorization more likely. This suggests an alternative surrogate based on a sufficiently small-capacity model trained on the same limited dataset, whose restricted capacity keeps it in the generalization regime. We leave the exploration of such alternatives to future work.

D.2 Other Emperiment Details

D.2.1 Figure 2.

We consider several data and training configurations in this figure and describe each below.

Panel (a): Synthetic dataset. We consider a mixture of $K = 4$ isotropic Gaussians in $d = 16$ dimensions and draw $N = 256$ training samples. In this setting, both the population and empirical Bayes-optimal denoisers in equation 2 are available in closed form. We train a two-hidden-layer MLP with width 512 and SiLU activations using full-batch gradient descent with learning rate 5×10^{-3} for 1.5×10^5 iterations. At each iteration, every training sample is perturbed using a freshly sampled noise level, drawn log-uniformly from $[0.002, 2]$, together with fresh Gaussian noise. To evaluate the geometry in the training trajectory, we use a fixed bank of noisy training inputs at $\sigma = 0.5$ and obtain (a_k, b_k) and the relative residual energy $\|\boldsymbol{r}_k\|^2/\|\boldsymbol{d}_k\|^2$ by least-squares projection onto $\mathrm{span}\{\boldsymbol{d}_{\mathrm{pop}}, \boldsymbol{d}_{\mathrm{emp}}\}$. The test loss used to identify the minimum-test-loss checkpoint is evaluated on 8192 fresh samples at the same noise level.

Panel (b): DDPM on CelebA. For this experiment, we use a DDPM (Ho et al., 2020) implementation based on the public codebase of (Zou) and train on the CelebA dataset (Liu et al., 2015). We first downsample all images to 32×32 resolution and then randomly select a subset of $N = 4096$ images for the limited-data training run. To construct the surrogate population denoiser, we use the same model and training configuration but train on the full available CelebA dataset for 1500 epochs.

To evaluate the training trajectory, we use a fixed set of noisy training inputs at diffusion timestep $t = 200$ and track the evolution of a_k, b_k, and the relative residual energy throughout training.

Panel (c): JiT on ImageNet. For this experiment, we use the JiT-B/16 configuration (Li & He, 2026), following the official implementation, and train on a randomly selected subset of $N = 5000$ ImageNet images (Deng et al., 2009). The officially released JiT-B/16 checkpoint trained on all ImageNet dataset is used as the surrogate population denoiser. We note that our limited-data model is trained unconditionally. Thus, when evaluating the trajectory, we use the null class for the reference model as well to ensure consistency. We evaluate the trajectory at $t = 0.7$ in the JiT parameterization.

Panels (d)-(f): EDM on CIFAR-10, FFHQ, and AFHQ. For these experiments, we use the official EDM implementation (Karras et al., 2022). We follow the corresponding preprocessing protocol, VP configuration, and default training hyperparameters for each dataset. For each dataset, we deterministically select the first $N = 2048$ images as the limited-data training subset. For the surrogate population denoiser, we use the officially released VP checkpoint trained on the full dataset. We evaluate the training trajectory at $\sigma = 0.3$ for all three experiments shown in this figure.

D.2.2 FIGURE 3.

For the synthetic experiment in panel (a), the setup is described in the first paragraph of Appendix C. For panel (b), we use the same JiT-B/16 model from Figure 2 to evaluate the bound in Theorem 1. Unlike the synthetic setting, this practical training run follows discrete stochastic optimization rather than the gradient-flow dynamics assumed in our analysis, so the quantity $\rho(\tau)$ cannot be directly evaluated from its theoretical definition. We therefore use a fixed value $\rho(\tau) = 0.3$ when computing the bound shown in the figure.

D.2.3 FIGURE 4.

For this experiment, we use the same EDM model trained on the FFHQ subset with $N = 2048$ from Figure 2, together with the corresponding model trained with DAO. The computation of a_k, b_k, and the test loss follows the same procedure as in the FFHQ experiment for Figure 2.

D.2.4 TABLE 1.

We evaluate DAO under the EDM setup on CIFAR-10 (Krizhevsky et al., 2009) and FFHQ (Karras et al., 2021). For CIFAR-10, we consider subset sizes $N = 2048$ and $N = 4096$. For FFHQ, we consider $N = 1024$ and $N = 2048$. We train both the standard baseline and DAO for a fixed budget of 30M training images in all settings except CIFAR-10 with $N = 4096$, for which we train for 40M images. The baseline follows the standard EDM configuration, while DAO follows the implementation described in Appendix D.1, using the same hyperparameter settings as stated there across all experiments.

Memorization ratio. We compute the memorization ratio at the end of training by generating 10K samples and extracting features for both generated and training images using the encoder of Pizzi et al. (2022). A generated sample is marked as memorized if its maximum cosine similarity to any training image exceeds 0.6. The reported memorization ratio is the fraction of generated samples satisfying this criterion.

FID and surrogate checkpoint selection. To evaluate generation quality, we generate 50K samples from checkpoints saved throughout training and compute FID accordingly. For the standard baseline, the best FID is attained at 4.096M training images for CIFAR-10 with $N = 2048$, at either 7.168M or 8.192M images for CIFAR-10 with $N = 4096$, at 2.560M images for FFHQ with $N = 1024$, and at 5.120M images for FFHQ with $N = 2048$. We thus use these checkpoints as the frozen surrogate population denoisers for the corresponding DAO runs. DAO typically matches or improves upon the baseline FID around these checkpoints and continues improving afterward. Its best FID is attained at 6.144M, 11.264M, 8.192M or 10.752M, and 14.336M training images for CIFAR-10 with $N = 2048$ and $N = 4096$, and FFHQ with $N = 1024$ and $N = 2048$, respectively.