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

Shucheng Li SHUCHENG.LI@CS.OX.AC.UK
Department of Computer Science, University of Oxford, Oxford, UK

Iolo Jones IOLO.JONES@CS.OX.AC.UK
Department of Computer Science, University of Oxford, Oxford, UK

Alexander Tong † ATONG@AITHYRA.AT
AITHYRA, Research Institute for Biomedical AI, Vienna, AT

Michael M. Bronstein † MICHAEL.BRONSTEIN@CS.OX.AC.UK
Department of Computer Science, University of Oxford, Oxford, UK
AITHYRA, Research Institute for Biomedical AI, Vienna, AT

Abstract

Distribution Matching Distillation (DMD) aligns noised distributions across scales to compress diffusion models. While Distribution Matching Distillation (DMD) is theoretically pairing-agnostic, we identify an emergent copying phenomenon: high-dimensional students spontaneously reproduce the teacher's original noise–data pairings. This behavior, absent in low-dimensional settings, is not an artifact of auxiliary losses or memorization. Instead, we argue that copying arises from the constrained geometric freedom of the student model during high-dimensional distillation.

Keywords: Diffusion Models, Distillation, High-dimensional Geometry, Learning Dynamics

1. Introduction

Diffusion models achieve state-of-the-art generation results across multiple modalities [4, 5, 7, 9, 11, 13] but suffer from high inference-time sampling latency. Distribution Matching Distillation (DMD) [14, 15] addresses this by training single-step generators via distribution-level alignment. While reproducibility, a phenomenon that multiple diffusion models with different architectures trained on the same dataset reproduce the same noise–data pairings, is well-documented [16], the pairings learned during distillation remain underexplored. In high-dimensional distribution matching distillation, we identify a non-trivial *copying* phenomenon: a student model faithfully reproducing teacher pairings despite a pairing-indifferent objective. Our key contributions are as follows:

- **Theoretical derivations and development of a quantitative metric.** We prove the DMD objective is pairing-indifferent, confirming copying is a non-trivial dynamical behavior. We define pairing inefficiency Δ_E, a scale-invariant measure of copying across datasets of various scales for fair comparison.
- **Geometric Characterization of Copying.** We demonstrate a sharp transition in copying between dimensionality regimes. We provide evidence that copying is not due to teacher

. † Equal supervision.

© S. Li, I. Jones, A.T. † & M.M.B. †.

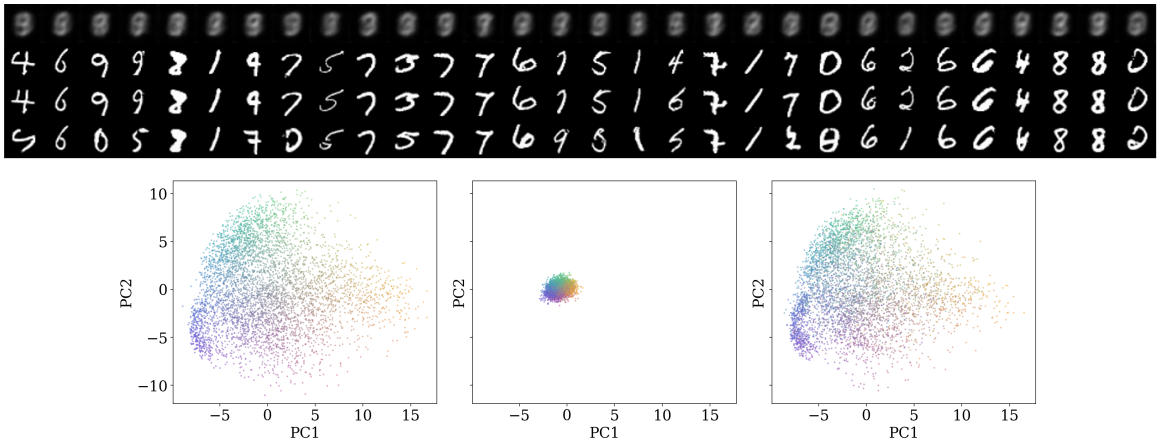


Figure 1: **Significant copying in high-dimensional settings.** The distilled student on unconditional MNIST aligns closely with the teacher. **Top:** Image quadruples from random seeds z: teacher 1-step $\Phi_1(z)$, 8-step $\Phi_8(z)$, 32-step $\Phi_{32}(z)$, and student 1-step $G(z)$. **Bottom:** PCA projection of 2000 triplets $(\Phi_8(z), \Phi_1(z), G(z))$ onto 2 leading PCs of $\Phi_8(z)$. Matching colors indicate identical noise seed z; $G(z)$ occupies the same manifold location as the multi-step teacher $\Phi_8(z)$. Pairing inefficiency is low ($\Delta_E \approx 0.0367$).

overfitting and memorization or the use of adversarial objectives but rather is governed by the data manifold geometry.

2. Background

2.1. Diffusion Models

Diffusion models map a data distribution p_{data} to a noise distribution $p_\varepsilon \approx \mathcal{N}(0, \sigma^2(T)\mathbf{I})$ via a forward stochastic process. In the Variance Exploding (VE) framework [6] with schedule $\sigma(t) = t$, the forward SDE is defined as $\mathrm{d}x_t = \sqrt{2t}\mathrm{d}\mathbf{W}_t$ over $0 \leqslant t \leqslant T$. Sampling reverses this process by solving the Probability Flow ODE:

$$\mathrm{d}x_t = -ts(x_t, t)\mathrm{d}t, \tag{1}$$

where $s(x_t, t) := \nabla_x \log p_t(x_t)$ is the Stein score. In practice we usually apply Tweedie's formula to reparametrize the intractable $s(x_t, t)$ by a neural network, $s(x_t, t) \approx (x_{0,\theta}(x_t, t) - x_t)/t^2$, where $x_{0,\theta}$ the neural network estimates the posterior expectation $\mathbb{E}[x_0|x_t]$ under the forward path measure.

2.2. Distribution Matching Distillation (DMD)

Distribution Matching Distillation (DMD) [15] distills the teacher score model $s(x_t, t)$ into a single-step generator $G_\theta(z)$ by aligning the student's noisy probability path $p_{\theta,t} := \text{Law}(G_\theta(z) + t\varepsilon)$ with the teacher's path p_t. This is achieved by minimizing the **Distribution Matching (DM) loss**:

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

The DM gradient is given by:

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

where a learnable student score s_ψ approximates $\nabla_x \log(p_{\theta,t}(x))$, the intractable score of the current student's noised probability path. During distillation, s_ψ is updated each iteration via denoising score matching, while G_θ is updated every 5 iterations using the gradient in Equation 3. Before distillation, we initialize both G_θ and s_ψ from the learned teacher score model $s(x_t,t)$ to ensure a stable starting point. In particular, we set $G_\theta(z) := z + T^2 s(z,T)$ and $s_\psi(x_t,t) = s(x_t,t)$.

3. The Copying Behavior

We identify a non-trivial *copying* phenomenon in high-dimensional DMD: the student faithfully reproduces the teacher's noise–data pairings, despite it is distilled using an objective (Equation 2) that only requires matching the teacher's distribution. As shown in Figure 1 for unconditional MNIST, the student single-step result $G_\theta(z)$ aligns closely with the teacher K-step result $\Phi_8(z)$, where $\Phi_K(z)$ is obtained by following K deterministic Euler steps of Equation 1 from $t = T$ back to $t = 0$ (with optional second-order corrections). The reason such behavior is unexpected is described in subsection 3.2.

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 metric $\Delta_E(\Phi_K, G_\theta)$ called *pairing inefficiency*, which quantifies the "cost of sub-optimality" in the student's learned noise–data mapping. It measures the percentage of L^2 energy wasted by following the teacher's specific seeds ($G_\theta(z) \to \Phi_K(z)$) instead of adopting the most efficient path (the Optimal Transport plan) to achieve the same output distribution.

Definition 1 (Pairing Inefficiency) *Let the initial noise be $z \sim p_\varepsilon = \mathcal{N}(0, T^2\mathbf{I})$. Let $p_\Phi = (\Phi_K)_\# p_\varepsilon$ and $p_G = G_{\theta\#} p_\varepsilon$ be the distributions learned by the teacher and student, respectively. The **optimal transport (OT) energy** and the **distillation transport (DT) energy** are defined as:*

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

$$E_{\text{DT}}(\Phi_K, G_\theta) := \int \|\Phi_K(z) - G(z)\|_2^2 \, \mathrm{d}p_\varepsilon(z), \tag{5}$$

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

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

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

3.2. Copying is Not Required for Successful Distillation

Crucially, copying is not a theoretical necessity of the DMD objective. The distribution matching loss L_{DM} is **pairing-indifferent**: it penalizes only the discrepancy between student and teacher distributions, and does not enforce pointwise alignment $G_\theta(z) \approx \Phi_K(z)$ for all $z \sim \mathcal{N}(0, T^2\mathbf{I})$.

Lemma 2 *Let G_θ and $G_{\theta'}$ be two student generators. Whenever $G_\theta(z) \overset{d}{=} G_{\theta'}(z)$, their DM losses are equal $L_{DM}(\theta) = L_{DM}(\theta')$, even though generally $\nabla_\theta L_{DM}(\theta) \neq \nabla_\theta L_{DM}(\theta')$. Consequently, stochastic optimization can lead to G_θ and $G_{\theta'}$ toward pairings with vastly different inefficiencies $\Delta_E(G, \Phi_K)$ despite achieving identical distributional fidelity $G_\theta(z) \overset{d}{=} G_{\theta'}(z) \overset{d}{=} \Phi_K(z)$.*

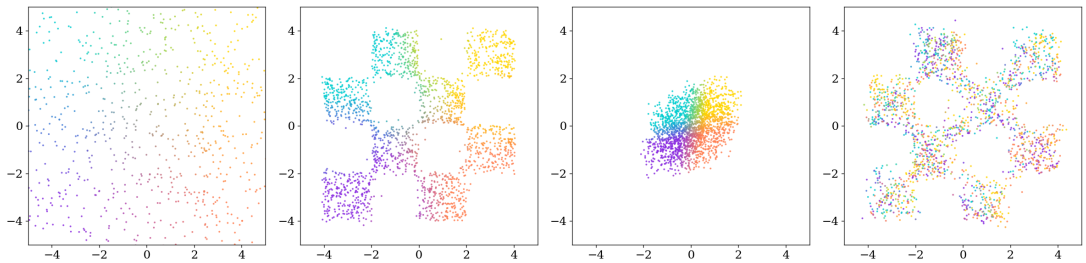


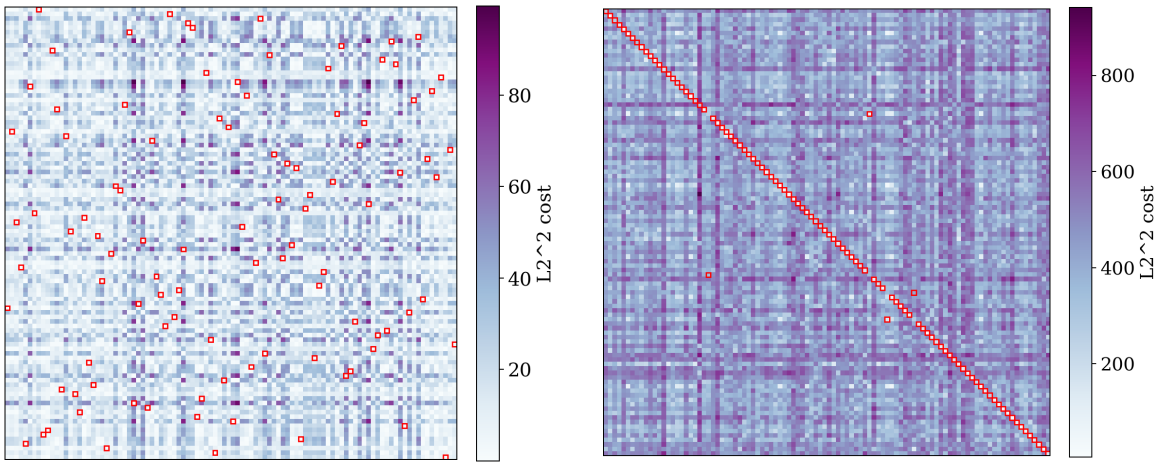
Figure 2: **Copying rarely occurs in low-dimensional settings.** The distilled student on a 2D chessboard dataset (embedded in $\mathbb{R}^4$) exhibits strong *remapping* rather than copying. Panels (left to right): initial noise z, teacher 8-step $\Phi_8(z)$, teacher 1-step $\Phi_1(z)$, and student 1-step $G(z)$. Matching colors represent identical noise seeds z. Points are projected onto the first two coordinates. We observe that while $\Phi_8(z) \overset{d}{=} G(z)$, $\Phi_8(z) \neq G(z)$ for general seed z, leading to high pairing inefficiency ($\Delta_E \approx 8.55$) in this setting.

3.3. Copying Rarely Occurs in Low Dimensions

We empirically observe that copying is weak in low-dimensional artificial settings. For a toy 2D 4×4 chessboard distribution embedded in $\mathbb{R}^4$, the distilled student successfully recovers p_{data} but exhibits significant remapping (Figures 2, 3), resulting in high pairing inefficiency ($\Delta_E \approx 8.55$). Similar remapping is observed across other low-dimensional synthetic datasets (Appendix B).

3.4. Copying Frequently Occurs in High Dimensions

In stark contrast, copying consistently emerges in high-dimensional tasks. We distill a single-step student from an EDM teacher [6] trained for 8192 iterations on unconditional MNIST. The student is highly prone to copying (Figures 1, 3), achieving an extremely low pairing inefficiency $\Delta_E \approx 0.0367$. Similar strong copying is observed across other high dimensional datasets, and equally profound in conditional generation settings (Appendix B). Such selection of copying solutions over other valid remappings suggests an emergent property of high-dimensional distillation dynamics.



(a) 2D Chessboard (b) Unconditional MNIST

Figure 3: **Dimensionality and Copying.** Heatmaps showing pairwise L_2^2 distances between teacher $\{\Phi_K(z_i)\}$ and student $\{G(z_j)\}$ samples . Red boxes indicate OT pairings; the diagonal indicates DT pairings. The chessboard student remaps pairings ($\Delta_E \approx 8.55$), whereas the MNIST student copies them ($\Delta_E \approx 0.0367$).

4. Analyzing the Mechanisms Behind Copying

In this section, we analyze the mechanisms driving the copying behavior. For clarity, we focus on the unconditional MNIST distillation setting.

4.1. Removal of Intuitive Explanations

We confirm that copying occurs without adversarial training or regression. The student model in Yin et al. [14] incorporates an auxiliary diffusion adversarial loss,

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

where $x_t' = G_\theta(z) + t\varepsilon$ and $x_t = x + t\varepsilon$. While an adversarial objective (or regression objective in [15]) could implicitly enforce copying, our experiments show copying persists even when these objectives are absent. This confirms that adversarial training is not a prerequisite for copying.

We also demonstrate that student copying is not triggered by the teacher trivially memorizing the training dataset. We define the memorization distance ratio for a generated point y as

$$r(y) := \|y - x^1(y)\| / \|y - x^2(y)\|,$$

where $x^1(y)$ and $x^2(y)$ are its nearest and second-nearest training neighbors. A point is considered memorized by the teacher if its memorization distance ratio is smaller than some threshold $r(y) < r_{\text{thres}}$. As shown in Figure 5, teacher models exhibit no signs of memorization on random samples.

4.2. Copying and Geometric Complexity

We view the DM distillation dynamics as a two-stage process (Figure 4). In the first stage, the DM loss increases as the surrogate student score s_ψ deviates from the teacher to track the true student score $\nabla_x \log p_{\theta,t}$. In the second stage, s_ψ provides accurate signals for the generator G_θ to minimize the discrepancy with the teacher. We conjecture that copying occurs in the second stage when the student has limited degrees of freedom to deform while preserving the teacher's target distribution (see Appendix D).

4.2.1. MICRO-LEVEL: BOUNDARY POINTS ARE MORE LIKELY COPIED

We observe that teacher samples distant from the training data bulk are more susceptible to copying. We define the relative displacement $\delta(z) = \|G(z) - \Phi_K(z)\| - \|\Phi_1(z) - \Phi_K(z)\|$ and the average distance to the training set $D(z) := \mathrm{Avg}_{x\in\mathrm{train}}(\|\Phi_K(z) - x\|)$. As shown in Figure 6, $\delta(z)$ negatively correlates with $D(z)$. Samples with high $D(z)$ reside near sparse or extremal regions of the manifold where remapping while maintaining distributional fidelity is geometrically constrained.

4.2.2. MACRO-LEVEL: TEACHER CONVERGENCE INCREASES COPYING

Copying behavior scales with teacher training iterations. We distilled students from snapshots of an unconditional MNIST teacher at various stages. Later-stage teachers yield students with significantly lower pairing inefficiency (Figures 7). As the teacher resolves finer geometric structures, it imposes stronger constraints that reduce the student's flexibility for remapping.

4.3. Related Work

Prior works show that independently trained diffusion models often reproduce similar noise–data pairings [1, 16]. Our work differs by studying copying that emerges during *distillation* under a distribution-matching objective. DMD and VSD [13, 15] match distributions, theoretically allowing latent remapping. In contrast, flow-based distillation [3, 10, 12] provides trajectory-level supervision. We show that even without such trajectories, high-dimensional geometry enforces alignment. The manifold hypothesis suggests images concentrate near low-dimensional structures [8]. These geometric constraints influence copying behavior during distillation.

5. Conclusion

We investigated the unexpected *copying* phenomenon in distribution matching distillation, wherein student models spontaneously reproduce teacher noise–data pairings despite receiving only distribution-level supervision. Our experiments demonstrate that copying is not an artifact of auxiliary adversarial objectives or teacher memorization. Instead, our empirical evidence suggests that copying is an emergent property that arises when high-dimensional geometric constraints limit the student's degrees of freedom to deform its distribution while remaining aligned with the teacher's score.

Acknowledgements

S.L. is supported by the CSC-PAG Oxford Scholarship, jointly awarded by the China Scholarship Council and the University of Oxford. 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

[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 *Advances in Neural Information Processing Systems*, 2025.

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

[3] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, 2025.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[5] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[6] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[7] 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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[8] Jakiw Pidstrigach. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems*, 2022.

[9] 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*, pages 10684–10695, 2022.

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

[11] 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.

[12] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[13] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems*, 2023.

[14] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 2024.

[15] 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*, pages 6613–6623, 2024.

[16] Huijie Zhang, Jinfan Zhou, Yifu Lu, Minzhe Guo, Peng Wang, Liyue Shen, and Qing Qu. The emergence of reproducibility and consistency in diffusion models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60558–60590. PMLR, 21–27 Jul 2024.

Appendix A. Additional Figures

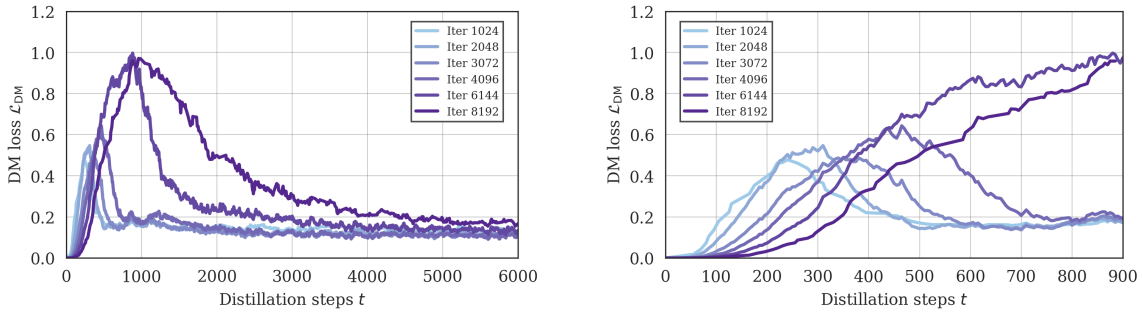


Figure 4: **Two Stages of Distribution Matching Distillation.** DM loss evolution for unconditional MNIST students initialized from various teacher snapshots (1024–8192 iterations). All cases exhibit a characteristic increase-decrease evolution. The right panel zooms into the first 900 iterations, highlighting the rapid initial dynamics.

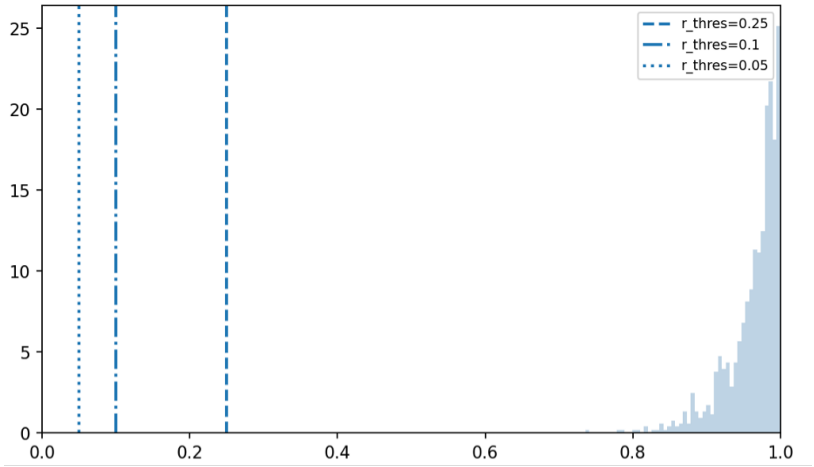


Figure 5: **The teacher model has not memorized any of the training datapoints.** Distribution of memorization distance ratios $r(\Phi_8(z)) := \|x^2(\Phi_8(z)) - \Phi_8(z)\|/\|x^1(\Phi_8(z)) - \Phi_8(z)\|$ for the teacher model trained for 8192 iterations on unconditional MNIST.

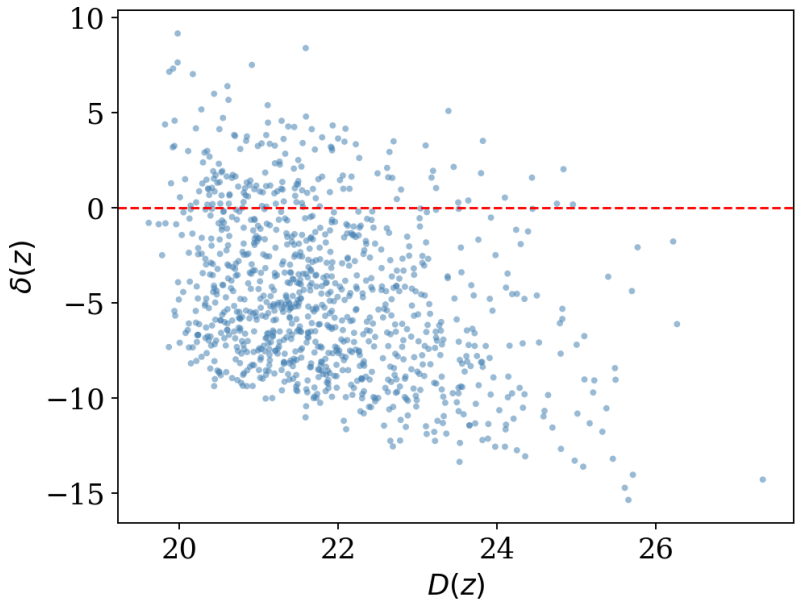
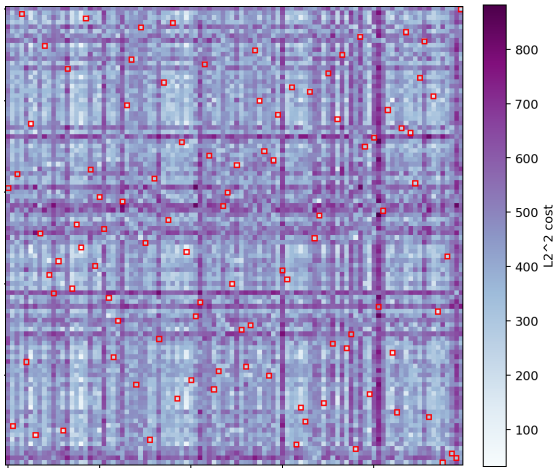


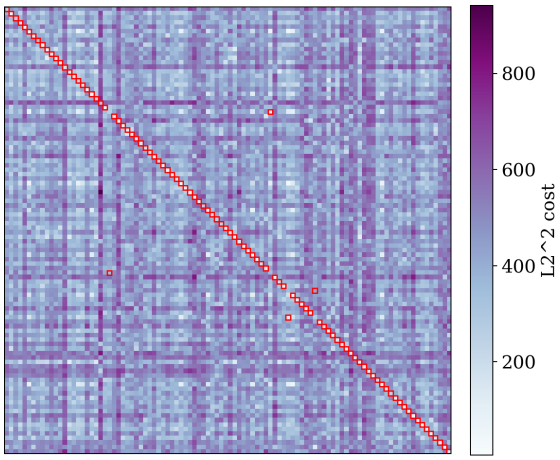
Figure 6: **Boundary points are more likely copied.** Student relative displacement $\delta(z)$ versus average distance to training set $D(z)$. Points below the dashed line indicate alignment; larger negative values signify stronger copying in sparse manifold regions.

(a) Partially trained teacher results (1k iterations)

(b) Converged teacher results (8k iterations)



(c) Heatmap: 1k Iterations ($\Delta_E \approx 1.05$)



(d) Heatmap: 8k Iterations ($\Delta_E \approx 0.0367$)

Figure 7: **Longer trained teachers are more likely copied. Top panels (a, b):** Comparison of generation triplets across different teacher training stages. The converged teacher (8k) shows much tighter alignment between the student and the multi-step teacher. **Bottom panels (c, d):** Distance heatmaps (described in Figure 3) reveal a sharp transition from remapping (off-diagonal noise) to copying (clean diagonal) as the teacher resolves the data manifold geometry.

Appendix B. Copying and Remapping Behaviors Across Diverse Datasets

Consistent with the 2D chessboard experiments, we observe that DMD students distilled on low-dimensional synthetic manifolds exhibit *substantial remapping*. In these regimes, the student frequently identifies noise–data pairings that deviate significantly from the teacher's trajectories, often resulting in $\|G(z) - \Phi_8(z)\| \gg \|\Phi_1(z) - \Phi_8(z)\|$ for a given seed z.

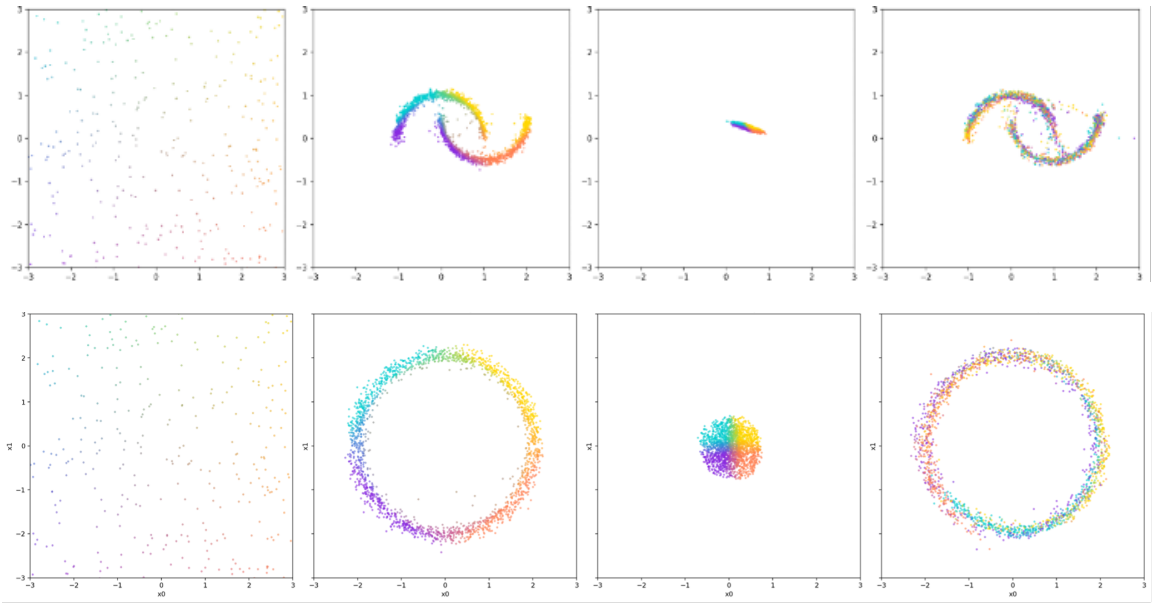


Figure 8: **Remapping behavior on low-dimensional synthetic datasets.** Each row depicts initial noise z, teacher 8-step samples $\Phi_8(z)$, teacher 1-step samples $\Phi_1(z)$, and student samples $G(z)$. Matching colors denote identical seeds. **Top:** 2D double-moons manifold embedded in $\mathbb{R}^4$. **Bottom:** Gaussian mixture with 32 components in $\mathbb{R}^4$. In both cases, the student recovers the distribution but fails to align with teacher trajectories, indicating significant remapping freedom.

In contrast, student learning high-dimensional datasets exhibits *significantly stronger copying*, which is further intensified in *conditional* generation settings (Figures 9–11). We conjecture that class-conditioning allows the teacher to resolve finer geometric structures within each sub-manifold. This imposes more rigid constraints during distillation, effectively exhausting the student's capacity to remap pairings while maintaining distributional fidelity.

Notably, copying is not restricted to natural images. To isolate the effect of dimensionality from semantic content, we constructed a synthetic high-dimensional dataset by mapping 16D Gaussian noise through a random two-layer MLP into $\mathbb{R}^{32\times 32}$, followed by superimposing class-specific whitened stripes. This creates a distribution supported on a 16D manifold in a 1024D ambient space. As shown in Figures 10 and 11, the student on this "MLP-manifold" dataset exhibits copying behavior indistinguishable from that observed on MNIST or ImageNet.

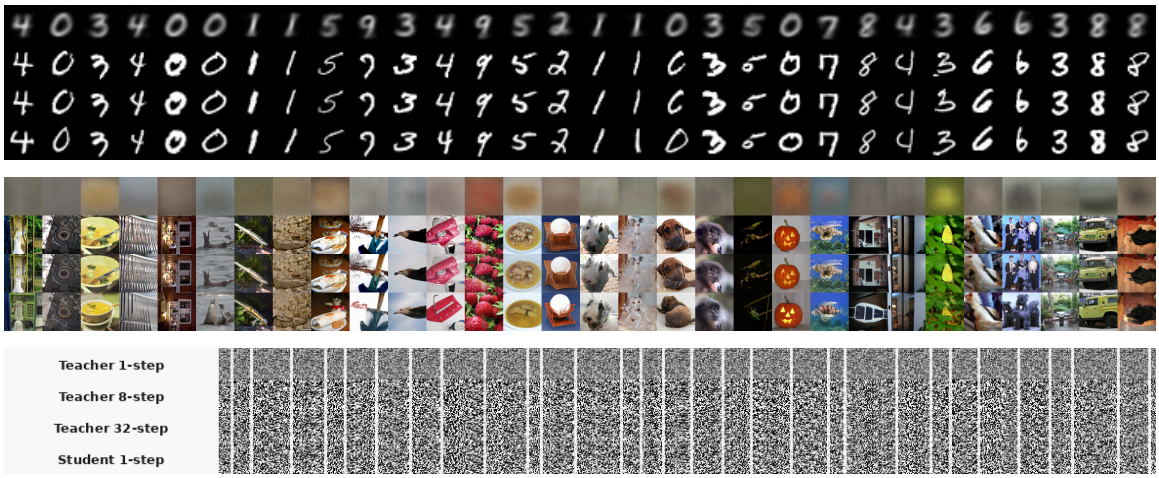


Figure 9: **Copying behavior in conditional high-dimensional regimes.** Panels (top to bottom): Conditional MNIST, ImageNet-64, and a synthetic MLP-manifold. Within each panel, rows show $\Phi_1(z)$, $\Phi_8(z)$, $\Phi_{32}(z)$, and $G(z)$ for identical seeds. The student $G(z)$ visually aligns with the multi-step teacher $\Phi_{32}(z)$ across all modalities.

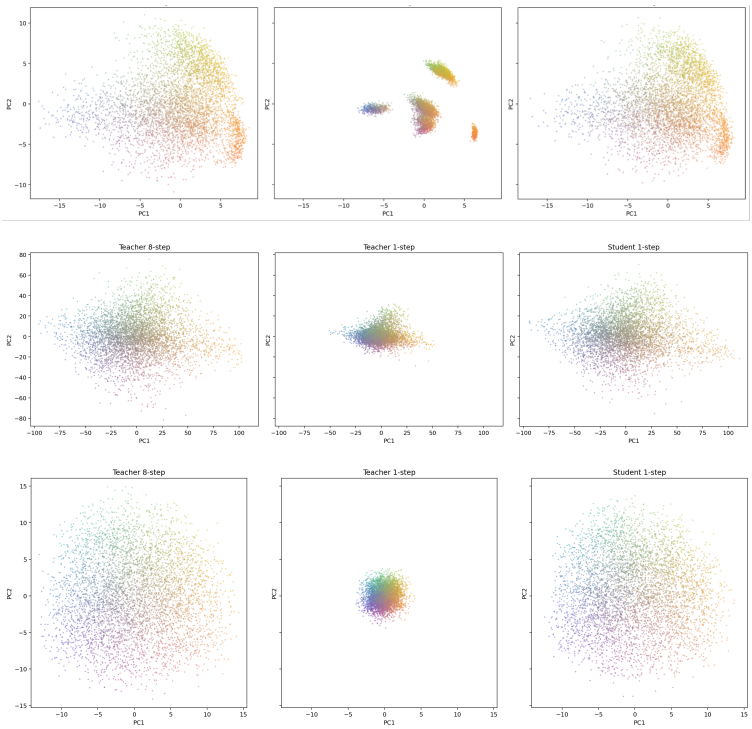


Figure 10: **Principal Component Visualization of Alignment.** Rows (top to bottom): Conditional MNIST, ImageNet-64, and synthetic MLP-manifold. Matching colors denote identical seeds, and conditional labels are uniformly randomly assigned to all seeds. We visualize 2000 triplets $(\Phi_8(z), \Phi_1(z), G(z))$ projected onto the top two PCs of $\Phi_8(z)$. Consistent student copying is observed by observing color alignment between $G(z)$ and $\Phi_8(z)$.

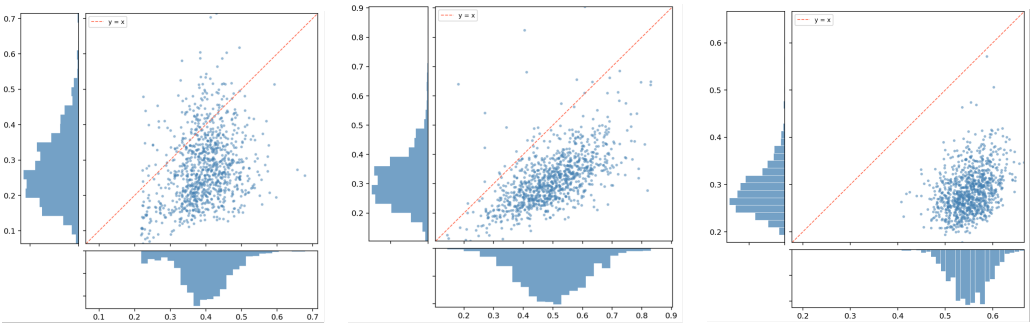


Figure 11: **Pointwise Displacement Correlation.** Comparing $\|G(z) - \Phi_8(z)\|$ (vertical) against $\|\Phi_1(z) - \Phi_8(z)\|$ (horizontal). Across MNIST (left), ImageNet-64 (middle), and the MLP-manifold (right), the student's proximity to the converged teacher (Φ_8) is consistently higher than that of the single-step teacher (Φ_1), indicating a strong pointwise copying effect.

Appendix C. Copying Behavior Across Teacher Training Stages

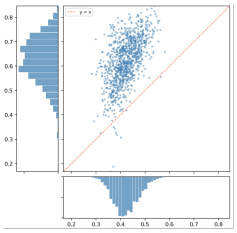
We analyze the evolution of copying behavior by distilling students from teacher snapshots at 1024, 2048, 4096, and 8192 training iterations on unconditional MNIST. Each student is distilled for 50k iterations to ensure convergence. Our results confirm that students initialized from later-stage teacher checkpoints exhibit progressively stronger alignment with the teacher's original noise–data pairings.

(a) Teacher: 1024 iterations

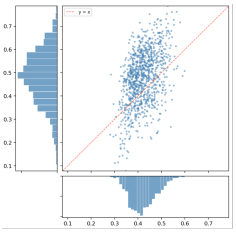
(b) Teacher: 2048 iterations

(c) Teacher: 4096 iterations

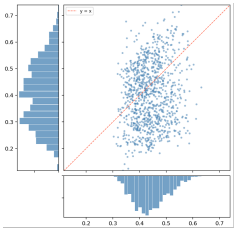
(d) Teacher: 8192 iterations



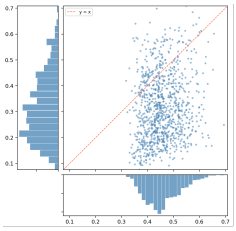
$\Delta_E \approx 1.05$



$\Delta_E \approx 0.389$



$\Delta_E \approx 0.237$



$\Delta_E \approx 0.0367$

Figure 12: **Evolution of copying during teacher training. Top (a–d):** Qualitative comparison of distillation results from different teacher snapshots on unconditional MNIST. Each quintuple shows $\Phi_1(z)$, $\Phi_8(z)$, $G(z)$, and nearest neighbors x^1, x^2 from the training set. Alignment between $G(z)$ and $\Phi_8(z)$ improves as training progresses. **Bottom:** Quantitative alignment plots showing $\|G(z) - \Phi_8(z)\|$ (vertical) vs. $\|\Phi_1(z) - \Phi_8(z)\|$ (horizontal). The sharp decrease in pairing inefficiency Δ_E from 1.05 to 0.0367 highlights that teacher convergence is a primary driver of student copying.

Appendix D. Derivations

Proof of Lemma 2: 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 $\nabla_\theta L_{\mathrm{DM}}(\theta) \neq \nabla_\theta L_{\mathrm{DM}}(\theta')$ for $\theta' = \theta + \pi$ for $\theta \notin \{0, \pi\}$.

Lemma 3 (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 as $N \to \infty$. ■

Intuition for the Geometric Conjecture

To illustrate the intuition that higher degree of copying arises from limited geometric freedom, 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[1], 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.

1. 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.

Appendix E. 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. [14], which employs the ADM architecture [2] 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 [6] 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 the 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 $\nabla_x \log((G_\theta(z) * \mathcal{N}(0, t^2\mathbf{I}))(x))$ while the generator is updated via the distribution matching loss every five iterations. We utilize the Karras noise schedule [6] 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.