On the Memorization of Consistency Distillation for Diffusion Models

Bingqing Jiang[1,2] **Difan Zou**[1,2*]
[1]School of Computing and Data Science, The University of Hong Kong
[2]Shenzhen Loop Area Institute
bingqjiang@gmail.com dzou@hku.hk

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 [1], 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 [2, 3, 4, 5, 6, 7]. 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 [8]. Given their growing importance, understanding how diffusion models balance memorization and generalization has become a fundamental question [9, 10, 11, 12, 13]. Recent studies show that memorization in diffusion models is governed by training dynamics [1, 14, 15]. 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 [16].

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 [17, 18, 19, 20, 21]. This raises a critical question that has received little attention to date:

How does distillation affect the memorization properties of diffusion models?

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

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 [22]. If memorization in diffusion models depends sensitively on training dynamics and time scales, as prior work suggests, then distillation may further reshape, suppress, or even amplify memorization inherited from the teacher. Yet, the effect of distillation on memorization remains largely uncharacterized.

Recent progress toward few-step or even single-step diffusion generation has largely relied on distillation-based approaches [23, 24, 25, 26, 27], among which consistency models have emerged as a representative and widely adopted framework [17]. 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 [28, 29], 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 examined memorization in diffusion models [30, 10, 31]. The denoising score matching objective admits empirical minimizers that reproduce training samples, highlighting the risk of memorization under weak regularization or limited data [9, 32]. Memorization and generalization can exhibit sharp transitions with dataset size, model capacity, and training dynamics [33, 34, 15]. Exact high-dimensional learning curves have been derived for random-feature diffusion models [14], revealing implicit dynamical regularization that separates the onset of generalization from memorization [1]. Extraction studies further show that explicit or surrogate conditioning can amplify memorization risks [35]. However, existing work primarily studies diffusion models trained from scratch, leaving distilled models underexplored. We investigate how 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 [17, 36]. Subsequent work extends this framework to improve quality–speed trade-offs, stabilize training, and provide theoretical guarantees on estimation, discretization, and convergence [19, 37, 38, 39, 40, 41]. 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 $\{\mathbf{x}_\nu\}_{\nu=1}^n$ [1, 42, 43]:

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

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 [17]. 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.

4 Memorization and Generation Quality in Consistency Distillation

4.1 Experimental Setup

Datasets. We evaluate consistency distillation on unconditional CIFAR-10 [44], class-conditional ImageNet [45], and text-to-image generation with Stable Diffusion v1.5 [46]. For CIFAR-10, we

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	18.23	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

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 [13, 10]. 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 $3\times$, $4\times$, or $5\times$ 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 [9].*

Training configuration. For CIFAR-10 and ImageNet, we use pre-trained EDM models as teachers [6] and initialize students from the same checkpoints before consistency distillation [17]. We use LPIPS [47] 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 [48], 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 [49] 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 [10]. 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.

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

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	Repeated classes	Non-repeated classes	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

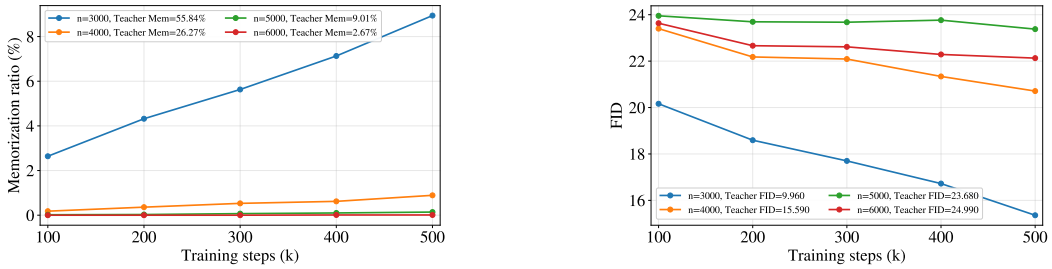


Figure 1: **Student dynamics during consistency distillation on CIFAR-10. Left**: memorization ratio. **Right**: FID. Consistency distillation reliably reduces memorization, while FID improvement depends on the memorization state of the teacher.

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

To further understand the dynamics of consistency distillation, we examine how memorization and generation quality evolve 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.

Table 5: Memorization and held-out coverage in the localized-memorization settings. Anchor coverage reports covered/supported anchor cells. Rare and PRDC coverage are evaluated on held-out images in DINOv2 feature space. SSCD memorization is measured against repeated training images.

Dataset	Model	SSCD Mem(%)	Anchor coverage	Rare coverage(%)	PRDC coverage(%)
CIFAR-10	Teacher	74.54	98/98	57.50	62.72
	CD student	33.51	98/98	58.33	57.39
ImageNet	Teacher	7.35	180/180	98.00	98.89
	CD student	0.00	179/180	96.00	95.33

4.5 Memorization and Held-Out Coverage

Lower similarity to training images does not by itself establish that a model preserves the underlying distribution: a reduction in coverage could also lower measured memorization. We therefore complement the localized-memorization experiments with a held-out coverage analysis. For CIFAR-10, we use the disjoint test set. For ImageNet, we use 450 non-overlapping held-out images balanced across the 50 classes containing repeated training examples. We generate 10,000 samples per model and evaluate coverage in DINOv2 feature space. For this analysis, SSCD memorization measures the fraction of evaluated samples whose similarity to a repeated training image exceeds 0.6. We consider three complementary coverage measures. First, repeated training images define class-conditional anchor cells. A cell is supported if it contains held-out images and covered if it also contains generated samples; we report the number of covered cells among the supported cells. Repeated training images define the cells but do not count as held-out support. Second, held-out PRDC coverage measures the fraction of held-out images whose 5-nearest-neighbor neighborhood contains a generated sample. Third, rare-region coverage applies the same criterion to low-density regions of the held-out data. Table 5 shows that the substantial reduction in repeated-image memorization is accompanied by retention of all supported anchor cells on CIFAR-10 and 179 of 180 supported cells on ImageNet. Rare-region coverage changes from 57.50% to 58.33% on CIFAR-10 and from 98.00% to 96.00% on ImageNet. These observations suggest that the reduction in memorization is not solely attributable to wholesale loss of the regions associated with repeated training examples. Thus, these results do not establish that distributional coverage is unchanged. Rather, they indicate that substantially lower recall of repeated training images can coexist with broad retention of the evaluated held-out regions.

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 [17], 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 [14, 1, 50]. For tractability, we omit the EMA target and stop-gradient in Eq. (5) and analyze the induced symmetric local consistency penalty [39, 41]. 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 [12, 1, 14], 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 $\mathbf{h}(\mathbf{x}) = \sigma\Big(\frac{\mathbf{W}\mathbf{x}}{\sqrt{d}}\Big) \in \mathbb{R}^p$. 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 $\boldsymbol{\delta x}(\mathbf{x}, t', t) = \widehat{\mathbf{x}}_t^{\phi} - \mathbf{x}_{t'} = \Delta t\,\mathbf{v}_\phi(\mathbf{x}, t')$. 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 [14, 1], 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).

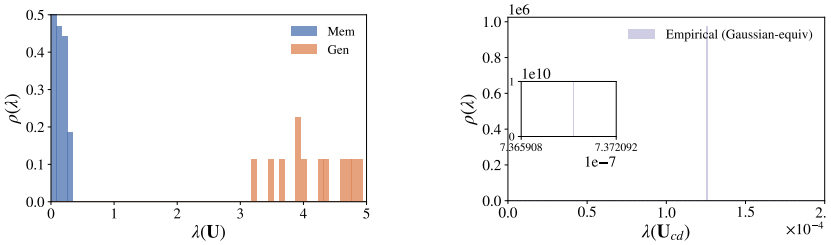
Lemma 5.1 (Orthogonal second-moment decomposition of $\Delta\mathbf{h}$)**.** *Under Assumptions C.1 and C.2, define the scalar coefficients* $a_1(\mathbf{x}, \boldsymbol{\delta x}) = \frac{\mathbb{E}_\zeta[\Delta h_i\,\Delta g_i|\mathbf{x},\boldsymbol{\delta x}]}{\mathbb{E}_\zeta\big[\Delta g_i^2|\mathbf{x},\boldsymbol{\delta x}\big]}$ *and* $a_0(\mathbf{x}, \boldsymbol{\delta x}) = \frac{\mathbb{E}_\zeta\big[\Delta h_i^2|\mathbf{x},\boldsymbol{\delta x}\big]}{\mathbb{E}_\zeta\big[\Delta g_i^2|\mathbf{x},\boldsymbol{\delta x}\big]} - a_1(\mathbf{x}, \boldsymbol{\delta x})^2$. *Then the conditional second moment of the feature increment admits the decomposition*

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

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. (22) and Eq. (24) 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 [1].



(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.

Lemma 5.2 ([1])**.** *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 [1]) *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 [1]) *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),* $\mu_1(t')=\Gamma_{t'}\,\mathbb{E}_{Z\sim\mathcal{N}(0,\Gamma_{t'}^2)}[\sigma'(Z)]$.

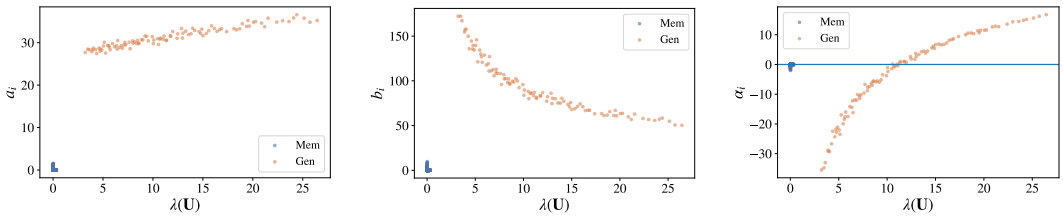
Combining Lemma 5.1 with the teacher characterization in Lemma 5.2, we obtain the following structural characterization of the consistency distillation curvature.

Theorem 5.3. *Let $\{\mathbf{x}_\nu\}_{\nu=1}^n \subset \mathbb{R}^d$ be training samples with $\mathrm{Cov}(\mathbf{x}_\nu) = \boldsymbol{\Sigma} = \mathbf{I}_d$. At a fixed diffusion time $t' > 0$, draw $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}_d)$ and define the OU forward sample, and assume the small-noise one-step regime and the deterministic limits $a_1(t')$, $a_0(t')$ from Lemma 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. (34). 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



(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 $\rho_\Sigma(\lambda) = \delta(\lambda - 1)$.

system for understanding how consistency distillation redistributes curvature. As established in [1], 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

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

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 $\mathcal{I}_{\mathrm{mem}} = \{i : \lambda_i(\mathbf{U}) < \lambda_{\mathrm{th}}\}, \mathcal{I}_{\mathrm{gen}} = \{i : \lambda_i(\mathbf{U}) \ge \lambda_{\mathrm{th}}\}$ [1].

(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

Table 6: Last-layer spectral diagnostics in practical EDM/CD U-Nets. The localized setting compares mean directional CD energy across operationally defined Gen and Mem modes.

Setting	Statistic	Result
Localized CIFAR-10	Mean Gen ρ_i / mean Mem ρ_i	1.87
Standard 5k CIFAR-10	Spearman correlation of ρ_i and λ_i	0.994

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 $\mathrm{Share}^{+}_{\mathrm{mem}} = \frac{\sum_{i\in\mathcal{I}_{\mathrm{mem}}} \max(\alpha_i,0)}{\sum_{i=1}^{p} \max(\alpha_i,0)}$. This metric captures global positive-curvature allocation and is robust to isolated atypical modes. Empirically, $\mathrm{Share}^{+}_{\mathrm{mem}} \approx 9.5 \times 10^{-3}$, 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.

5.5 Last-Layer Spectral Diagnostics in U-Nets

To examine whether the mode-dependent curvature allocation suggested by the random-feature analysis has a counterpart in practical networks, we perform a last-layer diagnostic on the CIFAR-10 EDM/CD NCSN++ U-Nets. For each adjacent Heun pair, we compute the final-layer feature change ϕ_{cd} and construct its second-moment operator $U_{\mathrm{cd}}^{\mathrm{UNet}} = \mathbb{E}\big[\phi_{\mathrm{cd}}\phi_{\mathrm{cd}}^{\top}\big]$. For a unit-norm teacher-curvature direction u_i with associated curvature λ_i, we measure $\rho_i = \mathbb{E}\Big[\big(u_i^{\top}\phi_{\mathrm{cd}}\big)^2\Big] = u_i^{\top} U_{\mathrm{cd}}^{\mathrm{UNet}} u_i$. This quantity measures the CD feature-change energy along each teacher-curvature direction. We evaluate two complementary settings. In the localized CIFAR-10 setting, repeated classes 0–1 and non-repeated classes 2–9 provide disjoint sets for discovering Mem and Gen modes, respectively. We compare the mean ρ_i across these two mode groups. In the standard 5,000-image CIFAR-10 setting, we impose no explicit Mem/Gen partition and instead measure the Spearman rank correlation between ρ_i and λ_i. As reported in Table 6, the mean directional CD energy of Gen modes is 1.87 times that of Mem modes in the localized setting. In the standard setting, a Spearman correlation of 0.994 indicates strong rank alignment between directional CD energy and teacher curvature. These last-layer diagnostics provide preliminary empirical evidence consistent with the mode-dependent curvature allocation suggested by the random-feature analysis.

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.

Acknowledgments

We sincerely thank the reviewers and the area chair for their careful reading, constructive feedback, and thoughtful suggestions, which helped us improve the clarity and presentation of this work. This work is supported in part by NSFC 62306252, Hong Kong GRF award 17312526, Guangdong NSF 2024A1515012444, and the central fund from HKU.

References

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

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

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

[4] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

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

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

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

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

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

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

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

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

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

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

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

[16] Hanyu Wang, Yujin Han, and Difan Zou. On the discrepancy and connection between memorization and generation in diffusion models. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[31] Zekai Zhang, Xiao Li, Xiang Li, Lianghe Shi, Meng Wu, Molei Tao, and Qing Qu. Generalization of diffusion models arises with a balanced representation space. *arXiv preprint arXiv:2512.20963*, 2025.

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

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

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

[35] Yunhao Chen, Shujie Wang, Difan Zou, and Xingjun Ma. SIDE: Surrogate conditional data extraction from diffusion models, 2025.

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

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

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

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

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

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

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

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

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

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

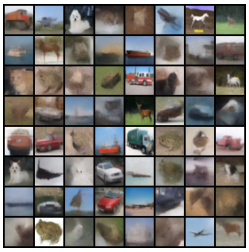
[48] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

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

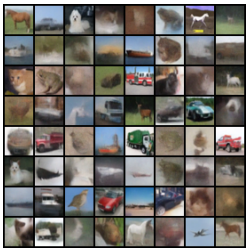
[50] Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[51] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

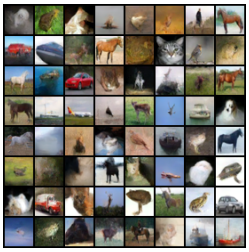
[52] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.



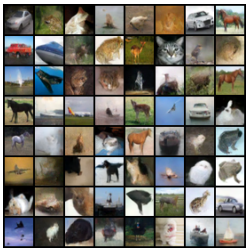
(a) PD, 10%-mem teacher



(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) [18], 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.

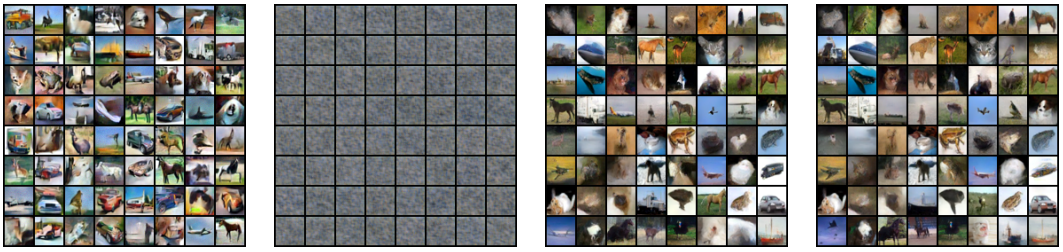
A critical hyperparameter in PD is the *initial noise resolution*, controlled by `start_scales`. Setting `start_scales=4096` corresponds to initializing distillation from a very fine-grained discretization of the diffusion process, i.e., a large number of teacher sampling steps. This choice ensures that 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.

Table 7: 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.

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-suppression effect extends beyond this trajectory-based setting [51]. 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 [6]. 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 [17]. 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 [4], following the standard architectural choice in prior consistency distillation work [17]. For ImageNet, we use the class-conditional DDPM++ architecture [52], consistent with the ImageNet setup in Consistency Models [17]. 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 [17] 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 [17, 6]. No data augmentation or additional regularization is used, in order to avoid potential confounding effects in the assessment of memorization [9].

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 [48], 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

For the finite-dimensional spectral diagnostic, we use a ridge-regularized counterpart of the non-isotropic response operator derived from the local probability-flow ODE:

$$\mathbf{A}_\gamma = \mathbf{S} - \mu_1^2 \, \mathbf{S} \, (\mathbf{U} + \gamma \mathbf{I})^{-1} \, \mathbf{S}, \tag{13}$$

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 unregularized inverse $\mathbf{U}^{-1}$ arises in the theoretical derivation, whereas $\gamma > 0$ is introduced only in the finite-dimensional diagnostic. Ridge regularization is absent from the CD training objective and Theorem 5.3, and does not affect the trained models or their reported memorization rates.

In the setting considered here, the spectrum of the teacher curvature $\mathbf{U}$ is highly ill-conditioned: a large fraction of its eigenvalues concentrate near zero, corresponding to memorization-associated directions, while larger eigenvalues correspond to generalization-associated 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 strong 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-associated components to dominate the inverse-weighted contribution even when the input direction itself lies in a generalization-associated region. This effect follows from inverse spectral weighting and is not merely a floating-point numerical artifact. In the inverse-curvature quadratic form

$$b_i = (\mathbf{S}\mathbf{u}_i)^\top (\mathbf{U} + \gamma \mathbf{I})^{-1} (\mathbf{S}\mathbf{u}_i), \tag{14}$$

where $\mathbf{u}_i$ is an eigenvector of $\mathbf{U}$, the contribution from memorization directions can overwhelm that from generalization directions. As a result, response ratios $r_i = \mu_1^2 b_i / (a_i + \varepsilon)$, where $a_i = \mathbf{u}_i^\top \mathbf{S}\mathbf{u}_i$ and $\varepsilon > 0$ stabilizes the denominator, can become large even for modes associated with large $\lambda(\mathbf{U})$, producing negative directional responses dominated by near-null contributions.

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 index set of memorization-associated modes, defined by small eigenvalues of $\mathbf{U}$. We define the inverse-weighted memorization leakage as

$$\mathrm{fracBmem}_i = \frac{\sum_{k \in \mathcal{M}} \frac{y_{ki}^2}{\lambda_k + \gamma}}{\sum_{k=1}^{p} \frac{y_{ki}^2}{\lambda_k + \gamma}}. \tag{15}$$

This quantity measures how much of the inverse-curvature quadratic form for mode i originates from memorization directions. Without ridge regularization, the median $\mathrm{fracBmem}_i$ over generalization modes is 96.48%, indicating severe cross-subspace leakage. At $\gamma = 0$, the within-Gen Spearman correlation between the directional response and teacher curvature is 0.955, but all Gen modes have negative directional responses. This behavior makes an interpretation based solely on response signs unreliable.

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

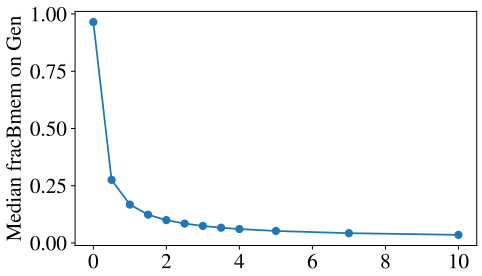


Figure 6: **Gen-to-Mem leakage after inverse-curvature weighting.** Median $\mathrm{fracBmem}$ on Gen (Eq. (15)) versus ridge γ. We choose $\gamma^\star$ by the minimal-sufficient rule in Eq. (16) with tolerance $\tau = 0.1$.

1. It bounds the maximum amplification of small-eigenvalue directions by $1/\gamma$.
2. It reduces inverse-weighted leakage from memorization directions in the evaluated diagnostic.

Ridge regularization does not define the Mem/Gen partition or create the structural coupling between $\mathbf{S}$ and $\mathbf{U}$. Nevertheless, it attenuates the inverse-weighted contribution of near-null directions and can therefore change the magnitudes and signs of individual mode-wise responses.

Rather than selecting γ to enforce a particular sign of the response, we adopt a leakage-based criterion:

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

where GEN denotes the index set of generalization-associated modes and τ is a small constant (e.g., 0.1 or 0.2). This criterion controls the median inverse-weighted memorization contribution over Gen modes, rather than bounding the leakage of every individual mode.

Fig. 6 reports the ridge sweep of the leakage statistic $\mathrm{fracBmem}$ (Eq. (15)) 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 the median leakage over the evaluated range.

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

$$\gamma^\star = \min\Big\{\gamma > 0 :\ \mathrm{median}_{i\in\mathrm{GEN}}\big[\mathrm{fracBmem}_i(\gamma)\big] \le \tau\Big\}. \tag{16}$$

In our experiments, setting $\tau = 0.1$ yields $\gamma^\star = 1.991$, rounded to $\gamma = 2$. For reference, 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 diagnostic ($\gamma = 0$) yields near-total median Gen-to-Mem leakage after inverse-curvature weighting. Increasing γ rapidly reduces this effect; beyond $\gamma \approx 2$, the leakage curve flattens.

Sensitivity to the leakage tolerance. We vary τ to assess whether the aggregate patterns depend on the diagnostic regularization. For each selected $\gamma^\star$, we evaluate the directional curvature correction $\alpha_i = \mathbf{u}_i^\top \mathbf{A}_{\gamma^\star}\mathbf{u}_i$. We classify $\alpha_i \approx 0$ when $|\alpha_i| \le 10^{-8}$, and define Gen positive mass as

$$\frac{\sum_{i\in\mathrm{GEN}}[\alpha_i]_+}{\sum_i[\alpha_i]_+}, \qquad [x]_+ = \max(x, 0).$$

Table 8 shows that positive directional curvature corrections remain concentrated on Gen modes across the tested leakage tolerances. All modes in the highest-curvature Gen quartile exhibit positive corrections, while Gen modes account for 99.05%–99.08% of the total positive response mass. Furthermore, the within-Gen Spearman correlation between α_i and teacher curvature remains between 0.9974 and 0.9976, indicating consistently strong rank alignment. Together, these results show that the concentration of positive response mass on Gen modes and its within-Gen rank alignment with teacher curvature are robust across the evaluated regularization range.

Table 8: Sensitivity of the spectral diagnostic to the leakage tolerance τ. Sign fractions are reported as percentages. "Top Gen quartile" denotes the positive-response fraction among the highest-curvature quartile of Gen modes.

τ	$\gamma^\star$	Gen$+/-$ (%)	Top Gen quartile positive (%)	Mem$+/-/\approx 0$ (%)	Gen positive mass (%)
0.20	0.787	40/60	100	1.10/11.81/87.10	99.08
0.15	1.160	42/58	100	1.35/11.55/87.10	99.08
0.125	1.475	45/55	100	1.45/11.45/87.10	99.06
0.10	1.991	48/52	100	1.68/11.23/87.10	99.05

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 $\lambda_{\text{th}} = 2$, with $\lambda_i(\mathbf{U}) < \lambda_{\text{th}}$ classified as Mem and $\lambda_i(\mathbf{U}) \geq \lambda_{\text{th}}$ 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 $\Gamma_d^2 = \mathrm{Var}(g_i \,|\mathbf{x}) = \frac{\|\mathbf{x}\|_2^2}{d}$, $\Delta_d^2 = \mathrm{Var}(\Delta g_i \,|\mathbf{x}, \boldsymbol{\delta x}) = \frac{\|\boldsymbol{\delta x}\|_2^2}{d}$, $c_d = \mathrm{Cov}(g_i, \Delta g_i \,|\mathbf{x}, \boldsymbol{\delta x}) = \frac{\mathbf{x}^\top \boldsymbol{\delta x}}{d}$. 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 $\mathbf{\Sigma} = I_d$ 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

$$\mathbb{E}[RX \mid \mathbf{x}, \boldsymbol{\delta x}] = 0. \tag{17}$$

Moreover,

$$\begin{aligned}\mathbb{E}[R^2 \mid \mathbf{x}, \boldsymbol{\delta x}] &= \mathbb{E}\big[(Y - a_1 X)^2 \mid \mathbf{x}, \boldsymbol{\delta x}\big] \\ &= \mathbb{E}[Y^2 \mid \mathbf{x}, \boldsymbol{\delta x}] - a_1(\mathbf{x}, \boldsymbol{\delta x})^2\, \mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}] \\ &= a_0(\mathbf{x}, \boldsymbol{\delta x})\, \mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}], \end{aligned} \tag{18}$$

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 $i \in \{1, \dots, p\}$, 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 (17) applied coordinate-wise and independence across i, the cross terms vanish: $\mathbb{E}[\Delta\mathbf{g}\mathbf{R}^\top \mid \mathbf{x}, \boldsymbol{\delta x}] = 0$ and $\mathbb{E}[\mathbf{R}\Delta\mathbf{g}^\top \mid \mathbf{x}, \boldsymbol{\delta x}] = 0$. Furthermore, since the coordinates are i.i.d. and R_i has conditional second moment $\mathbb{E}[R_i^2 \mid \mathbf{x}, \boldsymbol{\delta x}] = \mathbb{E}[R^2 \mid \mathbf{x}, \boldsymbol{\delta x}]$, we have $\mathbb{E}[\mathbf{R}\mathbf{R}^\top \mid \mathbf{x}, \boldsymbol{\delta x}] = \mathbb{E}[R^2 \mid \mathbf{x}, \boldsymbol{\delta x}]\, \mathbf{I}_p$. Using (18) and $\mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}] = \mathbb{E}[\Delta g_i^2 \mid \mathbf{x}, \boldsymbol{\delta x}]$ 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

$$Y = -\sigma'(G)\Delta G - \rho, \qquad \rho = \frac{1}{2}\, \sigma''(G + \theta\Delta G)\, (\Delta G)^2. \tag{19}$$

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 (19) 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,

$$a_1(\mathbf{x}, \boldsymbol{\delta x}) = \frac{\mathbb{E}[YX \mid \mathbf{x}, \boldsymbol{\delta x}]}{\mathbb{E}[X^2 \mid \mathbf{x}, \boldsymbol{\delta x}]} = -\frac{\mathbb{E}[\sigma'(G) Z^2]}{\kappa(t')^2} + o(1), \tag{20}$$

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

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

with $G \sim \mathcal{N}(0, 1)$. Combining (20) and (21) yields

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

Limit of a_0. Similarly, using (19),

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

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

$$\mathbb{E}[\sigma'(G)^2 Z^2] = \gamma(t')^2\, \mathbb{E}\big[\sigma'(G)^2 G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big)\, \mathbb{E}\big[\sigma'(G)^2\big], \tag{23}$$

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

$$a_0(t') = \frac{\gamma(t')^2 \mathbb{E}_G\big[\sigma'(G)^2 G^2\big] + \big(\kappa(t')^2 - \gamma(t')^2\big) \mathbb{E}_G\big[\sigma'(G)^2\big]}{\kappa(t')^2} - a_1(t')^2. \tag{24}$$

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

$$\mathbf{U}_{\mathrm{cd}} = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_\xi\big[a_1(\mathbf{x}, \delta\mathbf{x})^2\, \Delta\mathbf{g}\, \Delta\mathbf{g}^\top\big] + \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_\xi\left[a_0(\mathbf{x}, \delta\mathbf{x})\, \frac{\|\delta\mathbf{x}\|_2^2}{d}\right] \mathbf{I}_p. \tag{25}$$

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 (25) at leading order, obtaining

$$\mathbf{U}_{\mathrm{cd}} = a_1(t')^2\, \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_\xi\big[\Delta\mathbf{g}\, \Delta\mathbf{g}^\top\big] + \beta(t', \Delta t)\, \mathbf{I}_p + o(\Delta t^2), \tag{26}$$

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

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\delta\mathbf{x}\,\delta\mathbf{x}^\top] = \Delta t^2\,\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[(\mathbf{x} + \mathbf{s}_\phi)(\mathbf{x} + \mathbf{s}_\phi)^\top\right]. \tag{27}$$

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

$$\mathbf{s}_\phi(\mathbf{x}, t') = -\mu_1(t')\,\frac{\mathbf{W}^\top}{\sqrt{d}}\,\mathbf{U}^{-1}\,\mathbf{h}(\mathbf{x}), \tag{28}$$

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

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[\mathbf{x}\,\mathbf{h}(\mathbf{x})^\top\right] \simeq \mu_1(t')\,\mathbf{B}^\top, \tag{29}$$

while by definition of $\mathbf{U} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[\mathbf{h}(\mathbf{x})\,\mathbf{h}(\mathbf{x})^\top\right]$, plugging (28) into the cross and score terms and using (29) gives

$$\begin{aligned}
&\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{x}\mathbf{s}_\phi^\top] \simeq -\mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}, \quad \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{s}_\phi\mathbf{x}^\top] \simeq -\mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}, \\
&\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{s}_\phi\mathbf{s}_\phi^\top] = \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\Big(\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{h}\mathbf{h}^\top]\Big)\mathbf{U}^{-1}\mathbf{B} = \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}.
\end{aligned} \tag{30}$$

Moreover, the isotropic OU marginal implies

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi[\mathbf{x}\mathbf{x}^\top] \simeq \mathbf{I}_d. \tag{31}$$

Combining (30) and (31) yields:

$$\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[(\mathbf{x} + \mathbf{s}_\phi)(\mathbf{x} + \mathbf{s}_\phi)^\top\right] \simeq \mathbf{I}_d - \mu_1(t')^2\,\mathbf{B}^\top\mathbf{U}^{-1}\mathbf{B}. \tag{32}$$

Substituting (32) into (27) 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. (26). 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. (32), we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[\frac{\|\delta\mathbf{x}\|_2^2}{d}\right] &= \frac{\Delta t^2}{d}\,\mathrm{Tr}\left(\frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\left[(\mathbf{x} + \mathbf{s}_\phi)(\mathbf{x} + \mathbf{s}_\phi)^\top\right]\right) \\
&\simeq \Delta t^2\left(1 - \frac{\mu_1(t')^2}{d}\mathrm{Tr}\big(\mathbf{U}^{-1}\mathbf{S}\big)\right).
\end{aligned} \tag{33}$$

Table 9: Stochastic sampling comparisons between EDM teachers and two-step consistency-distilled students on CIFAR-10. The student achieves lower memorization and better FID across all dataset sizes. This comparison controls for the broad sampling regime, but not for NFE.

Dataset size	Model	FID	ℓ_2 Mem (%)	SSCD Mem (%)	SSCD p95
3000	Teacher	16.76	14.60	39.08	0.8868
3000	Student (2-step)	12.95	3.21	29.55	0.7792
4000	Teacher	18.44	16.25	39.92	0.8973
4000	Student (2-step)	16.57	1.76	27.07	0.7579
5000	Teacher	21.81	12.50	32.24	0.8796
5000	Student (2-step)	19.52	0.60	15.10	0.7064
6000	Teacher	22.97	13.26	33.12	0.8901
6000	Student (2-step)	21.23	0.48	11.11	0.6783

Therefore, the isotropic shift is

$$\beta(t', \Delta t) = a_0(t')\Delta t^2 \left(1 - \frac{\mu_1(t')^2}{d} \mathrm{Tr}\big(\mathbf{U}^{-1}\mathbf{S}\big)\right). \tag{34}$$

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

$$\mathbf{U}_{\mathrm{cd}} = \Delta t^2 \, a_1(t')^2 \Big(\mathbf{S} - \mu_1(t')^2 \, \mathbf{S}\,\mathbf{U}^{-1}\,\mathbf{S}\Big) \; + \; \beta(t', \Delta t)\, \mathbf{I}_p, \tag{35}$$

This completes the proof.

□

D Additional Results

D.1 Disentangling Model and Sampler Effects

To assess whether the observed reduction in memorization can be explained by the sampling procedure or inference budget, we complement the deterministic comparison in Table 1 with two additional analyses. We first compare teachers and students under stochastic sampling across dataset sizes, and then compare them at the same number of function evaluations (NFE).

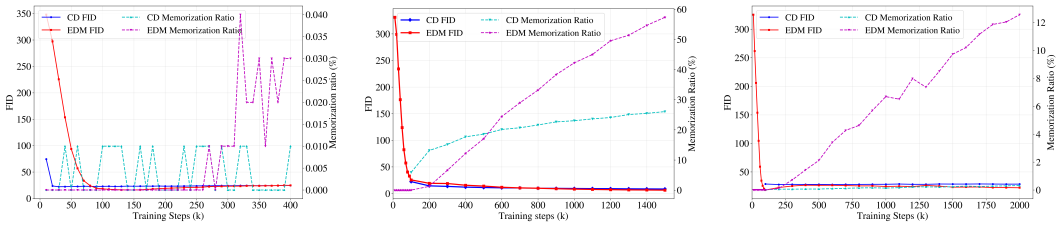
Comparisons under stochastic sampling. We compare two-step consistency-distilled students with their EDM teachers under stochastic sampling on CIFAR-10. As shown in Table 9, the students exhibit lower memorization under both ℓ_2 and SSCD metrics, together with lower SSCD p95 and better FID, across all four dataset sizes. Thus, the reduction persists when both models use stochastic inference and cannot be attributed solely to a contrast between deterministic and stochastic sampling.

Controlling for sampling budget. To examine whether fewer sampling steps alone account for the reduction, we compare the CD student with the same diffusion teacher at matched NFE in the 5,000-image CIFAR-10 setting. Each configuration is evaluated using 50,000 generated samples. We focus on stochastic sampling at 8 and 12 NFE, where the teacher achieves reasonable generation quality. At smaller budgets of 1–4 NFE, the teacher exhibits severe quality degradation, making low memorization rates difficult to distinguish from a failure to generate plausible samples.

Table 10 shows that the student achieves lower memorization across all three indicators while also improving FID at both sampling budgets. At 8 NFE, the SSCD memorization rate decreases from 22.350% to 18.512%, while FID improves from 23.85 to 19.13. At 12 NFE, the corresponding memorization rate decreases from 28.926% to 19.942%, with FID improving from 22.79 to 18.50. These results show that the observed memorization reduction persists at matched NFE and is accompanied by improved FID. It therefore cannot be explained solely by fewer sampling steps or poorer generation quality as measured by FID, providing further evidence that CD reduces memorization inherited from the teacher.

Table 10: Teacher–student comparisons at matched NFE on the 5,000-image CIFAR-10 setting. Both models use stochastic sampling, with 50,000 generated samples per configuration. Lower values are better for all reported metrics.

NFE	Model	FID	ℓ_2 Mem (%)	SSCD Mem (%)	SSCD p95
8	Teacher	23.85	5.464	22.350	0.7755
8	CD student	**19.13**	**1.316**	**18.512**	**0.7277**
12	Teacher	22.79	9.702	28.926	0.8463
12	CD student	**18.50**	**2.178**	**19.942**	**0.7371**



(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.

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

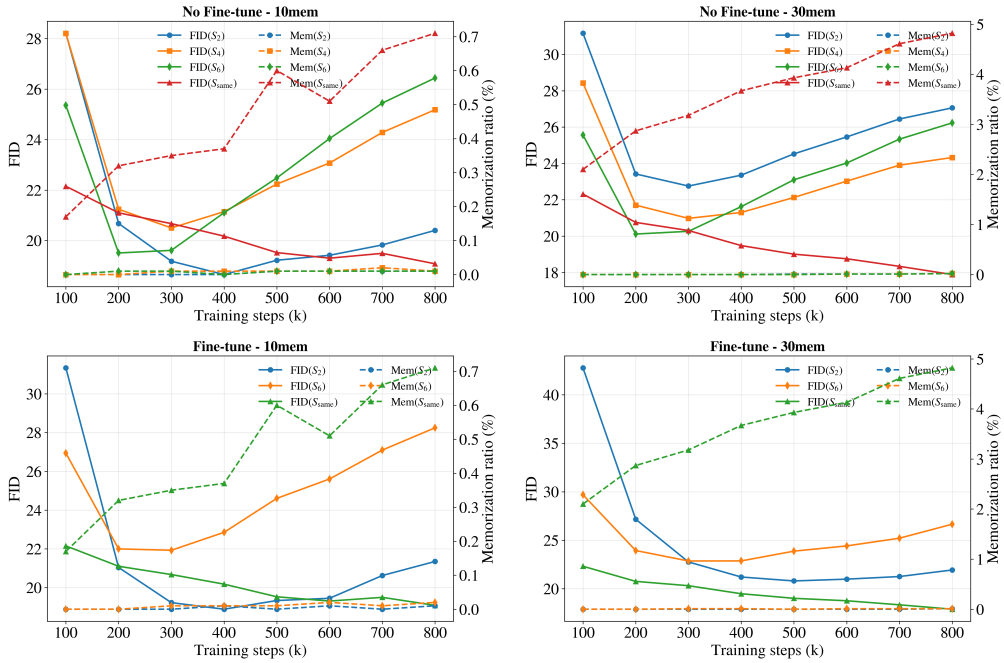


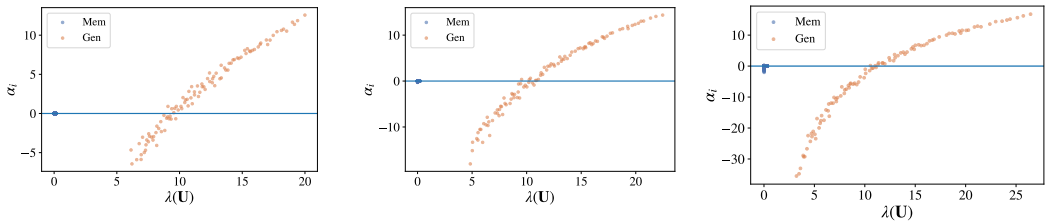
Figure 8: **Effect of student–teacher architectural mismatch on memorization and sample quality under two-step consistency distillation.** The figure reports FID (left y-axis, solid lines) and memorization ratio (right y-axis, dashed lines) as functions of training steps for different student–teacher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

= 22.31 (2000k steps). The corresponding two-step consistency-distilled student closely matches this behavior, achieving FID = 28.67 with a memorization ratio again near zero. Overall, two-step consistency distillation behaves robustly across model capacities: it preserves non-memorizing behavior when the teacher does not memorize, and substantially suppresses memorization when increased capacity induces it, while maintaining competitive sample quality.

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 S2, S4, and S6 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., $\mathrm{S2}/\mathrm{S4}/\mathrm{S6}$ 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



(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 $\alpha_i \simeq 0$.

student can faithfully realize under its parameterization. When the student has sufficient expressivity and a well-aligned parameterization (the $\mathrm{S}_{\mathrm{same}}$ setting), it can reproduce a larger portion of the teacher mapping—including a small but non-negligible fraction of teacher-specific, memorization-associated behavior. When the student is mismatched or starts far from the teacher solution, the distillation signal becomes harder to realize precisely, and the learned solution is biased toward more conservative, distribution-level smoothing, which suppresses instance-level memorization.

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

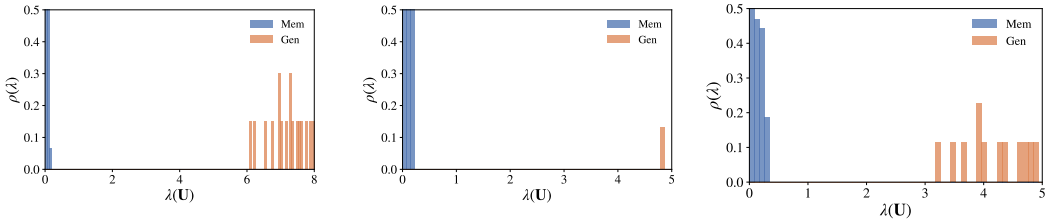
Consequently, these results highlight a tradeoff controlled by architectural compatibility. Exact architectural matching enables stable quality improvement but allows partial transfer of teacher memorization, whereas mismatch or random initialization strongly suppresses memorization but can suffer from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

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 $\mathbf{U}$. When $\psi_p = 32$ and $\psi_n = 16$, Gen modes occupy a well-separated, high-eigenvalue



(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.

Table 11: 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

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 11, 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.

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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction refer to the contributions and scope in the following sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a discussion of limitations in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions are clarified and all the proofs are included in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide experimental details in both the main paper and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We use the publicly available dataset and give detailed instructions for the implementation of our method. We plan to release the code after the acceptance of the paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: These details can be found in the experiment section and in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes. We report 95th-percentile statistics where appropriate to reduce the influence of outliers and provide additional evidence of experimental stability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include these information in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have a discussion in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit the creators and original owners of all codebases, datasets, benchmarks, and models used in this paper through citations. We also respect their licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Not applicable.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.