ON THE MEMORIZATION OF CONSISTENCY DISTILLATION FOR DIFFUSION MODELS

Bingqing Jiang & Difan Zou*
The University of Hong Kong
bingqingjiang@connect.hku.hk, dzou@hku.hk

ABSTRACT

Diffusion models play a central role in modern generative modeling, and understanding how they balance memorization and generalization is critical for their reliability and practical use. Recent work has shown that memorization in diffusion models is shaped by training dynamics, with generalization and memorization emerging at different stages of training. However, in practice, pretrained diffusion models are often distilled—an extra training phase whose impact on memorization is not well understood. In this work, we analyze how distillation reshapes memorization behavior in diffusion models, and take the prevalent consistency distillation as a representative framework. Empirically, we show that when applied to a teacher model that has memorized data, distillation significantly reduces transferred memorization in the student while simultaneously improving overall sample quality. To explain this, we further provide a theoretical analysis using a random feature neural network model (Bonnaire et al., 2025), showing that consistency distillation suppresses unstable feature directions associated with memorization while preserving stable, generalizable modes. Our findings show the potential of distillation beyond acceleration, enhancing generalization and long-term trustworthiness.

1 INTRODUCTION

Diffusion models have become a central paradigm in modern generative modeling due to their strong empirical performance, stable training dynamics, and flexibility across data modalities Song & Ermon (2019; 2020); Song et al. (2021b); Ho et al. (2020); Karras et al. (2022); Song et al. (2021a). By modeling generation as a gradual denoising process, diffusion models achieve high sample fidelity and robust generalization, making them a cornerstone of modern generative systems Podell et al. (2024). Given their growing importance, understanding how diffusion models balance memorization and generalization has become a fundamental question Gu et al. (2025); Wen et al. (2024); Jeon et al. (2024); Li et al. (2023); Somepalli et al. (2023). Recent studies show that memorization in diffusion models is governed by training dynamics Bonnaire et al. (2025); George et al. (2025); Pham et al. (2025). In particular, models typically achieve high generative quality before memorization emerges at later training stages, indicating that memorization is a dynamic, time-dependent phenomenon.

However, existing analyses of memorization have largely focused on diffusion models trained from scratch. This setting does not fully capture modern deployment pipelines, where diffusion models are often further trained via distillation to improve sampling efficiency and reduce computational cost Song et al. (2023); Salimans & Ho (2022); Kim et al. (2024); Luo et al. (2023); Geng et al. (2025). This raises a critical question that has received little attention to date:

How does distillation affect the memorization properties of diffusion models?

At a high level, distillation is not a passive model compression step, but an entirely new training process with its own objective, data distribution, and optimization dynamics Xiang et al. (2025). Intuitively, if memorization in diffusion models depends sensitively on training dynamics and time scales, as prior work suggests, then distillation may further reshape, suppress, or even amplify memorization inherited from the teacher. Yet, the effect of distillation on memorization remains largely uncharacterized.

*Corresponding author.

Recent progress toward few-step or even single-step diffusion generation has largely relied on distillation-based approaches Tee et al. (2024); Luo et al. (2024); Geng et al. (2023); Yin et al. (2024b); Starodubcev et al. (2025), among which consistency models have emerged as a representative and widely adopted framework Song et al. (2023). In this work, we study this question in the context of consistency distillation. Rather than explicitly supervising entire sampling trajectories or relying on large collections of teacher-generated samples Yin et al. (2024a); Park et al. (2025), consistency-based methods train a student model via an additional optimization procedure that enforces local agreement between neighboring points along the probability flow ODE. From the perspective of learning dynamics, consistency distillation introduces a new and nontrivial training phase beyond the original diffusion training. This additional optimization stage operates under a distinct objective and data distribution, and therefore has the potential to further reshape the balance between memorization and generalization established during the teacher's training. The main contributions of this paper can be summarized as follows:

- **Consistency distillation mitigates memorization.** We show that consistency distillation can reduce memorization inherited by the student model, even when the teacher exhibits strong overfitting. This effect holds across a wide range of settings, demonstrating that distillation actively reshapes memorization behavior rather than passively inheriting it.
- **Distillation improves generalization and performance.** Beyond reducing memorization, we find that consistency distillation can also improve sample quality. When the teacher model operates in a moderate memorization regime, the distilled student can even surpass the teacher in generative performance, indicating that memorization reduction does not come at the expense of utility.
- **How consistency distillation reshapes training dynamics?** We provide a mechanistic understanding of consistency distillation using a tractable Random Feature Neural Network (RFNN) model. Our analysis shows that consistency distillation reshapes training dynamics by concentrating updates on statistically stable feature directions, while rendering memorization-associated modes dynamically negligible. This structured update dynamics preserves generalizable representations and explains the empirical reduction of memorization under consistency distillation.

2 Definitions and Preliminaries

2.1 Generative Score Matching

Diffusion models define a generative mechanism by gradually transforming data drawn from an unknown target distribution P_0 on $\mathbb{R}^d$ into Gaussian noise through a continuous-time stochastic process. A standard formulation uses the Ornstein–Uhlenbeck (OU) stochastic differential equation

$$\mathrm{d}\mathbf{x}_t = -\mathbf{x}_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t, \tag{1}$$

where $\mathbf{B}_t$ denotes a standard Wiener process. This forward diffusion induces a family of intermediate distributions $\{P_t\}_{t\geq 0}$ that smoothly interpolate between P_0 and the standard Gaussian distribution $\mathcal{N}(0, I_d)$ as $t \to \infty$. The closed-form solution of (1) yields

$$\mathbf{x}_t = e^{-t}\mathbf{x}_0 + \sqrt{\Delta_t}\,\boldsymbol{\xi}, \quad \Delta_t = 1 - e^{-2t}, \tag{2}$$

with $\boldsymbol{\xi} \sim \mathcal{N}(0, I_d)$ independent of $\mathbf{x}_0$. Sampling from the target distribution is achieved by reversing the forward process in time, which takes the form

$$-\mathrm{d}\mathbf{x}_t = \big[\mathbf{x}_t + 2\nabla_{\mathbf{x}} \log P_t(\mathbf{x})\big]\mathrm{d}t + \sqrt{2}\,\mathrm{d}\widetilde{\mathbf{B}}_t, \tag{3}$$

where $\widetilde{\mathbf{B}}_t$ is a Wiener process evolving backward in time, and $\nabla_{\mathbf{x}} \log P_t(\mathbf{x})$ is the score function of the forward marginal at time t. Then generation reduces to learning the score $\nabla_{\mathbf{x}} \log P_t(\mathbf{x})$ for all relevant diffusion times.

The score function can be characterized as the minimizer of a denoising score-matching objective. In practice, the score is restricted to a parametrized family $\{\mathbf{s}_\theta(\cdot, t)\}_\theta$, typically implemented by a neural network, and the expectation over P_0 is replaced by an empirical average over a finite training set $\{\mathbf{x}_\nu\}_{\nu=1}^n$ Vincent (2011); Hyvärinen & Dayan (2005); Bonnaire et al. (2025):

$$\mathcal{L}_t(\theta) = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(0,I_d)}\Big[\big\|\sqrt{\Delta_t}\,\mathbf{s}_\theta(\mathbf{x}_{\nu,t}, t) + \boldsymbol{\xi}\big\|_2^2\Big], \tag{4}$$

where $\mathbf{x}_{\nu,t} = e^{-t}\mathbf{x}_\nu + \sqrt{\Delta_t}\boldsymbol{\xi}$.

2.2 Consistency Distillation

Consistency distillation trains fast generative models by transferring the local probability-flow dynamics of a pretrained diffusion model into a time-consistent student mapping Song et al. (2023). The student is trained to produce consistent predictions along short segments of the teacher-induced flow. We adopt consistency distillation as our focus, and further rationale is given in Appendix B.

Recall that the diffusion forward process admits an equivalent probability flow ODE $\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = h(t)\,\mathbf{x}_t - \frac{1}{2}g^2(t)\,\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$, which generates the same marginal distributions $\{P_t\}$ as the forward SDE. Given a pretrained teacher score model $\mathbf{s}_\phi(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, the PF-ODE induces a deterministic velocity field $\mathbf{v}_\phi(\mathbf{x}, t) = h(t)\,\mathbf{x} - \frac{1}{2}g^2(t)\,\mathbf{s}_\phi(\mathbf{x}, t)$. Under the OU forward in (1), the drift and diffusion coefficients are given by $f(\mathbf{x}, t) = -\mathbf{x}$ and $g(t) = \sqrt{2}$. The associated probability flow ODE therefore admits the simplified form

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = -\mathbf{x}_t - \mathbf{s}_\phi(\mathbf{x}_t, t). \tag{5}$$

In practice, the PF-ODE is discretized on a decreasing sequence of times $T = t_0 > t_1 > \cdots > t_K = 0$. Given a sample $\mathbf{x}_{t_{k+1}}$ at time t_{k+1}, we adopt an explicit Euler ODE solver and obtain the following single-step update:

$$\widehat{\mathbf{x}}^{\phi}_{t_k} = \mathbf{x}_{t_{k+1}} + (t_k - t_{k+1})\Big(-\mathbf{x}_{t_{k+1}} - \mathbf{s}_\phi(\mathbf{x}_{t_{k+1}}, t_{k+1})\Big). \tag{6}$$

We refer to $\widehat{\mathbf{x}}^{\phi}_{t_k}$ as the *teacher-induced one-step target*. Let $f_{\boldsymbol{\theta}}(\mathbf{x}, t)$ denote a student consistency model that maps a noisy input $\mathbf{x}$ at time t to a common representation, typically corresponding to an estimate of the clean data. The objective of consistency distillation enforces *time consistency* across neighboring discretization points:

$$L_{\mathrm{CD}}(\boldsymbol{\theta}) = \mathbb{E}_{k, \mathbf{x}_{t_{k+1}}}\left[\left\| f_\theta\big(\mathbf{x}_{t_{k+1}}, t_{k+1}\big) - f_\theta\big(\widehat{\mathbf{x}}^{\phi}_{t_k}, t_k\big)\right\|_2^2\right]. \tag{7}$$

This objective transfers the local dynamics of the teacher PF-ODE to the student without requiring the student to explicitly approximate the score function.

3 Memorization and Sample Quality in Consistency Distillation

Experimental Setup. We evaluate the memorization behavior and sample quality of consistency-distilled models on real-world image generation benchmarks. All experiments are conducted on the CIFAR-10 dataset (Krizhevsky et al., 2009), and no data augmentation is applied during training to avoid ambiguity in the assessment of memorization Gu et al. (2025). To study the effect of dataset size, we construct training sets by uniformly subsampling $n \in \{3000, 4000, 5000, 6000, 20000\}$ images from CIFAR-10 and conduct experiments separately for each setting. All distillation experiments use a pre-trained EDM diffusion model as the teacher (Karras et al., 2022). Student models are initialized from the same pretrained diffusion models and fine-tuned via consistency distillation Song et al. (2023). We use LPIPS (Zhang et al., 2018) as the metric in the consistency loss, and generate teacher one-step targets using Heun's second-order solver with $N = 18$ discretization steps.

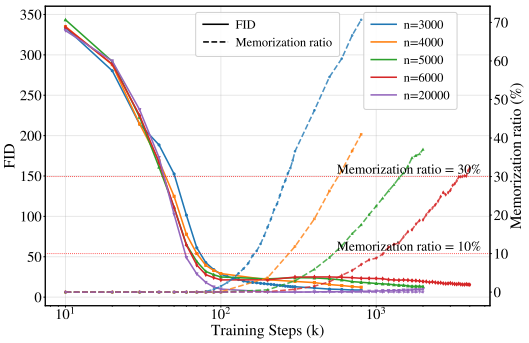
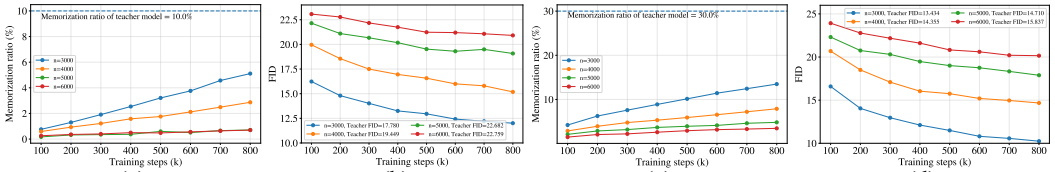


Figure 1: **FID–memorization dynamics of EDM teacher models.** Solid lines show FID (left axis) and dashed lines show memorization ratio (right axis) as a function of training steps for different dataset sizes n. Across all n, extended training improves FID but is eventually accompanied by a sharp increase in memorization. Smaller datasets enter the high-memorization regime earlier, while larger n delay this effect.

Evaluation metrics. We evaluate sample quality using Fréchet Inception Distance (FID) (Heusel et al., 2017), computed on 50,000 generated samples. To quantify memorization, we follow the



(a) (b) (c) (d)

Figure 2: **FID and memorization behavior under two-step consistency distillation. (a,b)** Moderate-memorization regime, where the teacher model exhibits a 10% memorization ratio. **(c,d)** Severe-memorization regime, with a teacher memorization ratio of 30%. In each regime, we report both memorization ratio and generative quality (FID) of distilled student models as training progresses.

memorization ratio defined in (Yoon et al., 2023). Specifically, a generated sample is deemed memorized if its ℓ_2 distance to the nearest training image is less than $1/3$ of the distance to the second nearest neighbor. The memorization ratio is defined as the fraction of memorized samples among $50{,}000$ generated images. Additional implementation details are provided in Appendix C.1.

Time-window effect of teacher models. We examine the training dynamics of EDM teacher models. Fig. 1 reveals a pronounced *time-window effect* in which improvements in sample quality precede the onset of memorization Bonnaire et al. (2025); Gu et al. (2025). During early training, FID decreases rapidly and then saturates while memorization remains negligible. Beyond a dataset-dependent critical point, further optimization yields diminishing FID gains accompanied by a sharp increase in memorization. This transition occurs earlier for smaller datasets, whereas increasing the dataset size systematically extends the low-memorization regime and can substantially suppress memorization when sufficiently large. Motivated by this behavior, we define memorization ratios of approximately 10% and 30% as representative *moderate* and *severe* regimes, respectively, and study how consistency distillation reshapes learning dynamics under these two teacher conditions. Unless otherwise stated, student models are evaluated using *two-step* sampling, which consistently outperforms one-step sampling in sample quality Song et al. (2023); Kim et al. (2024).

3.1 Moderate Memorization Case

Fig. 2a shows the memorization ratio of generated samples during distillation. First, *all distilled students remain far below the teacher's memorization level*: even at late training, memorization stays well under 10% for every n. Second, memorization tends to *increase with distillation steps* for fixed n, indicating a quality–memorization trade-off as the student is further fine-tuned. Importantly, this increase is moderate in absolute terms: for example, for $n = 3000$ it rises from below 1% early in training to around 5% by 800k steps, while for $n \geq 5000$ it remains below 1% throughout. Thus, in the moderate-memorization regime, consistency distillation can substantially reduce training-set copying relative to the teacher model. Fig. 2b reports FID as a function of consistency distillation training steps for different training-set sizes n. Across all n, FID decreases monotonically with training steps, and the student eventually *outperforms its teacher* in every case. By 800k steps, the student improves over the teacher by several FID points, with smaller but still consistent gains for larger n. This confirms that the reduced memorization observed under consistency distillation does not come at the expense of sample quality; instead, the student acquires useful information from the teacher that is directly relevant to the generation process.

3.2 Severe Memorization Case

In Fig. 2c, two observations mirror the moderate regime but at a more challenging scale. First, distilled students *consistently remain far below the teacher's memorization level* of 30% across all dataset size n, demonstrating a large absolute reduction in training-set copying despite the teacher being strongly memorizing. Second, for each fixed n, memorization increases with training steps, indicating a persistent quality–memorization tension as the student becomes more aligned with the teacher. Importantly, increasing the dataset size substantially suppresses this growth: the smallest dataset ($n = 3000$) shows the largest rise (reaching on the order of 10%–15% by 800k), whereas larger datasets keep memorization at much lower levels. Similarly, Fig. 2d shows that FID decreases steadily during distillation, indicating that consistency distillation continues to improve sample quality even under a severely memorizing teacher. While the magnitude of improvement depends on the dataset size n, this effect is largely explained by differences in teacher optimization level. Although all teachers are selected at a comparable memorization ratio, larger datasets require substantially more

training steps to reach this regime, resulting in teachers that are more fully optimized. Consequently, students trained from such teachers require a larger distillation budget to match or exceed teacher quality. Additional results are provided in Appendix E.1.

4 THEORETICAL ANALYSIS

In this section, we provide a theoretical explanation for why consistency distillation can reduce memorization without degrading sample quality.

4.1 ONE-STEP CONSISTENCY OBJECTIVE.

Our analysis considers a time-local regime with fixed t' and $\Delta t \to 0$, and focuses on a *one-step* consistency distillation objective. While consistency distillation is originally defined across multiple diffusion times (Song et al., 2023), the one-step formulation captures the leading-order training signal induced by the teacher probability-flow dynamics. It therefore provides a principled local approximation to the objective in Eq. (7), capturing the shared leading-order consistency constraint across diffusion times George et al. (2025); Bonnaire et al. (2025); Li et al. (2025). In this setting, the training objective compares the student outputs evaluated at two nearby inputs connected by a single teacher-induced step:

$$L_{\mathrm{CD}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}\left[\left\| f_{\boldsymbol{\theta}}\big(\mathbf{x}_{\nu,t'}(\boldsymbol{\xi})\big) - f_{\boldsymbol{\theta}}\big(\widehat{\mathbf{x}}^{\phi}_{\nu,t}(\boldsymbol{\xi})\big)\right\|_2^2\right], \tag{8}$$

where $\mathbf{x}_{\nu,t} = e^{-t}\mathbf{x}_\nu + \sqrt{\Delta_t}\boldsymbol{\xi}$ and $\widehat{\mathbf{x}}^{\phi}_{\nu,t}(\boldsymbol{\xi})$ is the teacher-induced one-step target in (6).

4.2 RANDOM FEATURE PARAMETERIZATION

Following prior theoretical studies of diffusion learning dynamics Li et al. (2023); Bonnaire et al. (2025); George et al. (2025), we parameterize both the teacher and the student using a RFNN. An RFNN is a two-layer neural network in which the first-layer weights $\mathbf{W} \in \mathbb{R}^{p\times d}$ are drawn i.i.d. from a Gaussian distribution and kept fixed, while only the second-layer weights are learned. We work in an asymptotic regime where d, p, and n jointly diverge to infinity, while the ratios p/d and n/d remain fixed. The teacher and student share the same frozen random features matrix $\mathbf{W}_\phi = \mathbf{W}_\theta = \mathbf{W} \in \mathbb{R}^{p\times d}$, $W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and use the same elementwise activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Define the feature map $\mathbf{h}(\mathbf{x}) = \sigma\left(\frac{\mathbf{W}\mathbf{x}}{\sqrt{d}}\right) \in \mathbb{R}^p$. Then the teacher score is modeled as $\mathbf{s}_\phi(\mathbf{x}) = \frac{1}{\sqrt{p}}\,\mathbf{A}_\phi\,\mathbf{h}(\mathbf{x})$ with $\mathbf{A}_\phi \in \mathbb{R}^{d\times p}$ is fixed, while the student consistency mapping is parameterized as $\mathbf{f}_\theta(\mathbf{x}) = \frac{1}{\sqrt{p}}\,\mathbf{B}_\theta\,\mathbf{h}(\mathbf{x})$ with $\mathbf{B}_\theta \in \mathbb{R}^{d\times p}$ is trainable. At fixed reference time t' and step size Δt, the one-step distillation loss compares the student outputs evaluated at two nearby inputs generated by a teacher one-step update. Let $\Delta\mathbf{h}_\nu(\boldsymbol{\xi}) = \mathbf{h}\big(\mathbf{x}_{\nu,t'}(\boldsymbol{\xi})\big) - \mathbf{h}\big(\widehat{\mathbf{x}}^{\phi}_{\nu,t}(\boldsymbol{\xi})\big)$ denote the feature increment induced by the teacher one-step update. Under the RFNN parameterization, the output difference of the student model can be written as $\mathbf{f}_\theta\big(\mathbf{x}_{\nu,t'}(\boldsymbol{\xi})\big) - \mathbf{f}_\theta\big(\widehat{\mathbf{x}}^{\phi}_{\nu,t}(\boldsymbol{\xi})\big) = \frac{1}{\sqrt{p}}\,\mathbf{B}_\theta\,\Delta\mathbf{h}_\nu(\boldsymbol{\xi})$. Substituting this expression into the one-step consistency distillation loss (8), we obtain

$$L_{\mathrm{CD}}(\mathbf{B}_\theta) = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}\left[\left\| \frac{1}{\sqrt{p}}\,\mathbf{B}_\theta\,\Delta\mathbf{h}_\nu(\boldsymbol{\xi})\right\|_2^2\right] = \frac{1}{p}\,\mathrm{Tr}\big(\mathbf{B}_\theta\,\mathbf{U}_{\mathrm{cd}}\,\mathbf{B}_\theta^\top\big), \tag{9}$$

where the consistency distillation curvature matrix $\mathbf{U}_{\mathrm{cd}} \in \mathbb{R}^{p\times p}$ is defined as

$$\mathbf{U}_{\mathrm{cd}} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}\big[\Delta\mathbf{h}_\nu(\boldsymbol{\xi})\,\Delta\mathbf{h}_\nu(\boldsymbol{\xi})^\top\big]. \tag{10}$$

Consequently, the spectrum of $\mathbf{U}_{\mathrm{cd}}$ fully characterizes the curvature of the one-step consistency distillation objective.

4.3 SPECTRAL STRUCTURE OF CONSISTENCY DISTILLATION

We define the teacher-induced perturbation by $\boldsymbol{\delta x}(\mathbf{x}, t', t) = \widehat{\mathbf{x}}^{\phi}_t - \mathbf{x}_{t'} = \Delta t\,\mathbf{v}_\phi(\mathbf{x}, t')$. For notational simplicity, we will abbreviate $\boldsymbol{\delta x}(\mathbf{x}, t)$ as $\boldsymbol{\delta x}$ and $\mathbf{x}_{\mathbf{t}'}$ as $\mathbf{x}$. Under the RFNN parameterization, the objective of consistency distillation reduces to a quadratic form whose curvature is determined by the second moment of the nonlinear feature increment $\Delta\mathbf{h} = \mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{x} + \boldsymbol{\delta x})$. A central

difficulty is that $\Delta\mathbf{h}$ is a nonlinear transformation of random features. To make this problem tractable in the high-dimensional regime, we adopt the Gaussian equivalence principle for RFNN George et al. (2025); Bonnaire et al. (2025), which replaces the full feature interaction by an equivalent low-dimensional Gaussian characterization. Based on the assumptions in Appendix D.1, the following lemma characterizes the leading-order contribution to the curvature matrix U_{cd} defined in (10).

Lemma 4.1 (Orthogonal second-moment decomposition of $\Delta\mathbf{h}$)**.** *Under Assumptions D.1 and D.2, define the scalar coefficients* $a_1(\mathbf{x}, \boldsymbol{\delta x}) = \frac{\mathbb{E}_\zeta[\Delta h_i\, \Delta g_i | \mathbf{x}, \boldsymbol{\delta x}]}{\mathbb{E}_\zeta[\Delta g_i^2 | \mathbf{x}, \boldsymbol{\delta x}]}$ *and* $a_0(\mathbf{x}, \boldsymbol{\delta x}) = \frac{\mathbb{E}_\zeta[\Delta h_i^2 | \mathbf{x}, \boldsymbol{\delta x}]}{\mathbb{E}_\zeta[\Delta g_i^2 | \mathbf{x}, \boldsymbol{\delta x}]} - a_1(\mathbf{x}, \boldsymbol{\delta x})^2$*. Then the conditional second moment of the feature increment admits the decomposition*

$$\mathbb{E}_\zeta\left[\Delta\mathbf{h}\Delta\mathbf{h}^\top | \mathbf{x}, \boldsymbol{\delta x}\right] = a_1(\mathbf{x}, \boldsymbol{\delta x})^2 \mathbb{E}_\zeta\left[\Delta\mathbf{g}\Delta\mathbf{g}^\top | \mathbf{x}, \boldsymbol{\delta x}\right] + a_0(\mathbf{x}, \boldsymbol{\delta x}) \mathbb{E}_\zeta\left[\Delta g_i^2 |\, \mathbf{x}, \boldsymbol{\delta x}\right] \mathbf{I}_p, \tag{11}$$

where the conditional expectations $\mathbb{E}_\zeta[\cdot \mid \mathbf{x}, \boldsymbol{\delta x}]$ *are taken with respect to an auxiliary Gaussian variable* ζ *representing the joint Gaussian law of the coordinate pairs* $(g_i, \Delta g_i)$ *induced by a generic row* $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$*, as specified in Assumption D.1.*

Assume further the small-noise one-step regime of Assumption D.3. In the isotropic setting $\boldsymbol{\Sigma} = \mathbf{I}_d$ *and for the one-step OU probability-flow update. let* $\gamma(t')$ *and* $\kappa(t')$ *be the deterministic limits, then* $a_1(\mathbf{x}, \boldsymbol{\delta x})$ *and* $a_0(\mathbf{x}, \boldsymbol{\delta x})$ *concentrate to deterministic limits* $a_1(t')$ *and* $a_0(t')$ *given by Eq. (36) and Eq. (39) in Appendix D.2. See Appendix D.2 for the proof.*

Lemma 4.1 implies that, to leading order, the only source of non-isotropic structure in $\mathbf{U}_{\mathrm{cd}}$ arises from the rank-one term $\Delta\mathbf{g}\,\Delta\mathbf{g}^\top$. All remaining contributions collapse to an isotropic shift. Consequently, the learning geometry induced by consistency distillation is entirely governed by how the teacher-induced perturbation $\boldsymbol{\delta x}$ is embedded into the random feature space through $\Delta\mathbf{g} = \mathbf{W}\boldsymbol{\delta x}/\sqrt{d}$. To make this dependence explicit, we relate the one-step teacher update to the geometry of the random feature space. The following lemma recalls a closed-form characterization of the trained teacher top layer, adapted from Bonnaire et al. (2025).

Lemma 4.2 (Bonnaire et al. (2025))**.** *For RFNN score-matching trained by gradient flow with zero initialization at fixed* t' *on (4), the converged teacher top layer satisfies*

$$\mathbf{A}_\phi = -\frac{\sqrt{p}}{\sqrt{\Delta_{t'}}}\, \mathbf{V}^\top\, \mathbf{U}^{-1}, \tag{12}$$

where $\mathbf{U} = \frac{1}{n}\sum_{\nu=1}^n \mathbb{E}_\xi\left[\sigma\Big(\frac{\mathbf{W}\mathbf{x}_{\nu,t'}(\boldsymbol{\xi})}{\sqrt{d}}\Big)\, \sigma\Big(\frac{\mathbf{W}\mathbf{x}_{\nu,t'}(\boldsymbol{\xi})}{\sqrt{d}}\Big)^\top\right]$, $\mathbf{V} = \frac{1}{n}\sum_{\nu=1}^n \mathbb{E}_\xi\left[\sigma\Big(\frac{\mathbf{W}\mathbf{x}_{\nu,t'}(\boldsymbol{\xi})}{\sqrt{d}}\Big)\boldsymbol{\xi}^\top\right]$*, and* $\Delta_{t'}$ *is the OU forward variance at time* t'*. Assume further that the data distribution* P_x *has zero mean, covariance* $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ *with bounded spectrum and sub-Gaussian tails. In the proportional high-dimensional limit* $n, p, d \to \infty$*,* $\psi_p = \frac{p}{d}$*,* $\psi_n = \frac{n}{d}$*, the Gaussian equivalence results hold:*

1. *(*Lemma C.1 in Bonnaire et al. (2025)*) The empirical spectral distribution of* $\mathbf{U}$ *coincides, in the large-dimensional limit, with that of the Gaussian-equivalent matrix*

$$\mathbf{U} = \frac{1}{n}\mathbf{G}\mathbf{G}^\top + b_{t'}^2 \frac{\mathbf{W}\mathbf{W}^\top}{d} + s_{t'}^2 \mathbf{I}_p, \tag{13}$$

   where $\mathbf{G} = e^{-t'} a_{t'}\, \frac{\mathbf{W}\mathbf{X}'}{\sqrt{d}} + v_{t'}\, \boldsymbol{\Omega}$*. Here* $\mathbf{X}' \in \mathbb{R}^{d\times n}$ *has i.i.d. columns* $\mathbf{x}'_\nu \sim \mathcal{N}(0, \boldsymbol{\Sigma})$*,* $\boldsymbol{\Omega} \in \mathbb{R}^{p\times n}$ *has i.i.d.* $\mathcal{N}(0,1)$ *entries independent of* $(\mathbf{W}, \mathbf{X}')$*, and scalars* $a_{t'}, b_{t'}, v_{t'}, s_{t'}$ *depend on* $(t', \sigma, \boldsymbol{\Sigma})$*.*
2. *(*Lemma C.4 in Bonnaire et al. (2025)*) The cross-covariance matrix* $\mathbf{V}$ *admits the deterministic equivalent*

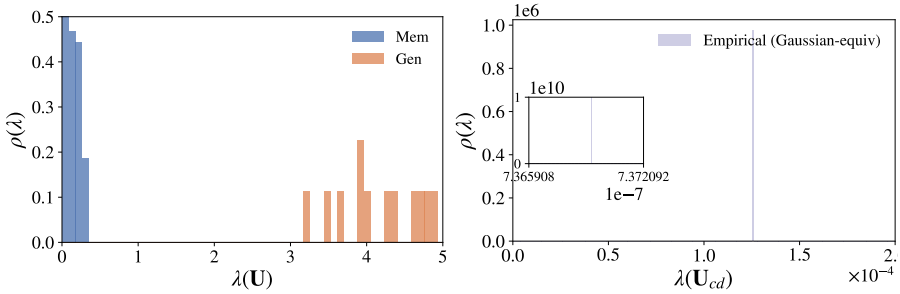
$$\mathbf{V} = \mu_1(t')\frac{\sqrt{\Delta_{t'}}}{\Gamma_{t'}}\frac{\mathbf{W}}{\sqrt{d}}, \tag{14}$$

   where $\mu_1(t') = \mathbb{E}_{u\sim\mathcal{N}(0,1)}\big[\sigma(\Gamma_{t'}u)u\big]$*,* $\Gamma_{t'}^2 = e^{-2t'\frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{d}} + \Delta_{t'}$*. By Gaussian integration by parts (Stein's lemma),* $\mu_1(t') = \Gamma_{t'}\, \mathbb{E}_{Z\sim\mathcal{N}(0,\Gamma_{t'}^2)}[\sigma'(Z)]$*.*

Combining Lemma 4.1 with the teacher characterization in Lemma 4.2, we obtain the following structural characterization of the consistency distillation curvature.

Theorem 4.3. *Let* $\{\mathbf{x}_\nu\}_{\nu=1}^n \subset \mathbb{R}^d$ *be training samples with* $\mathrm{Cov}(\mathbf{x}_\nu) = \boldsymbol{\Sigma} = \mathbf{I}_d$*. At a fixed diffusion time* $t' > 0$*, draw* $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}_d)$ *and define the OU forward sample, and assume the small-noise one-step regime and the deterministic limits* $a_1(t')$*,* $a_0(t')$ *from Lemma 4.1. Then, as* $\Delta t \to 0$*,*

$$\mathbf{U}_{\mathrm{cd}} = \Delta t^2 a_1(t')^2\Big(\mathbf{S} - \mu_1(t')^2\, \mathbf{S}\mathbf{U}^{-1}\mathbf{S}\Big) + \beta(t', \Delta t)\mathbf{I}_p, \tag{15}$$



(a) Spectral density of $\mathbf{U}$. (b) Spectral density of $\mathbf{U}_{\mathrm{cd}}$.

Figure 3: **Spectral density of the teacher and consistency distillation curvature operators. (a)** The teacher curvature operator U exhibits a separated spectrum, with low-eigenvalue modes associated with memorization and high-eigenvalue modes associated with generalization. **(b)** The consistency distillation curvature U_{cd} shows sharp spectral atoms: a dominant spike at $\lambda = \beta$ induced by the isotropic shift acting on the nullspace of S, while the remaining nontrivial eigenvalues are concentrated in a low-dimensional subspace. Both panels are computed with $\psi_p = 32$, $\psi_n = 4$, $t' = 0.01$, $\Delta t = 0.001$, and $\rho_\Sigma(\lambda) = \delta(\lambda - 1)$. See Appendix C.2 for details.

where $\mathbf{S} = \frac{\mathbf{W}\mathbf{W}^\top}{d}$ *and the isotropic shift is* $\beta(t', \Delta t) = a_0(t')\, a_1(t')^2\, \Delta t^2 t'^2\, \frac{1}{d}\, \mathrm{Tr}\big(\mathbf{U}^{-1}\mathbf{S}\big)$. *See Appendix D.3 for the proof.*

Theorem 4.3 provides an explicit decomposition of the curvature induced by one-step consistency distillation. The resulting matrix $\mathbf{U}_{\mathrm{cd}}$ consists of two qualitatively distinct components: an isotropic shift $\beta\mathbf{I}_p$, and a structured, non-isotropic term $\mathbf{A} = \mathbf{S} - \mu_1(t')^2\, \mathbf{S}\mathbf{U}^{-1}\mathbf{S}$. Since $\mathbf{U}$ appears explicitly inside the structured term $\mathbf{A}$ via $\mathbf{U}^{-1}$, the teacher eigen-geometry provides the natural coordinate system for understanding how consistency distillation redistributes curvature. As established in Bonnaire et al. (2025), the empirical spectral density of $\mathbf{U}$ exhibits a characteristic two-bulk structure in the overparameterized regime. This is visible in Fig. 3a, where modes with relatively large eigenvalues $\lambda_i(\mathbf{U})$ align with generalization-dominated directions, whereas small eigenvalues correspond predominantly to memorization-dominated directions. This separation will be inherited by the structured deformation $\mathbf{A} = \mathbf{S} - \mu_1^2\mathbf{S}\mathbf{U}^{-1}\mathbf{S}$ through the dependence on $\mathbf{U}^{-1}$.

We now turn to the empirical spectral density of $\mathbf{U}_{\mathrm{cd}}$ in Fig. 3b. Theorem 4.3 predicts a *sharp spectral spike* induced by the isotropic term, and the origin is purely algebraic: the random-feature Gram operator $\mathbf{S} = \frac{1}{d}\mathbf{W}\mathbf{W}^\top$ has rank at most d, hence $\mathbb{R}^p = \ker(\mathbf{S}) \oplus \mathrm{Im}(\mathbf{S}), \dim(\mathrm{Im}(\mathbf{S})) \le d \ll p$. For $\mathbf{v} \in \ker(\mathbf{S})$, we have $\mathbf{S}\mathbf{v} = 0$ and therefore $\mathbf{A}\mathbf{v} = 0$, implying $\mathbf{U}_{\mathrm{cd}}\mathbf{v} = \beta\mathbf{v}$. Thus, $\mathbf{U}_{\mathrm{cd}}$ has an eigenvalue exactly at $\lambda = \beta$ with multiplicity at least $p - d$, which explains the prominent spike in Fig. 3b. Beyond this atom, all remaining eigenvalues are confined to the low-dimensional subspace $\mathrm{Im}(\mathbf{S})$, where $\mathbf{U}_{\mathrm{cd}}|_{\mathrm{Im}(\mathbf{S})} = \beta\mathbf{I} + \Delta t^2 a_1(t')^2\ \mathbf{A}|_{\mathrm{Im}(\mathbf{S})}\,, \dim(\mathrm{Im}(\mathbf{S})) \le d$. Hence there are at most d non-isotropic eigenvalues beyond the atom at β. Moreover, because the prefactor $\Delta t^2 a_1(t')^2$ is of order Δt^2 and the two terms in $\mathbf{A} = \mathbf{S} - \mu_1^2\mathbf{S}\mathbf{U}^{-1}\mathbf{S}$ can partially cancel within $\mathrm{Im}(\mathbf{S})$, the spectrum of the structured component is typically highly compressed, appearing as a thin spike in Fig. 3b.

4.4 Empirical Validation of Spectral Filtering

To assess whether the non-isotropic consistency distillation term suppresses memorization-associated directions while preserving those relevant for generalization, we analyze its action along the teacher eigenmodes $\{\mathbf{u}_i\}_{i=1}^p$ of the curvature matrix $\mathbf{U}$.

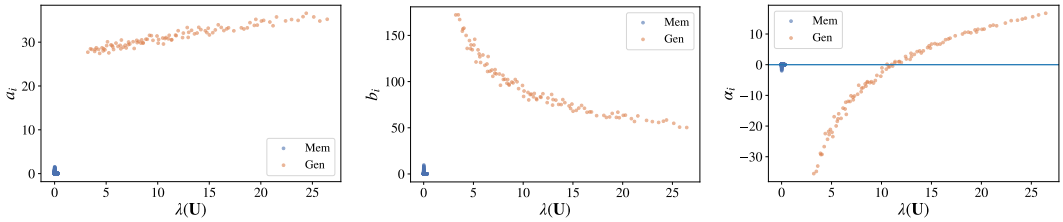
Per-mode decomposition of the non-isotropic consistency distillation response. Recall that the leading non-isotropic component of the consistency distillation curvature takes the form $\mathbf{A} = \mathbf{S} - \mu_1(t')^2\, \mathbf{S}\mathbf{U}^{-1}\mathbf{S}$. For each teacher eigenmode $\mathbf{u}_i$, we introduce the quadratic forms

$$a_i = \mathbf{u}_i^\top \mathbf{S}\, \mathbf{u}_i \ \ge\ 0, \quad b_i = \mathbf{u}_i^\top \mathbf{S}\mathbf{U}^{-1}\mathbf{S}\, \mathbf{u}_i = (\mathbf{S}\mathbf{u}_i)^\top \mathbf{U}^{-1}(\mathbf{S}\mathbf{u}_i) \ \ge\ 0, \tag{16}$$

which respectively measure (i) the *visibility* of mode $\mathbf{u}_i$ under the random-feature metric $\mathbf{S}$, and (ii) the strength of its *resolvent-mediated subtraction* via $\mathbf{U}^{-1}$. The resulting signed response along $\mathbf{u}_i$ is defined as

$$\alpha_i = \mathbf{u}_i^\top \mathbf{A}\, \mathbf{u}_i = a_i - \mu_1(t')^2\, b_i, \tag{17}$$

which quantifies the *net non-isotropic consistency distillation update* assigned to that mode: $\alpha_i < 0$ corresponds to suppression, while $\alpha_i > 0$ indicates net retention.



(a) Visibility a_i. (b) Resolvent term b_i. (c) Signed response α_i.

Figure 4: **Mode-wise spectral effects of non-isotropic consistency distillation.** Each point corresponds to a teacher eigenmode $\mathbf{u}_i$, plotted against its curvature eigenvalue $\lambda_i(\mathbf{U})$, with modes partitioned into memorization-associated (Mem) and generalization-associated (Gen) subspaces. **(a)** The visibility $a_i = \mathbf{u}_i^\top \mathbf{S}\mathbf{u}_i$ is uniformly small for Mem modes and significantly larger for Gen modes. **(b)** The resolvent term $b_i = \mathbf{u}_i^\top \mathbf{S}\mathbf{U}^{-1}\mathbf{S}\mathbf{u}_i$ decreases with $\lambda_i(\mathbf{U})$ within the Gen subspace. **(c)** The resulting response $\alpha_i = a_i - \mu_1(t')^2 b_i$ is negligible for Mem modes, while positive updates concentrate in high-curvature Gen modes. Results use $\psi_p = 32$, $\psi_n = 4$, $t' = 0.01$, and $\rho_\Sigma(\lambda) = \delta(\lambda - 1)$.

Mem/Gen partition. Following the established spectral geometry of $\mathbf{U}$, we classify modes into memorization-associated subspaces (Mem) and generalization-associated subspaces (Gen) according to a fixed threshold λ_{th}, defined as $\mathcal{I}_{\text{mem}} = \{i : \lambda_i(\mathbf{U}) < \lambda_{\text{th}}\}, \mathcal{I}_{\text{gen}} = \{i : \lambda_i(\mathbf{U}) \geq \lambda_{\text{th}}\}$.

(I) Visibility a_i and resolvent structure b_i. Figures 4a and 4b visualize the two constituent terms of α_i. Mem modes have uniformly small a_i, reflecting that these teacher eigenmodes are highly sample-specific and largely orthogonal to the representational subspace spanned by random features. In contrast, Gen modes attain substantially larger a_i, as their smoother, shared structure across samples aligns more strongly with the random-feature span, leading to increasing visibility with $\lambda_i(\mathbf{U})$. As for resolvent term b_i, within the Gen subspace, lower-λ modes incur larger b_i, reflecting stronger subtraction induced by $\mathbf{U}^{-1}$, whereas higher-λ Gen modes are progressively less affected. In contrast, Mem modes exhibit uniformly small b_i, as their weak alignment with the random-feature subspace implies that $\mathbf{S}\mathbf{u}_i$ carries little energy and is therefore minimally amplified by the resolvent.

(II) Net per-mode response α_i. Fig. 4c summarizes the net per-mode response α_i across the spectrum. Mem modes are tightly concentrated near zero, and any isolated Mem modes with $\alpha_i > 0$ remain negligible due to their uniformly small visibility a_i. In contrast, substantial positive responses occur almost exclusively within the Gen subspace and increase with $\lambda_i(\mathbf{U})$. Notably, the non-isotropic term is not uniformly enhancing within Gen: modes near the lower edge of the Gen bulk are often suppressive ($\alpha_i < 0$), whereas the dominant positive contribution is carried by higher-curvature Gen modes with larger $\lambda_i(\mathbf{U})$. Consequently, even if some weaker Gen directions are attenuated, the effective learning signal is governed by the high-eigenvalue Gen spectrum that encodes the most consequential generative structure. A more comprehensive analysis is provided in Appendix E.2.

(III) Global allocation of positive updates. To quantify how positive non-isotropic updates are distributed across subspaces, we aggregate $\max(\alpha_i, 0)$ and define

$$\text{Share}^+_{\text{mem}} = \frac{\sum_{i \in \mathcal{I}_{\text{mem}}} \max(\alpha_i, 0)}{\sum_{i=1}^{p} \max(\alpha_i, 0)}.$$

This metric captures the global allocation of positive curvature injection and is robust to isolated atypical modes. Empirically, we find $\text{Share}^+_{\text{mem}} \approx 9.5 \times 10^{-3}$, indicating that nearly all positive non-isotropic updates are assigned to generalization-associated directions. Thus, at the level of global training dynamics, the non-isotropic component of consistency distillation effectively shifts positive learning signal away from memorization-associated modes.

5 Conclusion and Discussion

In summary, consistency distillation substantially reduces memorization in diffusion models, including cases with strongly overfitted teacher models. Furthermore, sample quality is preserved and can be further improved when the teacher exhibits a moderate level of memorization. Our theoretical analysis shows that consistency distillation actively reshapes training dynamics by suppressing memorization-associated directions while preserving generalization-relevant updates, rather than passively inheriting teacher behavior.

REFERENCES

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.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Sam Buchanan, Druv Pai, Yi Ma, and Valentin De Bortoli. On the edge of memorization in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.

Yiding Chen, Yiyi Zhang, Owen Oertell, and Wen Sun. Convergence of consistency model with multistep sampling under general data assumptions. In *Forty-second International Conference on Machine Learning*, 2025.

Zehao Dou, Minshuo Chen, Mengdi Wang, and Zhuoran Yang. Theory of consistency diffusion models: Distribution estimation meets fast sampling. In *Forty-first International Conference on Machine Learning*, 2024.

Zhengyang Geng, Ashwini Pokle, and J Zico Kolter. One-step diffusion distillation via deep equilibrium models. *Advances in Neural Information Processing Systems*, 2023.

Zhengyang Geng, Ashwini Pokle, Weijian Luo, Justin Lin, and J Zico Kolter. Consistency models made easy. In *The Thirteenth International Conference on Learning Representations*, 2025.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves. *arXiv preprint arXiv:2502.00336*, 2025.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding memorization in generative models via sharpness in probability landscapes. *arXiv preprint arXiv:2412.04140*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 2022.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Naoki Murata, Yuki Mitsufuji, and Stefano Ermon. On the equivalence of consistency-type models: Consistency models, consistent diffusion models, and fokker-planck regularization. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. *Advances in Neural Information Processing Systems*, 2023.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 2023.

Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 2024.

Geon Yeong Park, Sang Wan Lee, and Jong Chul Ye. Inference-time diffusion model distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: Emergence of diffusion models from associative memory networks. In *New Frontiers in Associative Memories*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 2020.

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*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Nikita Starodubcev, Denis Kuznedelev, Artem Babenko, and Dmitry Baranchuk. Scale-wise distillation of diffusion models. *arXiv preprint arXiv:2503.16397*, 2025.

Joshua Tian Jin Tee, Kang Zhang, Hee Suk Yoon, Dhananjaya Nagaraja Gowda, Chanwoo Kim, and Chang D. Yoo. Physics informed distillation for diffusion models. *Transactions on Machine Learning Research*, 2024.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 2011.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. In *Advances in Neural Information Processing Systems*, 2024.

Fu-Yun Wang, Zhengyang Geng, and Hongsheng Li. Stable consistency tuning: Understanding and improving consistency models. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Qianlong Xiang, Miao Zhang, Yuzhang Shang, Jianlong Wu, Yan Yan, and Liqiang Nie. Dkdm: Data-free knowledge distillation for diffusion models with any architecture. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Ruofeng Yang, Bo Jiang, Cheng Chen, and Shuai Li. Improved discretization complexity analysis of consistency models: Variance exploding forward process and decay discretization scheme. In *Forty-second International Conference on Machine Learning*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024b.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 workshop on structured probabilistic inference {\&} generative modeling*, 2023.

Chen Zeno, Hila Manor, Greg Ongie, Nir Weinberger, Tomer Michaeli, and Daniel Soudry. When diffusion models memorize: Inductive biases in probability flow of minimum-norm shallow neural nets. In *Forty-second International Conference on Machine Learning*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

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. *arXiv preprint arXiv:2512.20963*, 2025.

(a) PD, 10%-mem teacher (b) PD, 30%-mem teacher (c) CD, 10%-mem teacher (d) CD, 30%-mem teacher

Figure 5: Qualitative comparison between progressive distillation and consistency distillation under different teacher memorization levels trained on 6000 data points. PD samples exhibit noticeably degraded visual fidelity compared with CD, especially under limited data and higher teacher memorization.

A Related Work

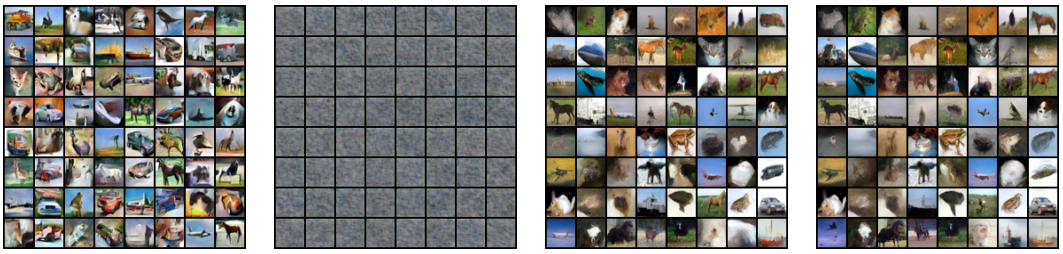
Memorization in Diffusion Models. Recent studies have investigated memorization in diffusion models Carlini et al. (2023); Wen et al. (2024); Zhang et al. (2025). A key motivation is that the denoising score matching objective admits empirical minimizers that reproduce training samples, implying that memorization is theoretically expected in weakly regularized or small-data regimes Gu et al. (2025); Baptista et al. (2025). Subsequent work shows that memorization and generalization undergo sharp transitions as a function of dataset size, model capacity, and training dynamics, including phase-transition and crossover phenomena Buchanan et al. (2025); Zeno et al. (2025); Pham et al. (2025). A precise high-dimensional analysis is given by George et al. George et al. (2025), who derive exact learning curves for diffusion models with random-feature parameterizations. Building on this framework, Bonnaire et al. Bonnaire et al. (2025) show that diffusion training dynamics induce an implicit form of dynamical regularization, creating a growing time window between the onset of generalization and the emergence of memorization in overparameterized regimes. While prior work has primarily focused on diffusion models trained from scratch, the memorization behavior of distilled diffusion models remains comparatively underexplored. In this work, we analyze how the additional training stage introduced by consistency distillation reshapes memorization relative to standard diffusion training and establish a comprehensive theoretical analysis.

Consistency Distillation. Consistency distillation accelerates diffusion model sampling by enforcing self-consistency across diffusion times, enabling efficient few-step generation via distillation from pretrained diffusion models Song et al. (2023); Lai et al. (2023). Subsequent work extends this framework to improve quality–speed trade-offs, stabilize training, and provide theoretical guarantees on estimation, discretization, and convergence Kim et al. (2024); Wang et al. (2024; 2025); Dou et al. (2024); Yang et al. (2025); Chen et al. (2025). However, these studies primarily emphasize efficiency and sample quality, while the impact of consistency distillation on memorization remains largely unexplored, motivating our investigation.

B Rationale for Adopting Consistency Distillation

B.1 Motivation for Consistency Distillation

To motivate our choice of distillation framework, we first consider comparing with progressive distillation (PD) Salimans & Ho (2022), a widely adopted approach for accelerating diffusion sampling. PD iteratively reduces the number of sampling steps by training a student model to match the composition of multiple teacher DDIM steps, and has been shown to be effective in standard, data-rich regimes. In each distillation stage, the student is initialized from the teacher and trained using a deterministic target constructed by composing two teacher DDIM transitions and analytically inverting a single student step. This procedure is repeated while halving the sampling steps, yielding progressively faster samplers.



(a) Continuous CD, 10%-mem teacher (b) Continuous CD, 30%-mem teacher (c) Discrete CD, 10%-mem teacher (d) Discrete CD, 30%-mem teacher

Figure 6: Comparison between continuous-time and discrete consistency distillation under limited data. Continuous-time CD exhibits blurred or failed generations, while the discrete objective remains stable under the same training budget.

In our experiments, PD is implemented following the canonical discrete-time formulation. The student time index is sampled from a fixed discrete grid, and training proceeds by matching one student step to two teacher steps. We train PD models for $600{,}000$ iterations with batch size 64, Adam optimization (weight decay 0), gradient clipping 1.0, and a linearly decayed learning rate. Unless otherwise specified, we employ the perceptual LPIPS loss, which is commonly used to stabilize distillation under aggressive step reduction.

A critical hyperparameter in PD is the *initial noise resolution*, controlled by `start_scales`. Setting `start_scales=4096` corresponds to initializing distillation from a very fine-grained discretization of the diffusion process, i.e., a large number of teacher sampling steps. This choice ensures that the initial teacher accurately resolves high-noise dynamics and provides well-defined multi-step trajectories for the student to imitate. While such a setting is computationally demanding, it represents a favorable configuration for PD and is commonly adopted to avoid compounding discretization errors in early distillation stages.

Empirical fragility of PD under limited data and memorizing teachers. Despite this favorable configuration, PD exhibits limited robustness in the regime we consider. With 6000 training samples, the final PD model distilled from a 10%-memorization teacher attains an FID of 45.42 with a memorization ratio of 0.79%. When distilled from a 30%-memorization teacher, the final PD model attains an FID of 44.05 with a memorization ratio of 3.74%. These results indicate a substantial degradation in sample quality, even though the memorization ratios remain moderate.

This degradation is not merely quantitative. As shown in left two panels of Fig. 5, PD samples under both 10% and 30% memorization teachers exhibit visibly reduced visual fidelity, including blurred structures and weakened object coherence. Notably, this behavior persists despite careful hyperparameter choices and a large initial discretization scale, suggesting that PD is sensitive to the combined effects of limited data, teacher memorization, and aggressive step reduction.

Why we focus on consistency distillation. In contrast, consistency distillation demonstrates markedly stronger robustness in the same regime. Rather than matching composed multi-step trajectories, consistency distillation trains the student to produce self-consistent predictions across neighboring noise levels. This objective avoids explicit inversion of multi-step DDIM transitions and reduces the dependence on long teacher trajectories, which can be particularly fragile when the teacher exhibits memorization or when data is scarce.

Empirically, this robustness translates into a substantially improved quality–memorization tradeoff. With 6000 training samples, CD distilled from a 10%-memorization teacher achieves an FID of 21.19 with a memorization ratio of 0.56%. Under a 30%-memorization teacher, CD (two-step) achieves an FID of 20.60 with a memorization ratio of 3.17%. As illustrated in right two panels of Fig. 5, CD preserves significantly higher visual fidelity than PD under both teacher settings, even when the teacher itself exhibits substantial memorization.

B.2 Motivation for the Discrete Formulation of Consistency Distillation

The original formulation of consistency distillation admits a continuous-time extension, obtained as the infinite-step limit of the discrete objective Song et al. (2023). Under suitable smoothness assumptions on the consistency function, the metric, and the teacher score, the rescaled discrete consistency loss converges to a continuous-time objective defined along the probability flow ODE. In the commonly used stop-gradient setting, this limit yields a *pseudo-objective* whose gradient matches that of the discrete loss as the number of time steps tends to infinity.

Concretely, letting $f_\theta(x_t, t)$ denote the student consistency model and $s_\phi(x_t, t)$ the teacher score, the continuous-time consistency distillation objective takes the form

$$\mathcal{L}_{\mathrm{CD}}^{\mathrm{cont}}(\theta) \;=\; \mathbb{E}_{x\sim p_{\mathrm{data}},\, t}\left[\lambda(t)\left\|\partial_t f_\theta(x_t,t) - t\,\nabla_x f_\theta(x_t,t)\, s_\phi(x_t,t)\right\|^2\right], \tag{18}$$

where $x_t \sim \mathcal{N}(x, t^2 I)$ and $\lambda(t)$ is a bounded weighting function. This objective is minimized if and only if the student model matches the ground-truth consistency function induced by the teacher probability flow.

While Eq. (18) provides a principled characterization of consistency distillation in the continuous-time limit, it introduces nontrivial practical challenges. At finite training resolution, optimizing this objective requires implicitly estimating directional derivatives of the student network along the teacher-induced flow, which involves Jacobian–vector products. Such derivative-based signals are sensitive to model curvature and to noise in the teacher score, and this sensitivity is amplified when the teacher exhibits memorization.

These effects become particularly pronounced in low-data regimes. With limited data, the student observes fewer distinct diffusion trajectories, and the continuous-time objective aggregates derivative information along these trajectories, increasing variance and compounding optimization noise. Empirically, this leads to unstable optimization behavior: under 5000 training samples, the continuous-time consistency objective already produces noticeably blurred generations under a 10%-memorization teacher in Fig. 6a, and fails to generate coherent samples under a 30%-memorization teacher in Fig. 6b.

In contrast, the discrete consistency distillation objective enforces consistency through explicit, finite differences between model predictions at neighboring noise levels. Each update depends only on forward evaluations of the student model at a small number of discrete time points, avoiding the need to estimate derivatives along the probability flow. As a result, the discrete objective is less sensitive to high-curvature, memorization-associated directions in the teacher dynamics. Under identical data scale, architecture, and training budget, the discrete formulation yields stable and visually coherent samples for both 10% and 30%-memorization teachers in Figs. 6c and 6d.

C Detailed Experimental Setup

C.1 Experimental Setup in Section 3

Backbone architecture. All CIFAR-10 experiments use the NCSN++ backbone Song et al. (2021b), following the standard architectural choice in prior consistency distillation work Song et al. (2023). We employ the same backbone for the EDM teachers and the corresponding consistency models, so architectural differences do not confound comparisons.

Consistency model parameterization and boundary condition. We represent a consistency model as

$$f_\theta(x,t) = c_{\mathrm{skip}}(t)\, x + c_{\mathrm{out}}(t)\, F_\theta(x,t), \tag{19}$$

where the coefficients are chosen to satisfy the boundary constraint at the minimum time ε. Using $\sigma_{\mathrm{data}} = 0.5$, we set

$$c_{\mathrm{skip}}(t) = \frac{\sigma_{\mathrm{data}}^2}{(t-\varepsilon)^2 + \sigma_{\mathrm{data}}^2}, \qquad c_{\mathrm{out}}(t) = \frac{\sigma_{\mathrm{data}}(t-\varepsilon)}{\sqrt{\sigma_{\mathrm{data}}^2 + t^2}}, \tag{20}$$

which guarantees $c_{\mathrm{skip}}(\varepsilon) = 1$ and $c_{\mathrm{out}}(\varepsilon) = 0$. This adjustment slightly departs from the original EDM coefficient choice and is needed to ensure the consistency boundary condition for $\varepsilon > 0$.

Teacher models. We distill EDM teachers trained on CIFAR-10 following the training protocol of Karras et al. Karras et al. (2022). To avoid ambiguity in the assessment of memorization, teacher models are trained without data augmentation.

Consistency distillation and optimization. For consistency distillation, the student network is initialized from the corresponding pretrained EDM weights. Training uses Rectified Adam (RAdam) with no warm-up, no learning-rate decay, and no weight decay. We maintain an exponential moving average (EMA) of the online model parameters, consistent with the EDM setup.

Schedules for consistency training. When training consistency models, we use a time-step schedule $N(k)$ and an EMA schedule $\mu(k)$ of the form

$$N(k) = \left\lfloor \frac{c\,k}{K}\left((s_1+1)^2 - s_0^2\right) + s_0^2 \right\rfloor - 1, \tag{21}$$

$$\mu(k) = \exp\left(\frac{s_0 \log \mu_0}{N(k)}\right), \tag{22}$$

where $k \in \{1, \ldots, K\}$ indexes training iterations, K is the total number of iterations, s_0 and s_1 are the starting and ending discretization budgets, and μ_0 is the initial EMA decay factor.

C.2 Experimental Setup in Section 4

C.2.1 Ridge Regularization in Non-Isotropic Consistency Distillation

Under consistency distillation, the local probability-flow ODE induces a non-isotropic response operator of the form

$$\mathbf{A} = \mathbf{S} - \mu_1^2\,\mathbf{S}\,(\mathbf{U} + \gamma\mathbf{I})^{-1}\,\mathbf{S}, \tag{23}$$

where $\mathbf{U} \in \mathbb{R}^{p\times p}$ denotes the teacher curvature matrix, $\mathbf{S} = \frac{1}{d}\mathbf{W}\mathbf{W}^\top$ is the random feature covariance, and $\mu_1 = \mathbb{E}[\sigma'(Z)]$. The inverse operator $(\mathbf{U} + \gamma\mathbf{I})^{-1}$ arises unavoidably from the PF-ODE linearization and governs how teacher curvature directions are transferred to the student.

Empirically and theoretically, the spectrum of the teacher curvature $\mathbf{U}$ is highly ill-conditioned: a large fraction of its eigenvalues concentrate near zero, corresponding to memorization-dominated directions, while a small number of large eigenvalues correspond to generalization-relevant modes. Denoting the eigendecomposition $\mathbf{U} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, the inverse curvature weights each mode by $(\lambda_k + \gamma)^{-1}$. Without ridge regularization ($\gamma = 0$), this induces an extreme amplification of small-eigenvalue directions,

$$(\mathbf{U})^{-1} = \sum_{k=1}^{p} \frac{1}{\lambda_k}\mathbf{v}_k\mathbf{v}_k^\top, \tag{24}$$

causing memorization subspaces to dominate the response even when the input direction itself lies in a generalization-relevant region. This phenomenon is not a benign numerical artifact. In the non-isotropic response energy

$$b_i = (\mathbf{S}\mathbf{u}_i)^\top(\mathbf{U} + \gamma\mathbf{I})^{-1}(\mathbf{S}\mathbf{u}_i), \tag{25}$$

where $\mathbf{u}_i$ is an eigenvector of $\mathbf{U}$, the contribution from memorization directions can overwhelm that from generalization directions purely due to inverse spectral weighting. As a result, response ratios $r_i = \mu_1^2 b_i/(a_i + \varepsilon)$ become large even for modes associated with large $\lambda(\mathbf{U})$, leading to spurious over-subtraction signals.

To make this effect explicit, decompose $\mathbf{S}\mathbf{u}_i = \sum_{k=1}^{p} y_{ki}\mathbf{v}_k$ in the eigenbasis of $\mathbf{U}$. Then

$$b_i = \sum_{k=1}^{p} \frac{y_{ki}^2}{\lambda_k + \gamma}. \tag{26}$$

Let $\mathcal{M}$ denote the memorization subspace, defined by small eigenvalues of $\mathbf{U}$. We define the inverse-weighted memorization leakage as

$$\mathrm{fracBmem}_i = \frac{\sum_{k\in\mathcal{M}} \frac{y_{ki}^2}{\lambda_k+\gamma}}{\sum_{k=1}^{p} \frac{y_{ki}^2}{\lambda_k+\gamma}}. \tag{27}$$

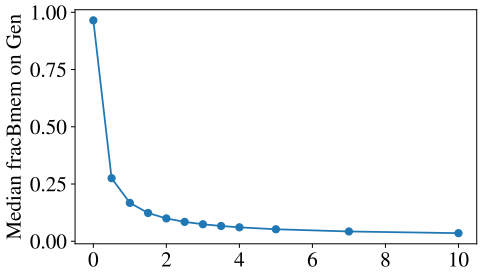


Figure 7: **Gen-to-Mem leakage after inverse-curvature weighting.** Median $\mathrm{fracBmem}$ on Gen (Eq. (27)) versus ridge γ. We choose $\gamma^\star$ by the minimal-sufficient rule in Eq. (29) with tolerance $\tau = 0.1$.

This quantity directly measures how much of the PF-ODE response energy of mode i originates from memorization directions under the inverse curvature metric. Without ridge regularization, $\mathrm{fracBmem}_i \approx 1$ even for generalization modes, indicating severe cross-subspace leakage. This behavior invalidates a naive interpretation of the non-isotropic response as purely suppressive or amplifying.

Introducing a ridge term $\gamma > 0$ modifies the inverse curvature to $(\mathbf{U} + \gamma\mathbf{I})^{-1}$, which has two principled effects:

1. It bounds the maximum amplification of small-eigenvalue directions, preventing memorization modes from dominating the response.
2. It restores a meaningful separation between memorization and generalization subspaces by suppressing inverse-weighted leakage.

Importantly, ridge regularization does not alter the qualitative structure of the PF-ODE operator; it only controls the conditioning of the inverse curvature, which is unavoidable in consistency distillation.

Rather than selecting γ to enforce a particular sign of the response, we adopt a geometrically meaningful criterion:

$$\mathrm{median}_{i\in\mathrm{GEN}}\big[\mathrm{fracBmem}_i\big] \;\le\; \tau, \tag{28}$$

where GEN denotes the generalization subspace and τ is a small constant (e.g., 0.1 or 0.2). This criterion ensures that, for generalization modes, the PF-ODE response is not dominated by memorization directions.

Fig. 7 reports the ridge sweep of the proposed leakage statistic $\mathrm{fracBmem}$ (Eq. (27)) on the generalization subspace: $\gamma \mapsto \mathrm{median}_{i\in\mathrm{GEN}}\big[\mathrm{fracBmem}_i(\gamma)\big]$. As γ increases, the inverse-curvature weighting $(\lambda + \gamma)^{-1}$ caps the amplification of small-eigenvalue directions, yielding a monotone reduction of memorization leakage into GEN after applying $(\mathbf{U} + \gamma\mathbf{I})^{-1}$.

To make the ridge choice reproducible, we select the *minimal sufficient* regularization level $\gamma^\star$ that enforces a target leakage tolerance τ:

$$\gamma^\star = \min\Big\{\gamma > 0 :\; \mathrm{median}_{i\in\mathrm{GEN}}\big[\mathrm{fracBmem}_i(\gamma)\big] \le \tau\Big\}. \tag{29}$$

In our experiments, setting $\tau = 0.1$ yields $\gamma^\star \approx 2.0$, since the GEN median leakage drops from 0.123 at $\gamma = 1.5$ to 0.0997 at $\gamma = 2.0$, while larger ridge values produce diminishing returns in leakage reduction. This choice controls cross-subspace contamination induced by the inverse curvature metric, without tuning γ to force a particular sign pattern of the response.

Fig. 7 further shows that the unregularized operator ($\gamma = 0$) yields near-total GEN-to-MEM leakage after $(\mathbf{U} + \gamma\mathbf{I})^{-1}$. Increasing γ rapidly suppresses this effect; beyond $\gamma \approx 2$, the leakage curve flattens, indicating that $\gamma^\star$ captures the main conditioning benefit while avoiding excessive attenuation of the overall non-isotropic structure.

C.2.2 NUMERICAL SETUP FOR COMPUTING THE SPECTRUM OF $\mathbf{U}_{\mathrm{cd}}$

We compute the empirical spectrum of the one-step consistency distillation curvature operator under the Gaussian-equivalent RFNN model described in Section 4. All reported metrics are evaluated for the full one-step curvature $\mathbf{U}_{\mathrm{cd}}$, which admits the small-step approximation

$$\mathbf{U}_{\mathrm{cd}} = \Delta t^2\, a_1(t')^2 \Big(\mathbf{S} - \mu_1(t')^2\, \mathbf{S}\mathbf{U}^{-1}\mathbf{S}\Big) + \beta(t', \Delta t)\, \mathbf{I}.$$

All experiments are conducted with ambient dimension fixed to $d = 100$. The number of random features and effective training samples scale linearly with d as $p = \psi_p d$ and $n = \psi_n d$, where we use $\psi_p = 32$ and $\psi_n = 4$ throughout, corresponding to $p = 3200$ and $n = 400$. Curvature metrics are evaluated at a fixed diffusion time $t' = 10^{-2}$ under the OU forward process, and the consistency distillation update is approximated using a single Euler step with step size $\Delta t = 10^{-3}$. The RFNN uses the $\tanh(\cdot)$ activation function, and the data covariance is assumed isotropic, $\Sigma = \mathbf{I}_d$, corresponding to $\rho_\Sigma(\lambda) = \delta(\lambda - 1)$.

Teacher-dependent constants (a_t, b_t, v_t, s_t^2) are estimated via Monte Carlo sampling under the OU forward process using 2×10^5 samples. The Gaussian-equivalent curvature matrices are then constructed as

$$\mathbf{S} = \frac{1}{d}\mathbf{W}\mathbf{W}^\top, \qquad \mathbf{U} = \frac{1}{n}\mathbf{G}\mathbf{G}^\top + b_t^2\mathbf{S} + s_t^2\mathbf{I},$$

where $\mathbf{W} \in \mathbb{R}^{p\times d}$ and the auxiliary random matrices used to form $\mathbf{G}$ have i.i.d. standard Gaussian entries. To avoid trace-based closure approximations, PF-ODE constants are estimated directly by sampling $x \sim p_{t'}$ and computing $\eta = \mathbb{E}[x^\top s_\phi(x)]/d$ and $\upsilon = \mathbb{E}[\|s_\phi(x)\|_2^2]/d$ using 5×10^4 Monte Carlo samples. These estimates define $\gamma = 1 + \eta$ and $\kappa^2 = 1 + 2\eta + \upsilon$, which are used to compute the closed-form coefficients $a_1(t')$ and $a_0(t')$ appearing in $\mathbf{U}_{\mathrm{cd}}$.

The eigenvalue spectrum of $\mathbf{U}_{\mathrm{cd}}$ is computed via dense eigendecomposition. To isolate the continuous spectral component, eigenvalues below $\varepsilon_{\mathrm{atom}} = 10^{-50}$ are discarded, and all reported histograms and summary statistics are computed over the remaining non-zero spectral support.

C.2.3 NUMERICAL SETUP FOR RFNN-BASED NON-ISOTROPIC CD DIAGNOSTICS

We describe the numerical setup used to evaluate RFNN-based diagnostics for the non-isotropic one-step consistency distillation operator.

All experiments are conducted under isotropic data covariance $\rho_\Sigma(\lambda) = \delta(\lambda - 1)$ with input dimension fixed to $d = 100$. The random feature and sample dimensions scale linearly with d as $p = \psi_p d$ and $n = \psi_n d$, where we use $\psi_p = 32$ and $\psi_n = 4$ throughout, corresponding to $p = 3200$ and $n = 400$. We consider the OU forward process at a fixed diffusion time $t' = 0.01$ and use a single Euler step of size $\Delta t = 10^{-3}$ in all reported experiments.

Random features are constructed by sampling $\mathbf{W} \in \mathbb{R}^{p\times d}$ with i.i.d. $\mathcal{N}(0, 1)$ entries and setting $\mathbf{B} = \mathbf{W}/\sqrt{d}$, yielding the metric $\mathbf{S} = \mathbf{B}\mathbf{B}^\top$. Teacher-dependent constants (a_t, b_t, v_t, s_t^2) and $\mu_1 = \mathbb{E}[\sigma'(Z)]$ for $\sigma = \tanh$ are estimated via Monte Carlo sampling using 5×10^5 samples. The Gaussian-equivalent teacher curvature is constructed as

$$\mathbf{U} = \frac{1}{n}\mathbf{G}\mathbf{G}^\top + b_t^2\mathbf{S} + s_t^2\mathbf{I}_p, \qquad \mathbf{G} = e^{-t'}\, a_t\, (\mathbf{B}\mathbf{X}') + v_t\, \mathbf{\Omega},$$

where $\mathbf{X}' \in \mathbb{R}^{d\times n}$ and $\mathbf{\Omega} \in \mathbb{R}^{p\times n}$ have i.i.d. standard Gaussian entries. The non-isotropic channel operator is defined as

$$\mathbf{A} = \mathbf{S} - \mu_1^2\, \mathbf{S}(\mathbf{U} + \gamma\mathbf{I})^{-1}\mathbf{S},$$

with ridge parameter $\gamma = 2$ used throughout to ensure numerical stability.

Diagnostics are evaluated along the eigenmodes $\{\mathbf{u}_i\}_{i=1}^p$ of $\mathbf{U}$. Memorization- and generalization-associated modes are separated using a fixed spectral threshold $\lambda_{\mathrm{th}} = 2$, with $\lambda_i(\mathbf{U}) < \lambda_{\mathrm{th}}$ classified as Mem and $\lambda_i(\mathbf{U}) \geq \lambda_{\mathrm{th}}$ as Gen. For each mode we compute

$$a_i = \mathbf{u}_i^\top \mathbf{S}\mathbf{u}_i, \quad b_i = \mathbf{u}_i^\top \mathbf{S}(\mathbf{U} + \gamma\mathbf{I})^{-1}\mathbf{S}\mathbf{u}_i, \quad \alpha_i = a_i - \mu_1^2 b_i.$$

D PROOFS

D.1 ASSUMPTIONS

Assumption D.1 (Gaussian-equivalent random features)**.** $\mathbf{W} \in \mathbb{R}^{p\times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries and define $g(\mathbf{x}) = \frac{\mathbf{W}\mathbf{x}}{\sqrt{d}}$, $h(\mathbf{x}) = \sigma(g(\mathbf{x}))$. For a given perturbation $\boldsymbol{\delta x} \in \mathbb{R}^d$, set $\Delta g = g(\mathbf{x} + \boldsymbol{\delta x}) - g(\mathbf{x}) = \frac{\mathbf{W}\boldsymbol{\delta x}}{\sqrt{d}}$, $\Delta h = h(\mathbf{x}) - h(\mathbf{x} + \boldsymbol{\delta x})$ and $\mathbf{w}_i^\top$ denote the i-th row of $\mathbf{W}$. Conditionally on $(\mathbf{x}, \boldsymbol{\delta x})$ and with respect to the randomness of $\mathbf{w}_i \sim \mathcal{N}(0, I_d)$, the coordinate pair $(g_i, \Delta g_i) = \left(\frac{\mathbf{w}_i^\top \mathbf{x}}{\sqrt{d}}, \frac{\mathbf{w}_i^\top \boldsymbol{\delta x}}{\sqrt{d}}\right)$ is centered jointly Gaussian with $\Gamma_d^2 = \mathrm{Var}(g_i \mid \mathbf{x}) = \frac{\|\mathbf{x}\|_2^2}{d}$, $\Delta_d^2 = \mathrm{Var}(\Delta g_i \mid \mathbf{x}, \boldsymbol{\delta x}) = \frac{\|\boldsymbol{\delta x}\|_2^2}{d}$, $c_d = \mathrm{Cov}(g_i, \Delta g_i \mid \mathbf{x}, \boldsymbol{\delta x}) = \frac{\mathbf{x}^\top \boldsymbol{\delta x}}{d}$. Moreover, the pairs $\{(g_i, \Delta g_i)\}_{i=1}^p$ are i.i.d. across i.

Assumption D.2 (Activation moment and smoothness conditions)**.** The activation $\sigma : \mathbb{R} \to \mathbb{R}$ is measurable and satisfies $\mathbb{E}_\zeta\left[(\sigma(g_i) - \sigma(g_i + \Delta g_i))^2 \mid \mathbf{x}, \boldsymbol{\delta x}\right] < \infty$. In addition, σ is almost everywhere differentiable and $\mathbb{E}_{G\sim\mathcal{N}(0,1)}\left[\sigma'(G)^2\right] < \infty$. For sharper control of higher-order terms, we assume $\sigma \in C^2$ with $\mathbb{E}_{G\sim\mathcal{N}(0,1)}\left[\sigma''(G)^2\right] < \infty$.

Assumption D.3 (Small-noise one-step regime)**.** As $d \to \infty$ and $\Delta t \to 0$, the perturbation satisfies $\frac{\|\boldsymbol{\delta x}\|_2^2}{d} \longrightarrow 0$, $\Gamma_d^2 = \frac{\|\mathbf{x}\|_2^2}{d} \longrightarrow \Gamma^2 \in (0, \infty)$, so that $\Delta_d^2 = \mathrm{Var}(\Delta g_i) = \frac{\|\boldsymbol{\delta x}\|_2^2}{d} \longrightarrow 0$. In the isotropic setting $\boldsymbol{\Sigma} = I_d$ and for the one-step OU probability flow update, we further assume the existence of deterministic limits $\eta(t') = \lim_{d\to\infty} \frac{\mathbf{x}^\top \mathbf{s}_\phi(\mathbf{x}, t')}{d}$ and $\upsilon(t') = \lim_{d\to\infty} \frac{\|\mathbf{s}_\phi(\mathbf{x}, t')\|_2^2}{d}$ holding in probability. Consequently, we have

$$\Gamma_d^2 \to 1, c_d = \frac{\mathbf{x}^\top \boldsymbol{\delta x}}{d} \to -\Delta t\, \gamma(t'), \Delta_d^2 \to \Delta t^2\, \kappa(t')^2,$$

where $\gamma(t') = 1 + \eta(t'), \kappa(t')^2 = 1 + 2\eta(t') + \upsilon(t')$.

D.2 PROOF OF LEMMA 4.1

Proof. Fix $(\mathbf{x}, \boldsymbol{\delta x}) \in \mathbb{R}^d \times \mathbb{R}^d$. Let $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$ and define the single-coordinate Gaussian pair

$$G = \frac{\mathbf{w}^\top \mathbf{x}}{\sqrt{d}}, \qquad \Delta G = \frac{\mathbf{w}^\top \boldsymbol{\delta x}}{\sqrt{d}}.$$

By Assumption D.1, conditionally on $(\mathbf{x}, \boldsymbol{\delta x})$, $(G, \Delta G)$ is centered jointly Gaussian with

$$\mathrm{Var}(G) = \Gamma_d^2, \qquad \mathrm{Var}(\Delta G) = \Delta_d^2, \qquad \mathrm{Cov}(G, \Delta G) = c_d.$$

Let $Y = \sigma(G) - \sigma(G + \Delta G)$ and $X = \Delta G$. By Assumption D.2, $Y \in L^2$ and $X \in L^2$. Assume $\Delta_d^2 > 0$ so that $\mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}] = \Delta_d^2 > 0$.

Define the projection coefficient

$$a_1(\mathbf{x}, \boldsymbol{\delta x}) = \frac{\mathbb{E}[YX \mid \mathbf{x}, \boldsymbol{\delta x}]}{\mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}]}, \qquad R = Y - a_1(\mathbf{x}, \boldsymbol{\delta x})\, X.$$

Then $a_1(\mathbf{x}, \boldsymbol{\delta x})$ is the L^2-projection coefficient of Y onto $\mathrm{span}\{X\}$, hence

$$\mathbb{E}[RX \mid \mathbf{x}, \boldsymbol{\delta x}] = 0. \tag{30}$$

Moreover,

$$\begin{aligned}
\mathbb{E}[R^2 \mid \mathbf{x}, \boldsymbol{\delta x}] &= \mathbb{E}\left[(Y - a_1 X)^2 \mid \mathbf{x}, \boldsymbol{\delta x}\right] \\
&= \mathbb{E}[Y^2 \mid \mathbf{x}, \boldsymbol{\delta x}] - a_1(\mathbf{x}, \boldsymbol{\delta x})^2\, \mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}] \\
&= a_0(\mathbf{x}, \boldsymbol{\delta x})\, \mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}],
\end{aligned} \tag{31}$$

where $a_0(\mathbf{x}, \boldsymbol{\delta x})$ is defined in Lemma 4.1. Now lift from a single coordinate to the full vector. For each $i \in \{1, \ldots, p\}$, let $(g_i, \Delta g_i)$ be i.i.d. copies of $(G, \Delta G)$ under the conditional law induced by $\mathbf{w}_i \sim \mathcal{N}(0, \mathbf{I}_d)$, and set

$$\Delta h_i = \sigma(g_i) - \sigma(g_i + \Delta g_i), \qquad R_i = \Delta h_i - a_1(\mathbf{x}, \boldsymbol{\delta x})\, \Delta g_i.$$

Writing $\Delta\mathbf{h} = a_1\Delta\mathbf{g} + \mathbf{R}$ and expanding second moments gives

$$\mathbb{E}[\Delta\mathbf{h}\Delta\mathbf{h}^\top|\,\mathbf{x},\boldsymbol{\delta x}] = a_1^2\,\mathbb{E}[\Delta\mathbf{g}\Delta\mathbf{g}^\top|\,\mathbf{x},\boldsymbol{\delta x}] + a_1\,\mathbb{E}[\Delta\mathbf{g}\mathbf{R}^\top|\,\mathbf{x},\boldsymbol{\delta x}] + a_1\,\mathbb{E}[\mathbf{R}\Delta\mathbf{g}^\top|\,\mathbf{x},\boldsymbol{\delta x}] + \mathbb{E}[\mathbf{R}\mathbf{R}^\top|\,\mathbf{x},\boldsymbol{\delta x}].$$

By (30) applied coordinate-wise and independence across i, the cross terms vanish: $\mathbb{E}[\Delta\mathbf{g}\mathbf{R}^\top \mid \mathbf{x},\boldsymbol{\delta x}] = 0$ and $\mathbb{E}[\mathbf{R}\Delta\mathbf{g}^\top \mid \mathbf{x},\boldsymbol{\delta x}] = 0$. Furthermore, since the coordinates are i.i.d. and R_i has conditional second moment $\mathbb{E}[R_i^2 \mid \mathbf{x},\boldsymbol{\delta x}] = \mathbb{E}[R^2 \mid \mathbf{x},\boldsymbol{\delta x}]$, we have $\mathbb{E}[\mathbf{R}\mathbf{R}^\top \mid \mathbf{x},\boldsymbol{\delta x}] = \mathbb{E}[R^2 \mid \mathbf{x},\boldsymbol{\delta x}]\,\mathbf{I}_p$. Using (31) and $\mathbb{E}[X^2\,|\mathbf{x},\boldsymbol{\delta x}] = \mathbb{E}[\Delta g_i^2 \mid \mathbf{x},\boldsymbol{\delta x}]$ yields the exact decomposition (11).

Assume now the regime of Assumption D.3 with $\boldsymbol{\Sigma} = \mathbf{I}_d$ and the one-step PF-ODE update (5). Let $G_d = G/\Gamma_d$ so that $G_d \sim \mathcal{N}(0,1)$. Define $Z = \frac{\Delta G}{\Delta t}$, we have $\Delta G = \Delta t Z$. By Assumption D.3, the joint Gaussian parameters satisfy

$$\Gamma_d^2 \to 1, \qquad \mathrm{Var}(Z) = \frac{\Delta_d^2}{\Delta t^2} \to \kappa(t')^2, \qquad \mathrm{Cov}(G,Z) = \frac{c_d}{\Delta t} \to \gamma(t').$$

Hence (G,Z) converges in distribution to a centered jointly Gaussian pair with

$$G \sim \mathcal{N}(0,1), \qquad \mathbb{E}[Z^2] = \kappa(t')^2, \qquad \mathbb{E}[GZ] = \gamma(t').$$

We next compute the leading-order limits of $a_1(\mathbf{x},\boldsymbol{\delta x})$ and $a_0(\mathbf{x},\boldsymbol{\delta x})$. Write $X = \Delta G$ and $Y = \sigma(G) - \sigma(G+\Delta G)$ as above. Using the mean-value form of Taylor's theorem, for each realization there exists $\theta \in (0,1)$ such that

$$\sigma(G+\Delta G) = \sigma(G) + \sigma'(G)\Delta G + \frac{1}{2}\,\sigma''(G+\theta\Delta G)\,(\Delta G)^2.$$

Thus

$$Y = -\sigma'(G)\Delta G - \rho, \qquad \rho = \frac{1}{2}\,\sigma''(G+\theta\Delta G)\,(\Delta G)^2. \tag{32}$$

Under the smoothness conditions in Assumption D.2 and since $\mathbb{E}[(\Delta G)^4] = O(\Delta_d^4) = O(\Delta t^4)$, we have $\mathbb{E}[\rho^2] = O(\Delta t^4)$ and hence $\mathbb{E}|\rho\,\Delta G| = o(\Delta t^2)$.

Limit of a_1. Using (32) and $X = \Delta G$,

$$\begin{aligned}\mathbb{E}[YX \mid \mathbf{x},\boldsymbol{\delta x}] &= \mathbb{E}\big[Y\Delta G \mid \mathbf{x},\boldsymbol{\delta x}\big] \\ &= -\mathbb{E}\big[\sigma'(G)(\Delta G)^2 \mid \mathbf{x},\boldsymbol{\delta x}\big] - \mathbb{E}\big[\rho\,\Delta G \mid \mathbf{x},\boldsymbol{\delta x}\big] \\ &= -\Delta t^2\,\mathbb{E}\big[\sigma'(G)Z^2 \mid \mathbf{x},\boldsymbol{\delta x}\big] + o(\Delta t^2),\end{aligned} \tag{33}$$

while

$$\mathbb{E}[X^2 \mid \mathbf{x},\boldsymbol{\delta x}] = \mathbb{E}[(\Delta G)^2 \mid \mathbf{x},\boldsymbol{\delta x}] = \Delta t^2\,\mathbb{E}[Z^2 \mid \mathbf{x},\boldsymbol{\delta x}] = \Delta t^2\,\kappa(t')^2 + o(\Delta t^2).$$

Therefore,

$$a_1(\mathbf{x},\boldsymbol{\delta x}) = \frac{\mathbb{E}[YX \mid \mathbf{x},\boldsymbol{\delta x}]}{\mathbb{E}[X^2 \mid \mathbf{x},\boldsymbol{\delta x}]} = -\frac{\mathbb{E}[\sigma'(G)Z^2]}{\kappa(t')^2} + o(1), \tag{34}$$

where expectations on the right-hand side are taken under the limiting joint Gaussian law of (G,Z).

Since (G,Z) is jointly Gaussian with $\mathbb{E}[GZ] = \gamma(t')$ and $\mathbb{E}[Z^2] = \kappa(t')^2$, we have

$$\mathbb{E}[Z^2 \mid G] = \big(\mathbb{E}[Z \mid G]\big)^2 + \mathrm{Var}(Z \mid G) = \gamma(t')^2\,G^2 + \big(\kappa(t')^2 - \gamma(t')^2\big).$$

Hence

$$\mathbb{E}[\sigma'(G)Z^2] = \mathbb{E}\big[\sigma'(G)\,\mathbb{E}[Z^2 \mid G]\big] = \gamma(t')^2\,\mathbb{E}\big[\sigma'(G)G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big)\,\mathbb{E}\big[\sigma'(G)\big], \tag{35}$$

with $G \sim \mathcal{N}(0,1)$. Combining (34) and (35) yields

$$a_1(t') = -\,\frac{\gamma(t')^2\mathbb{E}_G\big[\sigma'(G)\,G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big)\mathbb{E}_G\big[\sigma'(G)\big]}{\kappa(t')^2}. \tag{36}$$

Limit of a_0. Similarly, using (32),

$$\begin{aligned}\mathbb{E}[Y^2 \mid \mathbf{x},\boldsymbol{\delta x}] &= \mathbb{E}\big[\sigma'(G)^2(\Delta G)^2 \mid \mathbf{x},\boldsymbol{\delta x}\big] + 2\,\mathbb{E}\big[\sigma'(G)\Delta G\,\rho \mid \mathbf{x},\boldsymbol{\delta x}\big] + \mathbb{E}[\rho^2 \mid \mathbf{x},\boldsymbol{\delta x}] \\ &= \Delta t^2\,\mathbb{E}\big[\sigma'(G)^2Z^2 \mid \mathbf{x},\boldsymbol{\delta x}\big] + o(\Delta t^2),\end{aligned} \tag{37}$$

where the $o(\Delta t^2)$ term follows from Cauchy–Schwarz together with $\mathbb{E}[\rho^2] = O(\Delta t^4)$. Dividing by $\mathbb{E}[X^2] = \Delta t^2 \kappa(t')^2 + o(\Delta t^2)$ gives

$$\frac{\mathbb{E}[Y^2]}{\mathbb{E}[X^2]} = \frac{\mathbb{E}[\sigma'(G)^2 Z^2]}{\kappa(t')^2} + o(1).$$

Using again $\mathbb{E}[Z^2 \mid G] = \gamma(t')^2 G^2 + (\kappa(t')^2 - \gamma(t')^2)$ yields

$$\mathbb{E}[\sigma'(G)^2 Z^2] = \gamma(t')^2\, \mathbb{E}\big[\sigma'(G)^2 G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big)\, \mathbb{E}\big[\sigma'(G)^2\big], \tag{38}$$

with $G \sim \mathcal{N}(0,1)$. Combining these expressions with the definition $a_0 = \frac{\mathbb{E}[Y^2]}{\mathbb{E}[X^2]} - a_1^2$, we have

$$a_0(t') = \frac{\gamma(t')^2 \mathbb{E}_G\big[\sigma'(G)^2 G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big)\mathbb{E}_G\big[\sigma'(G)^2\big]}{\kappa(t')^2} - a_1(t')^2. \tag{39}$$

This completes the proof. □

D.3 Proof of Theorem 4.3

Proof. Lemma 4.1 gives, for each $(\mathbf{x}, \delta\mathbf{x})$,

$$\mathbb{E}_\zeta\big[\Delta\mathbf{h}\,\Delta\mathbf{h}^\top \mid \mathbf{x}, \delta\mathbf{x}\big] = a_1(\mathbf{x},\delta\mathbf{x})^2\, \mathbb{E}_\zeta\big[\Delta\mathbf{g}\,\Delta\mathbf{g}^\top \mid \mathbf{x}, \delta\mathbf{x}\big] + a_0(\mathbf{x},\delta\mathbf{x})\, \mathbb{E}_\zeta\big[\Delta g_i^2 \mid \mathbf{x}, \delta\mathbf{x}\big]\ \mathbf{I}_p,$$

with $\mathbb{E}_\zeta[\Delta g_i^2 \mid \mathbf{x}, \delta\mathbf{x}] = \|\delta\mathbf{x}\|_2^2/d$. Taking $\mathbb{E}_\xi$ and then averaging over ν yields

$$\mathbf{U}_{\mathrm{cd}} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\big[a_1(\mathbf{x},\delta\mathbf{x})^2\, \Delta\mathbf{g}\,\Delta\mathbf{g}^\top\big] + \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[a_0(\mathbf{x},\delta\mathbf{x})\, \frac{\|\delta\mathbf{x}\|_2^2}{d}\right] \mathbf{I}_p. \tag{40}$$

Under the small-noise one-step regime and the concentration hypotheses of Lemma 4.1, we may replace $a_1(\mathbf{x},\delta\mathbf{x}) \to a_1(t')$ and $a_0(\mathbf{x},\delta\mathbf{x}) \to a_0(t')$ in (40) at leading order, obtaining

$$\mathbf{U}_{\mathrm{cd}} = a_1(t')^2\, \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\big[\Delta\mathbf{g}\,\Delta\mathbf{g}^\top\big] + \beta(t',\Delta t)\, \mathbf{I}_p + o(\Delta t^2), \tag{41}$$

where $\beta(t',\Delta t) = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[\frac{\|\delta\mathbf{x}\|_2^2}{d}\right]$.

Since $\Delta\mathbf{g} = \mathbf{W}\,\delta\mathbf{x}/\sqrt{d}$, then we have

$$\Delta\mathbf{g}\,\Delta\mathbf{g}^\top = \frac{1}{d}\,\mathbf{W}\,\delta\mathbf{x}\,\delta\mathbf{x}^\top\mathbf{W}^\top, \quad \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi[\Delta\mathbf{g}\,\Delta\mathbf{g}^\top] = \frac{1}{d}\,\mathbf{W}\Big(\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi[\delta\mathbf{x}\,\delta\mathbf{x}^\top]\Big)\mathbf{W}^\top.$$

Using $\delta\mathbf{x} = -\Delta t(\mathbf{x} + \mathbf{s}_\phi(\mathbf{x},t'))$, we obtain

$$\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi[\delta\mathbf{x}\,\delta\mathbf{x}^\top] = \Delta t^2\, \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\big[(\mathbf{x}+\mathbf{s}_\phi)(\mathbf{x}+\mathbf{s}_\phi)^\top\big]. \tag{42}$$

From (12) and the deterministic equivalent (14), the converged teacher score can be expressed as

$$\mathbf{s}_\phi(\mathbf{x},t') = -\mu_1(t')\, \frac{\mathbf{W}^\top}{\sqrt{d}}\, \mathbf{U}^{-1}\, \mathbf{h}(\mathbf{x}), \tag{43}$$

where $\mathbf{U} = (1/n)\sum_\nu \mathbb{E}_\xi[\mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x})^\top]$. Let $\mathbf{B} = \mathbf{W}/\sqrt{d}$ so that $\mathbf{S} = \mathbf{B}\mathbf{B}^\top$. Under Gaussian equivalence and $\Sigma = I_d$), the OU marginal of $\mathbf{x}$ is asymptotically isotropic, and Gaussian integration by parts yields the identity

$$\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\big[\mathbf{x}\,\mathbf{h}(\mathbf{x})^\top\big] \simeq \mu_1(t')\,\mathbf{B}^\top, \tag{44}$$

while by definition of $\mathbf{U} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[\mathbf{h}(\mathbf{x})\,\mathbf{h}(\mathbf{x})^\top\right]$, plugging (43) into the cross and score terms and using (44) gives

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{x}\mathbf{s}_\phi^\top] \simeq -\mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}, \quad \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{s}_\phi\mathbf{x}^\top] \simeq -\mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B},$$
$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{s}_\phi\mathbf{s}_\phi^\top] = \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\Big(\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{h}\mathbf{h}^\top]\Big)\mathbf{U}^{-1}\mathbf{B} = \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}. \tag{45}$$

Moreover, the isotropic OU marginal implies

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{x}\mathbf{x}^\top] \simeq \mathbf{I}_d. \tag{46}$$

Combining (45) and (46) yields:

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[(\mathbf{x}+\mathbf{s}_\phi)(\mathbf{x}+\mathbf{s}_\phi)^\top\right] \simeq \mathbf{I}_d - \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}. \tag{47}$$

Substituting (47) into (42) gives

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\delta\mathbf{x}\,\delta\mathbf{x}^\top] \simeq \Delta t^2\Big(\mathbf{I}_d - \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}\Big).$$

Therefore,

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\Delta\mathbf{g}\,\Delta\mathbf{g}^\top] \simeq \Delta t^2\,\frac{1}{d}\,\mathbf{W}\Big(\mathbf{I}_d - \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}\Big)\mathbf{W}^\top = \Delta t^2\Big(\mathbf{S} - \mu_1(t')^2\,\mathbf{S}\,\mathbf{U}^{-1}\,\mathbf{S}\Big).$$

Now we begin to analyze the second term in (55). Using $\delta\mathbf{x} = \Delta t\,t'\,\mathbf{s}_\phi(\mathbf{x},t')$ and $\mathbf{s}_\phi(\mathbf{x},t') = (1/\sqrt{p})\mathbf{A}_\phi\mathbf{h}(\mathbf{x})$, we obtain

$$\frac{\|\delta\mathbf{x}\|_2^2}{d} = \Delta t^2 t'^2\,\frac{1}{d}\frac{1}{p}\,\mathbf{h}(\mathbf{x})^\top\mathbf{A}_\phi^\top\mathbf{A}_\phi\,\mathbf{h}(\mathbf{x}). \tag{48}$$

Averaging over ξ and ν and using $\mathbf{U} = \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi[\mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x})^\top]$ yields

$$\begin{aligned}\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\left[\frac{\|\delta\mathbf{x}\|_2^2}{d}\right] &= \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\left[\frac{1}{d}\left\|\Delta t\,t'\,\frac{1}{\sqrt{p}}\mathbf{A}_\phi\mathbf{h}(\mathbf{x})\right\|_2^2\right] \\ &= \Delta t^2 t'^2\,\frac{1}{dp}\,\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\left[\mathbf{h}(\mathbf{x})^\top\mathbf{A}_\phi^\top\mathbf{A}_\phi\,\mathbf{h}(\mathbf{x})\right].\end{aligned} \tag{49}$$

Using the standard quadratic-form identity $\mathbb{E}\left[\mathbf{h}^\top\mathbf{M}\mathbf{h}\right] = \mathrm{Tr}\left(\mathbf{M}\,\mathbb{E}[\mathbf{h}\mathbf{h}^\top]\right), \mathbf{M}\in\mathbb{R}^{p\times p}$, we obtain

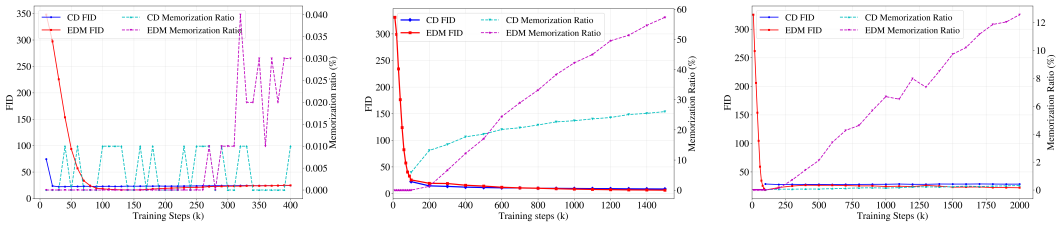
$$\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\left[\mathbf{h}(\mathbf{x})^\top\mathbf{A}_\phi^\top\mathbf{A}_\phi\,\mathbf{h}(\mathbf{x})\right] = \mathrm{Tr}\left(\mathbf{A}_\phi^\top\mathbf{A}_\phi\cdot\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\left[\mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x})^\top\right]\right) = \mathrm{Tr}\left(\mathbf{A}_\phi^\top\mathbf{A}_\phi\,\mathbf{U}\right). \tag{50}$$

Therefore, we finally obtain

$$\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\left[\frac{\|\delta\mathbf{x}\|_2^2}{d}\right] = \Delta t^2 t'^2\,\frac{1}{dp}\,\mathrm{Tr}\left(\mathbf{A}_\phi^\top\mathbf{A}_\phi\,\mathbf{U}\right). \tag{51}$$

Invoking Lemma 4.2, $\mathbf{A}_\phi = -(\sqrt{p}/\sqrt{\Delta_{t'}})\,\mathbf{V}^\top\mathbf{U}^{-1}$, we have $\mathbf{A}_\phi^\top\mathbf{A}_\phi = \frac{p}{\Delta_{t'}}\,\mathbf{U}^{-1}\mathbf{V}\mathbf{V}^\top\mathbf{U}^{-1}$. Using the Gaussian-equivalent form of $\mathbf{V}$ from Lemma 4.2, we get

$$\mathbf{V}\mathbf{V}^\top = \left(\mu_1(t')\frac{\sqrt{\Delta_{t'}}}{\Gamma_{t'}}\right)^2\frac{\mathbf{W}\mathbf{W}^\top}{d} = \left(\mu_1(t')\frac{\sqrt{\Delta_{t'}}}{\Gamma_{t'}}\right)^2\mathbf{S}, \tag{52}$$



(a) Base channel width decreased from 128 to 64.

(b) Base channel width increased from 128 to 192.

(c) Reduced depth from 4 residual blocks to 2.

Figure 8: **Effect of teacher model capacity on memorization and consistency distillation.** Each panel reports the evolution of FID (left axis) and memorization ratio (right axis) as a function of training steps for the teacher diffusion model (EDM) and the corresponding consistency-distilled (CD) student. Model capacity is varied along three axes: network width (channels) and depth (number of ResNet blocks). In all cases, the teacher used for distillation is selected from the terminal training checkpoint of the corresponding run. Reducing model capacity—either by decreasing width or depth—suppresses memorization throughout training, and the CD student distilled from such teachers exhibits near-zero memorization. Conversely, increasing model capacity leads to earlier and stronger memorization in the teacher, while consistency distillation consistently yields students with substantially reduced memorization at comparable FID.

and therefore

$$\frac{1}{p}\,\mathrm{Tr}(\mathbf{A}_\phi^\top \mathbf{A}_\phi\,\mathbf{U}) = \frac{1}{\Delta_{t'}}\,\mathrm{Tr}\big(\mathbf{U}^{-1}\mathbf{V}\mathbf{V}^\top\big) = \left(\frac{\mu_1(t')}{\Gamma_{t'}}\right)^2 \mathrm{Tr}\big(\mathbf{U}^{-1}\mathbf{S}\big)\,. \tag{53}$$

Combining (48)–(53) gives

$$\begin{aligned}\beta(t',\Delta t) &= a_0(t')\,\Delta t^2 t'^2\,\left(\frac{\mu_1(t')}{\Gamma_{t'}}\right)^2 \frac{1}{d}\,\mathrm{Tr}\big(\mathbf{U}^{-1}\mathbf{S}\big)\\ &= a_0(t')\,a_1(t')^2\,\Delta t^2 t'^2\,\frac{1}{d}\,\mathrm{Tr}\big(\mathbf{U}^{-1}\mathbf{S}\big)\,.\end{aligned} \tag{54}$$

Putting together the non-isotropic and isotropic contributions, we finally obtain

$$\mathbf{U}_{\mathrm{cd}} = \Delta t^2\,a_1(t')^2\Big(\mathbf{S} - \mu_1(t')^2\,\mathbf{S}\,\mathbf{U}^{-1}\,\mathbf{S}\Big) \;+\; \beta(t',\Delta t)\,\mathbf{I}_p, \tag{55}$$

This completes the proof.

□

E Additional Results

E.1 Additional Results in Section 3

E.1.1 Effect of Model Capacity on Memorization and Distillation

We further study how the model capacity affects memorization behavior and how effectively consistency distillation mitigates such effects. All results in this subsection are obtained using the *two-step* consistency distillation objective. All experiments are conducted on a randomly sampled subset of 6000 training images and trained under an identical optimization schedule, while the total number of training steps varies across configurations according to their respective training setups. For each architectural configuration, the teacher model used for distillation is selected from the terminal checkpoint of the corresponding training trajectory.

We first vary the model capacity by changing the *base channel width* of the U-Net backbone while keeping the depth fixed. When the base channel width is reduced from 128 to 64, the teacher model exhibits negligible memorization throughout training. At the terminal training step (400k steps), the teacher achieves FID = 24.8 with a memorization ratio below 0.1%. Applying two-step consistency

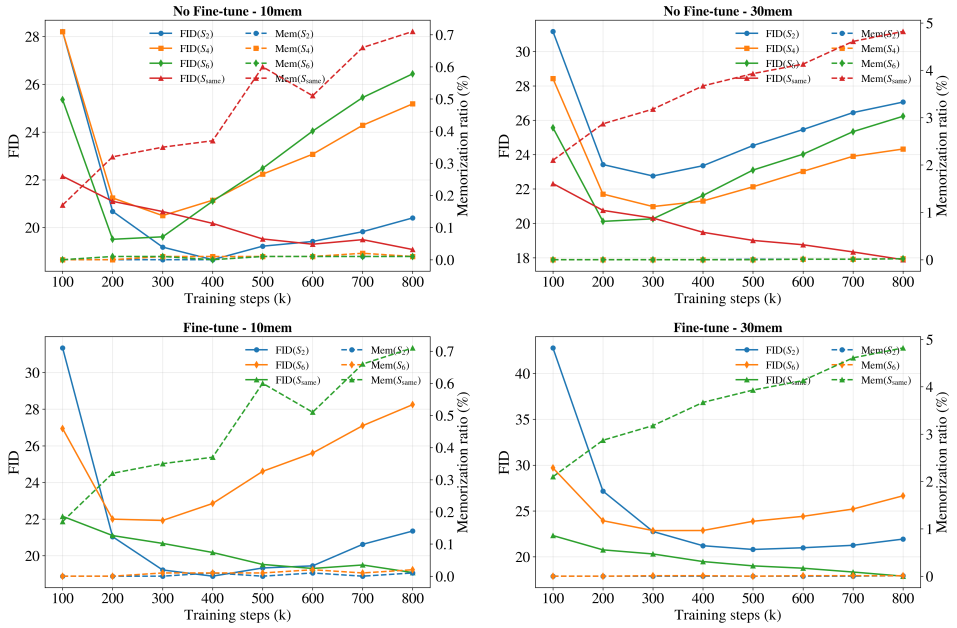


Figure 9: **Effect of student–teacher architectural mismatch on memorization and sample quality under two-step consistency distillation.** The figure reports FID (left y-axis, solid lines) and memorization ratio (right y-axis, dashed lines) as functions of training steps for different student–teacher architectural configurations. Results are shown for both 10%-memorization and 30%-memorization teachers. Across all settings, the student memorization ratio remains close to zero regardless of initialization strategy or architectural mismatch. However, when the student is randomly initialized (No Fine-tune), FID degrades at later training stages, indicating reduced optimization stability. In contrast, inheriting shared teacher parameters substantially stabilizes training and mitigates late-stage FID degradation, even when student and teacher architectures differ.

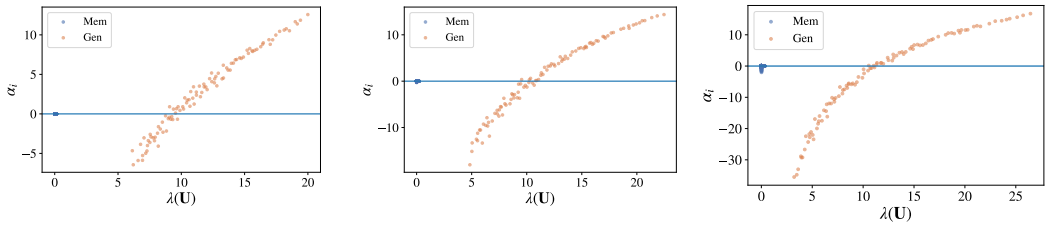
distillation to this teacher yields a student model with comparable sample quality (FID = 24.58) and a memorization ratio statistically indistinguishable from zero.

In contrast, increasing the base channel width to 192 substantially alters the training dynamics. In this high-capacity regime, the teacher model exhibits early onset and steady growth of memorization, reaching a memorization ratio of approximately 57.34% at the terminal checkpoint (1500k steps), while achieving FID = 6.34. Despite this pronounced memorization in the teacher, the two-step consistency-distilled student reduces the memorization ratio to 26.05%, while preserving comparable sample quality with FID = 8.45.

We further examine architectural depth by reducing the *number of residual blocks per resolution level* from 4 to 2, while keeping the base channel width fixed with 128. This shallower teacher model exhibits weak memorization, with the terminal memorization ratio of approximately 12.56% and FID = 22.31 (2000k steps). The corresponding two-step consistency-distilled student closely matches this behavior, achieving FID = 28.67 with a memorization ratio again near zero. Overall, two-step consistency distillation behaves robustly across model capacities: it preserves non-memorizing behavior when the teacher does not memorize, and substantially suppresses memorization when increased capacity induces it, while maintaining competitive sample quality.

E.1.2 Effect of student–teacher architectural mismatch and initialization

Fig. 9 investigates the impact of architectural mismatch and initialization strategy on two-step consistency distillation, under both 10%- and 30%-memorization teachers. All experiments use two-step consistency distillation on a randomly sampled subset of 5000 training images. We vary the depth of the student network while keeping the teacher architecture fixed, and compare two initialization strategies: (i) *no fine-tuning*, where the student is randomly initialized, and (ii) *fine-tuning*, where the student inherits all compatible parameters from the teacher, with unmatched parameters initialized



(a) $\psi_p = 32,\ \psi_n = 16.$ (b) $\psi_p = 32,\ \psi_n = 8.$ (c) $\psi_p = 32,\ \psi_n = 4.$

Figure 10: **Effect of the ψ_p/ψ_n ratio on mode-wise non-isotropic CD response.** Each point corresponds to a teacher eigenmode u_i, plotted against its curvature eigenvalue $\lambda_i(\mathbf{U})$, with modes partitioned into MEM and GEN by a fixed spectral threshold. Across panels, the feature dimension is fixed at $\psi_p = 32$, while the effective sample ratio ψ_n decreases from left to right. As ψ_p/ψ_n decreases, the fraction of GEN modes with positive signed response $\alpha_i > 0$ increases, indicating that a broader portion of the generalization-associated spectrum receives positive non-isotropic updates, while MEM modes remain tightly concentrated near $\alpha_i \simeq 0$.

randomly. We denote by $\mathrm{S}2$, $\mathrm{S}4$, and $\mathrm{S}6$ students whose network depth is defined by 2, 4, and 6 residual blocks per stage, respectively. $\mathrm{S}_{\mathrm{same}}$ corresponds to the setting where the student architecture exactly matches that of the teacher, whose residual blocks per stage is fixed with 4. All experiments use the same two-step consistency objective and identical training protocols.

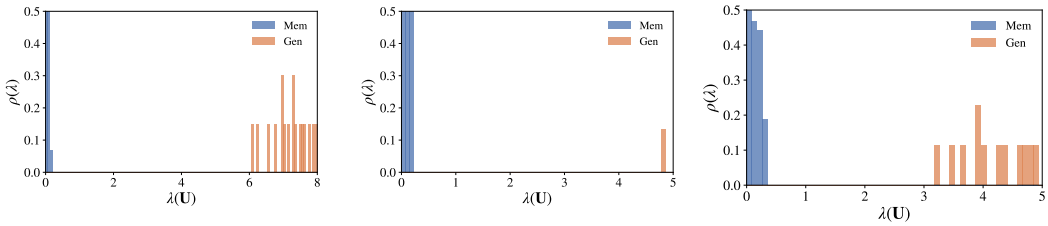
A key observation is that the student memorization behavior depends strongly on *architectural match*. When the student exactly matches the teacher architecture ($\mathrm{S}_{\mathrm{same}}$), its memorization ratio increases steadily over training, although it remains substantially lower than that of the teacher. In contrast, when the student is either randomly initialized or architecturally mismatched (e.g., $\mathrm{S}2/\mathrm{S}4/\mathrm{S}6$ distilled from a fixed teacher), its memorization ratio stays near zero throughout. This separation suggests that consistency distillation does not universally "erase" teacher behaviors; rather, it transfers what the student can faithfully realize under its parameterization. When the student has sufficient expressivity and a well-aligned parameterization (the $\mathrm{S}_{\mathrm{same}}$ setting), it can reproduce a larger portion of the teacher mapping—including a small but non-negligible fraction of teacher-specific, memorization-associated behavior. When the student is mismatched or starts far from the teacher solution, the distillation signal becomes harder to realize precisely, and the learned solution is biased toward more conservative, distribution-level smoothing, which suppresses instance-level memorization.

The FID trends are consistent with this interpretation. Only the architecture-matched student ($\mathrm{S}_{\mathrm{same}}$) exhibits monotonic improvement in FID over training, indicating that the consistency objective can be optimized in a stable manner when the student can closely approximate the teacher-induced consistency function. By contrast, for randomly initialized or mismatched students, FID improves early but degrades at later stages. This late-stage degradation occurs despite near-zero memorization, and is indicative of an optimization bias: when exact teacher-matching is unattainable under the student parameterization, continued minimization of the consistency loss increasingly favors overly smooth and low-diversity solutions, which harms sample quality while keeping memorization low.

Consequently, these results highlight a tradeoff controlled by architectural compatibility. Exact architectural matching enables stable quality improvement but allows partial transfer of teacher memorization, whereas mismatch or random initialization strongly suppresses memorization but can suffer from late-stage quality degradation. A more detailed theoretical characterization of how architectural realizability and optimization dynamics jointly shape memorization transfer under consistency distillation is left for future work.

E.2 Additional Results in Section 4

To understand how the data–feature scaling ratio ψ_p/ψ_n influences the mode-wise behavior of non-isotropic consistency distillation, we examine how the signed response α_i distributes across teacher curvature modes under different ψ_p/ψ_n configurations. Fig. 10 shows that, for fixed ψ_p, decreasing ψ_n (and hence increasing the ratio ψ_p/ψ_n) systematically enlarges the subset of generalization-associated (Gen) modes with positive signed response $\alpha_i > 0$. While Gen modes at very small



(a) $\psi_p = 32,\ \psi_n = 16.$ (b) $\psi_p = 32,\ \psi_n = 8.$ (c) $\mathbf{U}$ ($\psi_p = 32,\ \psi_n = 4.$

Figure 11: **Spectral separation of the teacher curvature matrix $\mathbf{U}$ under different ψ_p/ψ_n ratios.** Shown are the empirical spectral densities $\rho(\lambda)$ of $\mathbf{U}$, with modes partitioned into memorization-associated (MEM) and generalization-associated (GEN) subspaces by a fixed threshold. For $\psi_p = 32$ and $\psi_n = 16$ (left), the GEN spectrum is well separated and concentrated at significantly larger eigenvalues than the MEM spectrum, yielding a clear spectral gap. As ψ_n decreases (middle to right), the GEN spectrum shifts leftward and becomes less separated from MEM modes. This progressive loss of spectral separation provides a geometric explanation for the change in the sign distribution of α_i: configurations with stronger MEM/GEN separation admit a larger fraction of GEN modes with $\alpha_i > 0$, facilitating more effective non-isotropic transfer.

curvature eigenvalues may remain suppressive, a progressively larger portion of the Gen spectrum receives positive non-isotropic updates as ψ_n decreases. Across all configurations, memorization-associated (Mem) modes remain tightly concentrated near $\alpha_i \approx 0$, indicating that positive transfer induced by the non-isotropic CD term is selectively allocated within the Gen subspace rather than leaking into memorization-dominated directions.

Fig. 11 further shows that this behavior is closely tied to the spectral geometry of the teacher curvature matrix $\mathbf{U}$. When $\psi_p = 32$ and $\psi_n = 16$, Gen modes occupy a well-separated, high-eigenvalue region, while Mem modes remain concentrated near the origin, yielding a clear spectral gap. In this regime, the resolvent term $\mathbf{S}\mathbf{U}^{-1}\mathbf{S}$ primarily suppresses low-eigenvalue directions, so that most Gen modes satisfy $\alpha_i > 0$. As ψ_n decreases, the Gen spectrum shifts toward smaller eigenvalues and the spectral separation from Mem modes weakens, increasing the fraction of Gen modes that fall into the suppressive regime. This progressive loss of spectral separation directly explains the change in the sign distribution of α_i, linking the mode-wise effect of non-isotropic consistency distillation to the underlying curvature spectrum of $\mathbf{U}$.