On the Memorization of Consistency Distillation for Diffusion Models

Bingqing Jiang
bingqingjiang@connect.hku.hk
The University of Hong Kong

Difan Zou
dzou@hku.hk
The University of Hong Kong

Abstract

Diffusion models are central to modern generative modeling, and understanding how they balance memorization and generalization is critical for reliable deployment. 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, deployed diffusion models are often further distilled, introducing an additional training phase whose impact on memorization remains unclear. In this work, we analyze how distillation reshapes memorization behavior in diffusion models, taking consistency distillation as a representative framework. Empirically, we show that when applied to a teacher model that has memorized data, consistency distillation significantly reduces transferred memorization in the student while preserving, and sometimes improving, sample quality. To explain this behavior, we 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 suggest that distillation can serve not only as an acceleration tool, but also as a mechanism for improving the memorization–generalization trade-off.

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. This view is also supported by evidence that trained denoising scores are smoother than closed-form empirical optima, which helps explain why practical diffusion models can generalize instead of simply reproducing training samples (Wang et al., 2024b).

However, existing analyses of memorization have largely focused on diffusion models trained from scratch. This setting does not fully capture modern deployment pipelines, where a pretrained diffusion model is often further distilled to improve sampling efficiency and reduce computational cost (Song et al., 2023; Salimans & Ho, 2022; Kim et al., 2024; Luo et al., 2023b; 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 merely a passive model compression step, but a new training process with its own objective, data distribution, and optimization dynamics (Xiang et al., 2025). If memorization in diffusion models depends sensitively on training dynamics and time scales, as prior work suggests, then distillation may further reshape,

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

suppress, or even amplify memorization inherited from the teacher. Yet, the effect of distillation on memorization remains largely uncharacterized.

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., 2024c; 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 nontrivial training phase beyond the original diffusion training. This additional optimization stage operates under a distinct objective and data distribution, and can therefore alter the balance between memorization and generalization established during the teacher's training. The main contributions can be summarized as follows:

- 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.
- 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.
- 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. 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., 2025), who derive exact learning curves for diffusion models with random-feature parameterizations. Building on this framework, (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. Beyond direct duplication, recent extraction studies show that explicit or surrogate conditioning can amplify memorization risks in diffusion models (Chen et al., 2025a). While prior work has primarily focused on diffusion models trained from scratch, the memorization behavior of distilled diffusion models remains underexplored. In this work, we analyze how the additional training stage introduced by consistency distillation reshapes memorization relative to standard diffusion training and provide a theoretical explanation for the observed behavior.

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., 2024a; 2025; Dou et al., 2024; Yang et al., 2025; Chen et al., 2025b). However, these studies primarily emphasize efficiency and sample quality. The effect of consistency distillation on memorization remains largely unexplored, which is the focus of our study.

3. Definitions and Preliminaries

3.1. 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

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

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

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

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 $\{\mathbf{x}_\nu\}_{\nu=1}^n$ (Bonnaire et al., 2025; Vincent, 2011; Hyvärinen & Dayan, 2005):

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

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

3.2. 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 A.

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

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = -\mathbf{x}_t - \mathbf{s}_\phi(\mathbf{x}_t, t). \tag{3}$$

In practice, we discretize the time interval into an increasing sequence $\epsilon = t_0 < t_1 < \cdots < t_K = T$. 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:

$$\widehat{\mathbf{x}}^{\phi}_{t_k} = \mathbf{x}_{t_{k+1}} + (t_k - t_{k+1})\Big(-\mathbf{x}_{t_{k+1}} - \mathbf{s}_\phi(\mathbf{x}_{t_{k+1}}, t_{k+1})\Big). \tag{4}$$

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:

$$L_{\mathrm{CD}}(\boldsymbol{\theta}) = \mathbb{E}_{k,\mathbf{x}_{t_{k+1}}}\left[\left\| f_\theta\big(\mathbf{x}_{t_{k+1}}, t_{k+1}\big) - \mathrm{sg}\Big[f_{\theta^-}\big(\widehat{\mathbf{x}}^{\phi}_{t_k}, t_k\big)\Big]\right\|_2^2\right], \tag{5}$$

where θ^- denotes the EMA target parameters and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. This objective transfers the local dynamics of the teacher PF-ODE to the student without requiring the student to explicitly approximate the score function.

Table 1. Memorization and generation quality comparison on unconditional CIFAR-10 under different data settings.

Setting	Model	FID	l_2 Mem	SSCD Mem / p95
3000-data	Teacher	21.78	4.88%	21.68% / 0.8164
	Student	20.82	0.01%	2.83% / 0.5623
4000-data	Teacher	21.52	3.73%	18.31% / 0.7983
	Student	20.87	0.01%	0.67% / 0.5128
5000-data	Teacher	23.82	5.94%	21.59% / 0.8270
	Student	23.03	0.02%	1.66% / 0.5342
6000-data	Teacher	24.57	4.35%	17.65% / 0.8027
	Student	23.68	0.01%	0.37% / 0.5010

Table 2. Memorization and generation quality comparison on class-conditional ImageNet under different data settings.

Setting	Model	FID	SSCD Mem / p95
5000-data	Teacher	16.89	17.4% / 0.7011
	Student	17.70	2.67% / 0.5713
7000-data	Teacher	20.14	30.25% / 0.7309
	Student	28.65	2.65% / 0.5552
10000-data	Teacher	27.38	18.57% / 0.6972
	Student	28.65	1.68% / 0.4852

Table 3. Memorization and generation quality comparison on Stable Diffusion v1.5 under different extra-data settings.

Setting	Model	CLIP	SSCD-mean	SSCD-max	SSCD-p95
3× extra	Teacher	0.2274	0.3612	0.6046	0.5451
	Student	0.2209	0.3380	0.5774	0.5064
4× extra	Teacher	0.2237	0.3390	0.5539	0.5109
	Student	0.2151	0.3002	0.5194	0.4795
5× extra	Teacher	0.2125	0.3003	0.6271	0.4764
	Student	0.2042	0.2685	0.5494	0.4204

4. Memorization and Generation Quality in Consistency Distillation

4.1. Experimental Setup

Datasets. We evaluate consistency distillation on unconditional CIFAR-10 (Krizhevsky et al., 2009), class-conditional ImageNet (Deng et al., 2009), and text-to-image generation with Stable Diffusion v1.5 (Rombach et al., 2022). For CIFAR-10, we consider multiple reduced-data regimes by uniformly subsampling $n \in \{3000, 4000, 5000, 6000\}$ training images. For ImageNet, we study class-conditional generation on three reduced-data subsets with 5k, 7k and 10k training images. For text-to-image generation, we adopt a memorization-oriented protocol adapted from prior work (Somepalli et al., 2023; Wen et al., 2024). Specifically, we fine-tune Stable Diffusion v1.5 on a dataset consisting of 200 image–prompt pairs, each repeated 10 times to induce memorization, together with additional COCO image–prompt pairs to preserve generalization. The amount of added COCO data is set to 3×, 4×, or 5× the size of the original 200-pair subset. The resulting fine-tuned model serves as the teacher for subsequent consistency distillation. *Following prior memorization studies, no data augmentation is applied throughout to avoid ambiguity in memorization assessment (Gu et al., 2025).*

Training configuration. For CIFAR-10 and ImageNet, we use pre-trained EDM models as teachers (Karras et al., 2022) and initialize students from the same checkpoints before consistency distillation (Song et al., 2023). We use LPIPS (Zhang et al., 2018) as the metric in the consistency loss, using Heun's second-order solver with 18 discretization steps on CIFAR-10 and 40 discretization steps on ImageNet. In both settings, the distilled student is evaluated with 1-step generation unless otherwise noted. For Stable Diffusion v1.5, we apply latent consistency distillation to the fine-tuned memorizing teacher with Huber loss following the official LCM setup (Luo et al., 2023a), where teacher targets are constructed with a DDIM-based ODE solver on a 50-step DDIM discretization and the student is evaluated with 4-step sampling. All experiments are conducted on 8 NVIDIA H100 GPUs.

Evaluation metrics. We evaluate generation quality using FID on CIFAR-10 and ImageNet, and CLIP score on Stable Diffusion v1.5. For memorization, we report the standard ℓ_2-based memorization ratio (Yoon et al., 2023) and SSCD-based semantic metrics. On CIFAR-10, both metrics are used: under the ℓ_2 criterion, a sample is memorized if its nearest-training distance is less than one-third of its second-nearest distance. On ImageNet, we use only SSCD features, as pixel-space matching is less reliable for complex natural images (Wen et al., 2024). For CIFAR-10 and ImageNet, we use an SSCD threshold of 0.6 to compute memorization ratios and also report the 95th-percentile (p95) similarity to reduce sensitivity to the threshold choice. For Stable Diffusion v1.5, we report SSCD mean, maximum, and p95 similarity over generated samples. Additional implementation details are provided in Appendix B.1 and B.2.

Table 4. Results under realistic localized memorization settings on CIFAR-10 and ImageNet. Memorization statistics are reported as SSCD memorization ratio / p95 similarity to training data.

Dataset	Model	FID	Rep. cls.	Non-rep. cls.	Overall
CIFAR-10	Teacher	12.02	74.54% / 0.8983	0 / 0.4782	13.91% / 0.8255
	Student	11.39	33.51% / 0.7158	0 / 0.4747	5.69% / 0.6139
ImageNet	Teacher	25.22	7.35% / 0.6373	0 / 0.2308	0.37% / 0.2497
	Student	24.84	0 / 0.4118	0 / 0.2357	0 / 0.2416

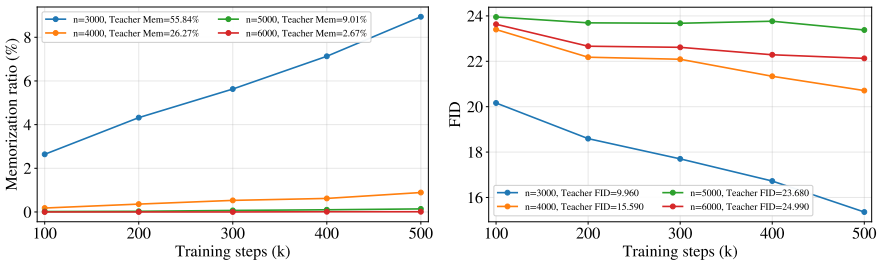


Figure 1. **Student dynamics during consistency distillation on CIFAR-10. Left**: memorization ratio. **Right**: FID. Consistency distillation reliably reduces memorization, while FID improvement depends on the memorization state of the teacher.

4.2. Memorization under Matched Generation Quality

We first ask whether the student memorizes less simply because it generates worse samples. To avoid this confound, we compare teacher–student pairs with comparable FID or CLIP score and then examine their memorization behavior. As shown in Table 1, teachers and students on CIFAR-10 achieve comparable FID across all data settings, yet the student consistently exhibits much lower memorization under both l_2- and SSCD-based criteria, including a markedly reduced upper tail of SSCD similarity. This suggests that consistency distillation suppresses both overall memorization and severe near-duplicate generations. The same trend holds for class-conditional ImageNet in Table 2, where the student remains close to the teacher in FID but shows a much lower SSCD memorization ratio, indicating that the effect is not specific to CIFAR-10. Finally, Table 3 shows that, for Stable Diffusion v1.5 under different extra-data settings, teacher and student remain broadly comparable in CLIP score while the student consistently reduces SSCD mean, maximum similarity, and p95 similarity, covering both average similarity and high-similarity cases indicative of direct recall. Additional results on sampler effects, model capacity, and student–teacher architectural mismatch are provided in Appendix D.1–D.3.

4.3. Localized Memorization under Global Generalization

We next evaluate consistency distillation in a more realistic regime where memorization is localized rather than global. In practice, memorization is often concentrated in a small subset of repeated or overexposed samples, while the model continues to generalize over the rest of the data distribution. To reflect this structure, we repeat only a small subset of classes and keep the rest non-repeated: on CIFAR-10, we repeat 50 images from each of classes 0 and 1 five times and sample 1000 non-repeated images from classes 2–9; on ImageNet, we use 50 repeated classes with 5 anchor images per class repeated 10 times, and 50 distinct non-repeated images for each of the remaining 950 classes. As shown in Table 4, the teacher exhibits clear memorization on repeated classes but negligible memorization on non-repeated classes, confirming that memorization is indeed localized in this setting. Under the same setup, the distilled student substantially reduces memorization on the repeated classes for both CIFAR-10 and ImageNet, while leaving the non-repeated classes essentially unchanged and maintaining comparable FID. These results suggest that consistency distillation selectively reduces localized memorization while preserving broader generalization.

4.4. Training Dynamics of the Student model during Consistency Distillation

Finally, we examine how memorization and generation quality evolve during consistency distillation on CIFAR-10, where both the EDM teacher and student are trained for 500k steps for a fair comparison. Figure 1 reports the student memorization ratio and FID over distillation steps. Across all settings, consistency distillation substantially suppresses memorization: although student memorization may gradually increase, it remains well below the teacher level throughout. At the same time, the FID trends depend on the teacher regime: when teacher memorization is moderate, distillation reduces memorization and eventually yields better FID than the teacher; when the teacher

is severely memorization-dominated, the student still memorizes much less but has worse FID. Thus, consistency distillation reliably lowers memorization, while improving FID only when the teacher is not severely memorizing. A theoretical explanation for the degraded student generation quality in the severe-teacher regime is provided in Appendix D.4.

5. Theoretical Analysis

We now turn to the mechanism behind the empirical findings. The goal is to explain why consistency distillation can reduce memorization while preserving feature directions that support sample quality.

5.1. 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 defined across multiple diffusion times (Song et al., 2023), the one-step formulation isolates the leading-order local consistency constraint induced by the teacher probability-flow dynamics. It therefore provides a principled local approximation to Eq. (5), capturing the shared leading-order update geometry underlying neighboring-step consistency relations (George et al., 2025; Bonnaire et al., 2025; Li et al., 2025). For tractability, we omit the EMA target and stop-gradient in Eq. (5) and analyze the induced symmetric local consistency penalty (Dou et al., 2024; Chen et al., 2025b). Under this view, finite discretization mainly affects how accurately this local constraint is realized in practice, while the leading-order mechanism is already determined by the consistency objective itself. This interpretation is also consistent with the discretization study in Appendix D.5. In this setting, the training objective compares the student outputs evaluated at two nearby inputs connected by a single teacher-induced step:

$$L_{\mathrm{CD}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}\left[\left\| f_{\boldsymbol{\theta}}\big(\mathbf{x}_{\nu,t'}(\boldsymbol{\xi})\big) - f_{\boldsymbol{\theta}}\big(\widehat{\mathbf{x}}^{\phi}_{\nu,t}(\boldsymbol{\xi})\big)\right\|_2^2\right], \tag{6}$$

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 (4).

5.2. 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 (6), we obtain

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

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

$$\mathbf{U}_{\mathrm{cd}} = \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_{\boldsymbol{\xi}}\big[\Delta\mathbf{h}_\nu(\boldsymbol{\xi})\,\Delta\mathbf{h}_\nu(\boldsymbol{\xi})^\top\big]. \tag{7}$$

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

5.3. 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 C.1, the following lemma characterizes the leading-order contribution to the curvature matrix U_{cd} defined in (7).

Lemma 5.1 (Orthogonal second-moment decomposition of $\Delta\mathbf{h}$)**.** *Under Assumptions C.1 and C.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*

$$\mathbb{E}_\zeta\left[\Delta\mathbf{h}\Delta\mathbf{h}^\top | \mathbf{x}, \boldsymbol{\delta x}\right] = a_1(\mathbf{x}, \boldsymbol{\delta x})^2 \mathbb{E}_\zeta\left[\Delta\mathbf{g}\Delta\mathbf{g}^\top | \mathbf{x}, \boldsymbol{\delta x}\right] + a_0(\mathbf{x}, \boldsymbol{\delta x}) \mathbb{E}_\zeta\left[\Delta g_i^2 |\, \mathbf{x}, \boldsymbol{\delta x}\right] \mathbf{I}_p, \tag{8}$$

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 C.1.*

Assume further the small-noise one-step regime of Assumption C.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. (20) and Eq. (22) in Appendix C.2. See Appendix C.2 for the proof.*

Lemma 5.1 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 5.2 ((Bonnaire et al., 2025))**.** *For RFNN score-matching trained by gradient flow with zero initialization at fixed* t' *on (2), the converged teacher top layer satisfies*

$$\mathbf{A}_\phi = -\frac{\sqrt{p}}{\sqrt{\Delta_{t'}}}\, \mathbf{V}^\top\, \mathbf{U}^{-1}, \tag{9}$$

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*

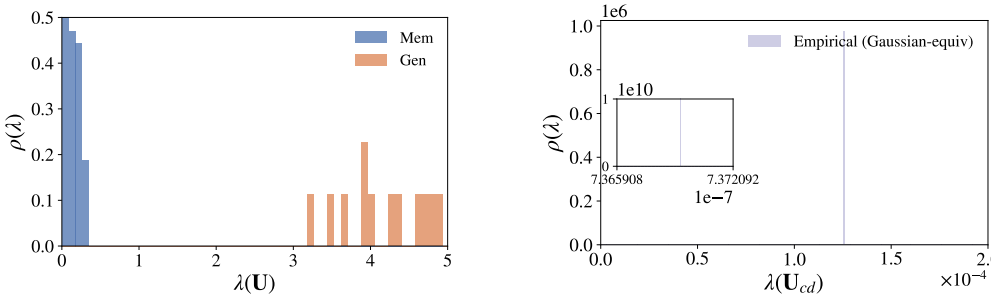
$$\mathbf{U} = \frac{1}{n}\mathbf{G}\mathbf{G}^\top + b_{t'}^2 \frac{\mathbf{W}\mathbf{W}^\top}{d} + s_{t'}^2 \mathbf{I}_p,$$

   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*

$$\mathbf{V} = \mu_1(t')\frac{\sqrt{\Delta_{t'}}}{\Gamma_{t'}}\frac{\mathbf{W}}{\sqrt{d}}, \tag{10}$$

   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 5.1 with the teacher characterization in Lemma 5.2, we obtain the following structural characterization of the consistency distillation curvature.



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

Figure 2. **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 B.3 for details.

Theorem 5.3. *Let $\{\mathbf{x}_\nu\}_{\nu=1}^n \subset \mathbb{R}^d$ be training samples with $\mathrm{Cov}(\mathbf{x}_\nu) = \mathbf{\Sigma} = \mathbf{I}_d$. 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 5.1. Then, as $\Delta t \to 0$,*

$$\mathbf{U}_{\mathrm{cd}} = \Delta t^2 a_1(t')^2 \Big(\mathbf{S} - \mu_1(t')^2\, \mathbf{S}\mathbf{U}^{-1}\mathbf{S}\Big) + \beta(t', \Delta t)\mathbf{I}_p, \tag{11}$$

where $\mathbf{S} = \mathbf{W}\mathbf{W}^\top/d$, and the isotropic shift $\beta(t', \Delta t)$ is defined in Eq. (32). See Appendix C.3 for the proof.

Theorem 5.3 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. 2a, 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 $\mathbf{U}^{-1}$.

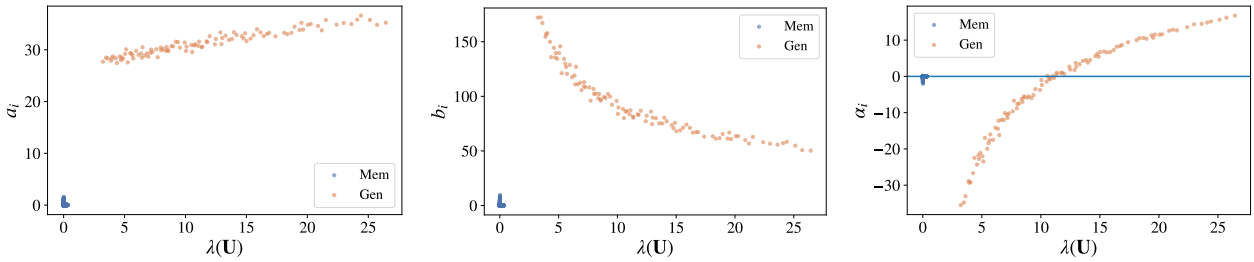
We now turn to the empirical spectral density of $\mathbf{U}_{\mathrm{cd}}$ in Fig. 2b. Theorem 5.3 predicts a *sharp spectral spike* induced by the isotropic term, and the origin is purely algebraic: the random-feature Gram operator $\mathbf{S} = \frac{1}{d}\mathbf{W}\mathbf{W}^\top$ has rank at most 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. 2b. 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. 2b.

5.4. Empirical Validation of Spectral Filtering

To test this mechanism, we 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

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



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

Figure 3. **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 Mem and 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 $\rho_\Sigma(\lambda) = \delta(\lambda - 1)$.

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

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 $\mathcal{I}_{\text{mem}} = \{i : \lambda_i(\mathbf{U}) < \lambda_{\text{th}}\}, \mathcal{I}_{\text{gen}} = \{i : \lambda_i(\mathbf{U}) \geq \lambda_{\text{th}}\}$ (Bonnaire et al., 2025).

(I) Visibility a_i and resolvent structure b_i. Figures 3a and 3b visualize the two constituent terms of α_i. Mem modes have uniformly small a_i, indicating that these sample-specific teacher eigenmodes are largely orthogonal to the random-feature span. In contrast, Gen modes attain larger a_i because their smoother shared structure aligns more strongly with this span, with visibility increasing with $\lambda_i(\mathbf{U})$. For b_i, lower-λ Gen modes incur stronger $\mathbf{U}^{-1}$-induced subtraction, whereas higher-λ Gen modes are less affected. In contrast, Mem modes again show uniformly small b_i, since their weak random-feature alignment leaves little energy in $\mathbf{S}\mathbf{u}_i$ for resolvent amplification.

(II) Net per-mode response α_i. Fig. 3c 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 D.4.

(III) Global allocation of positive updates.

To quantify the subspace allocation of positive non-isotropic updates, 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 global positive-curvature allocation and is robust to isolated atypical modes. Empirically, $\text{Share}^+_{\text{mem}} \approx 9.5 \times 10^{-3}$, indicating that almost all positive non-isotropic updates are assigned to generalization-associated directions. Thus, consistency distillation shifts positive learning signal away from memorization-associated modes at the global-dynamics level.

6. Conclusion and Discussion

In summary, consistency distillation substantially reduces memorization in diffusion models, including cases with strongly overfitted teacher models. At the same time, sample quality is often preserved and can improve when the teacher exhibits a moderate level of memorization. Our theoretical analysis suggests that this behavior arises because consistency distillation actively reshapes the training geometry: it suppresses memorization-associated directions while preserving generalization-relevant updates, rather than passively inheriting teacher behavior.

References

Baptista, R., Dasgupta, A., Kovachki, N. B., Oberai, A., and Stuart, A. M. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

Bonnaire, T., Urfin, R., Biroli, G., and Mezard, M. 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.

Buchanan, S., Pai, D., Ma, Y., and Bortoli, V. D. On the edge of memorization in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.

Chen, Y., Wang, S., Zou, D., and Ma, X. SIDE: Surrogate conditional data extraction from diffusion models, 2025a.

Chen, Y., Zhang, Y., Oertell, O., and Sun, W. Convergence of consistency model with multistep sampling under general data assumptions. In *Forty-second International Conference on Machine Learning*, 2025b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

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

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

Geng, Z., Pokle, A., Luo, W., Lin, J., and Kolter, J. Z. Consistency models made easy. In *The Thirteenth International Conference on Learning Representations*, 2025.

George, A. J., Veiga, R., and Macris, N. Denoising score matching with random features: Insights on diffusion models from precise learning curves. *arXiv preprint arXiv:2502.00336*, 2025.

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020.

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

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

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 2022.

Kim, D., Lai, C.-H., Liao, W.-H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., and Ermon, S. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lai, C.-H., Takida, Y., Uesaka, T., Murata, N., Mitsufuji, Y., and Ermon, S. 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.

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

Li, P., Li, Z., Zhang, H., and Bian, J. On the generalization properties of diffusion models. *Advances in Neural Information Processing Systems*, 2023.

Luo, S., Tan, Y., Huang, L., Li, J., and Zhao, H. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023a.

Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., and Zhang, Z. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 2023b.

Luo, W., Huang, Z., Geng, Z., Kolter, J. Z., and Qi, G.-j. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 2024.

Park, G. Y., Lee, S. W., and Ye, J. C. Inference-time diffusion model distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Pham, B., Raya, G., Negri, M., Zaki, M. J., Ambrogioni, L., and Krotov, D. Memorization to generalization: Emergence of diffusion models from associative memory networks. In *New Frontiers in Associative Memories*, 2025.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

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

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Starodubcev, N., Kuznedelev, D., Babenko, A., and Baranchuk, D. Scale-wise distillation of diffusion models. *arXiv preprint arXiv:2503.16397*, 2025.

Tee, J. T. J., Zhang, K., Yoon, H. S., Gowda, D. N., Kim, C., and Yoo, C. D. Physics informed distillation for diffusion models. *Transactions on Machine Learning Research*, 2024.

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

Wang, F.-Y., Huang, Z., Bergman, A. W., Shen, D., Gao, P., Lingelbach, M., Sun, K., Bian, W., Song, G., Liu, Y., Wang, X., and Li, H. Phased consistency models. In *Advances in Neural Information Processing Systems*, 2024a.

Wang, F.-Y., Geng, Z., and Li, H. Stable consistency tuning: Understanding and improving consistency models. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025.

Wang, H., Han, Y., and Zou, D. On the discrepancy and connection between memorization and generation in diffusion models. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024b.

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

Xiang, Q., Zhang, M., Shang, Y., Wu, J., Yan, Y., and Nie, L. Dkdm: Data-free knowledge distillation for diffusion models with any architecture. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

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

Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., and Freeman, B. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 2024a.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T., and Park, T. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T., and Park, T. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024c.

Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 workshop on structured probabilistic inference $\{\backslash\&\}$ generative modeling*, 2023.

Zeno, C., Manor, H., Ongie, G., Weinberger, N., Michaeli, T., and Soudry, D. When diffusion models memorize: Inductive biases in probability flow of minimum-norm shallow neural nets. In *Forty-second International Conference on Machine Learning*, 2025.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

Zhang, Z., Li, X., Li, X., Shi, L., Wu, M., Tao, M., and Qu, Q. 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 4. 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. Rationale for Adopting Consistency Distillation

A.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.

In these experiments, the teacher remains an EDM model trained on CIFAR-10, and we construct teachers with 10% and 30% memorization rates under the l_2 metric for both PD and CD distillation. 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.

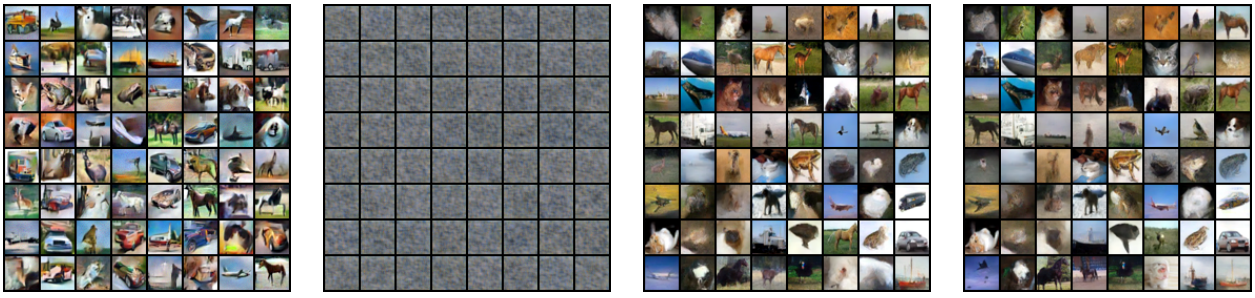
Empirical fragility of PD under limited data and memorizing teachers. Despite this favorable configuration, PD exhibits limited robustness in the regime we consider. With 6000 training samples, the final PD model distilled from a 10%-memorization teacher attains an FID of 45.42 with a memorization ratio 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. 4, 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.

Extension beyond trajectory-based distillation. While PD is the primary non-CD baseline in our study because it is the most directly comparable distillation method to CD, we also evaluate DMD to examine whether the memorization-

Table 5. Results for PD and DMD before and after EDM-based refinement under limited-data CIFAR-10 settings. Refinement substantially improves sample quality while memorization remains well below teacher levels.

Method Group	Model	FID	l_2 Mem	SSCD Mem	SSCD p95
PD	Teacher	22.68	10.00%	26.93%	0.8586
	PD	43.66	0.46%	3.63%	0.5664
	PD+refine	10.88	0.00%	0.32%	0.4822
DMD	Teacher	14.29	26.57%	49.12%	0.9220
	DMD	46.91	3.93%	7.17%	0.6390
	DMD+refine	15.59	0.55%	3.71%	0.5617



(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 5. 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.

suppression effect extends beyond this trajectory-based setting (Yin et al., 2024b). We find that DMD likewise exhibits reduced memorization relative to the teacher, suggesting that the effect is not unique to CD. At the same time, its raw sample quality is substantially worse. To distinguish genuine memorization reduction from trivial degradation, we further refine both PD and DMD outputs using the same pretrained EDM prior by perturbing the one-step outputs to a high noise level and reconstructing them with the standard EDM PF-ODE solver (Karras et al., 2022). After this prior-guided reconstruction, sample quality improves substantially for both methods, while memorization remains clearly below teacher levels. These results suggest that the observed memorization reduction is not merely a byproduct of poor raw sample quality.

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. 4, CD preserves significantly higher visual fidelity than PD under both teacher settings, even when the teacher itself exhibits substantial memorization.

A.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

$$\mathcal{L}_{\mathrm{CD}}^{\mathrm{cont}}(\theta) \;=\; \mathbb{E}_{x\sim p_{\mathrm{data}},\, t}\Big[\lambda(t)\,\Big\|\partial_t f_\theta(x_t,t) - t\,\nabla_x f_\theta(x_t,t)\,s_\phi(x_t,t)\Big\|^2\Big], \tag{12}$$

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. (12) 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. 5a, and fails to generate coherent samples under a 30%-memorization teacher in Fig. 5b.

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. 5c and 5d.

B. Detailed Experimental Setup

B.1. Additional Experimental Details for CIFAR-10 and ImageNet

Backbone architecture. For CIFAR-10, we use the NCSN++ backbone (Song et al., 2021b), following the standard architectural choice in prior consistency distillation work (Song et al., 2023). For ImageNet, we use the class-conditional DDPM++ architecture (Dhariwal & Nichol, 2021), consistent with the ImageNet setup in Consistency Models (Song et al., 2023). In both cases, we employ the same backbone family 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

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

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

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

which guarantees $c_{\mathrm{skip}}(\varepsilon) = 1$ and $c_{\mathrm{out}}(\varepsilon) = 0$. This follows the consistency-model parameterization in (Song et al., 2023) and ensures the required boundary condition when $\varepsilon > 0$.

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 setup in (Song et al., 2023; Karras et al., 2022). No data augmentation or additional regularization is used, in order to avoid potential confounding effects in the assessment of memorization (Gu et al., 2025).

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

$$N(k) = \left\lfloor \frac{c\,k}{K}\big((s_1+1)^2 - s_0^2\big) + s_0^2 \right\rfloor - 1,$$
$$\mu(k) = \exp\left(\frac{s_0 \log \mu_0}{N(k)}\right),$$

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.

B.2. Additional Experimental Details for Stable Diffusion v1.5

Latent-space formulation. Following Latent Consistency Models (Luo et al., 2023a), we perform consistency distillation in the latent space rather than in pixel space. Let x denote an image and c its text prompt. Using the pretrained VAE encoder of Stable Diffusion v1.5, we first map the image into latent space,

$$z = E(x),$$

where $E(\cdot)$ is the fixed encoder. Distillation is then performed entirely in latent space rather than pixel space.

Following the standard latent diffusion parameterization, a noisy latent at time t is written as

$$z_t = \alpha_t z + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

where (α_t, σ_t) are determined by the diffusion noise schedule.

Consistency model parameterization. We parameterize the latent consistency model as

$$f_\theta(z_t, c, t) = c_{\text{skip}}(t)\, z_t + c_{\text{out}}(t)\left(\frac{z_t - \sigma_t \hat{\epsilon}_\theta(z_t, c, t)}{\alpha_t}\right),$$

where $\hat{\epsilon}_\theta$ is the student noise predictor. This is the standard ϵ-prediction parameterization used in LCM-style distillation for latent diffusion models. In our experiments, the student is initialized from the fine-tuned Stable Diffusion v1.5 teacher before consistency distillation.

Teacher trajectory construction. To construct distillation targets, we follow the official LCM implementation and approximate the PF-ODE trajectory of the teacher with a DDIM-based solver. Given a discretization pair (t_{n+1}, t_n) on a 50-step DDIM grid, we first sample

$$z_{t_{n+1}} = \alpha_{t_{n+1}} z + \sigma_{t_{n+1}} \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

and then obtain the previous latent by a DDIM ODE step,

$$\hat{z}^{\Psi}_{t_n} = \Psi_{\text{DDIM}}(z_{t_{n+1}}, t_{n+1}, t_n, c),$$

where Ψ_{DDIM} denotes the DDIM-based solver applied to the teacher model. Thus, the target is defined by an earlier point on the same reverse-time teacher trajectory, rather than by matching the teacher prediction at exactly the same noisy input.

Consistency distillation objective. Let θ^- denote the EMA parameters of the student. We optimize the latent consistency objective

$$\mathcal{L}_{\text{LCD}} = \mathbb{E}_{z,c,n,\epsilon}\left[\ell_{\text{Huber}}\Big(f_\theta(z_{t_{n+1}}, c, t_{n+1}),\ f_{\theta^-}(\hat{z}^{\Psi}_{t_n}, c, t_n)\Big)\right],$$

where $\ell_{\text{Huber}}(\cdot, \cdot)$ is the Huber loss. No auxiliary reconstruction, perceptual, or adversarial losses are introduced.

Components kept fixed. Throughout the Stable Diffusion v1.5 experiments, the VAE, text encoder, tokenizer, and prompt-conditioning pipeline are inherited directly from the pretrained backbone. We do not modify the conditioning architecture or introduce prompt engineering specifically designed to amplify memorization. Across different COCO mixing ratios, the backbone architecture, latent consistency distillation setup, solver choice for teacher target construction, and 4-step evaluation protocol are all kept fixed. This ensures that the main source of variation is the memorization pressure induced during teacher fine-tuning, followed by the same consistency distillation procedure.

B.3. Experimental Setup in Section 5

B.3.1. Ridge Regularization in Non-Isotropic Consistency Distillation

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

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

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,

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

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

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

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

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

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

$$\mathrm{fracBmem}_i = \frac{\sum_{k \in \mathcal{M}} \frac{y_{ki}^2}{\lambda_k + \gamma}}{\sum_{k=1}^{p} \frac{y_{ki}^2}{\lambda_k + \gamma}}. \tag{13}$$

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.

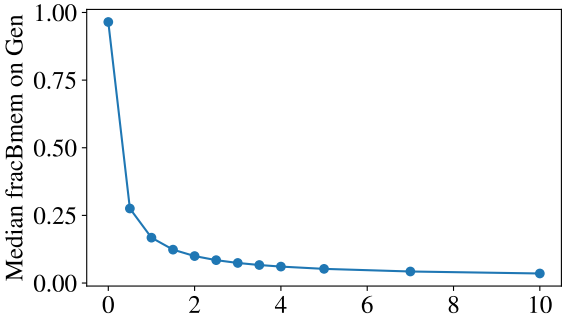


Figure 6. **Gen-to-Mem leakage after inverse-curvature weighting.** Median fracBmem on Gen (Eq. (13)) versus ridge γ. We choose $\gamma^\star$ by the minimal-sufficient rule in Eq. (14) with tolerance $\tau = 0.1$.

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:

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

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. 6 reports the ridge sweep of the proposed leakage statistic $\mathrm{fracBmem}$ (Eq. (13)) 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:

$$\gamma^\star = \min\Big\{\gamma > 0 :\; \mathrm{median}_{i\in\mathrm{GEN}}\big[\mathrm{fracBmem}_i(\gamma)\big] \le \tau\Big\}. \tag{14}$$

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. 6 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.

B.3.2. 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 5. 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.

B.3.3. 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

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

C. Proofs

C.1. Assumptions

Assumption C.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 C.2 (Activation moment, symmetry, and smoothness conditions)**.** The activation $\sigma : \mathbb{R} \to \mathbb{R}$ is odd, $\sigma(-z) = -\sigma(z)$, measurable, and almost everywhere differentiable. Moreover, it satisfies $\mathbb{E}_\zeta\left[(\sigma(g_i) - \sigma(g_i + \Delta g_i))^2 \,\middle|\, \mathbf{x}, \boldsymbol{\delta x}\right] < \infty$. 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 C.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 $\gamma(t') = 1 + \eta(t'), \kappa(t')^2 = 1 + 2\eta(t') + \upsilon(t')$.

C.2. Proof of Lemma 5.1

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 C.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 C.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

$$\mathbb{E}[RX \mid \mathbf{x}, \boldsymbol{\delta x}] = 0. \tag{15}$$

Moreover,

$$\begin{aligned} \mathbb{E}[R^2 \mid \mathbf{x}, \boldsymbol{\delta x}] &= \mathbb{E}\big[(Y - a_1 X)^2 \mid \mathbf{x}, \boldsymbol{\delta x}\big] \\ &= \mathbb{E}[Y^2 \mid \mathbf{x}, \boldsymbol{\delta x}] - a_1(\mathbf{x}, \boldsymbol{\delta x})^2\, \mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}] \\ &= a_0(\mathbf{x}, \boldsymbol{\delta x})\, \mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}], \end{aligned} \tag{16}$$

where $a_0(\mathbf{x}, \boldsymbol{\delta x})$ is defined in Lemma 5.1. 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

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

Since g_i and $g_i + \Delta g_i$ are centered Gaussian conditional on $(\mathbf{x}, \boldsymbol{\delta x})$, and since σ is odd by Assumption C.2, we have

$$\mathbb{E}[\sigma(g_i) \mid \mathbf{x}, \boldsymbol{\delta x}] = \mathbb{E}[\sigma(g_i + \Delta g_i) \mid \mathbf{x}, \boldsymbol{\delta x}] = 0.$$

Therefore,

$$\mathbb{E}[\Delta h_i \mid \mathbf{x}, \boldsymbol{\delta x}] = 0.$$

Moreover, Δg_i is centered conditional on $(\mathbf{x}, \boldsymbol{\delta x})$, and hence

$$\mathbb{E}[R_i \mid \mathbf{x}, \boldsymbol{\delta x}] = \mathbb{E}[\Delta h_i \mid \mathbf{x}, \boldsymbol{\delta x}] - a_1(\mathbf{x}, \boldsymbol{\delta x})\mathbb{E}[\Delta g_i \mid \mathbf{x}, \boldsymbol{\delta x}] = 0.$$

Since the rows of W are conditionally independent across i, for $i \neq j$,

$$\mathbb{E}[R_i R_j \mid \mathbf{x}, \boldsymbol{\delta x}] = \mathbb{E}[R_i \mid \mathbf{x}, \boldsymbol{\delta x}]\mathbb{E}[R_j \mid \mathbf{x}, \boldsymbol{\delta x}] = 0.$$

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 (15) 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 $\mathbb{E}[\mathbf{R}\Delta\mathbf{g}^\top|\,\mathbf{x},\boldsymbol{\delta x}] = 0$. 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 $\mathbb{E}[\mathbf{R}\mathbf{R}^\top|\,\mathbf{x},\boldsymbol{\delta x}] = \mathbb{E}[R^2 \mid \mathbf{x},\boldsymbol{\delta x}]\,\mathbf{I}_p$. Using (16) 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 (8).

Assume now the regime of Assumption C.3 with $\boldsymbol{\Sigma} = \mathbf{I}_d$ and the one-step PF-ODE update (3). 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 C.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

$$Y = -\sigma'(G)\Delta G - \rho, \qquad \rho = \frac{1}{2}\,\sigma''(G+\theta\Delta G)\,(\Delta G)^2. \tag{17}$$

Under the smoothness conditions in Assumption C.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 (17) and $X = \Delta G$,

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

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,

$$a_1(\mathbf{x},\boldsymbol{\delta x}) = \frac{\mathbb{E}[YX \mid \mathbf{x},\boldsymbol{\delta x}]}{\mathbb{E}[X^2 \mid \mathbf{x},\boldsymbol{\delta x}]} = -\frac{\mathbb{E}[\sigma'(G)Z^2]}{\kappa(t')^2} + o(1), \tag{18}$$

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

$$\mathbb{E}[\sigma'(G)Z^2] = \mathbb{E}\big[\sigma'(G)\,\mathbb{E}[Z^2 \mid G]\big] = \gamma(t')^2\,\mathbb{E}\big[\sigma'(G)G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big)\,\mathbb{E}\big[\sigma'(G)\big], \tag{19}$$

with $G \sim \mathcal{N}(0,1)$. Combining (18) and (19) yields

$$a_1(t') = -\,\frac{\gamma(t')^2\mathbb{E}_G\big[\sigma'(G)\,G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big)\mathbb{E}_G\big[\sigma'(G)\big]}{\kappa(t')^2}. \tag{20}$$

Limit of a_0. Similarly, using (17),

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

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

$$\mathbb{E}[\sigma'(G)^2 Z^2] = \gamma(t')^2\,\mathbb{E}\big[\sigma'(G)^2 G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big)\,\mathbb{E}\big[\sigma'(G)^2\big], \tag{21}$$

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

$$a_0(t') = \frac{\gamma(t')^2\mathbb{E}_G\big[\sigma'(G)^2 G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big)\mathbb{E}_G\big[\sigma'(G)^2\big]}{\kappa(t')^2} - a_1(t')^2. \tag{22}$$

This completes the proof. □

C.3. Proof of Theorem 5.3

Proof. Lemma 5.1 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

$$\mathbf{U}_{\mathrm{cd}} = \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\big[a_1(\mathbf{x},\delta\mathbf{x})^2\,\Delta\mathbf{g}\,\Delta\mathbf{g}^\top\big] + \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\left[a_0(\mathbf{x},\delta\mathbf{x})\,\frac{\|\delta\mathbf{x}\|_2^2}{d}\right]\mathbf{I}_p. \tag{23}$$

Under the small-noise one-step regime and the concentration hypotheses of Lemma 5.1, 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 (23) at leading order, obtaining

$$\mathbf{U}_{\mathrm{cd}} = a_1(t')^2\,\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\big[\Delta\mathbf{g}\,\Delta\mathbf{g}^\top\big] + \beta(t',\Delta t)\,\mathbf{I}_p + o(\Delta t^2), \tag{24}$$

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

$$\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi[\delta\mathbf{x}\,\delta\mathbf{x}^\top] = \Delta t^2\,\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_\xi\big[(\mathbf{x}+\mathbf{s}_\phi)(\mathbf{x}+\mathbf{s}_\phi)^\top\big]. \tag{25}$$

From (9) and the deterministic equivalent (10), the converged teacher score can be expressed as

$$\mathbf{s}_\phi(\mathbf{x},t') = -\mu_1(t')\,\frac{\mathbf{W}^\top}{\sqrt{d}}\,\mathbf{U}^{-1}\,\mathbf{h}(\mathbf{x}), \tag{26}$$

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

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\big[\mathbf{x}\,\mathbf{h}(\mathbf{x})^\top\big] \simeq \mu_1(t')\,\mathbf{B}^\top, \tag{27}$$

while by definition of $\mathbf{U} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\big[\mathbf{h}(\mathbf{x})\,\mathbf{h}(\mathbf{x})^\top\big]$, plugging (26) into the cross and score terms and using (27) gives

$$\begin{aligned}
&\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{x}\mathbf{s}_\phi^\top] \simeq -\mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}, \quad \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{s}_\phi\mathbf{x}^\top] \simeq -\mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B},\\
&\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{s}_\phi\mathbf{s}_\phi^\top] = \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\Big(\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{h}\mathbf{h}^\top]\Big)\mathbf{U}^{-1}\mathbf{B} = \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}.
\end{aligned} \tag{28}$$

Moreover, the isotropic OU marginal implies

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{x}\mathbf{x}^\top] \simeq \mathbf{I}_d. \tag{29}$$

Combining (28) and (29) yields:

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\big[(\mathbf{x}+\mathbf{s}_\phi)(\mathbf{x}+\mathbf{s}_\phi)^\top\big] \simeq \mathbf{I}_d - \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}. \tag{30}$$

Substituting (30) into (25) 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 analyze the isotropic shift in Eq. (24). Using the same OU PF-ODE displacement $\delta\mathbf{x} = -\Delta t\big(\mathbf{x} + \mathbf{s}_\phi(\mathbf{x}, t')\big)$, and the averaged second moment in Eq. (30), we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[\frac{\|\delta\mathbf{x}\|_2^2}{d}\right] &= \frac{\Delta t^2}{d}\,\mathrm{Tr}\left(\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[(\mathbf{x}+\mathbf{s}_\phi)(\mathbf{x}+\mathbf{s}_\phi)^\top\right]\right)\\
&\simeq \Delta t^2\left(1 - \frac{\mu_1(t')^2}{d}\mathrm{Tr}\big(\mathbf{U}^{-1}\mathbf{S}\big)\right).
\end{aligned} \tag{31}$$

Therefore, the isotropic shift is

$$\beta(t', \Delta t) = a_0(t')\Delta t^2\left(1 - \frac{\mu_1(t')^2}{d}\mathrm{Tr}\big(\mathbf{U}^{-1}\mathbf{S}\big)\right). \tag{32}$$

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

$$\mathbf{U}_{\mathrm{cd}} = \Delta t^2\,a_1(t')^2\Big(\mathbf{S} - \mu_1(t')^2\,\mathbf{S}\,\mathbf{U}^{-1}\,\mathbf{S}\Big) \;+\; \beta(t', \Delta t)\,\mathbf{I}_p, \tag{33}$$

This completes the proof.

□

Table 6. Comparison between stochastic EDM teachers and two-step consistency models under matched stochastic sampling settings on CIFAR-10. Since both methods use stochastic sampling, the memorization gap isolates the model effect rather than the sampler effect.

Dataset size	Sampler/Model	FID	l_2 Mem	SSCD Mem / p95
3000	Teacher stochastic	16.76	14.60%	39.08% / 0.8868
3000	Student 2-step	12.95	3.21%	39.55% / 0.7792
4000	Teacher stochastic	18.44	16.25%	39.92% / 0.8973
4000	Student 2-step	16.57	1.76%	27.07% / 0.7579
5000	Teacher stochastic	21.81	12.50%	32.24% / 0.8796
5000	Student 2-step	19.52	0.60%	15.10% / 0.7064
6000	Teacher stochastic	22.97	13.26%	33.12% / 0.8901
6000	Student 2-step	21.23	0.48%	11.11% / 0.6783

D. Additional Results

D.1. Disentangling Model and Sampler Effects

To determine whether the observed reduction in memorization should be attributed to the distilled model itself or to the sampling procedure, we consider two complementary comparisons. The first is reported in the main text: Table 1 compares the 1-step student with deterministic EDM sampling, which is the evaluation setting most directly aligned with our theoretical analysis of the one-step consistency objective. The results there already show substantial memorization suppression in the 1-step setting, indicating that the effect is not a consequence of multi-step inference.

We further examine this question under matched stochastic sampling. Specifically, we compare two-step consistency models with EDM teachers under stochastic sampling and otherwise identical settings. Because both models are evaluated with stochastic inference, the resulting memorization gap cannot be explained by differences between ODE and stochastic sampling. As shown in Table 6, the two-step consistency model consistently exhibits substantially lower memorization than the stochastic EDM teacher under both l_2 and SSCD metrics, while also achieving better FID across all settings. This comparison therefore isolates the contribution of the distilled model and provides further evidence that the observed memorization reduction is not induced by the sampler.

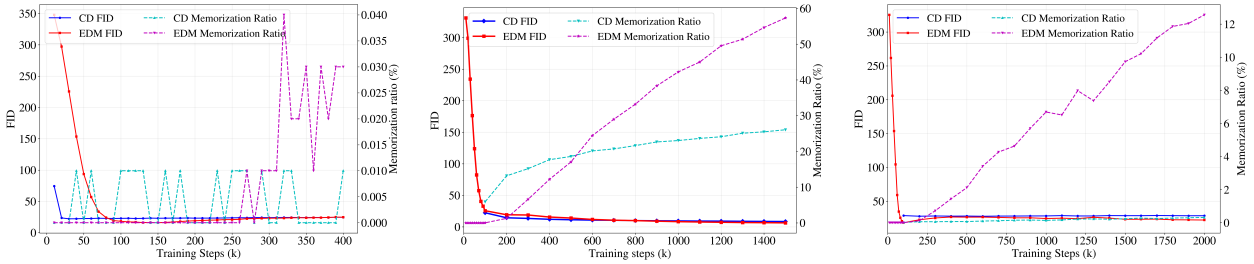
D.2. Effect of Model Capacity on Memorization and Distillation

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, as it yields better generation quality while exhibiting the same qualitative trends as the one-step setting. 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 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



(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 7. **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.

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.

D.3. Effect of student–teacher architectural mismatch and initialization

Fig. 8 examines the effect of architectural mismatch and initialization strategy on consistency distillation under teachers with approximately 10% and 30% memorization, measured by the l_2 metric. All experiments still 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 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.

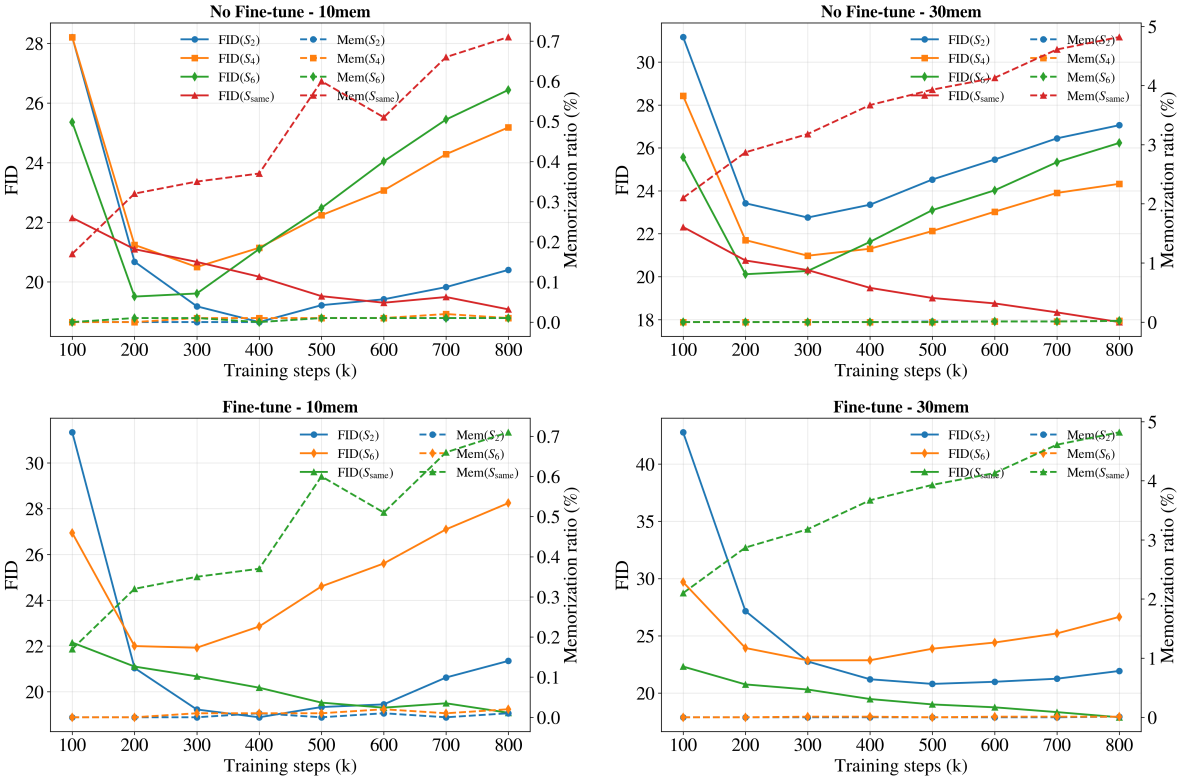


Figure 8. **Effect of student–teacher architectural mismatch on memorization and sample quality under two-step consistency distillation.** The figure reports FID (left y-axis, solid lines) and memorization ratio (right y-axis, dashed lines) as functions of training steps for different student–teacher 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.

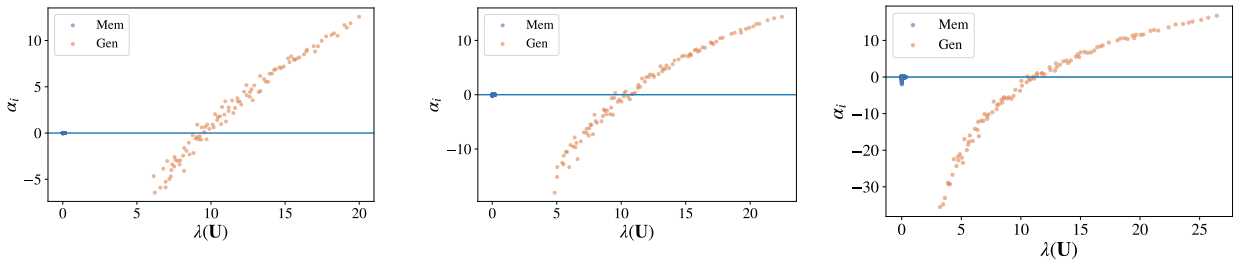
Consequently, these results highlight a tradeoff controlled by architectural compatibility. Exact architectural matching enables stable quality improvement but allows partial transfer of teacher memorization, whereas mismatch or random initialization strongly suppresses memorization but can suffer from 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.

D.4. Why Severely Memorizing Teachers Degrade Student Generation Quality

As observed in Section 4.4, distillation from severely memorizing teachers can lead to noticeably degraded student FID. In this subsection, we provide a theoretical explanation for this phenomenon through the lens of our mode-wise analysis. In particular, we use the data–feature scaling ratio ψ_p/ψ_n to characterize the memorization regime underlying this behavior.

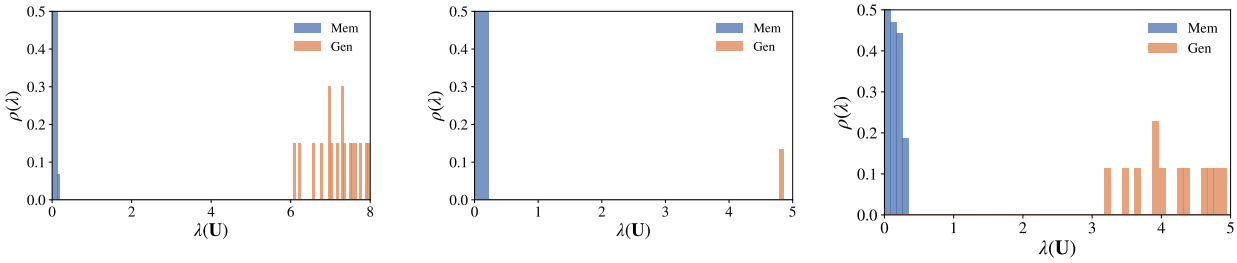
To understand how the ratio ψ_p/ψ_n influences the mode-wise behavior of non-isotropic consistency distillation, we examine how the signed response α_i distributes across teacher curvature modes under different ψ_p/ψ_n configurations. Fig. 9 shows that, for fixed ψ_p, decreasing ψ_n (and hence increasing the ratio ψ_p/ψ_n) weakens the positive non-isotropic response over the Gen spectrum. In particular, a larger fraction of low-curvature Gen modes falls into the suppressive regime with $\alpha_i < 0$, while positive responses become increasingly concentrated in the high-curvature tail of the Gen spectrum. Across all configurations, memorization-associated (Mem) modes remain tightly concentrated near $\alpha_i \approx 0$, indicating that the non-isotropic CD term does not allocate substantial positive response to memorization-dominated directions.

Fig. 10 further shows that this behavior is closely tied to the spectral geometry of the teacher curvature matrix $\mathbf{U}$.



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

Figure 9. **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, equivalently increasing ψ_p/ψ_n. As ψ_n decreases, a larger fraction of low-curvature GEN modes falls into the suppressive regime with negative signed response $\alpha_i < 0$. Positive responses become increasingly concentrated in the high-curvature tail of the GEN spectrum, while MEM modes remain tightly concentrated near $\alpha_i \simeq 0$.



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

Figure 10. **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 broader positive response over the GEN spectrum, whereas weaker separation causes more low-curvature GEN modes to fall into the suppressive regime.

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 a substantial portion of high-curvature Gen modes can still receive positive response. 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 explains why, under severe memorization regimes, consistency distillation can still suppress memorization but may fail to improve generation quality.

D.5. Effect of Discretization Granularity

To clarify whether the observed memorization reduction in consistency distillation is merely an artifact of coarse temporal discretization, we study the effect of the discretization granularity N used during distillation. This analysis is important for distinguishing an intrinsic filtering effect of the consistency objective from a purely temporal effect caused by insufficient resolution of the teacher trajectory.

As shown in Table 7, increasing N consistently improves the student performance while further reducing memorization. This observation suggests that the reduction of memorization is not simply due to a coarse discretization failing to capture the highly curved late-stage dynamics of the teacher. Instead, finer discretization makes the local consistency constraints more faithful to the underlying trajectory, yet does not restore the teacher's memorized behavior. Empirically, better trajectory resolution leads to both improved sample quality and lower memorization, which supports the view that the filtering effect arises from the consistency distillation objective itself rather than from an accidental temporal bias induced by an overly coarse schedule.

Table 7. Effect of the discretization granularity N in consistency distillation.

Setting	FID	l_2 Mem	SSCD Mem / p95
Teacher	22.68	10%	26.93% / 0.8586
Student $N = 12$	29.14	0.18%	4.88% / 0.5974
Student $N = 18$	23.76	0.10%	4.36% / 0.5892
Student $N = 36$	17.62	0.02%	2.48% / 0.5544

E. Broader Impacts

This work studies memorization and generalization in distilled diffusion models. On the positive side, understanding how distillation affects memorization may help develop generative models with improved reliability, reduced training-data leakage, and better deployment efficiency. On the negative side, diffusion models may still be misused for generating harmful, misleading, or privacy-sensitive content, and reduced memorization does not eliminate these risks. Our analysis is intended to support safer model development and should be considered together with appropriate data governance, model release policies, and misuse mitigation measures.

F. Limitations

Although our experiments cover multiple datasets, model settings, and distillation configurations, the observed memorization behavior may still depend on factors such as data scale, teacher memorization level, model capacity, sampling strategy, and the specific distillation objective. A broader study across more architectures, datasets, and deployment scenarios would further strengthen the generality of our findings. In addition, reducing memorization does not eliminate all risks associated with diffusion models. Distilled models may still be misused to generate harmful, misleading, or privacy-sensitive content, and lower memorization does not by itself guarantee safe deployment. Responsible deployment therefore requires appropriate data governance, model release policies, and misuse mitigation measures.