Why Are DMD Students Lazy? Understanding Copying Behavior in Few-Step Distillation

Shucheng Li[1,*] **Iolo Jones**[1] **Alexander Tong**[2,†] **Michael M. Bronstein**[1,2,†]

Abstract

Distribution Matching Distillation (DMD) compresses pretrained diffusion models into efficient few-step generators by aligning their noised distributions across all scales. In principle, such distribution-level supervision remains agnostic to the teacher's specific noise–data pairings; this provides the student the freedom to remap latent noise, a behavior consistently observed in low-dimensional settings. Surprisingly, we find that in high-dimensional settings, distilled students spontaneously reproduce the teacher's original noise–data pairings—a phenomenon we term copying. We demonstrate that copying is neither a byproduct of adversarial objectives nor a result of teacher memorization. Instead, our evidence suggests that copying is an emergent property arising from the limited geometric freedom of the student model during high-dimensional distillation.

1. Introduction

Diffusion models have fueled significant advances across images, videos, and text (Song et al., 2021a; Ho et al., 2020; Luo et al., 2023; Ho et al., 2022; Wang et al., 2023; Rombach et al., 2022). However, high-quality sampling through stochastic differential equations (SDEs) is computationally expensive. Distillation methods, such as Distribution Matching Distillation (DMD) (Yin et al., 2024b;a), bypass this bottleneck by training single-step students to match the teacher's distribution.

While recent studies have uncovered reproducibility (Zhang et al., 2024), where multiple *trained* diffusion models with different architectures produce the same noise–data pairings, the noise–data pairings of *distilled* students remain largely unexplored. In this work, we identify an unexpected behavior which we term **copying** for distillation in high-dimensional settings: students faithfully reproduce the teacher's noise–data pairings, despite being trained on an objective that is *pairing-indifferent*.

We demonstrate that copying is neither a trivial artifact of teacher memorization (Bonnaire et al., 2025) nor a consequence of auxiliary losses. Instead, we posit that copying emerges from the constrained degrees of freedom inherent in high-dimensional manifolds. Our experiments reveal that copying is most pronounced at the distribution boundaries and correlates with the convergence level of the teacher. These patterns emphasize that student distillation dynamics are fundamentally governed by the geometric structure of the teacher's mapping, providing a new perspective on the interaction between high-dimensional data geometry and generative distillation.

Our main contributions are as follows:

- **Observing and Quantifying Copying:** We establish copying as a non-trivial dynamical behavior by proving that the DMD objective is in principle invariant to the noise–data pairings learned by the student. We propose *Pairing Inefficiency* (Δ_E), a scale-invariant metric, allowing fair quantitative comparison of strengths of copying across datasets of disparate scales and resolutions.

[1]Department of Computer Science, University of Oxford, Oxford, United Kingdom.
[2]AITHYRA, Research Institute for Biomedical AI, Vienna, Austria.
[3]Corresponding Author: Shucheng Li shucheng.li@cs.ox.ac.uk.
† Equal Supervision.
Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

- **Analyzing triggering factors of copying:** We demonstrate a sharp empirical divergence in copying behavior between low- and high-dimensional settings. Through ablations, we rule out auxiliary objectives and teacher memorization as primary drivers for copying. We provide evidence on macroscopic and microscopic scales that characterize copying as a consequence of emergent high-dimensional geometric constraints.

2. Background

2.1. Diffusion Models

Diffusion models generate samples from an unknown data distribution p_{data} by reversing a learned stochastic process that transforms p_{data} into a prior noise distribution p_ε, typically approximately a scaled isotropic Gaussian $\mathcal{N}(0, \sigma^2\mathbf{I})$. This process is characterized by a pair of forward and backward stochastic differential equations (Song et al., 2021b).

In the Variance Exploding (VE) framework, the forward SDE evolves $x_0 \sim p_{\text{data}}$ into increasingly noisy latent variables x_t over the time interval $t \in [0, T]$ according to a noise schedule $\sigma(t)$:

$$\mathrm{d}x_t = \sqrt{2\dot{\sigma}(t)\sigma(t)}\mathrm{d}\mathbf{W}_t,$$

where $\mathbf{W}_t$ denotes a standard multi-dimensional Wiener process. The transition kernel for this process is given by $p_t(x_t|x_0) = \mathcal{N}(x_t; x_0, \sigma^2(t)\mathbf{I})$. The marginals $\{p_t\}_{t\in[0,T]}$ define a **probability path** starting at $p_0 = p_{\text{data}}$ and terminating at a distribution $p_T \overset{d}{\approx} p_\varepsilon := \mathcal{N}(0, \sigma^2(T)\mathbf{I})$, provided the noise variance scale $\sigma^2(T) \gg \sigma^2_{\text{data}}$ the variance scale of the data distribution.

To sample from p_{data}, the probability path can be inverted by initializing $x_T = z \sim \mathcal{N}(0, \sigma^2(T)\mathbf{I})$ and solving a corresponding backward SDE, or its equivalent deterministic Probability Flow ODE:

$$\mathrm{d}x_t = -\dot{\sigma}(t)\sigma(t)s(x_t, t)\mathrm{d}t, \tag{1}$$

where $s(x_t, t) := \nabla_x \log p_t(x_t)$ represents the Stein score function for $\{p_t\}$. Since the ground-truth score function is inaccessible, it is typically estimated by a learnable time-conditioned neural network $s_\theta(x_t, t)$. While other frameworks such as Variance Preserving (VP) or Flow Matching (Lipman et al., 2023) can be similarly formulated under this SDE template with differing drift and diffusion coefficients (Karras et al., 2022; Lai et al., 2025), we restrict attention to the VE framework with $\sigma(t) = t$ for ease of presentation.

Under the VE framework, since the transition kernel $p_t(x_t|x_0)$ is Gaussian, the Stein score $s(x_t, t)$ relates to the posterior expectation of the clean signal via Tweedie's formula: $s(x_t, t) = (\mathbb{E}[x_0|x_t] - x_t)/t^2$. This identity allows a reparametrization of score to stabilize learning. In particular, we typically train a neural network $x_{0,\theta}(x_t, t)$ to approximate $\mathbb{E}[x_0|x_t]$, then set

$$s_\theta(x_t, t) = \frac{x_{0,\theta}(x_t, t) - x_t}{t^2}.$$

In image dataset learning tasks, this denoiser $x_{0,\theta}$ is often implemented as a UNet with additional pre-conditioning to ensure learning stability (Yin et al., 2024a; Karras et al., 2022).

2.2. Distribution Matching Distillation

Distribution Matching Distillation (Yin et al., 2024b;a) aims to overcome the computational bottleneck of multi-step inference by distilling the teacher diffusion model $s(x_t, t)$ into a single-step student generator $G_\theta(z)$, by aligning the student probability path $p_{\theta,t}$, defined as

$$p_{\theta,t} := \textbf{Law}(G_\theta(z) + \sigma(t)\varepsilon)$$

where $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$ is independent of $z \sim \mathcal{N}(0, \sigma^2(T)\mathbf{I})$, with the teacher's probability path p_t across all noise levels $\sigma(t)$. This is achieved by minimizing a time-weighted integral of the Kullback-Leibler (KL) divergence, referred to as the **Distribution Matching (DM) loss**:

$$L_{\mathrm{DM}}(\theta) := \int w(t)\mathrm{KL}(p_{\theta,t}\|p_t)\,\mathrm{d}t. \tag{2}$$

The gradient of this objective with respect to the student parameters θ can be expressed in closed form:

$$\nabla_\theta L_{\mathrm{DM}}(\theta) = \mathbb{E}_{t,z,\varepsilon}\left[w(t)\left(s_\psi(x_t,t) - s(x_t,t)\right)\frac{\partial G_\theta(z)}{\partial\theta}\right], \tag{3}$$

where $x_t = G_\theta(z) + \sigma(t)\varepsilon$. Here, $s(x_t,t)$ is the pre-trained teacher score, and $s_\psi(x_t,t)$ is a learnable student score neural network that approximates the intractable student score $\nabla_x \log((G_\theta(z) * \mathcal{N}(0,t^2\mathbf{I}))(x))$ during distillation.

We initialize the student generator and student score with teacher's weights, specifically setting $G_\theta(z) := x_0(z,T)$, and $s_\psi(x_t,t) = (x_0(x_t,t) - x_t)/t^2$, providing a strong starting point for the distillation procedure. To maintain stability during distillation, the student score s_ψ and generator G_θ are updated via a joint training schedule. In each iteration, s_ψ is updated by minimizing a standard denoising score matching loss

$$\min_\psi \mathbb{E}_{t,z,\varepsilon}\left[\|s_\psi(x_t,t) + \frac{\varepsilon}{\sigma(t)}\|^2\right],$$

where $x_t = G_\theta(z) + \sigma(t)\varepsilon$, to ensure that s_ψ accurately tracks the score of the current student probability path $p_{\theta,t}$. The student generator G_θ is updated every 5 iterations using the gradient from Equation 3. Auxiliary regression or adversarial objectives may be added (Yin et al., 2024a;b).

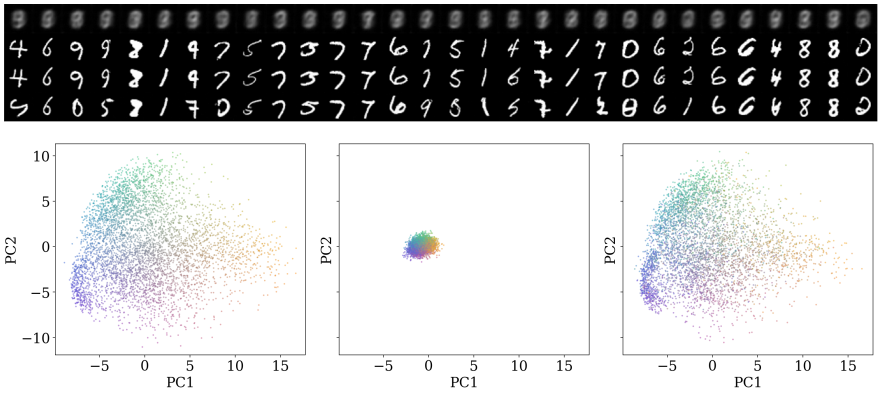


Figure 1. **Significant copying in high-dimensional settings.** The distilled student on unconditional MNIST exhibits strong copying behavior. **Top:** Thirty image quadruples generated from random initial noise seeds z. From top to bottom: teacher 1-step samples $\Phi_1(z)$, teacher 8-step samples $\Phi_8(z)$, teacher 32-step samples $\Phi_{32}(z)$, and student 1-step samples $G(z)$. **Bottom:** Visualization of 2000 triplets $(\Phi_8(z), \Phi_1(z), G(z))$ projected onto the two leading principal components of $\Phi_8(z)$. Points generated from the same noise seed are assigned the same color across panels, showing that $G(z)$ occupies the same manifold location as the multi-step teacher $\Phi_8(z)$. Pairing inefficiency is low at $\Delta_E \approx 0.0367$.

3. The Copying Behaviour

We observe an unexpected phenomenon of Distribution Matching Distillation (DMD) in high-dimensional settings: a student model tends to faithfully reproduce the teacher's noise-data pairings pointwise, even though it is trained exclusively to match the teacher's *distribution*. As illustrated in Figure 1, over the unconditional MNIST learning task, the student generator $G_\theta(z)$ aligns very closely with the multi-step teacher target $\Phi_K(z)$ obtained by taking K Euler steps[1] backwards with second-order corrections. We term this spontaneous alignment behavior **copying**.

[1]We fix $K = 8$ across all experiments, as this setting suffices for high-quality synthesis, with larger values offering only marginal improvements in sample quality.

3.1. Measuring Copying by Pairing Inefficiency

To formally quantify the strength of copying and compare it across different scales and dimensions, we introduce a scale-invariant measure $\Delta_E(\Phi_K, G_\theta)$ called pairing inefficiency.

Definition 3.1 (Pairing Inefficiency)**.** Let the initial noise be $z \sim p_\varepsilon = \mathcal{N}(0, \sigma^2(T)\mathbf{I})$. Let $p_\Phi = (\Phi_K)_\# p_\varepsilon$ and $p_G = G_\# p_\varepsilon$ be the distributions learned by the teacher and student. The **Optimal Transport (OT) energy** and the **Distillation Transport (DT) energy** are defined as:

$$E_{OT}(\Phi_K, G) := \min_{\pi \in \Gamma(p_\Phi, p_G)} \int \|x - y\|_2^2 \, d\pi(x, y) \tag{4}$$

$$E_{DT}(\Phi_K, G) := \int \|\Phi_K(z) - G(z)\|_2^2 \, dp_\varepsilon(z) \tag{5}$$

where $\Gamma(p_\Phi, p_G)$ is the class of all couplings with marginals p_Φ and p_G. We define the **pairing inefficiency** as

$$\Delta_E(\Phi_K, G) := \frac{E_{DT}(\Phi_K, G)}{E_{OT}(\Phi_K, G)} - 1. \tag{6}$$

A small inefficiency $\Delta_E \approx 0$ implies strong copying, while a larger inefficiency Δ_E implies the teacher pairings are *remapped*. In practice, we use a consistent Monte Carlo estimator $\Delta_E^{(N)}$ with $N = 1000$ samples. Formal justifications including the non-negativeness and scale-invariance of Δ_E are provided in Appendix C.

3.2. Copying is Not Required for Successful Distillation

Crucially, we emphasize that copying is not a theoretical necessity of the DMD objective. The Distribution Matching loss L_{DM} is *pairing-indifferent*: it only penalizes the discrepancy between the student distribution p_G and the teacher's p_Φ, remaining agnostic to the noise–data pairings the student learned.

Lemma 3.2. *Let G_θ and $G_{\theta'}$ be two student generators. Whenever $G_\theta(z) \overset{d}{=} G_{\theta'}(z)$, it follows that $L_{DM}(\theta) = L_{DM}(\theta')$, even though in general $\nabla_\theta L_{DM}(\theta) \neq \nabla_\theta L_{DM}(\theta')$. Consequently, the stochastic optimization dynamics can drive students to converge toward pairings with vastly different inefficiencies Δ_E despite achieving similar distributional fidelity.*

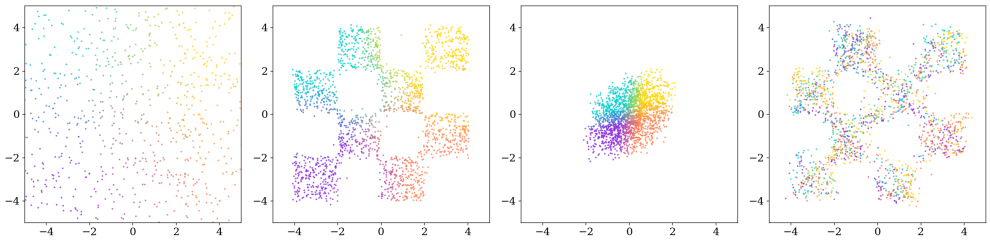
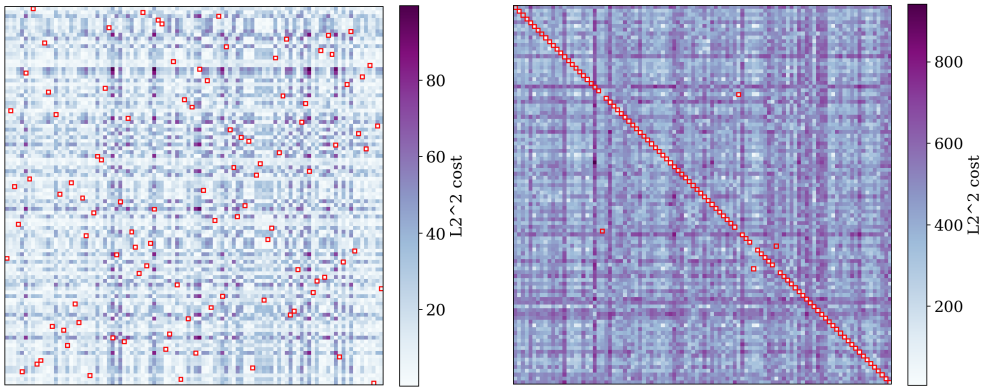


Figure 2. **Copying does not necessarily occur, and rarely occurs in low-dimensional space.** The distilled student on synthetic chessboard dataset exhibits strong remapping behavior. The dataset is a 2D chessboard dataset embedded in the first two coordinates of a four-dimensional ambient space. Panels from left to right show the initial Gaussian noise z, the teacher 8-step samples $\Phi_8(z)$, the teacher 1-step samples $\Phi_1(z)$, and the student one-step samples $G(z)$. Points generated from the same initial noise seed z are assigned the same color across panels. Visualization is obtained by projection onto the first two coordinates. The pairing inefficiency is high at $\Delta_E \approx 8.55$.

A particular consequence of Lemma 3.2 is that minimizing the distillation objective in Equation 2 does not require the student to satisfy $G_\theta(z) \approx \Phi_K(z)$ for most z hence achieving low pairing inefficiency $\Delta_E \approx 0$. A student could match the teacher's distribution perfectly while arbitrarily remapping or reflecting the noise-to-data manifold. The proof of this lemma, along with an intuitive example where a non-copying student attains a global minimum of Objective 2, is provided in Appendix C.



(a) 2D Chessboard 0.3 *(b)* Unconditional MNIST

Figure 3. **Copying is more pronounced in high-dimensional settings.** The heatmaps represent the pairwise squared L_2 distances between teacher-generated images $\{\Phi_K(z_i)\}_{i=1}^{100}$ and student-generated images $\{G(z_j)\}_{j=1}^{100}$ for the 2D chessboard dataset (left) and the unconditional MNIST dataset (right). The horizontal and vertical axes denote the teacher and student image indices, i and j, respectively. The Optimal Transport (OT) pairing is highlighted with red boxes, while the Distillation Transport (DT) pairing corresponds to the diagonal. In the chessboard experiment, students mainly remaps the teacher's pairings, resulting in high pairing inefficiency ($\Delta_E \approx 8.55$). Conversely, in the unconditional MNIST experiment, the student mainly copies the pairings, exhibiting very low pairing inefficiency ($\Delta_E \approx 0.0367$).

3.3. Copying Rarely Occurs in Low Dimensions

We have theoretically established that in principle a distilled student is not required to copy the teacher's noise-data pairings and to exhibit low pairing inefficiency.

Indeed, we empirically observe that on low-dimensional datasets, copying is rarely present and pairing inefficiency is high. For example, we define p_{data} as a two-dimensional 4×4 chessboard distribution embedded within the first two coordinates of $\mathbb{R}^4$. We first train a teacher diffusion model, parameterized by an MLP with five hidden layers of width 384, and subsequently distill a single-step student generator until convergence. As shown in Figure 2 and Figure 3, while the student successfully recovers the target distribution p_{data}, it exhibits significant remapping behavior with high pairing inefficiency $\Delta_E \approx 8.55$, confirming that the student has found a valid distributional fit that deviates from the teacher's original trajectories.

Similarly significant remapping is consistently observed across other standard low-dimensional synthetic datasets (see Appendix A).

3.4. Copying Frequently Occurs in High Dimensions

In stark contrast to low-dimensional settings, copying behavior emerges consistently in high-dimensional tasks. We train a clean-prediction teacher diffusion model using a standard UNet architecture on the unconditional MNIST dataset consisting of $70,000$ digit images. The teacher is trained from scratch for 8192 iterations, after which we distill a single-step student generator until convergence.

Compared to the toy settings, the distilled student on conditional MNIST is now visually substantially more prone to copying (see Figure 1 and Figure 3), achieving an extremely low pairing inefficiency $\Delta_E \approx 0.0367$.

We also observe that such strong copying in high dimensional settings is not exclusive to natural images; it is equally prevalent in artificial high-dimensional datasets lacking natural image structures, as well as in conditional generation settings. For a comprehensive overview of these cases, see Appendix A.

The consistent selection of copying solutions with $\Delta_E \approx 0$ over other valid remappings is thus an emergent property that warrants further investigation.

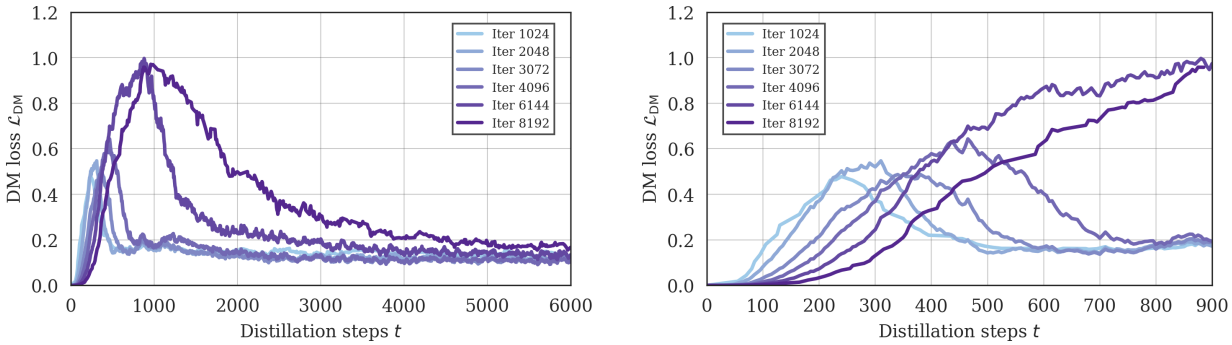


Figure 4. **Two stages of Distribution Matching Distillation.** We plot the distribution matching loss of all students (initiated at different teacher snapshots) on the unconditional MNIST dataset, throughout distillation. In all cases, the DM loss exhibits a characteristic two-stage increase-decrease evolution. Students are initialized from teacher checkpoints trained for 1024, 2048, 3072, 4096, 6144, and 8192 iterations. All students are distilled for 50K iterations. The left panel shows the loss evolution up to 6K distillation iterations, and the right panel zooms into the first 900 iterations.

4. Analyzing the Mechanisms Behind Copying

In this section, we analyze the mechanisms behind the copying behavior. For simplicity of discussion, we focus on the unconditional MNIST setting.

4.1. Removal of Intuitive Explanations

4.1.1. Copying Occurs Without Adversarial Objective

The models in Yin et al. (2024a) on which significant copying is observed were trained with an additional adversarial loss,

$$L_{\text{GAN}}(\gamma) := \mathbb{E}_{\substack{x\sim p_{\text{data}} \\ \varepsilon,z,t}} \left[-\log D_\gamma(x'_t) + \log D_\gamma(x_t)\right],$$

where $x'_t = G_\theta(z) + t\varepsilon$, $x_t = x + t\varepsilon$. This adversarial loss may implicitly contribute to the copying.

However, our experiments use neither regression loss nor the adversarial objective, the copying phenomenon nevertheless still emerges consistently. This demonstrates that neither target regression nor adversarial training is necessary for copying to occur.

4.1.2. Copying Occurs Without Memorization

One may suspect that the teacher model memorizes the training dataset, thereby encouraging the student to simply reproduce its outputs. For a generated datapoint y, we define its **memorization distance ratio**

$$r(y) := \frac{\|y - x^1(y)\|}{\|y - x^2(y)\|},$$

where $x^1(y)$ and $x^2(y)$ denote respectively the nearest and second-nearest neighbours of y in the training dataset. A datapoint y is considered memorized if $r(y) < r_{\text{thres}}$ for some prescribed threshold r_{thres} (Bonnaire et al., 2025). We evaluated the trained teacher models by computing memorization distance ratios for randomly generated samples. As illustrated in Figure 12, none of the generated samples exhibit signs of memorization.

4.2. Copying Follows Geometric Complexity

In this subsection, we present additional experiments and a more detailed analysis of when copying occurs and under what circumstances it becomes more significant.

We divide the distribution-matching distillation dynamics into two stages, as shown in Figure 4. In the first stage, the distribution-matching (DM) loss increases. This occurs because the initialized surrogate student score model $s_\psi(\cdot, t)$ gradually deviates from the teacher score $s(\cdot, t)$ and more accurately approximates and tracks the

score $\nabla_x \log \left(\left(G_\theta(z) * \mathcal{N}(0, t^2\mathbf{I}) \right)(x) \right)$ of the current student probability path. Once s_ψ begins consistently tracking the actual student score, distillation enters the second stage. At this point, the gradient in Equation 3 begins providing meaningful optimization signal for updating G_θ, so that s_ψ gradually matches s, progressively deforming student's one-step distribution toward the teacher's learned target distribution, decreasing the DM loss.

We conjecture that a higher degree of copying arises when, in the second stage of distillation, the student has limited freedom to deform its distribution while still preserving the target distribution implied by the teacher score. To illustrate this intuition, consider a target uniform distribution U with density $\rho = 1/4$ supported on the square with vertices $(\pm 1, \pm 1)$. Within the interior of the square, the student may continuously perturb or remap noise–data pairings while still approximately preserving the target density. In this regime, there exists substantial geometric flexibility for transport.

In contrast, near the boundary of the support, remapping becomes significantly more constrained. To preserve the uniform distribution while continuously deforming transport pairings near the edges, the student must satisfy additional geometric and continuity constraints induced by the boundary structure. Under limited expressiveness of a single-step generator[2], such constrained deformations may be substantially harder to realize during optimization. Consequently, the optimization dynamics may favour preserving the teacher's original noise–data pairings in these regions, leading to stronger copying behavior.

We provide one microscopic and one macroscopic piece of evidence in support of this conjecture.

4.2.1. Micro level: Boundary Points Are More Likely Copied

We observe that teacher samples further away from the bulk of training dataset are more likely to be copied by the student. In particular, Figure 11 shows that the student's relative displacement towards the teacher target, $\|G(z) - \Phi_K(z)\| - \|\Phi_1(z) - \Phi_K(z)\|$, is moderately negatively correlated with the average distance between the teacher target and the training dataset,

$$D(z) := \mathrm{Avg}_{x \in \mathrm{train}} \left(\|\Phi_K(z) - x\| \right).$$

Intuitively, teacher samples $\Phi_K(z)$ with high $D(z)$ typically lie near sparse or extremal parts of the learned data manifold, where remapping while preserving the induced distribution may require satisfying stronger geometric or continuity constraints. Consequently, optimization dynamics appear to favour copying in these regions.

We note that this more pronounced copying phenomenon at extremeties of data manifold may be closely intertwined with recent insights from Zhang et al. (2026), who demonstrated that trained generative models effectively capture local geometry over abundant data clusters while reverting to memorization on scarce, isolated points. We leave a formal characterization of this mechanistic correspondence and causality for future work.

4.2.2. Macro Level: Longer Trained Teachers Are More Likely Copied

We observe that, although the teacher models do not memorize, at a macroscopic level the degree of student copying increases with the number of iterations used to train the teacher model.

We train teacher diffusion models on the unconditional MNIST dataset for varying numbers of iterations, and initialize student generators from these teacher snapshots for subsequent distillation. We observe that students distilled from later-stage teacher checkpoints preserve a substantially higher proportion of the teacher's original noise–data pairings (see Figure 10, 9).

This observation is also consistent with our hypothesis. As training progresses, the teacher more accurately resolves the underlying target distribution, reducing excess diffusion-induced variability and progressively capturing finer geometric structure of the data manifold. These increasingly refined structures impose stronger constraints on the student during distillation, thereby reducing the flexibility available for large-scale remapping of noise–data pairings. Consequently, optimization dynamics become more favourable toward the copying behavior.

[2]The student model in the chessboard experiment in contrast, is highly expressive, and could easily learn a remapped distribution. See Figure 2 and Section 3.3.

4.3. Related Work

Memorization and reproducibility in diffusion models. Prior works studied memorization, generalization, and reproducibility in diffusion models, showing that memorization occurs on excessively small datasets (Zhang et al., 2024; Bamberger et al., 2026; Bonnaire et al., 2025), while reproducibility persists even in strongly generalizing regimes (Li et al., 2024; Kadkhodaie et al., 2024; Zhang et al., 2024). These works focus on similarities between independently trained diffusion models, whereas our work studies copying behavior arising during student distillation.

Distillation of diffusion and flow-based models. Existing distillation methods for diffusion models include DMD, VSD, Diff-Instruct (Yin et al., 2024b;a; Wang et al., 2023; Luo et al., 2023), while approaches to distill flow-based models include progressive distillation, consistency models, MeanFlow, and SplitMeanFlow (Salimans & Ho, 2022; Song et al., 2023; Geng et al., 2025; Guo et al., 2025). Unlike flow-based distillation, where trajectory-level supervision is directly available, distribution-matching distillation only provides distribution-level supervision, theoretically allowing freedom for remapping noise–data pairings.

Manifold structure of data distributions. The manifold hypothesis suggests that natural image distributions are concentrated near low-dimensional manifolds (Pidstrigach, 2022). Prior works studied geometric and topological challenges arising in score-based generative models under this setting, including numerical stability, convergence, and manifold-related artifacts (Pidstrigach, 2022; Bortoli, 2022; Potaptchik et al., 2025; Loaiza-Ganem et al., 2024). Our observations confirm that such geometric constraints may also influence copying behavior during distillation.

5. Conclusion

In this work, we investigate the unexpected copying behavior exhibited by distribution-matching distilled student models, where the student reproduces the teacher's noise–data pairings despite having access only to distribution-level supervision.

Through extensive experiments, we show that copying is not caused by regression or adversarial objectives or by memorization of the training dataset by the teacher model. Empirical evidence instead supports the hypothesis that copying emerges when the student has limited freedom during distillation to deform its distribution while remaining aligned with the data distribution induced by teacher's learned score.

6. Limitations and Future Work

While our experiments rule out several intuitive explanations and suggest a geometric driver for copying, our current insights remain primarily correlational and bounded in scope.

First, although empirical evidence supports that boundary constraints limit the student's deformation freedom, our framework lacks a formal analytical or topological proof. A more rigorous theoretical treatment is required. Future work could leverage high-dimensional geometric descriptors to dynamically track the student's score function during training (Jones, 2024), bypassing linear low-dimensional projections.

Second, our evaluation is restricted to compact datasets (MNIST, ImageNet-64). Validation on modern high-resolution text-to-image pipelines remains unproven, and scaling up these findings is a crucial next step. Finally, understanding which structural properties of high-dimensional data encourage copying, such as whether it relates to the teacher learning an approximate optimal transport map, offers a promising avenue for future investigation.

Acknowledgements

S.L. is supported by a China Scholarship Council (CSC) - PAG Oxford Scholarship. M.B. is partially supported by the EPSRC Turing AI World-Leading Research Fellowship No. EP/X040062/1 and EPSRC AI Hub No. EP/Y028872/1.

References

Bamberger, J., Jones, I., Duncan, D., Bronstein, M. M., Vandergheynst, P., and Gosztolai, A. Carré du champ flow matching: better quality-generalisation tradeoff in generative models. In *International Conference on Learning Representations*, 2026.

Bonnaire, T., Urfin, R., Biroli, G., and Mezard, M. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *Advances in Neural Information Processing Systems*, 2025.

Bortoli, V. D. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

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

Geng, Z., Deng, M., Bai, X., Kolter, J. Z., and He, K. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, 2025.

Guo, Y., Wang, W., Yuan, Z., Cao, R., Chen, K., Chen, Z., Huo, Y., Zhang, Y., Wang, Y., Liu, S., et al. Splitmeanflow: Interval splitting consistency in few-step generative modeling. *arXiv preprint arXiv:2507.16884*, 2025.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Jones, I. Manifold diffusion geometry: Curvature, tangent spaces, and dimension. *arXiv preprint arXiv:2411.04100*, 2024.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Lai, C.-H., Song, Y., Kim, D., Mitsufuji, Y., and Ermon, S. The principles of diffusion models. *arXiv preprint arXiv:2510.21890*, 2025.

Li, X., Dai, Y., and Qu, Q. Understanding generalizability of diffusion models requires rethinking the hidden gaussian structure. In *Advances in Neural Information Processing Systems*, 2024.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Loaiza-Ganem, G., Ross, B. L., Hosseinzadeh, R., Caterini, A. L., and Cresswell, J. C. Deep generative models through the lens of the manifold hypothesis: A survey and new connections. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

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

Pidstrigach, J. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems*, 2022.

Potaptchik, P., Azangulov, I., and Deligiannidis, G. Linear convergence of diffusion models under the manifold hypothesis. In Haghtalab, N. and Moitra, A. (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4668–4685. PMLR, 30 Jun–04 Jul 2025.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

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

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

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

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems*, 2023.

Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., and Freeman, W. T. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 2024a.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T., and Park, T. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b.

Zhang, H., Zhou, J., Lu, Y., Guo, M., Wang, P., Shen, L., and Qu, Q. The emergence of reproducibility and consistency in diffusion models. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60558–60590. PMLR, 21–27 Jul 2024.

Zhang, Z., Li, X., Li, X., Shi, L., Wu, M., Tao, M., and Qu, Q. Generalization of diffusion models arises with a balanced representation space. In *International Conference on Learning Representations*, 2026.

A. Copying and Remapping Behaviors of DMD Students on Other Datasets

Similar to the toy chessboard experiment, on other low-dimensional synthetic datasets, the distribution-matching distilled student also exhibits *substantial remapping* behavior, even with $\|G(z) - \Phi_8(z)\| \gg \|\Phi_1(z) - \Phi_8(z)\|$ for some z.

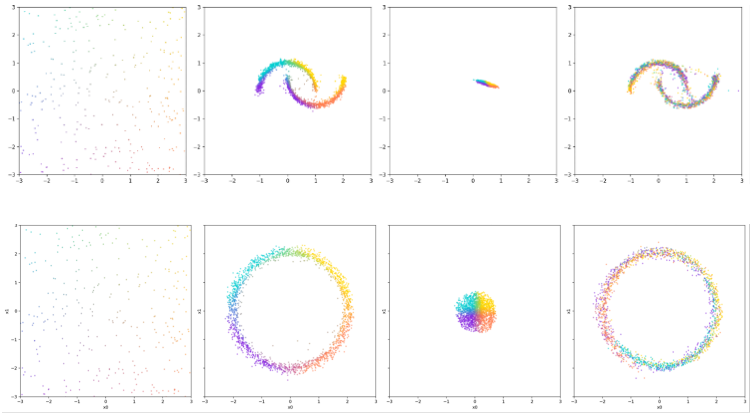


Figure 5. **Noise–data pairing results on low-dim additional artificial datasets.** In each row, panels from left to right show the initial Gaussian noise z, the teacher 8-step Euler samples $\Phi_8(z)$, the teacher 1-step Euler samples $\Phi_1(z)$, and the student one-step samples $G(z)$. Points generated from the same initial noise seed z are assigned the same color across panels. **Top row:** Distillation results on the 2D double-moons dataset embedded in the first two coordinates of a four-dimensional ambient space. **Bottom row:** Distillation results on a Gaussian mixture dataset consisting of 32 isotropic four-dimensional Gaussian components with shared standard deviation 0.2. The component means are arranged on a circle of radius 2 in the first two coordinates of the ambient space.

In contrast, high-dimensional datasets typically exhibit *significantly stronger copying behavior*, which becomes even more pronounced in *conditional* generation settings (Figure 6, 7, and 8). We conjecture that conditioning enables the teacher to resolve finer geometric structures within each class separately, thereby imposing stronger constraints on the student during distillation and reducing its freedom to remap noise–data pairings while preserving the target distribution.

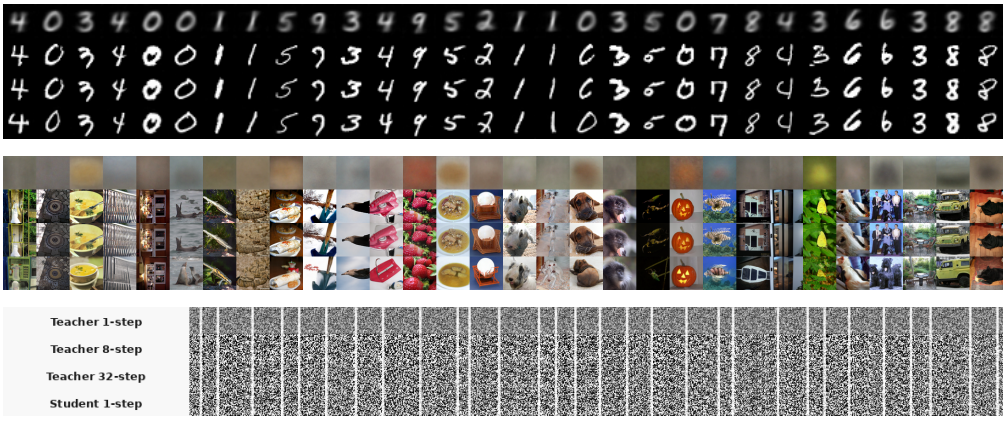


Figure 6. **Distillation results on conditional high-dimensional datasets.** Panels from top to bottom: conditional MNIST, conditional ImageNet64, and the conditional synthetic MLP-manifold dataset. Each panel contains thirty image quadruples generated from random initial noise seeds z and randomly assigned classes. From top to bottom within each panel are the teacher 1-step samples $\Phi_1(z)$, teacher 8-step samples $\Phi_8(z)$, teacher 32-step samples $\Phi_{32}(z)$, and student one-step samples $G(z)$. We use the teacher model provided by (Yin et al., 2024a) for the conditional ImageNet64 experiment, for deterministic teacher sampling we set Schurn = 0.

We also remark that the copying behavior is not restricted to natural image datasets. We construct a synthetic high-dimensional dataset by applying a randomly initialized two-layer MLP to 16-dimensional Gaussian noise, followed by superimposing whitened stripes at different spatial locations according to randomly assigned class labels $i = 0, \ldots, 9$. This produces a distribution in $\mathbb{R}^{32\times32\times1}$ supported on a 16-dimensional manifold with 10 conditional classes. We observe similarly strong copying behavior by the student on this dataset (Figure 6, 7, and 8).

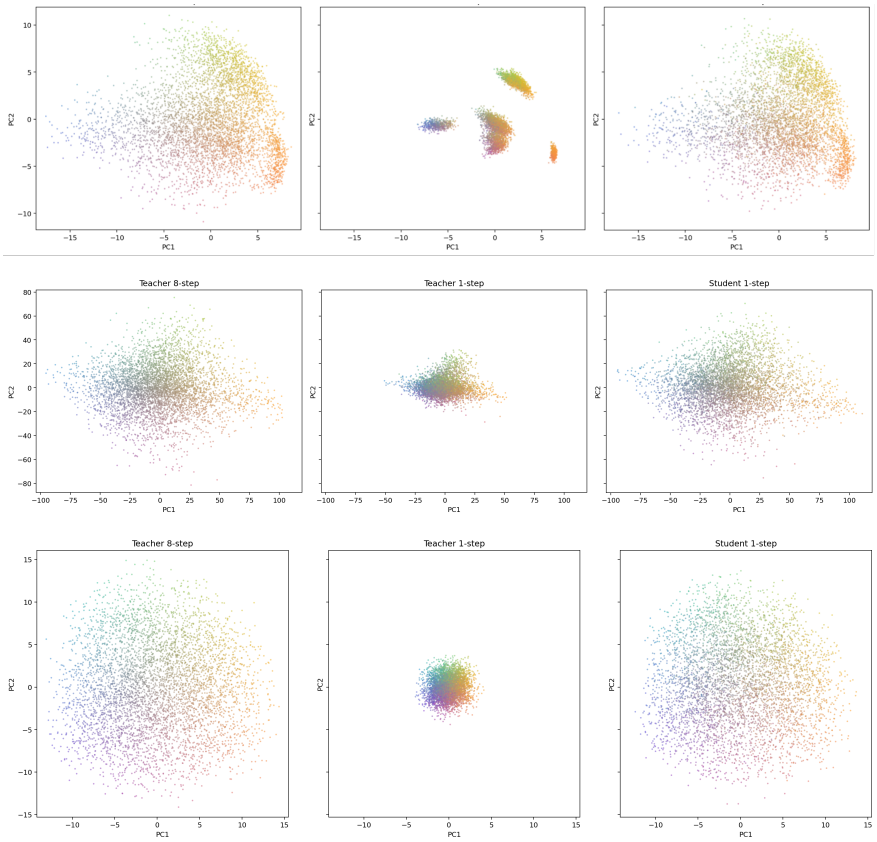


Figure 7. **Visualization of noise–data pairing relationships on conditional high-dimensional datasets.** Rows from top to bottom: conditional MNIST, conditional ImageNet64, and the conditional synthetic MLP-manifold dataset. Each row visualizes 2000 triplets $(\Phi_8(z), \Phi_1(z), G(z))$ projected onto the two leading principal components computed from the teacher 8-step samples $\Phi_8(z)$.

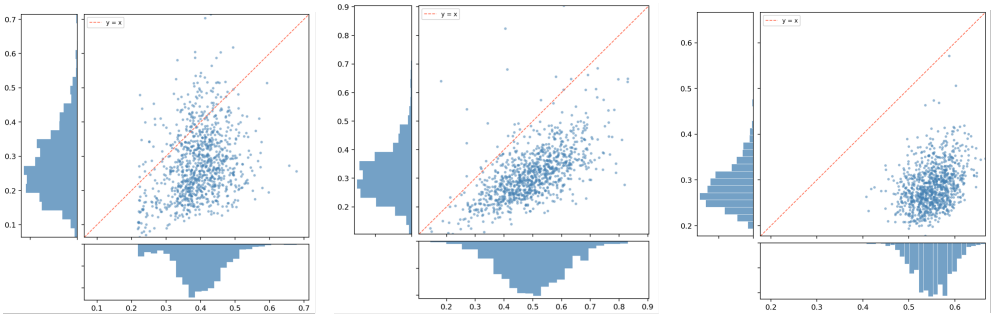


Figure 8. Comparison of student copying behavior on conditional high-dimensional datasets. **Left:** conditional MNIST. **Middle:** conditional ImageNet64. **Right:** conditional synthetic MLP-manifold dataset. In all plots, the horizontal axis shows $\|\Phi_1(z) - \Phi_8(z)\|$, while the vertical axis shows $\|G(z) - \Phi_8(z)\|$. Axes are proportionally rescaled where appropriate to reflect pixel-level or channel-wise discrepancy magnitudes. Across all three datasets, students distilled from sufficiently trained teachers exhibit strong copying behavior.

B. Copying on Teachers Trained for Various Lengths on Unconditional MNIST

We train teacher diffusion models on the unconditional MNIST dataset for 1024, 2048, 4096, and 8192 iterations, and distill corresponding student models initialized from these checkpoints. Each student is subsequently distilled for 50K iterations. We observe that students initialized from later-stage teacher checkpoints generally preserve a higher proportion of the teacher's original noise–data pairings.

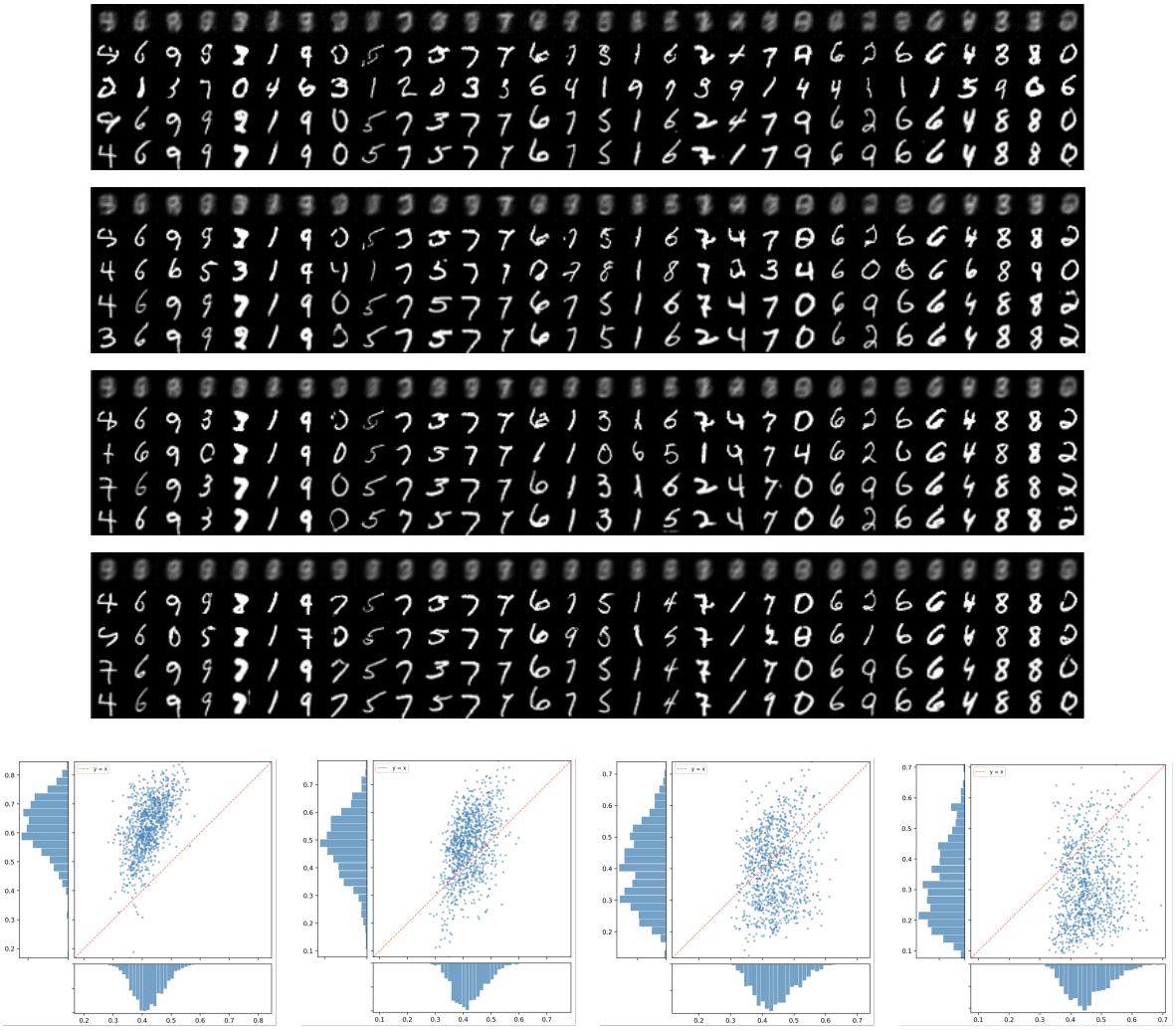
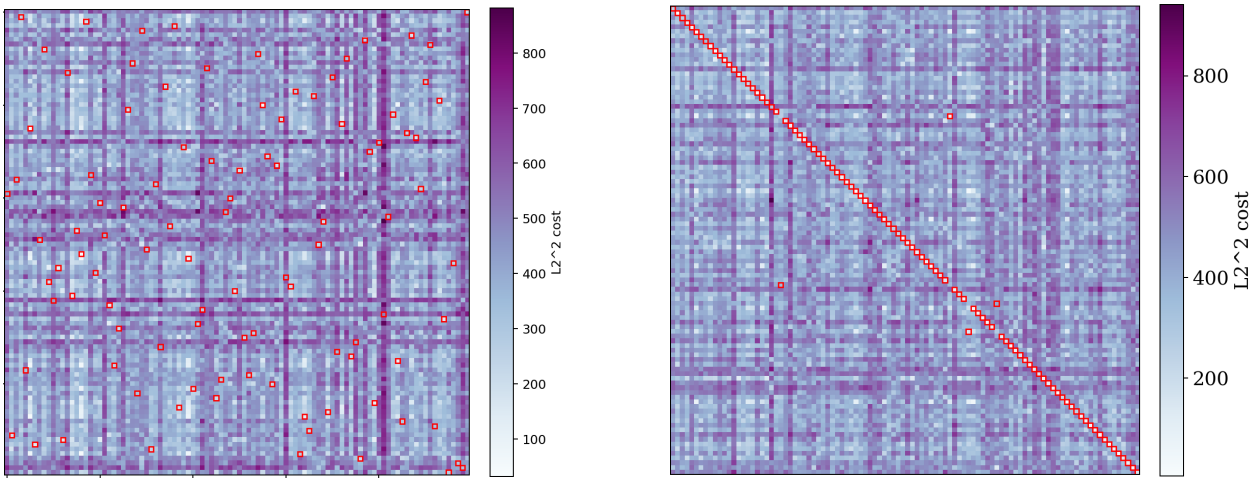


Figure 9. **Copying is more pronounced on longer-trained teachers.** Distillation results of teachers trained for various lengths on the unconditional MNIST dataset. **Top:** The four panels correspond to teachers trained for 1024, 2048, 4096, and 8192 iterations. For each panel, thirty image quintuples are generated from random initial noise seeds z. From top to bottom the five rows show the teacher 1-step samples $\Phi_1(z)$, teacher 8-step samples $\Phi_8(z)$, student one-step samples $G(z)$, closest training point $x^1(\Phi_8(z))$ and second closest training point $x^2(\Phi_8(z))$. **Bottom:** Comparison of copying behaviors for students distilled from teacher checkpoints trained for 1024, 2048, 4096, and 8192 iterations. In each plot the horizontal axis shows $\|\Phi_1(z) - \Phi_8(z)\|$ and the vertical axis shows $\|G(z) - \Phi_8(z)\|$, both axes rescaled by $1/32$. The pairing inefficiencies of students distilled at these snapshots are $\Delta_E = 1.05, 0.389, 0.237$ and 0.0367 respectively, showing stronger copying on teachers trained for larger number of iterations.



(a) Teacher 1k Iterations

(b) Teacher 8k Iterations

Figure 10. **Longer trained teachers are more likely copied.** Visualization of pairing inefficiency via distance heatmaps. Heatmaps display pairwise L_2 distances between teacher $\{\Phi_K(z_i)\}_{i=1}^{100}$ and student $\{G(z_j)\}_{j=1}^{100}$ outputs for students distilled from 1024-step (left) and 8192-step (right) teachers. Red boxes denote the Optimal Transport (OT) pairing, while the diagonal represents the Distillation Transport (DT) pairing. At 1024 teacher iterations, the student remaps the pairings with inefficiency $\Delta_E \approx 1.05$, whereas at 8192 iterations, the student exhibits strong copying behavior with $\Delta_E \approx 0.0367$.

C. Derivations

Lemma C.1. *Let G_θ and $G_{\theta'}$ be two single-step student generators, potentially with different architectures (or with the same architecture but different parameters). Whenever $G_\theta(z) \overset{d}{=} G_{\theta'}(z)$, it follows that $L_{DM}(\theta) = L_{DM}(\theta')$, even though in general $\nabla_\theta L_{DM}(\theta) \neq \nabla_\theta L_{DM}(\theta')$. Consequently, the stochastic optimization dynamics can drive students to converge toward different pairings with vastly different pairing inefficiencies Δ_E, despite achieving similar distributional fidelity.*

Proof. Given $G_\theta(z) \overset{d}{=} G_{\theta'}(z)$ equal in distribution, since $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$ is independent of $z \sim \mathcal{N}(0, \sigma^2(T)\mathbf{I})$, we have that

$$G_\theta(z) + \sigma(t)\varepsilon \overset{d}{=} G_{\theta'}(z) + \sigma(t)\varepsilon.$$

But by definition of the student probability path, this is just $p_{\theta,t} = p_{\theta',t}$ equal in distribution for all $t \in T$. Noting that KL-divergence is only dependent on the distribution supplied, it follows that

$$\begin{aligned} L_{\text{DM}}(\theta) &:= \int w(t)\text{KL}(p_{\theta,t} \| p_t)\, \mathrm{d}t. \\ &= \int w(t)\text{KL}(p_{\theta',t} \| p_t)\, \mathrm{d}t. \\ &= L_{\text{DM}}(\theta'). \end{aligned}$$

Finally, $\nabla_\theta L_{\text{DM}}(\theta) \neq \nabla_\theta L_{\text{DM}}(\theta')$ in general because $\partial G_\theta(z)/\partial\theta$ generally takes different values at different θ and θ'.

To illustrate how two student models can achieve similar (identical) distributional fits while learning pairings with vastly different efficiencies, consider the following geometric construction. Let the target distribution p_{data} be a singular measure in $\mathbb{R}^2$ supported on a slightly deformed circle $(1+\delta)x^2 + y^2 = 1$, where $\delta \ll 1$. This distribution is defined by pushing forward the isotropic Gaussian noise $(x, y) \sim N(0, \sigma^2 I)$ via the mapping T:

$$T(x, y) = \frac{(x, y)}{\sqrt{(1+\delta)x^2 + y^2}}.$$

Assume the teacher model has learned the distribution perfectly, such that $\Phi_K(x, y) := T(x, y)$. Now, consider a restricted class of student models G_θ which first applies a rotation θ to the noise space prior $\mathcal{N}(0, \sigma^2(T)\mathbf{I})$ then projects it onto the unit circle:

$$G_\theta(x, y) = \frac{(x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)}{\sqrt{x^2 + y^2}}.$$

Because the Gaussian noise source is rotationally invariant, G_θ generates the same output distribution p_{data} for any θ. Consequently, both $\theta = 0$ and $\theta = \pi$ are global optimizers of the distribution matching loss:

$$L_{\text{DM}}(\theta = 0) = L_{\text{DM}}(\theta = \pi) \approx 0.$$

However, these two solutions exhibit fundamentally different pairing efficiencies:

- At $\theta = 0$, the student's orientation aligns perfectly with the teacher's, resulting in $\Delta_E(\Phi_K, G_0) \approx 0$, manifesting clear copying behavior.
- At $\theta = \pi$, while the generated distribution is identical, the student reverses the orientation of the mapping relative to the teacher, resulting in a significantly larger $\Delta_E(\Phi_K, G_\pi) \gg 0$.

We note this example also works to show that in general $\nabla_\theta L_{\text{DM}}(\theta) \neq \nabla_\theta L_{\text{DM}}(\theta')$ for $\theta' = \theta + \pi$ for $\theta \notin \{0, \pi\}$.

□

Lemma C.2 (Properties of Pairing Inefficiency)**.** *The pairing inefficiency Δ_E is non-negative ($\Delta_E \geqslant 0$) and invariant under uniform scaling of all measures, i.e., $\Delta_E(c\Phi_K, cG) = \Delta_E(\Phi_K, G)$ for all $c > 0$, providing a robust measure for comparing copying behaviors across scales. Furthermore, the empirical estimator $\Delta_E^{(N)}$ is consistent, such that $\Delta_E^{(N)} \to \Delta_E$ almost surely as $N \to \infty$.*

Proof. Let π_{DT} denote the joint distribution of pairs $(\Phi_K(z), G(z))$ induced by the shared noise source $z \sim p_z$. By construction, π_{DT} is a valid coupling in the set of all couplings $\Gamma(p_\Phi, p_G)$. Since the optimal transport cost E_{OT} is defined as the infimum over this set, it follows that:

$$E_{DT}(\Phi_K, G) = \int \|x - y\|_2^2 \, d\pi_{DT}(x, y) \geqslant \inf_{\pi \in \Gamma} \int \|x - y\|_2^2 \, d\pi(x, y) = E_{OT}(\Phi_K, G).$$

Therefore,

$$\Delta_E = E_{DT}/E_{OT} - 1 \geqslant 0.$$

To show scale invariance, consider a constant $c > 0$. For any coupling $\pi \in \Gamma(p_\Phi, p_G)$, the scaled coupling $\pi_c = (c, c)_\# \pi$ belongs to $\Gamma(p_{c\Phi}, p_{cG})$. The scaled optimal transport cost is:

$$\begin{aligned} E_{OT}(c\Phi_K, cG) &= \min_{\pi_c \in \Gamma(p_{c\Phi}, p_{cG})} \int \|x - y\|_2^2 \, d\pi_c(x, y) \\ &= \min_{\pi \in \Gamma(p_\Phi, p_G)} \int \|cx - cy\|_2^2 \, d\pi(x, y) \\ &= c^2 \min_{\pi \in \Gamma(p_\Phi, p_G)} \int \|x - y\|_2^2 \, d\pi(x, y) \\ &= c^2 E_{OT}(\Phi_K, G) \end{aligned}$$

Similarly, $E_{DT}(c\Phi_K, cG) = \int \|cx - cy\|_2^2 \, d\pi_{DT}(x, y) = c^2 E_{DT}(\Phi_K, G)$. Because Δ_E is defined as a relative ratio, the factor c^2 cancels, yielding scale invariance

$$\Delta_E(c\Phi_K, cG) = \Delta_E(\Phi_K, G).$$

Finally, by Varadarajan's Theorem, the empirical measures $p_\Phi^{(N)}$ and $p_G^{(N)}$ converge weakly to p_Φ and p_G almost surely. Given that the images of Φ_K and G are bounded within a subset of compact support (due to the bounded non-linearities of the generators), the L^2 Wasserstein distance is continuous with respect to weak convergence. Consequently, the empirical estimator $\Delta_E^{(N)}$ converges almost surely to the population value Δ_E. □

D. Implementation Details

Network Architectures and Preconditioning. For our primary high-dimensional benchmarks, we utilize the conditional ImageNet-64 teacher model provided by Yin et al. (2024a), which employs the ADM architecture (Dhariwal & Nichol, 2021) with a base width of $C_{base} = 192$ and $label\_dim = 1000$. The student model is an architectural copy of this teacher, initialized directly from its pre-trained weights to observe the distillation phenomena. For the unconditional MNIST experiments (32×32, 1-channel), we adapt the same ADM architecture but reduce the capacity to $C_{base} = 64$ and set $label\_dim = 0$ to ensure strictly unconditional distillation. Both image-based models share a consistent UNet configuration, featuring a channel multiplier of $[1, 2, 3, 4]$, three residual blocks per resolution, and self-attention layers at resolutions of $32, 16$, and 8. EDM preconditioning (Karras et al., 2022) is applied to scale inputs, outputs, and skip connections by σ-dependent factors, maintaining unit variance throughout the distillation process. Finally, for the synthetic experiments on the 2D checkerboard manifold embedded in 4D space, we utilize a MLP architecture with 5 hidden layers of 384 hidden units, with 32-dimensional Fourier time embedding.

Distillation Dynamics and Optimization. Student models are initialized from their respective pre-trained teacher weights. A central feature of our distillation is the $1 : 5$ update ratio between the generator (G_θ) and the fake score model (s_ψ). The fake score model is optimized at every iteration to provide a high-quality surrogate of the evolving student's actual score $\nabla_x \log((G_\theta(z) * \mathcal{N}(0, t))(x))$ while the generator is updated via the distribution matching loss every five iterations. We utilize the Karras noise schedule (Karras et al., 2022) with parameters $\sigma_{\min} = 0.002, \sigma_{\max} = 80$, and $\rho = 7$. To ensure numerical stability and avoid training on uninformative noise levels at the extreme ends of the SDE trajectory, timesteps are sampled from a restricted interval of $[2\%, 98\%]$. All models are trained using the AdamW optimizer with a learning rate of 2×10^{-6}, weight decay of 0.01, and a 500-step linear warmup. Distillation for ImageNet-64 is conducted on a single NVIDIA H100 GPU for $50,000$ iterations using batch size of 32. MNIST and synthetic toy experiments are performed on the same GPU with batch sizes of 32 and 512 respectively.

E. Additional Figures

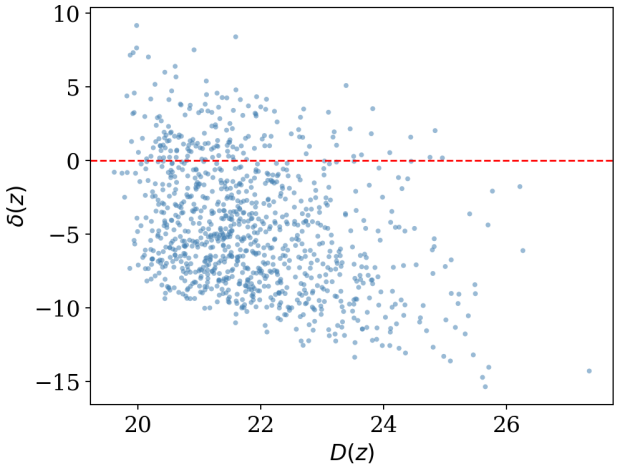


Figure 11. **Boundary Points are teacher's learned data manifold more likely copied.** For the student distilled from teacher trained on unconditional MNIST dataset for 8192 iterations, we plot with horizontal axis $D(z) = \text{Avg}_{x \in \text{train}}(\|\Phi_K(z) - x\|)$ the average distance to training set, and vertical axis $\delta(z) = \|G(z) - \Phi_K(z)\| - \|\Phi_1(z) - \Phi_K(z)\|$ the student relative displacement towards the teacher target. Points below the red dashed line indicate aligned pairs. A larger negative displacement indicates stronger alignment.

The figure below proves that the teacher model is not memorizing any of the training datapoints.

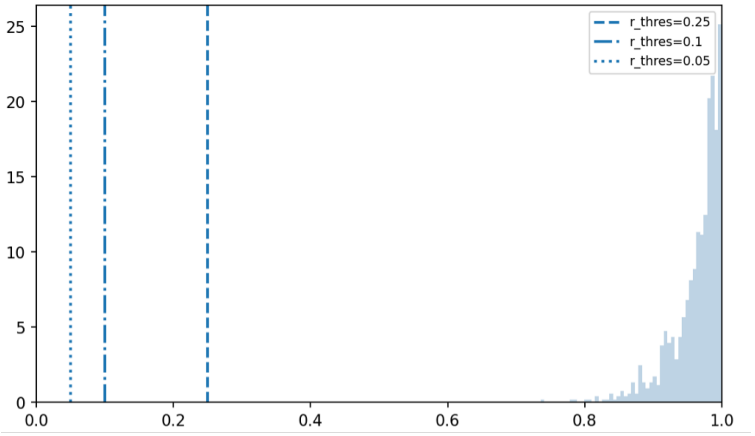


Figure 12. **Teacher model is not memorizing.** Distribution of memorization distance ratios $r(\Phi_8(z)) := \|x^2(\Phi_8(z)) - \Phi_8(z)\|/\|x^1(\Phi_8(z)) - \Phi_8(z)\|$ for the teacher model trained for 8192 iterations on unconditional MNIST.