DIFFUSION AS A MULTI-SCALE DISCRIMINATOR: RETHINKING NOISE SCALES FOR DIFFUSION MODEL DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose MuSD (Multi-Scale Discriminator), a simple framework for diffusion distillation that uses a frozen diffusion teacher to guide adversarial training. We train lightweight discriminator heads on the teacher's features at multiple resolutions, without training an auxiliary score network or using a separate pretrained feature extractor. Prior work that uses the teacher as the discriminator biases its noise levels toward high noise to preserve global coherence. We observe that fine-scale structure in teacher features survives mainly at low noise, whereas the coarse focus that features acquire at high noise is also available from coarse readouts at low noise. Building on this, we derive a Beta noise sampler as the maximum-entropy density under a corruption budget, which favors low noise while keeping positive density over the whole noise interval, and we show that both properties are needed. On CIFAR-10, AFHQv2, FFHQ, and ImageNet, MuSD achieves competitive or better one-step FID than the baselines with fewer training images and lower GPU memory use. It applies to both U-Net and Transformer teachers, including EDM2-S and DiT-XL/2, where the one-step student surpasses its 250-step teacher, and extends to two-step text-to-image generation with SD3.5-Medium.

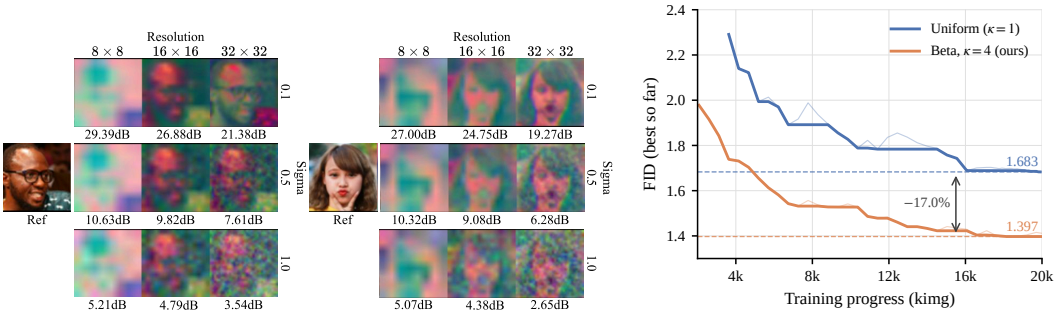
1 INTRODUCTION

Diffusion models achieve high-quality image synthesis but require repeated network evaluations for generation. Distillation reduces this cost by training one- or few-step generators. Beyond generation, pretrained diffusion models also provide representations useful for classification, segmentation, and dense prediction (Baranchuk et al., 2022; Li et al., 2023; Meng et al., 2024), which offer a source of supervision for diffusion distillation.

Many distillation methods introduce additional supervisory components, such as auxiliary score networks or pretrained feature extractors (Zhou et al., 2024; 2025; Yin et al., 2024b;a). Frozen pretrained networks have also served as GAN discriminator backbones (Sauer et al., 2021; Kumari et al., 2022), and LADD (Sauer et al., 2024) shows that a diffusion teacher can itself provide such features for few-step text-to-image distillation, biasing the discriminator toward high noise to preserve global coherence. This raises a key question: *which noise levels should be used to extract features for adversarial supervision?* Diffusion representations vary with the corruption level (Choi et al., 2022; Huang et al., 2026), so a noise distribution suited to diffusion training need not suit adversarial distillation.

Figure 1(a) shows that teacher features lose spatial structure as noise increases, and that fine-resolution features lose it first. Supervising fine readouts therefore calls for low noise, while coarse structure can be obtained from low-resolution teacher features without heavy corruption. Motivated by this, we propose MuSD (Multi-Scale Discriminator), which trains lightweight multi-resolution discriminator heads on a frozen teacher with a noise sampler biased toward low noise. We derive this sampler as the maximum-entropy density under a corruption budget, which favors low noise while keeping positive density over the whole noise interval. Figure 1(b) shows that it improves convergence and final FID over uniform sampling, and our ablations show that both the low-noise preference and the full support are needed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Teacher features across resolutions and noise levels.

(b) Best-so-far FID during training on conditional CIFAR-10.

Figure 1: **Noise sampling shapes adversarial supervision.** (a) Teacher features at three resolutions and $\sigma \in \{0.1, 0.5, 1.0\}$; each value is the PSNR against the same-resolution feature at $\sigma = 0.001$. Structure degrades as noise increases, and fastest at the finest resolution. (b) Low-noise-biased sampling converges faster and reaches a lower FID than uniform sampling over the same interval.

On standard one-step benchmarks, where score-based methods with auxiliary networks report the strongest results, MuSD achieves competitive or superior FID with fewer training images than SiDA (Zhou et al., 2025) and Uni-Instruct (Wang et al., 2026); with EDM2-S (Karras et al., 2024) at 512×512, it reduces peak GPU memory from 52.4 to $34.4\,\mathrm{GB}$ per GPU relative to SiDA at the same batch size. MuSD also extends to Transformer teachers: a one-step DiT-XL/2 (Peebles & Xie, 2023) student surpasses its 250-step teacher, and a two-step SD3.5-Medium (Stability AI, 2024) student runs about 28 times faster than its 40-step teacher.

Contributions.

1. **Noise-dependent supervision.** We analyze how noise shapes teacher features across scales and show that the discriminator's noise distribution should favor low noise while keeping full support.
2. **Maximum-entropy noise sampling.** We derive a Beta sampler as the maximum-entropy density under a corruption budget.
3. **Efficient distillation across backbones.** We achieve competitive FID at lower training cost with U-Net and Transformer teachers, including text-to-image generation.

2 Related Work

Efficient sampling and diffusion distillation. Many works reduce the computational cost of diffusion generation. DDIM (Song et al., 2021) and DPM-Solver (Lu et al., 2022) accelerate sampling without retraining the diffusion model, whereas distillation trains a student to generate samples in one or a few steps. DMD (Yin et al., 2024b) uses the difference between teacher and student scores, and SiD (Zhou et al., 2024) derives its objective from score identities; both train an auxiliary network to estimate the score of the evolving student distribution, and DMD2 (Yin et al., 2024a) and SiDA (Zhou et al., 2025) add adversarial supervision. Uni-Instruct (Wang et al., 2026) unifies such divergence-based objectives, and D2O (Zheng & Yang, 2025) shows that adversarial training alone can turn a pretrained diffusion model into a one-step generator, using a discriminator built on an external pretrained network. Our approach uses adversarial supervision from frozen teacher features without an auxiliary score network (Section 3).

Frozen features for adversarial supervision. Pretrained diffusion models provide representations useful beyond generation, for example for segmentation (Baranchuk et al., 2022) and classification (Li et al., 2023). In GANs, Projected GAN (Sauer et al., 2021) and Vision-aided GAN (Kumari et al., 2022) train lightweight heads on frozen pretrained networks, and multi-scale discriminators are well established (Wang et al., 2018). In diffusion distillation, ADD (Sauer et al., 2025) uses DINOv2 (Oquab et al., 2024) features, and SwD (Starodubcev et al., 2026) matches intermediate diffusion features with a patch-level Maximum Mean Discrepancy (MMD) (Gretton et al., 2012)

objective alongside scale-wise generation. Closest to our work, LADD (Sauer et al., 2024) uses the frozen diffusion teacher itself as the discriminator backbone and distills SD3 into a few-step text-to-image model with teacher-generated data. LADD finds that biasing the discriminator's noise levels toward low noise loses global coherence and therefore samples from a logit-normal distribution shifted toward high noise. MuSD takes the opposite direction: its multi-resolution readouts provide coarse supervision at low noise, so its maximum-entropy sampler favors low noise while keeping full support. MuSD is also trained on real images and evaluated on standard one-step benchmarks against score-based methods.

Noise-level sampling and weighting. The choice of noise levels is an important design decision in diffusion training. EDM (Karras et al., 2022) studies noise sampling, loss weighting, and network preconditioning together; P2 weighting (Choi et al., 2022) prioritizes noise levels associated with perceptually rich learning tasks, and Min-SNR weighting (Hang et al., 2023) balances conflicting optimization directions across timesteps. In adversarial learning, instance noise (Sønderby et al., 2017; Arjovsky & Bottou, 2017) and Diffusion-GAN (Wang et al., 2023) corrupt both real and generated samples so that their distributions overlap, which stabilizes training. In diffusion distillation, SwD (Starodubcev et al., 2026) reports slightly better feature matching within a low-to-mid noise range, while LADD favors high noise as discussed above. We derive the discriminator's noise distribution from a maximum-entropy principle and show that it should favor low noise while keeping full support over the noise interval.

3 Preliminaries

3.1 Diffusion Model Distillation

Diffusion models generate samples by iterative denoising, reversing a forward corruption process:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t; a_t\mathbf{x}_0, \sigma_t^2\mathbf{I}\big)\,, \tag{1}$$

where a_t and σ_t control the signal and noise levels. Let $p_{\boldsymbol{\phi}}(\cdot \mid c)$ denote the generation distribution of a pretrained teacher, conditioned on an optional class label or text prompt c. A student generates $\mathbf{x}_{\boldsymbol{\theta}} = G_{\boldsymbol{\theta}}(\mathbf{z}, c)$ from $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, inducing $p_{\boldsymbol{\theta}}(\cdot \mid c)$. Here, $G_{\boldsymbol{\theta}}$ denotes the complete one- or few-step sampling procedure in pixel or latent space. We omit c below for readability.

Applying the same forward process to teacher and student samples yields $p_{\boldsymbol{\phi},t}$ and $p_{\boldsymbol{\theta},t}$. For Gaussian corruption with $a_t \neq 0$, exact matching at any fixed noise level implies matching the clean distributions (Wang et al., 2023; Zhou et al., 2024):

$$p_{\boldsymbol{\theta},t} = p_{\boldsymbol{\phi},t} \implies p_{\boldsymbol{\theta}} = p_{\boldsymbol{\phi}}, \qquad a_t \neq 0. \tag{2}$$

In practice, distillation objectives combine discrepancies across noise levels, providing supervision at different corruption strengths (Luo et al., 2023; Zhou et al., 2024). Throughout, t indexes supervision noise rather than student generation steps.

3.2 Score-Based Distribution Matching

DMD (Yin et al., 2024b;a) and SiD (Zhou et al., 2024) match $p_{\boldsymbol{\theta},t}$ to $p_{\boldsymbol{\phi},t}$ through the teacher score $\mathbf{s}_{\boldsymbol{\phi}} = \nabla_{\mathbf{x}_t} \log p_{\boldsymbol{\phi},t}(\mathbf{x}_t)$ and the student score $\mathbf{s}_{\boldsymbol{\theta}} = \nabla_{\mathbf{x}_t} \log p_{\boldsymbol{\theta},t}(\mathbf{x}_t)$, minimizing a reverse KL divergence and a Fisher divergence across noise levels, respectively. Because the student score is unavailable, both estimate it with an auxiliary fake-score model $\mathbf{s}_{\boldsymbol{\psi}} \approx \mathbf{s}_{\boldsymbol{\theta}}$, trained by denoising score matching on current student samples.

3.3 Adversarial Distribution Matching

Adversarial learning (Goodfellow et al., 2014) trains a discriminator $\mathbf{D}_{\boldsymbol{\eta}}(\mathbf{x}_t, t) \in \mathbb{R}^n$ to distinguish noisy reference samples from student samples. The output may contain a single score ($n = 1$) or multiple scores, such as patch-level logits. Let $p_{\mathrm{ref},t}$ denote the noisy reference distribution; we use real images as references in all experiments. We minimize the following hinge adversarial losses with respect to the discriminator and generator parameters, respectively:

Adversarial learning: Match distributions through discrimination.

$$\mathcal{L}_D = \mathbb{E}_{t\sim\pi}\Big[\mathbb{E}_{\mathbf{x}_t\sim p_{\mathrm{ref},t}}\frac{1}{n}\mathbf{1}^\top\big[\mathbf{1}-\mathbf{D}_{\boldsymbol{\eta}}(\mathbf{x}_t,t)\big]_+ + \mathbb{E}_{\mathbf{x}_t\sim p_{\boldsymbol{\theta},t}}\frac{1}{n}\mathbf{1}^\top\big[\mathbf{1}+\mathbf{D}_{\boldsymbol{\eta}}(\mathbf{x}_t,t)\big]_+\Big], \tag{3}$$

$$\mathcal{L}_G = -\mathbb{E}_{t\sim\pi,\,\mathbf{x}_t\sim p_{\boldsymbol{\theta},t}}\left[\frac{1}{n}\mathbf{1}^\top\mathbf{D}_{\boldsymbol{\eta}}(\mathbf{x}_t,t)\right], \tag{4}$$

where $\mathbf{1}\in\mathbb{R}^n$ is the all-ones vector and $[\cdot]_+$ applies $\max(0,\cdot)$ elementwise.

The discriminator encourages each reference logit to be at least 1 and each generated logit to be at most -1, while the student learns to increase the mean score of its generated samples.

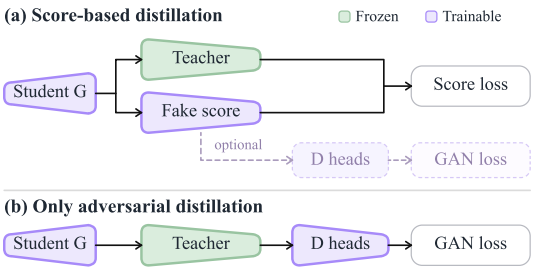


Figure 2: **Supervision structures.** (a) Score-based methods, with an optional adversarial branch. (b) Our frozen-feature discriminator.

Figure 2 contrasts the supervision structures. Our discriminator uses frozen teacher features: $\mathbf{D}_{\boldsymbol{\eta}}(\mathbf{x}_t,t) = h_{\boldsymbol{\eta}}(F_{\boldsymbol{\phi}}(\mathbf{x}_t,t))$, where $h_{\boldsymbol{\eta}}$ denotes trainable heads that output real values; the noise level enters only through the teacher features. Unlike score-based methods, this formulation does not require an auxiliary fake-score model. During student updates, the discriminator parameters are held fixed, while gradients propagate through the heads and the frozen backbone to the generator.

The sampling distribution $\pi(t)$ determines which noise-conditioned features provide supervision. We next examine this choice and introduce our noise-sampling strategy.

4 Method

Our method combines frozen diffusion features with a noise-sampling strategy designed for adversarial supervision. We use multi-resolution discriminator readouts to compare samples at different spatial scales. Fine readouts yield clean features mainly at low noise, since fine-scale structure is the first to be corrupted as noise increases, and coarse readouts provide at low noise the coarse focus that teacher features otherwise acquire only at high noise. We therefore favor lower-noise supervision under a corruption budget and derive a Beta sampling family.

4.1 Multi-Resolution Discriminator Readouts (MRD)

As illustrated in Figure 3, we extract intermediate features from a frozen teacher and process them with independent discriminator heads. All branches share the same corrupted input and teacher pass. For each selected teacher layer, we index the readout scales from the coarsest ($s=1$) to the finest ($s=S$), omitting the layer index for readability.

Each head produces a patch-level logit map:

$$\mathbf{D}_s = h_s(\widetilde{\mathbf{f}}_s), \qquad s=1,\dots,S, \tag{5}$$

where $\widetilde{\mathbf{f}}_s$ is the feature supplied to the head. The heads operate in parallel, and an aggregation head additionally combines their flattened logit maps. We apply the hinge objectives in Eqs. 3 and 4 to both branch and aggregated outputs, averaging patch losses within each branch and summing across branches.

Hierarchical backbones. For U-Net teachers (Karras et al., 2022), we directly use intermediate features at their native spatial resolutions.

Transformer backbones. Transformer teachers operate on a single token grid without a native spatial pyramid. For DiT-XL/2 (Peebles & Xie, 2023), we attach heads to the token grids of blocks 6, 12, and 18. These blocks already form a hierarchy in depth: deeper blocks vary smoothly over

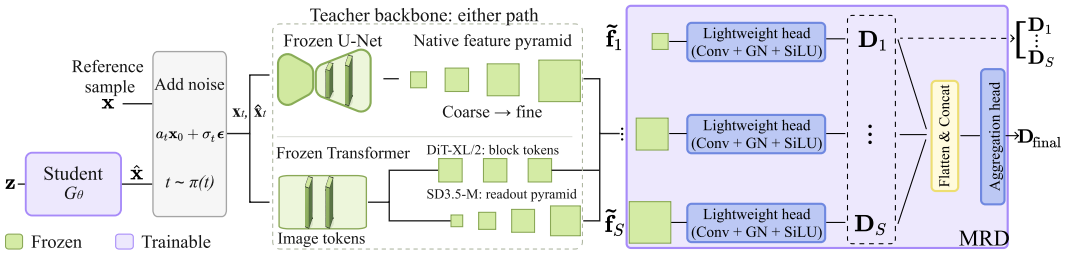


Figure 3: **Adversarial supervision from frozen diffusion features.** Reference and generated samples are corrupted at $t \sim \pi(t)$ and processed by a frozen teacher. For U-Net teachers, MRD uses native multi-resolution features; for SD3.5-M, it constructs a readout pyramid from intermediate image tokens. Trainable discriminator heads provide adversarial supervision.

object-level regions and act as coarse readouts, whereas earlier blocks reflect local texture (Appendix A.3.3). For SD3.5-M (Stability AI, 2024), we additionally construct a readout pyramid $\{\mathbf{f}_s\}_{s=1}^{S}$ from the token grid of each selected block using fixed low-pass filtering and downsampling. To reduce repeated low-frequency content, we use

$$\widetilde{\mathbf{f}}_s = \begin{cases} \mathbf{f}_1, & s = 1, \\ \mathbf{f}_s - \alpha\,\mathcal{U}(\mathbf{f}_{s-1}), & s = 2, \dots, S, \end{cases} \tag{6}$$

where $\mathcal{U}$ upsamples and filters the next-coarser feature to match the current resolution. We set $\alpha = 0.5$, retaining part of the shared low-frequency content at finer scales. The pyramid is constructed within the readout; the teacher operates at a single spatial resolution.

4.2 Maximum-Entropy Noise Sampling

We probe how well frozen SD3.5-M features at different noise levels and readout scales linearly reconstruct image structure at several spatial scales.

Observation 1: Coarse focus without heavy corruption. At high noise, the structure recoverable from teacher features becomes concentrated on coarse scales. Coarse readouts at low noise show a similar concentration while retaining more coarse structure than high-noise features.

The probing protocol and results are provided in Appendix A.1. As noise increases, fine-scale structure vanishes first, so fine readouts are informative mainly at low noise, and Section 5.4 shows that they are necessary, since coarse readouts alone degrade FID substantially. High noise reaches coarse focus by removing fine-scale structure, whereas coarse readouts reach it without increasing input corruption. Low noise can therefore receive most of the supervision: fine readouts require it, and coarse readouts retain coarse focus under it. High noise also plays a second role that coarse readouts do not replace: corrupting both reference and generated samples increases the overlap of their distributions, which stabilizes adversarial training (Arjovsky & Bottou, 2017; Sønderby et al., 2017; Wang et al., 2023). The sampler should therefore keep positive density over the whole interval.

Corruption cost. Let $t \in (t_{\min}, t_{\max})$ denote the supervision noise level (σ for EDM-family teachers and the diffusion timestep otherwise), with larger t meaning stronger corruption. The interval is set per teacher, with $t_{\min}$ the lowest level at which the teacher's features remain reliable (Appendix A.3.1). We measure corruption by

$$C(t) = -\log\left(\frac{t_{\max} - t}{t_{\max} - t_{\min}}\right). \tag{7}$$

C vanishes at $t_{\min}$ and diverges at $t_{\max}$. Under a uniform reference $T_0 \sim \mathcal{U}(t_{\min}, t_{\max})$, $C(t)$ is the surprisal of the event $T_0 \geq t$, and uniform sampling has expected cost 1.

Observation 1 motivates a preference for low noise but does not specify the shape of the distribution. We impose this preference as a budget b on the expected cost and assume nothing else; that is, among all densities that satisfy the budget, we choose the one with maximum entropy (Jaynes, 1957).

Sampling principle: Favor low noise while preserving coverage. We choose

$$\begin{aligned} \pi^\star &= \arg\max_{\pi} H(\pi), \\ \text{subject to} \quad & \mathbb{E}_{t\sim\pi}[C(t)] \le b, \qquad 0 < b < 1, \end{aligned} \tag{8}$$

where π is a density on $(t_{\min}, t_{\max})$ and $H(\pi) = -\int_{t_{\min}}^{t_{\max}} \pi(t)\log\pi(t)\,dt$.

Since $H(\pi) = \log(t_{\max} - t_{\min}) - \mathrm{KL}(\pi \,\|\, \mathcal{U})$, the solution is the density within the budget that is closest to uniform. Its only free parameter is the budget b; a smaller budget enforces a stronger preference for lower noise.

Proposition 1: Biased sampling. The unique maximizing density in Eq. 8 is

$$\pi^\star(t) = \frac{\kappa}{t_{\max} - t_{\min}} \left(\frac{t_{\max} - t}{t_{\max} - t_{\min}} \right)^{\kappa - 1}, \qquad \kappa = \frac{1}{b} > 1. \tag{9}$$

Equivalently, $(t - t_{\min})/(t_{\max} - t_{\min}) \sim \mathrm{Beta}(1, \kappa)$, and the budget is active: $\mathbb{E}_{\pi^\star}[C] = b$.

The proof is provided in Appendix A.2. We sample from this density as

$$t = t_{\min} + (t_{\max} - t_{\min}) \left(1 - v^{1/\kappa}\right), \qquad v \sim \mathcal{U}(0, 1). \tag{10}$$

Larger κ allocates more supervision to lower noise; every finite κ keeps positive density on the whole interval, and $\kappa = 1$ recovers uniform sampling. We use $\kappa \in \{3, 4\}$ (Appendix A.3.1) and study its effect in Section 5.4.

4.3 Text-Conditioned Supervision

Two-step text-to-image generation. For text-to-image synthesis, each step of the student predicts a clean sample: the first step predicts $\widehat{\mathbf{x}}_0$ from Gaussian noise and uses it to form an input at $t = 700$, from which the second step predicts the final clean sample. Both steps receive adversarial supervision against the reference distribution at their respective target timesteps.

We apply the same biased sampling rule with step-specific ranges: $t \in [700, 1000]$ for the first step and $t \in [1, 1000]$ for the second. Thus, sampling favors timesteps near each step's target while retaining coverage of higher noise levels. Reference and generated samples at each target timestep are further corrupted to the sampled supervision timestep before being passed to the discriminator.

Teacher CFG direction. Our noise sampling is designed for effective adversarial distribution matching. For text-to-image generation, faithfully following the text prompt is equally important. We therefore supplement the adversarial update with the frozen teacher's classifier-free guidance direction. This reuses the existing teacher and requires no additional network.

We apply guidance to the student's clean prediction at each generation step. Below, $\widehat{\mathbf{x}}_0$ denotes this clean prediction, with the step index omitted. For SD3.5-M, we use normalized rectified-flow time τ, with $a_\tau = 1 - \tau$ and $\sigma_\tau = \tau$. For each sample, we draw j uniformly from $\{500, \ldots, 980\}$, set $\tau = j/1000$, and compute

$$\begin{aligned} \mathbf{y} &= (1 - \tau)\,\mathrm{sg}(\widehat{\mathbf{x}}_0) + \tau\boldsymbol{\epsilon}, \\ \mathbf{d} &= -\tau\left[\mathbf{v}(\mathbf{y}, \tau, c) - \mathbf{v}(\mathbf{y}, \tau, \varnothing)\right], \end{aligned} \tag{11}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, sg denotes stop-gradient, and $\mathbf{v}$ is the frozen teacher's velocity prediction. Both teacher predictions are computed without backpropagation through the teacher. Since the teacher's clean prediction is $\mathbf{y} - \tau\mathbf{v}$, $\mathbf{d}$ is exactly the difference between its conditional and unconditional clean predictions.

Let $\mathbf{g}$ denote the gradient of $\mathcal{L}_G$ with respect to $\widehat{\mathbf{x}}_0$. We modify it as

$$\mathbf{g}' = \mathbf{g} - w\|\mathbf{g}\|_2 \frac{\mathbf{d}}{\max(\|\mathbf{d}\|_2, \varepsilon)}, \tag{12}$$

where $\varepsilon > 0$ prevents division by zero. The weight w controls the guidance strength relative to the adversarial gradient at the clean prediction. We use $w = 1$, giving the added component the same norm as the adversarial gradient whenever $\|\mathbf{d}\|_2 \ge \varepsilon$.

Table 1: One-step generation on CIFAR-10 and ImageNet 64×64 with EDM teachers. Mimgs denotes the number of training images seen by the student. †Initialized from a pretrained SiD generator; the reported budget excludes SiD training. Bold and underline mark the best and second-best one-step FID and IS.

		CIFAR-10 (uncond.)			CIFAR-10 (cond.)		ImageNet 64×64 (cond.)			
Family	Method	Mimgs	FID↓	IS↑	Mimgs	FID↓	Mimgs	FID↓	Prec.↑	Rec.↑
Teacher	EDM	–	1.97	–	–	1.79	–	1.36	–	–
Score-based	SiD, α=1.0	∼420	2.028	10.017	∼420	1.932	∼500	2.022	0.73	0.63
	SiD, α=1.2	∼410	1.923	9.980	∼720	1.710	∼940	1.524	0.74	0.63
	SiDA	∼220	1.516	10.323	∼300	1.436	∼250	1.353	0.74	0.63
	SiD^{2}A†	∼30	1.499	10.323	∼23	1.396	∼22	1.110	0.75	0.62
	Uni-Instruct	∼300	<u>1.46</u>	–	∼300	1.42	∼300	1.28	–	–
	Uni-Instruct (long)	>300	1.48	–	>300	<u>1.38</u>	>300	**1.02**	–	–
External disc.	D2O	∼5	1.66	10.11	∼5	1.58	∼5	1.42	0.77	0.59
	D2O-F	∼5	1.54	10.10	∼5	1.44	∼5	1.16	0.75	0.60
Teacher disc.	MuSD (Ours)	∼20	1.483	<u>10.498</u>	∼20	1.397	∼20	1.153	0.75	0.62
	MuSD (Ours)	∼50	**1.437**	**10.508**	∼50	**1.341**	∼50	<u>1.035</u>	0.75	0.62

Guidance noise is sampled independently of the discriminator's noise sampling. The teacher's conditional–unconditional difference provides an explicit text-conditioning signal, complementing the discriminator's supervision on distribution matching.

5 Experiments

5.1 Experimental Setup

For our experiments, we consider both pixel-space and latent diffusion teachers across convolutional and Transformer-based architectures. For pixel-space U-Net models, we use EDM (Karras et al., 2022) teachers trained on CIFAR-10 (Krizhevsky, 2009) at 32×32 resolution and on ImageNet (Deng et al., 2009), FFHQ (Karras et al., 2019), and AFHQv2 (Choi et al., 2020) at 64×64 resolution. For latent U-Net models, we use EDM2-S (Karras et al., 2024) trained on ImageNet at 512×512 resolution. We further evaluate Transformer-based teachers using DiT-XL/2 (Peebles & Xie, 2023) on ImageNet at 256×256 resolution. Finally, for text-to-image generation, we use SD3.5-Medium (SD3.5-M) (Stability AI, 2024) as the latent Transformer teacher and train a two-step student, as described in Sec. 4.3. For this setting, we use 0.5M images from the Fine-T2I (Ma et al., 2026) subset as the reference dataset. And we use 30K prompts from COCO-2014 (Lin et al., 2014) and MJHQ (Li et al., 2024) and report FID, PickScore (Kirstain et al., 2023), HPSv3 (Ma et al., 2025), and ImageReward (Xu et al., 2023), together with GenEval (Ghosh et al., 2023).

5.2 One-Step Generation Across Backbones

Pixel-space teachers. Table 1 compares MuSD with one-step generators distilled from EDM teachers (Karras et al., 2022). The score-based baselines are SiD (Zhou et al., 2024), SiDA and SiD^{2}A (Zhou et al., 2025), and Uni-Instruct (Wang et al., 2026); D2O and D2O-F (Zheng & Yang, 2025) rely on an external pretrained network for adversarial supervision. MuSD achieves the lowest FID on both CIFAR-10 settings and the second lowest on ImageNet 64×64, while using several times fewer training images than SiDA and Uni-Instruct. SiD^{2}A reports a smaller budget but starts from a pretrained SiD generator whose training is not counted, and D2O and D2O-F use fewer images but reach higher FIDs. With about 20M images, MuSD already matches strong baselines, consistent with the fast convergence in Figure 1(b). On FFHQ and AFHQv2 at 64×64, MuSD attains FIDs of 0.836 and 1.192 with 49M and 19M images, respectively (Appendix A.4.1).

Latent teachers and Transformer backbones. Table 2 evaluates latent-space teachers. With DiT-XL/2 (Peebles & Xie, 2023) on ImageNet 256×256, the one-step MuSD student trained on 50M images surpasses its 250-step teacher with CFG. It also outperforms HD (Cheng et al., 2025), which distills the stronger SiT-XL/2 teacher (Ma et al., 2024), and the one-step π-Flow (Chen et al., 2026). Readout heads are attached to blocks 6, 12, and 18 (Appendix A.3.2), so the frozen-feature discrim-

Table 2: One-step generation with latent teachers. Left: DiT-XL/2 on ImageNet 256×256. Right: EDM2-S on ImageNet 512×512. Teachers use multi-step sampling; our students use one step without CFG. †Initialized from a pretrained SiD generator.

Method	NFE	FID↓
DiT-XL/2 (w/ CFG)	250×2	2.27
SiT-XL/2 (w/ CFG)	250×2	2.06
HD (SiT teacher)	1	2.26
π-Flow	1	2.85
MuSD (Ours, DiT teacher)	1	**2.117**

Method	NFE	Mimgs	FID↓
EDM2-S	63	–	2.56
EDM2-S (w/ CFG)	63×2	–	2.23
SiD	1	192	2.707
SiDA	1	118	1.756
SiD^{2}A†	1	75	1.669
MuSD (Ours)	1	95	**1.654**

Table 3: Text-to-image generation with SD3.5-Medium. MuSD samples without CFG. PickScore (PS) is scaled by 100. FID is computed after resizing images to 256×256. Latency is measured on one NVIDIA B200 GPU (FP16, batch size 8), including text encoding and VAE decoding (Appendix A.3.5). Bold marks the best distilled model.

		COCO-30K				MJHQ-30K					
Method	Steps	PS↑	HPSv3↑	IR↑	FID↓	PS↑	HPSv3↑	IR↑	FID↓	GenEval↑	Latency (s/image)↓
SD3.5-M (w/ CFG)	40	22.4	10.2	1.00	16.3	21.6	9.9	0.97	9.5	0.69	1.092
SD3.5-M-Turbo	4	22.2	9.6	0.83	**17.6**	21.3	9.3	0.74	13.6	0.59	0.081
SwD	4	22.7	**11.7**	1.12	23.7	21.8	**10.7**	1.13	13.7	0.70	0.056
SwD	6	22.8	**11.7**	1.12	23.1	21.8	**10.7**	1.10	13.4	0.70	0.059
MuSD (Ours)	2	**23.2**	11.3	**1.20**	25.2	**22.0**	**10.7**	**1.20**	**12.7**	**0.73**	**0.039**

inator is not tied to U-Net features. With EDM2-S (Karras et al., 2024) on ImageNet 512×512, MuSD surpasses the teacher with CFG and SiDA while using fewer training images, and slightly outperforms SiD^{2}A, which starts from a pretrained SiD generator.

5.3 Two-Step Text-to-Image Generation

We distill SD3.5-Medium (Stability AI, 2024) into a two-step student (Section 4.3) and compare it with the 40-step teacher, SD3.5-M-Turbo (TensorArt Studios, 2025), and SwD (Starodubcev et al., 2026) in Table 3. MuSD samples without CFG. On both benchmarks, it attains the highest PickScore and ImageReward of all compared models, along with the highest GenEval score, and it exceeds the teacher on all three preference metrics with about 28 times lower latency. It also has the lowest MJHQ-30K FID among distilled models, while SwD attains a higher COCO-30K HPSv3 at 1.4–1.5 times the latency. The higher COCO-30K FID of MuSD comes from the teacher CFG direction: without it, FID drops to 15.3, the lowest in Table 3, at the cost of lower preference scores, so the direction trades distributional fidelity for preference and alignment, as CFG does in the teacher. Additional metrics and per-category GenEval scores are provided in Appendix A.4.3.

5.4 Noise Sampling and Readout Ablations

We ablate the two components of MuSD on conditional CIFAR-10 with 20M training images.

Noise sampling. Table 4 (left) and Figure 4 compare samplers that differ in support and budget. EDM sampling draws noise levels from the teacher's training distribution, similar to LADD's (Sauer et al., 2024) logit-normal distribution except that LADD shifts it toward higher noise. Samplers with full support, EDM and uniform sampling, reach similar FIDs, whereas those with reduced support perform worse: truncated EDM has a budget and entropy between those of $\kappa = 4$ and $\kappa = 5$ but no density above $\sigma = 0.5$ and little near $\sigma_{\min}$. Keeping the support suffices even with little high-noise supervision, as $\kappa = 5$ places only 3% above $\sigma = 0.5$ and matches $\kappa = 4$. The maximum-entropy sampler guarantees this support for every finite κ, and within it a preference for low noise further improves FID. Since $\kappa \in \{3, 4, 5\}$ performs nearly identically, we treat the Beta sampler as a representative choice rather than an optimal one.

Table 4: Ablations on conditional CIFAR-10 (20M images). Left: noise sampling with all discriminator losses, showing the support of each sampling distribution and its expected corruption cost $\mathbb{E}[C]$ on $\sigma \in [10^{-4}, 1]$. EDM sampling uses the teacher's training distribution, whose support extends beyond $\sigma = 1$, where C is not defined. Right: where the discriminator loss is applied, under EDM and Beta ($\kappa = 4$) sampling. With the aggregated output only, all branches are still trained through the aggregation head.

Noise sampling	Support	$\mathbb{E}[C]$	FID↓
EDM	$(0, \infty)$	–	1.645
Uniform (κ=1)	$[10^{-4}, 1]$	1.00	1.683
Truncated EDM	$[0, 0.5]$	0.24	1.827
Fixed	$\{10^{-3}\}$	0.001	1.762
Beta, κ=4 (ours)	$[10^{-4}, 1]$	0.25	**1.397**

Loss on	EDM	Beta
Aggregated only	1.684	1.435
16×16 branch only	1.706	1.456
8×8 branch only	2.239	1.904
All (MRD)	1.645	**1.397**

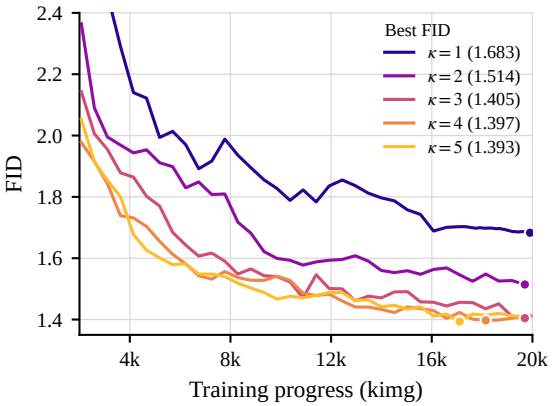


Figure 4: Effect of κ on conditional CIFAR-10.

Table 5: Training cost on ImageNet 512×512 with EDM2-S (per-GPU batch size 64). Auxiliary parameters are the fake-score network for SiDA and the discriminator heads for MuSD.

	SiDA	MuSD
Auxiliary params	280.2M	**6.1M**
Optimized params	560.4M	**286.3M**
GFLOPs per sample	∼2694	**1404**
Peak memory (GB)	52.4	**34.4**

Readout scales. Table 4 (right) varies where the discriminator loss is applied, under EDM and Beta sampling. Supervising only the 8×8 branch degrades FID substantially, so fine readouts, whose structure survives mainly at low noise (Section 4.2), are necessary; adding the 8×8 branch to the 16×16 branch gives a small further improvement, so coarse readouts complement finer ones rather than replace them. The Beta sampler improves FID in every configuration.

5.5 Training Efficiency

Table 5 compares the training cost of MuSD with SiDA (Zhou et al., 2025) for EDM2-S on ImageNet 512×512 at the same per-GPU batch size. SiDA trains a fake-score network of the same size as the teacher, whereas MuSD trains only lightweight heads on the frozen teacher features. As a result, MuSD roughly halves the number of optimized parameters and the training compute per sample, and reduces peak GPU memory by about a third.

6 Conclusion

We presented MuSD, a diffusion distillation framework that trains lightweight discriminator heads on frozen teacher features at multiple resolutions, without an auxiliary score network or an external feature extractor. Guided by an analysis of how noise shapes teacher features across scales, we derived a Beta noise sampler as the maximum-entropy density under a corruption budget, which favors low noise while keeping full support over the noise interval. Across U-Net and Transformer teachers, MuSD achieves competitive or better one-step FID with fewer training images and lower GPU memory, surpasses its 250-step DiT-XL/2 teacher in a single step, and extends to two-step text-to-image generation with SD3.5-Medium. Extending MuSD to larger text-to-image models and to other modalities such as video is a natural next step.

AI USE STATEMENT

The core research ideas and the proposed method were developed by the authors. Generative AI tools were not used to suggest the proposed method. In this work, we used generative AI tools to assist with manuscript editing, and literature discovery. We also used these tools to help check the mathematical correctness of our proofs and experimental results.

We take responsibility for the final content of this paper, including all claims, proofs, and reported results.

REPRODUCIBILITY STATEMENT

Section 4 specifies the discriminator readouts (Eqs. 5 and 6), the noise sampler (Eqs. 7–10), and the teacher CFG direction for text-to-image generation (Eqs. 11 and 12), and Appendix A.2 proves Proposition 1. Appendix A.3.1 lists the supervision interval and κ for each teacher, Appendix A.3.2 describes the discriminator heads and the teacher layers they read, and Appendix A.3.4 reports the full training configuration for every setting, including batch sizes, optimizers, learning rates, EMA, precision, and hardware. Appendix A.3.5 describes the evaluation protocol, including FID computation, text-to-image metrics, and latency measurement, and Appendix A.1 details the probing protocol. All teachers (EDM, EDM2-S, DiT-XL/2, and SD3.5-Medium) and datasets (CIFAR-10, ImageNet, FFHQ, AFHQv2, and Fine-T2I) are publicly available. We provide anonymized source code in the supplementary material.

REFERENCES

Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hk4_qw5xe.

Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SlxSY2UZQT.

Hansheng Chen, Kai Zhang, Hao Tan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. pi-flow: Policy-based few-step generation via imitation distillation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1vAte8dsap.

Hanbo Cheng, Peng Wang, Kaixiang Lei, Qi Li, Zhen Zou, Pengfei Hu, and Jun Du. From structure to detail: Hierarchical distillation for efficient diffusion model, 2025. URL https://arxiv.org/abs/2511.08930.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11472–11481, June 2022.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M. Susskind. BOOT: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7407–7417. IEEE, 2023.

Victor Shea-Jay Huang, Le Zhuo, Yi Xin, Zhaokai Wang, Fu-Yun Wang, Yuchi Wang, Renrui Zhang, Peng Gao, and Hongsheng Li. Tide: Temporal-aware sparse autoencoders for interpretable diffusion transformers in image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 435–443, 2026. URL https://arxiv.org/abs/2503.07050.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184. IEEE, 2024.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10641–10652. IEEE, 2022.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217. IEEE, 2023.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=2uAaGwlP_V.

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. URL https://openreview.net/forum?id=MLIs5iRq4w.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Xu Ma, Yitian Zhang, Qihua Dong, and Yun Fu. Fine-t2i: An open, large-scale, and diverse dataset for high-quality t2i fine-tuning, 2026. URL https://arxiv.org/abs/2602.09439.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025.

Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features. *Advances in Neural Information Processing Systems*, 37:55141–55177, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 87–103, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73016-0.

Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised MAP inference for image super-resolution. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=S1RP6GLle.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Stability AI. Stable diffusion 3.5. https://github.com/Stability-AI/sd3.5, 2024.

Nikita Starodubcev, Ilya Drobyshevskiy, Denis Kuznedelev, Artem Babenko, and Dmitry Baranchuk. Scale-wise distillation of diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Z06LNjqU1g.

TensorArt Studios. Stable diffusion 3.5 medium turbo. https://huggingface.co/tensorart/stable-diffusion-3.5-medium-turbo, 2025.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8798–8807. Ieee, 2018.

Yifei Wang, Weimin Bai, Debing Zhang, Weijian Luo, He Sun, et al. Uni-instruct: One-step diffusion model through unified diffusion divergence instruction. *Advances in Neural Information Processing Systems*, 38:57097–57132, 2026.

Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training GANs with diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=HZf7UbpWHuA.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Bowen Zheng and Tianming Yang. Revisiting diffusion models: From generative pre-training to one-step generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dUq579lefp.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=QhqQJqe0Wq.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lS2SGfWizd.

A Appendix

A.1 Probing Coarse Image Structure

We examine whether coarse readouts at low noise show the same focus on coarse structure as full-resolution features at high noise. We measure this focus by the relative predictability of spatial bands (Eq. 18) rather than by absolute scores, because corruption lowers all scores and the question is which bands remain predictable relative to the others. We fit linear ridge probes to frozen SD3.5-M features and predict spatial targets derived from clean VAE reconstructions. The probes operate on readout features before the trainable discriminator heads.

Reference data. We uniformly sample 1,536 examples without replacement from a 0.5M-image subset of Fine-T2I (Ma et al., 2026). Images are center-cropped to 1024×1024 and encoded with the SD3 VAE. One latent is sampled per image and reused across all conditions, together with its associated caption.

Teacher features. For each latent $\mathbf{x}_0 \in \mathbb{R}^{16\times128\times128}$, we form

$$\mathbf{x}_t = \left(1 - \frac{t}{1000}\right)\mathbf{x}_0 + \frac{t}{1000}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{13}$$

using the same noise realization per image across conditions. We evaluate $t \in \{50, 100, 200, 300, 500, 700, 800, 900, 950, 980\}$ without a timestep shift. The teacher receives the associated caption and timestep t. We extract image tokens from block 12 and reshape them into 64×64 grids with 1,536 channels. A fixed projection $\mathbf{P} \in \mathbb{R}^{1536\times64}$, with independent $P_{ij} \sim \mathcal{N}(0, 1/1536)$, reduces features to 64 channels and is shared across all conditions. Teacher inference uses bfloat16; features are converted to float32 before projection.

Spatial readouts. Following Section 4.1, $s = 1, 2, 3, 4$ indexes 8×8, 16×16, 32×32, and 64×64 readouts, respectively, with $S = 4$. For this probe, $\mathbf{f}_s$ and $\widetilde{\mathbf{f}}_s$ denote their 64-channel projected versions. We construct the pyramid using fixed BlurPool operators and apply Eq. 6 with $\alpha = 0.5$. The coarsest readout retains the low-pass residual, $\widetilde{\mathbf{f}}_1 = \mathbf{f}_1$. The full-resolution baseline is the unmodified $\mathbf{f}_S$, not the finest partial band-pass readout $\widetilde{\mathbf{f}}_S$. For probing only, all inputs are restored to 64×64 using the fixed feature upsampling operators and flattened into 262,144-dimensional vectors.

Reconstruction targets. We decode each clean latent into an RGB image $\mathbf{x}_0^{\mathrm{RGB}}$, clamped to $[-1, 1]$. Using r for target resolution, we define

$$\mathbf{T}_r = \mathrm{AdaptiveAvgPool}_{r\times r}(\mathbf{x}_0^{\mathrm{RGB}}), \qquad r \in \{8, 16, 32, 64, 128\}, \tag{14}$$

$$\mathbf{L}_r = \begin{cases} \mathbf{T}_8, & r = 8, \\ \mathbf{T}_r - \mathcal{B}(\mathbf{T}_{r/2}), & r \in \{16, 32, 64, 128\}, \end{cases} \tag{15}$$

where $\mathcal{B}$ is bilinear upsampling to $r \times r$ with `align_corners=False`. Thus, $\mathbf{L}_8$ is the coarse RGB thumbnail itself, while $\mathbf{L}_{16}$ and $\mathbf{L}_{32}$ contain changes relative to their upsampled next-coarser thumbnails. Each target has $3r^2$ entries and is fixed across conditions. Target resolution r is distinct from readout index s; the targets are spatial residuals, not exact Fourier bands.

Probe fitting and evaluation. We use a shared split of 921 training, 307 validation, and 308 test examples. For each condition, features are divided by their mean row norm over all 1,536 examples, without per-dimension standardization. Features and targets are then centered using training means only. With centered features $\mathbf{X}$ and training targets $\mathbf{Y}_{\mathrm{tr}}$, linear ridge regression solves

$$(\mathbf{K}_{\mathrm{tr,tr}} + n_{\mathrm{tr}}\lambda\mathbf{I})\,\mathbf{A} = \mathbf{Y}_{\mathrm{tr}}, \qquad \mathbf{K} = \mathbf{X}\mathbf{X}^{\mathsf{T}}, \quad n_{\mathrm{tr}} = 921. \tag{16}$$

Predictions use the centered cross-kernel and add back the training target mean. For each condition and target, validation R^2 selects $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Gram matrices are stored in float32 and promoted to float64 for regression; no additional prediction gain is fitted. Among the eight conditions in Table 6, validation selects the smallest tested $\lambda = 10^{-5}$ for all L_8

Table 6: **Reconstruction accuracy and relative band-wise predictability.** Each entry reports test R^2 followed by its normalized share $100P_r$ in parentheses (%), computed from the same unrounded scores. L_8 is the coarse RGB thumbnail; L_{16} and L_{32} are residuals relative to upsampled next-coarser thumbnails. Shares are normalized over these three targets after clipping negative scores to zero (Eq. 18); they describe relative predictability, not spectral energy. MRD uses $\alpha = 0.5$, with a low-pass residual at $s = 1$. Bold marks the two profiles compared in Observation 1. Percentages may not sum to 100 due to rounding.

			R^2 (normalized share, %)		
Feature	Native grid	t	L_8	L_{16}	L_{32}
Full-resolution baseline					
$\mathbf{f}_S$	64×64	50	0.858 (80.7)	0.180 (17.0)	0.0253 (2.4)
$\mathbf{f}_S$	64×64	300	0.838 (84.2)	0.149 (15.0)	0.0077 (0.8)
$\mathbf{f}_S$	64×64	900	0.745 (92.4)	0.061 (7.6)	−0.0034 (0.0)
$\mathbf{f}_S$	64×64	980	0.559 (**100.0**)	−0.001 (0.0)	−0.0040 (0.0)
MRD readouts at low noise					
$\tilde{\mathbf{f}}_1$	8×8	50	0.899 (**99.1**)	0.008 (0.9)	−0.0040 (0.0)
$\tilde{\mathbf{f}}_2$	16×16	50	0.929 (81.8)	0.206 (18.2)	−0.0035 (0.0)
$\tilde{\mathbf{f}}_3$	32×32	50	0.904 (72.9)	0.308 (24.8)	0.0288 (2.3)
$\tilde{\mathbf{f}}_4$	64×64	50	0.774 (84.9)	0.108 (11.9)	0.0291 (3.2)

probes and six of the eight L_{16} probes. Smaller values were not evaluated; comparisons therefore refer to the specified regularization grid. We report test scores

$$R^2 = 1 - \frac{\sum_{i \in \mathrm{te}} \|\mathbf{y}_i - \widehat{\mathbf{y}}_i\|_2^2}{\sum_{i \in \mathrm{te}} \|\mathbf{y}_i - \overline{\mathbf{y}}_{\mathrm{te}}\|_2^2}, \tag{17}$$

aggregating all target dimensions and test examples. The denominator uses the test target mean. Results use one fixed split, projection, and noise realization per image.

Normalized band-wise predictability. We summarize relative reconstruction scores over $\mathcal{R} = \{8, 16, 32\}$ as

$$P_r = \frac{[R_r^2]_+}{\sum_{j \in \mathcal{R}} [R_j^2]_+}, \qquad r \in \mathcal{R}, \tag{18}$$

where R_r^2 is the test score for $\mathbf{L}_r$ and $[u]_+ = \max(0, u)$. This profile measures relative linear predictability, not spectral energy. We use it as the primary measure of focus and report absolute R^2 for completeness.

Results and interpretation. The normalized profile of full-resolution features shifts toward L_8 as noise increases, from 80.7% at t=50 to 100.0% at t=980. The shift occurs because finer bands vanish first: L_{32} falls from 0.025 to below zero by t=900, and L_{16} from 0.180 to -0.001 at t=980, while L_8 also decreases, from 0.858 to 0.559. The 8×8 readout feature $\tilde{\mathbf{f}}_1$ at t=50 matches this high-noise profile (99.1%) without heavy corruption, which is the basis of Observation 1, and the 32×32 readout feature $\tilde{\mathbf{f}}_3$ recovers finer structure (R^2=0.308 on L_{16}). Each $\tilde{\mathbf{f}}_s$ is a fixed linear function of $\mathbf{f}_S$, so differences in absolute scores between $\tilde{\mathbf{f}}_s$ and $\mathbf{f}_S$ at equal t reflect finite-sample probe behavior rather than additional information. The evidence concerns linear recovery; L_{64} and L_{128} scores remain near zero.

A.2 Proof of Proposition 1

Normalized coordinate. Let

$$u = \frac{t - t_{\min}}{t_{\max} - t_{\min}} \in (0, 1),$$

and let ρ denote its density. The cost becomes $C(u) = -\log(1 - u)$, while

$$H(\pi) = H(\rho) + \log(t_{\max} - t_{\min}).$$

The expected cost is unchanged by this transformation. Since the entropy differs only by a fixed constant, it suffices to solve the problem on $(0, 1)$.

For notational simplicity, we also use π to denote the density in the normalized coordinate u.

Candidate density. Set $\kappa = 1/b > 1$ and consider

$$\pi^\star(u) = \kappa(1-u)^{\kappa-1}. \tag{19}$$

This density integrates to one and satisfies

$$\mathbb{E}_{\pi^\star}[C] = \kappa \int_0^1 w^{\kappa-1}(-\log w)\, dw = \frac{1}{\kappa} = b, \tag{20}$$

where $w = 1 - u$. Thus, the candidate meets the budget exactly. Moreover,

$$\log \pi^\star(u) = \log \kappa - (\kappa - 1)C(u), \tag{21}$$

so

$$H(\pi^\star) = -\log \kappa + (\kappa - 1)b. \tag{22}$$

Optimality and uniqueness. Take any feasible density π with $\mathbb{E}_\pi[C] \le b$. If $H(\pi) = -\infty$, then $H(\pi) < H(\pi^\star)$ trivially, since $H(\pi^\star)$ is finite. We therefore assume $H(\pi) > -\infty$. Expanding the KL divergence gives

$$\begin{aligned} H(\pi) &= H(\pi^\star) - D_{\mathrm{KL}}(\pi \| \pi^\star) \\ &\quad - (\kappa - 1)\,(b - \mathbb{E}_\pi[C]) \\ &\le H(\pi^\star). \end{aligned} \tag{23}$$

Both subtracted terms are nonnegative. Equality holds only when $\pi = \pi^\star$ almost everywhere. Thus, the maximizing probability distribution is unique. Transforming back to t yields Eq. 9.

Sampling. The candidate CDF is

$$F(u) = 1 - (1-u)^\kappa.$$

For $v \sim \mathcal{U}(0, 1)$, solving $F(u) = 1 - v$ gives $u = 1 - v^{1/\kappa}$. Transforming back to t yields Eq. 10. □

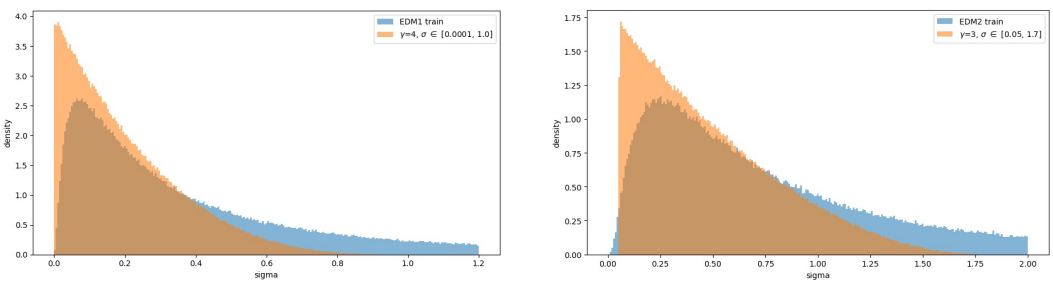
A.3 Implementation Details

A.3.1 Supervision Intervals

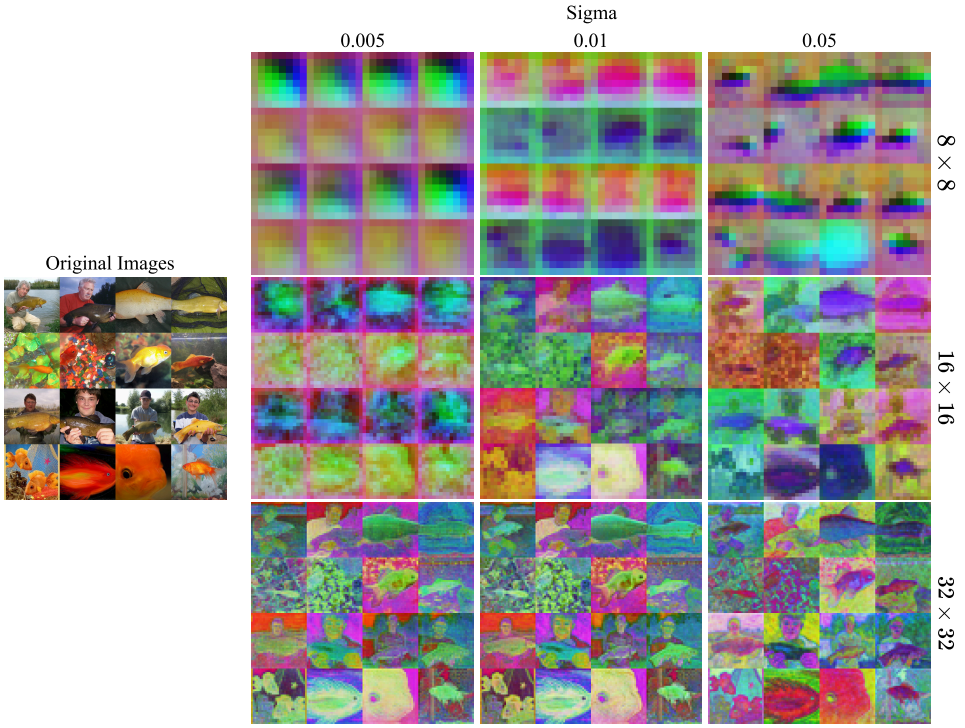
Section 4.2 defines the sampler over a generic noise index t. For EDM-family teachers, we apply it directly to the noise level, that is, $\sigma = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})\big(1 - v^{1/\kappa}\big)$ with $v \sim \mathcal{U}(0, 1)$; for DiT-XL/2 and SD3.5-M, we apply it to the diffusion timestep t. Table 7 lists the interval and κ for each teacher. For one-step students, the lower end of the interval is the lowest noise level at which the teacher's features remain reliable. For SD3.5-M, the two steps of the student use separate intervals that start at their target timesteps (Section 4.3).

Table 7: Supervision intervals and κ ($b = 1/\kappa$). EDM-family teachers use the noise level σ; the other teachers use the diffusion timestep t.

Teacher	Datasets	Interval	κ
EDM	CIFAR-10, ImageNet 64, FFHQ, AFHQv2	$\sigma \in [10^{-4}, 1]$	4
EDM2-S	ImageNet 512	$\sigma \in [0.05, 1.7]$	3
DiT-XL/2	ImageNet 256	$t \in [1, 1000]$	4
SD3.5-M (step 1)	Fine-T2I	$t \in [700, 1000]$	3
SD3.5-M (step 2)	Fine-T2I	$t \in [1, 1000]$	3



(a) EDM (b) EDM2



(c) EDM2-S features at very low noise

Figure 5: Supervision intervals for EDM-family teachers. (a) EDM training noise distribution and our Beta sampler on $\sigma \in [10^{-4}, 1]$. (b) EDM2 shifts its training noise distribution to higher noise; our Beta sampler uses $\sigma \in [0.05, 1.7]$. (c) At $\sigma = 0.005$ and 0.01, the frozen EDM2-S teacher produces grid-like feature artifacts, so we set $\sigma_{\min} = 0.05$.

EDM2 interval. Unlike EDM (Figure 5a), EDM2 shifts its log-normal training noise distribution to a higher mean (Figure 5b), so the teacher sees little training data at very low noise. Consistently, the intermediate features of the frozen EDM2-S teacher at $\sigma = 0.005$ and 0.01 show grid-like artifacts rather than image structure (Figure 5c). We therefore raise $\sigma_{\min}$ to 0.05 and extend $\sigma_{\max}$ to 1.7, in line with the higher noise levels emphasized during EDM2 training. We did not observe such artifacts with the Transformer teachers, DiT-XL/2 and SD3.5-M.

A.3.2 Readout Heads and Teacher Layers

Teacher features. For U-Net teachers, the finest head reads the decoder feature map one level below the output resolution, that is, at half the input resolution (16×16 for 32×32 inputs and 32×32 for 64×64 inputs), and coarser heads read decoder features at lower native resolutions, for example 8×8 on CIFAR-10. For SD3.5-M, heads are attached to blocks 10 and 12, and each block provides readout features at 64×64, 32×32, 16×16, and 8×8, constructed with fixed BlurPool low-pass

filtering and Eq. 6 with $\alpha = 0.5$ (Section 4.1). For DiT-XL/2, heads are attached to blocks 6, 12, and 18.

Heads. Each head is a two-layer stack of convolution, group normalization, and SiLU that outputs a patch-level logit map. An aggregation head combines the flattened logit maps of all branches. The hinge losses in Eqs. 3 and 4 are applied to every branch and to the aggregated output; the loss of each branch is averaged over its logits, and the losses of all branches are summed. All heads together contain 6.1M parameters with EDM2-S, about 2% of the 280M-parameter teacher, and 35.8M parameters with SD3.5-M, about 1.5% of the 2.47B-parameter teacher. The heads do not receive the noise level as an input; it enters only through the teacher features. For two-step text-to-image generation, the heads additionally receive an embedding of the generation step (first or second), which is added to each channel as a bias, similar to class-conditional normalization.

A.3.3 Readout Depths in Transformer Teachers

Figure 6 compares token-grid features across depth at $t = 1$ for the two Transformer teachers. In DiT-XL/2 (b), block 6 varies at the scale of individual tokens and reflects local texture, whereas block 18 varies smoothly over object-level regions, such as the silhouettes of the fox and the otter, with block 12 in between. The features thus become coarser with depth: the deepest block plays the role of the coarse readouts in the SD3.5-M pyramid, and the shallowest that of the fine readouts, so we read these blocks directly without constructing a pyramid. In SD3.5-M (a), by contrast, every block from 6 to 22 retains fine spatial detail, such as object boundaries and textures, at the full 64×64 token resolution, so depth alone does not provide coarse readouts; this may be because its joint attention mixes image and text tokens. We therefore construct the readout pyramid for SD3.5-M explicitly (Section 4.1).

A.3.4 Training Details

Tables 8 and 9 list the training configurations. For EDM-family teachers and DiT-XL/2, the one-step generator is initialized from the teacher weights, without an intermediate distillation stage. Data augmentation, dropout, and other network settings follow the EDM and EDM2 baselines. Training is stable with moderate batch sizes: we use a global batch size of 1024 for ImageNet 64×64 and 512×512, whereas SiD^{2}A (Zhou et al., 2025) uses 8192 and 2048, respectively.

Table 8: Training configurations for EDM-family teachers. The global batch size is the per-GPU batch size times the number of GPUs and gradient-accumulation steps.

	CIFAR-10	FFHQ 64	AFHQv2 64	ImageNet 64	ImageNet 512
Teacher	EDM	EDM	EDM	EDM	EDM2-S
Training images (M)	20 / 50	49	19	20 / 50	95
Global batch size	384	512	512	1024	1024
Batch size per GPU	16	64	64	64	64
GPUs	6×RTX 2080Ti	2×A6000	2×A6000	4×A6000	4×A6000
Gradient accumulation	4	4	4	4	4
Generator learning rate	5e-5	2e-5	2e-5	8e-6	5e-5
Discriminator learning rate	8e-5	4e-5	4e-5	4e-5	5e-5
Adam (β_1, β_2)	(0, 0.99)	(0, 0.99)	(0, 0.99)	(0, 0.99)	(0, 0.99)
Adam ϵ	1e-8	1e-6	1e-6	1e-6	1e-6
EMA half-life (M images)	0.5	0.5	0.5	2	2
EMA decay	0.999	0.999	0.999	0.999	0.999
FP16	No	Yes	Yes	Yes	Yes

Generator input noise level. The one-step generator reuses the teacher network, which expects a noisy input and its noise level. Following SiD (Zhou et al., 2024), the generator for EDM-family teachers maps $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to $G_\theta(\mathbf{z}) = f_\theta(\sigma_{\text{init}}\mathbf{z}, \sigma_{\text{init}})$, where f_θ is the denoiser initialized from the teacher and σ_{init} is fixed. We use $\sigma_{\text{init}} = 2.5$ or 5 for EDM and EDM2-S. For DiT-XL/2, the generator uses the fixed timestep $t_{\text{init}} = 500$, which corresponds to $\sigma \approx 3.44$ in the EDM parameterization.

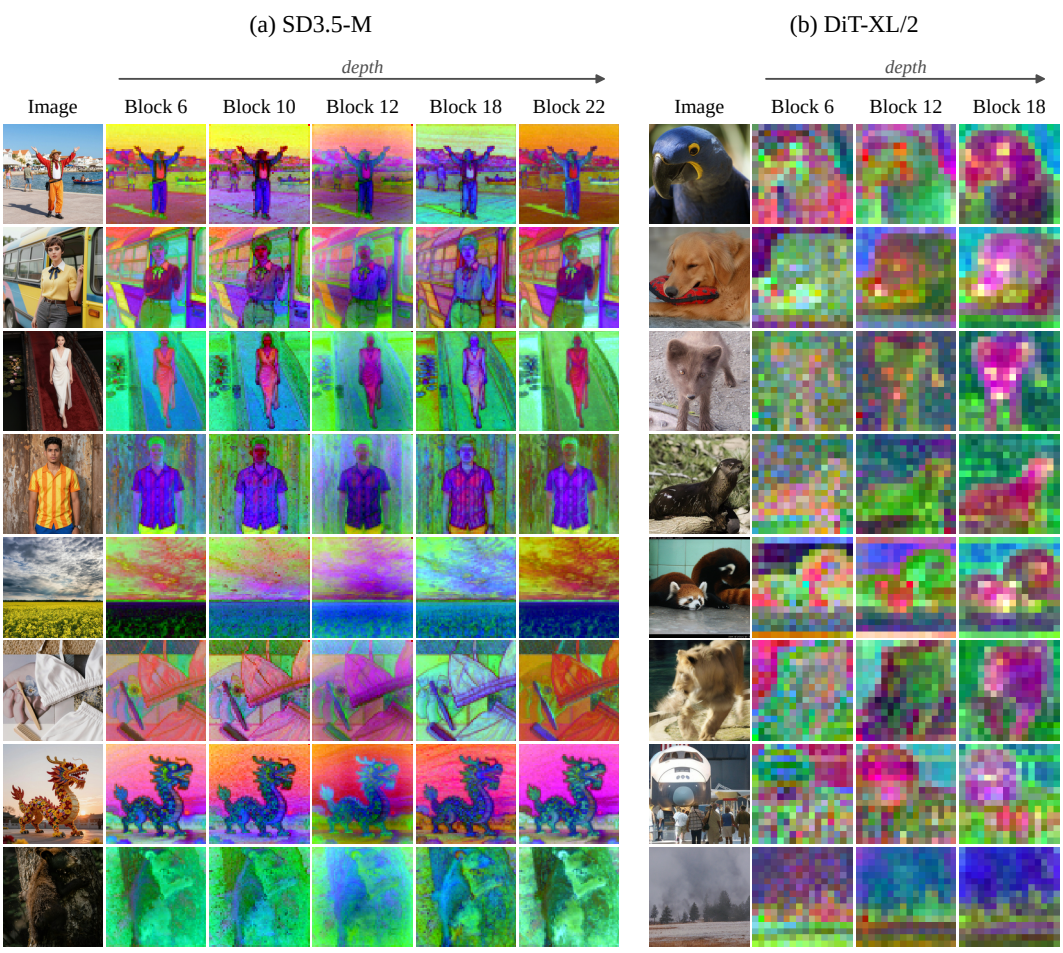


Figure 6: Token-grid features across depth at $t = 1$; each map shows its first three principal components. (a) SD3.5-M, blocks 6 to 22 (64×64 tokens): every block retains fine spatial detail. (b) DiT-XL/2 on ImageNet, blocks 6, 12, and 18 (16×16 tokens): deeper blocks vary smoothly over object-level regions, whereas block 6 reflects local texture.

For SD3.5-M, the two generation steps start at $t = 1000$ and $t = 700$ (Section 4.3). These generator inputs are separate from the supervision intervals in Table 7, which apply only to the discriminator.

Text-to-image generation. The two-step student is trained on a 0.5M-image subset of Fine-T2I (Ma et al., 2026) for one epoch. We train only LoRA adapters on the student, with rank 64, $\alpha = 64$, and no dropout, applied to the query, key, value, and output projections of the attention layers and to the two linear layers of the MLP in each block. Each step predicts the clean sample $\hat{\mathbf{x}}_0$: the first step predicts it from Gaussian noise at $t = 1000$ and forms the input at $t = 700$ from this prediction, and the second step predicts the final sample from $t = 700$. The teacher CFG direction uses $w = 1$ and $\tau = j/1000$ with j drawn uniformly from $\{500, \ldots, 980\}$ (Section 4.3).

A.3.5 Evaluation Protocol

Class-conditional and unconditional generation. FID is computed with 50K generated samples. For each run, we report the best FID among the evaluated checkpoints. Figures 1(b) and 4 show the best-so-far FID during training.

Text-to-image generation. We generate images for 30K prompts from COCO-2014 (Lin et al., 2014) and from MJHQ-30K (Li et al., 2024). FID is computed after resizing images to 256×256, following SwD (Starodubcev et al., 2026). PickScore is scaled by 100. GenEval follows the official

Table 9: Training configurations for Transformer teachers.

	DiT-XL/2 (ImageNet 256)	SD3.5-M (text-to-image)
Latent space	SD-VAE, 32×32×4	SD3.5 VAE, 16 channels
Trainable generator parameters	All	LoRA (rank 64, $\alpha = 64$)
Training images (M)	50	0.5 (one epoch)
Iterations	≈24.4K	1,953
Global batch size	2048	256
Batch size per GPU × GPUs × accumulation	$256 \times 4 \times 2$	$32 \times 2 \times 4$
Generator optimizer	AdamW, lr 1e-5	AdamW, lr 1e-4, ϵ 1e-8
Discriminator optimizer	Adam, lr 4e-5	Adam, lr 4e-5, ϵ 1e-8
Adam (β_1, β_2) / weight decay	(0, 0.99) / 0	(0, 0.99) / 0
Gradient clipping	1.0	1.0
EMA half-life (M images)	0.5	0.5
Precision	FP16 (generator)	FP16 (generator)
Hardware	4×B200	2×B200

protocol (Ghosh et al., 2023). Quality metrics for SD3.5-M, SD3.5-M-Turbo, and SwD are taken from SwD, whose evaluation protocol we follow.

Latency. Latency follows the protocol of SwD: FP16 precision, `torch.compile`, a TinyVAE decoder, batch size 8, and per-image time including text encoding, denoising, and VAE decoding. We measure all models, including the baselines, in the same setup on a single NVIDIA B200 GPU, so absolute values are lower than those reported by SwD on different hardware.

A.4 Additional Results

A.4.1 FFHQ and AFHQv2

Table 10 compares MuSD with BOOT (Gu et al., 2023), SiD (Zhou et al., 2024), SiDA and SiD^{2}A (Zhou et al., 2025), and D2O (Zheng & Yang, 2025) on FFHQ and AFHQv2 at 64×64. MuSD attains the lowest FID on both datasets while using fewer training images than the SiD family.

Table 10: One-step generation on FFHQ and AFHQv2 at 64×64 with EDM teachers. †Initialized from a pretrained SiD generator; Mimgs exclude SiD training.

	FFHQ			AFHQv2		
Method	NFE	Mimgs	FID↓	NFE	Mimgs	FID↓
EDM (teacher)	79	–	2.39	79	–	1.96
BOOT	1	–	9.00	–	–	–
SiD, α=1.0	1	498	1.710	1	305	1.628
SiD, α=1.2	1	453	1.550	1	372	1.711
SiDA, α=1.0	1	186	1.134	1	127	1.345
SiD^{2}A†, α=1.0	1	169	1.040	1	175	1.276
SiD^{2}A†, α=1.2	1	81	1.109	1	175	1.366
D2O	1	5	1.08	1	5	1.23
D2O-F	1	10	0.85	1	10	1.31
MuSD (Ours)	1	49	**0.836**	1	19	**1.192**

A.4.2 Readout Head Capacity

Table 11 varies the depth of each discriminator head on conditional CIFAR-10 with 20M training images. Both shallower and deeper heads perform worse than the default two-layer heads, supporting lightweight heads that read out the frozen teacher features rather than learn a separate discriminator.

Table 11: Discriminator head depth on conditional CIFAR-10 (20M images).

Head depth	FID↓
1 layer	1.419
2 layers (default)	**1.397**
3 layers	1.475

A.4.3 TEXT-TO-IMAGE DETAILS

Table 12 reports per-category GenEval scores of MuSD. MuSD performs well on single-object, two-object, and color prompts, while spatial position remains the most difficult category. Its CLIP-H scores are 0.359 on COCO-30K and 0.372 on MJHQ-30K.

Table 12: Per-category GenEval scores of MuSD (two steps, SD3.5-M teacher).

Single obj.	Two obj.	Counting	Colors	Position	Color attr.	Overall
1.00	0.97	0.70	0.80	0.34	0.60	0.73

Teacher CFG direction. Table 13 removes the teacher CFG direction ($w = 0$), leaving only the adversarial objective. The resulting two-step student attains a COCO-30K FID of 15.3, lower than every model in Table 3, including the 40-step teacher with CFG. The frozen-teacher adversarial objective alone thus matches the image distribution closely, even though the student is trained on Fine-T2I rather than COCO images. This comes with a trade-off: its PickScore, HPSv3, and ImageReward fall below those of the teacher, and adding the CFG direction raises these preference and alignment scores at the cost of distributional fidelity, similar to CFG in the teacher.

Table 13: Effect of the teacher CFG direction on COCO-30K with two steps. $w = 0$ removes the direction. PickScore (PS) is scaled by 100.

Method	PS↑	HPSv3↑	IR↑	FID↓
SD3.5-M (40 steps, w/ CFG)	22.4	10.2	1.00	16.3
MuSD, $w = 0$	22.2	8.3	0.57	**15.3**
MuSD, $w = 1$ (default)	**23.2**	**11.3**	**1.20**	25.2

Qualitative comparison. Figures 7–9 compare samples from MuSD, SwD, and the SD3.5-M teacher for the same prompts.

A.4.4 TWO-STEP SAMPLING

Figures 10 and 11 compare, for the same prompts and initial noise, the clean prediction after the first step of the two-step student with the final sample after the second step. The first step, supervised only at high noise ($700 < t < 1000$), already determines the overall layout in most cases. The second step, which starts from $t = 700$ and is supervised over the full interval, sharpens textures, removes grain-like artifacts, and restores fine structure such as graffiti lettering, the winding mountain road, and animal fur.

A.5 QUALITATIVE SAMPLES

Figures 12–17 show one-step samples (NFE $= 1$) from MuSD for all class-conditional and unconditional settings in Section 5.2. Text-to-image samples are shown in Appendices A.4.3 and A.4.4.



Figure 7: Qualitative text-to-image comparison with the SD3.5-M teacher. Rows show MuSD (two steps), SwD (four and six steps), and the teacher (40 steps with CFG), with per-image latency. The prompt is shown below each column; long prompts are truncated.

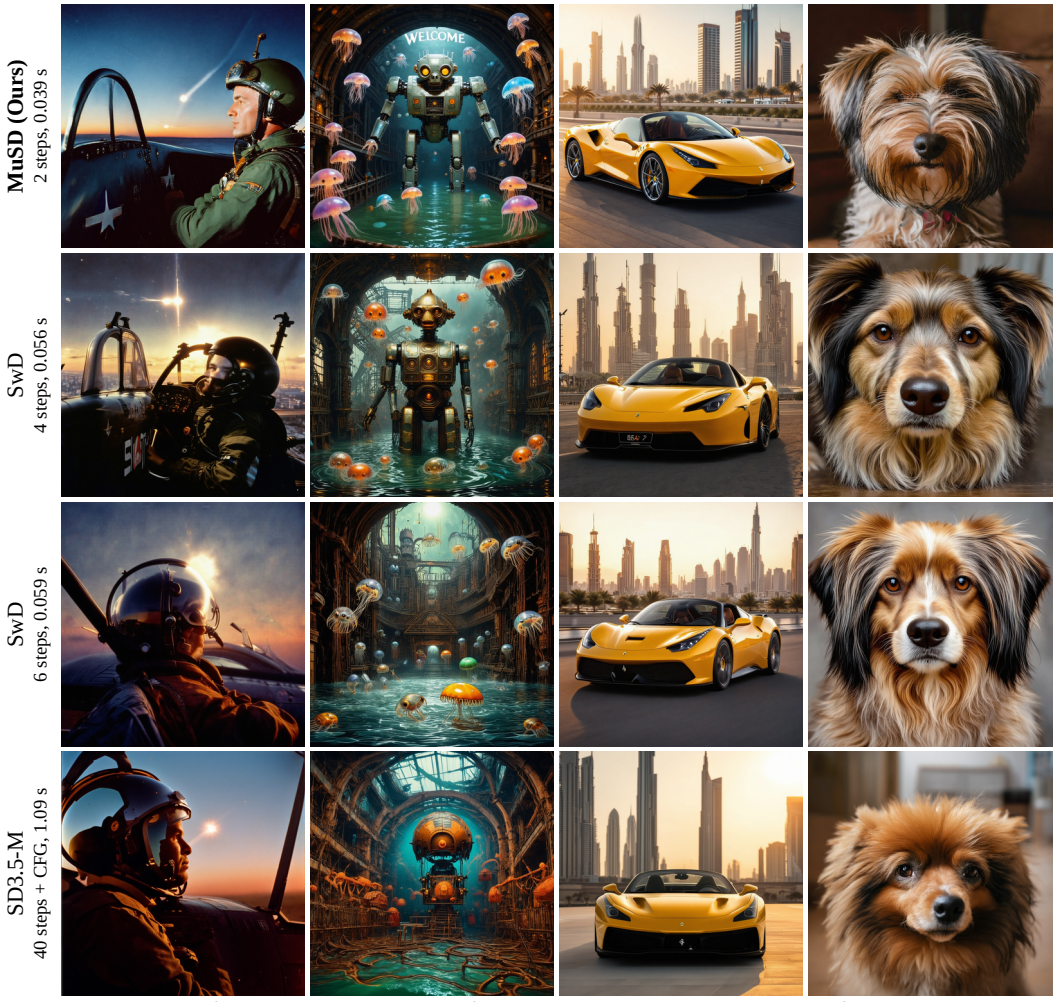


Figure 8: Additional text-to-image comparisons, arranged as in Figure 7.

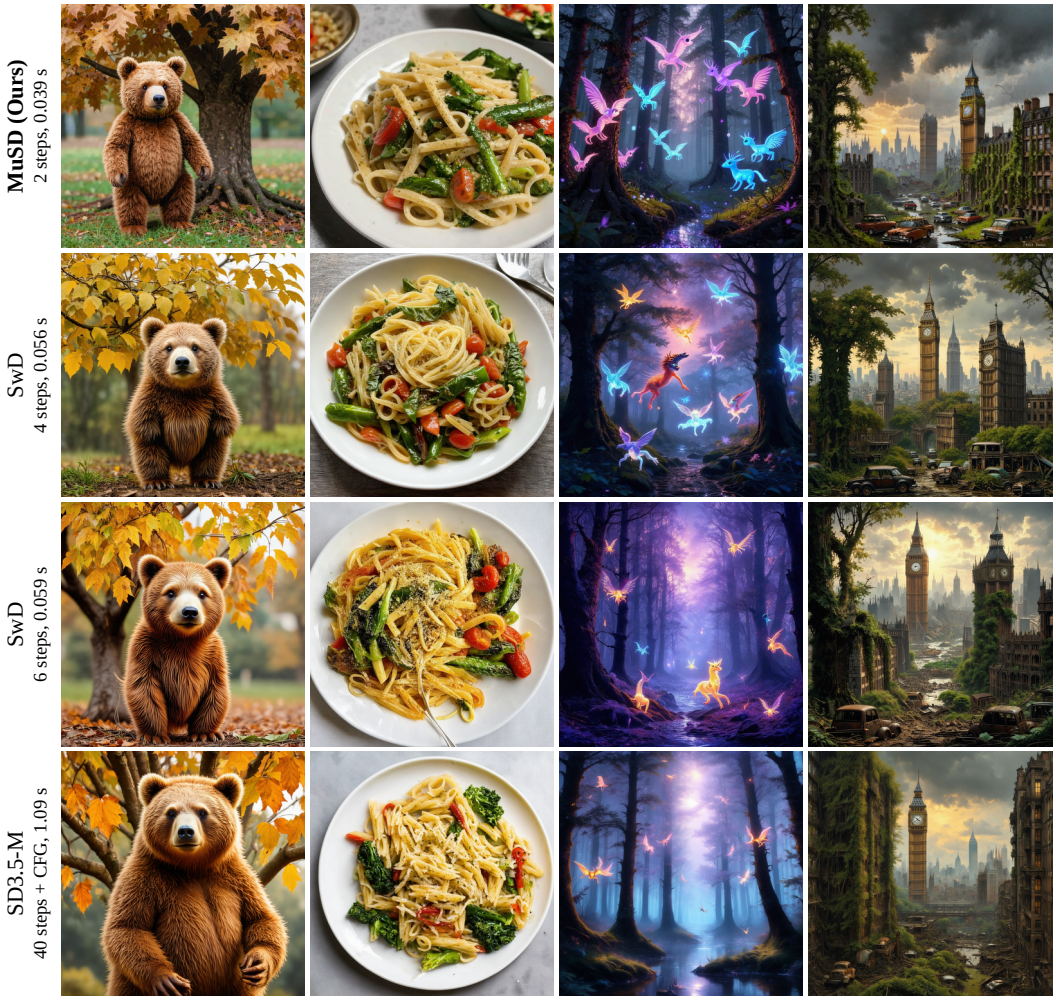


Figure 9: Additional text-to-image comparisons, arranged as in Figure 7.



Figure 10: Two-step sampling with MuSD (SD3.5-M teacher). Top row of each block: clean prediction after the first step ($t = 1000$). Bottom row: final sample after the second step ($t = 1000 \rightarrow 700$). Both rows use the same initial noise, and the prompt is shown below each column.

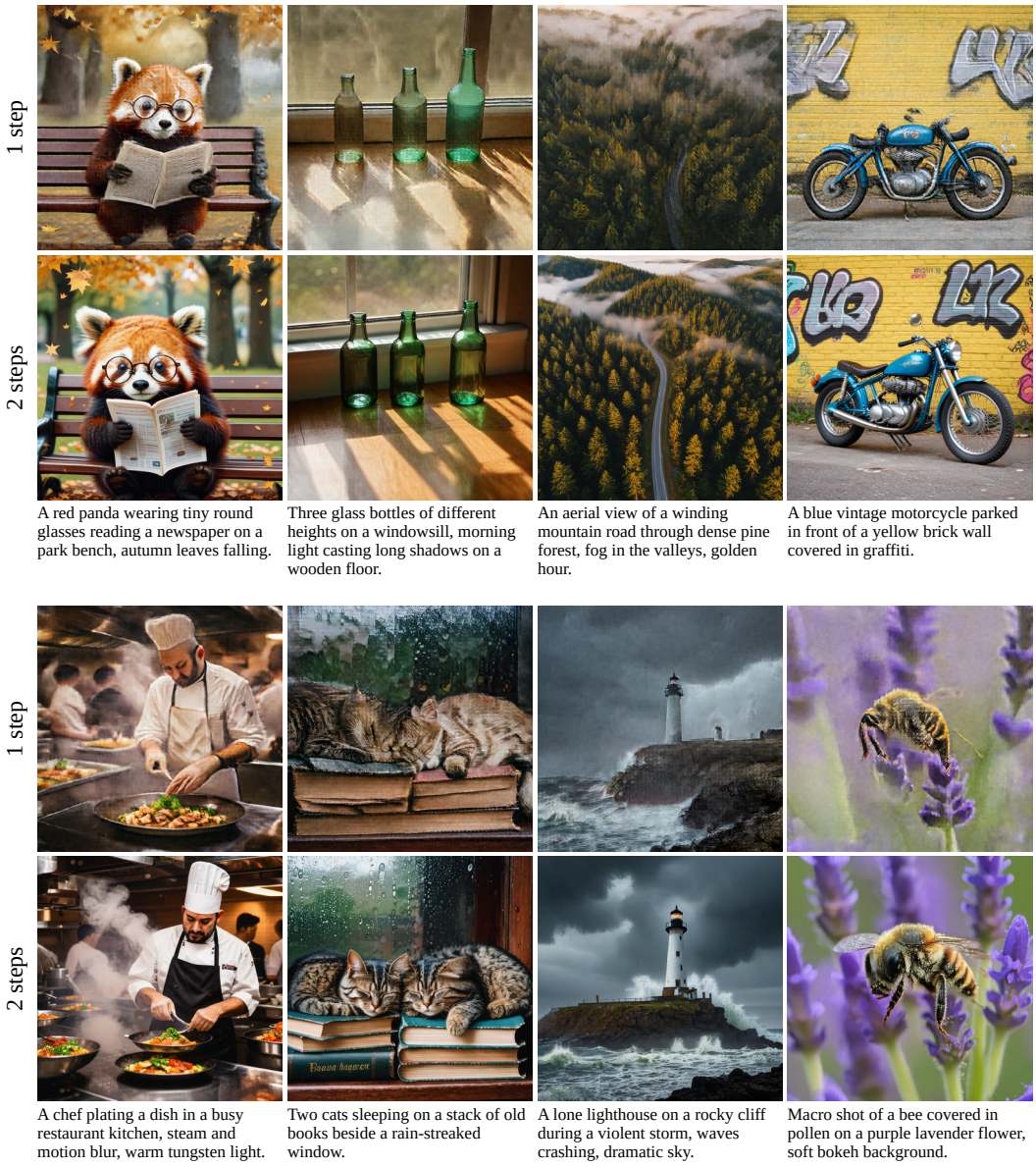


Figure 11: Additional two-step samples with MuSD, arranged as in Figure 10.



Figure 12: CIFAR-10 (unconditional). Samples from MuSD (NFE $= 1$, FID $= 1.437$).



Figure 13: CIFAR-10 (conditional). Samples from MuSD (NFE $= 1$, FID $= 1.341$).



Figure 14: ImageNet 64×64 (conditional). Samples from MuSD (NFE $= 1$, FID $= 1.035$).

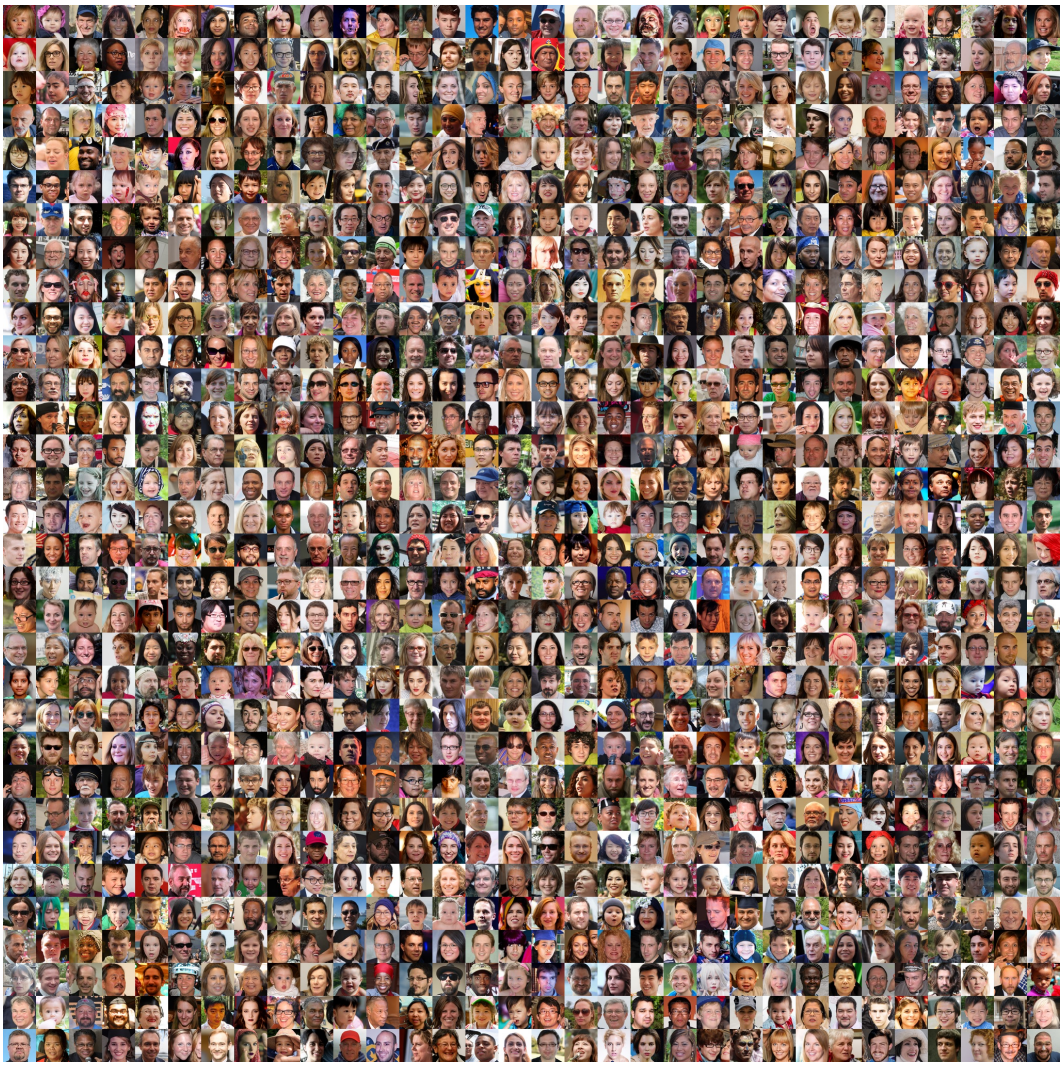


Figure 15: FFHQ 64×64. Samples from MuSD (NFE $= 1$, FID $= 0.836$).

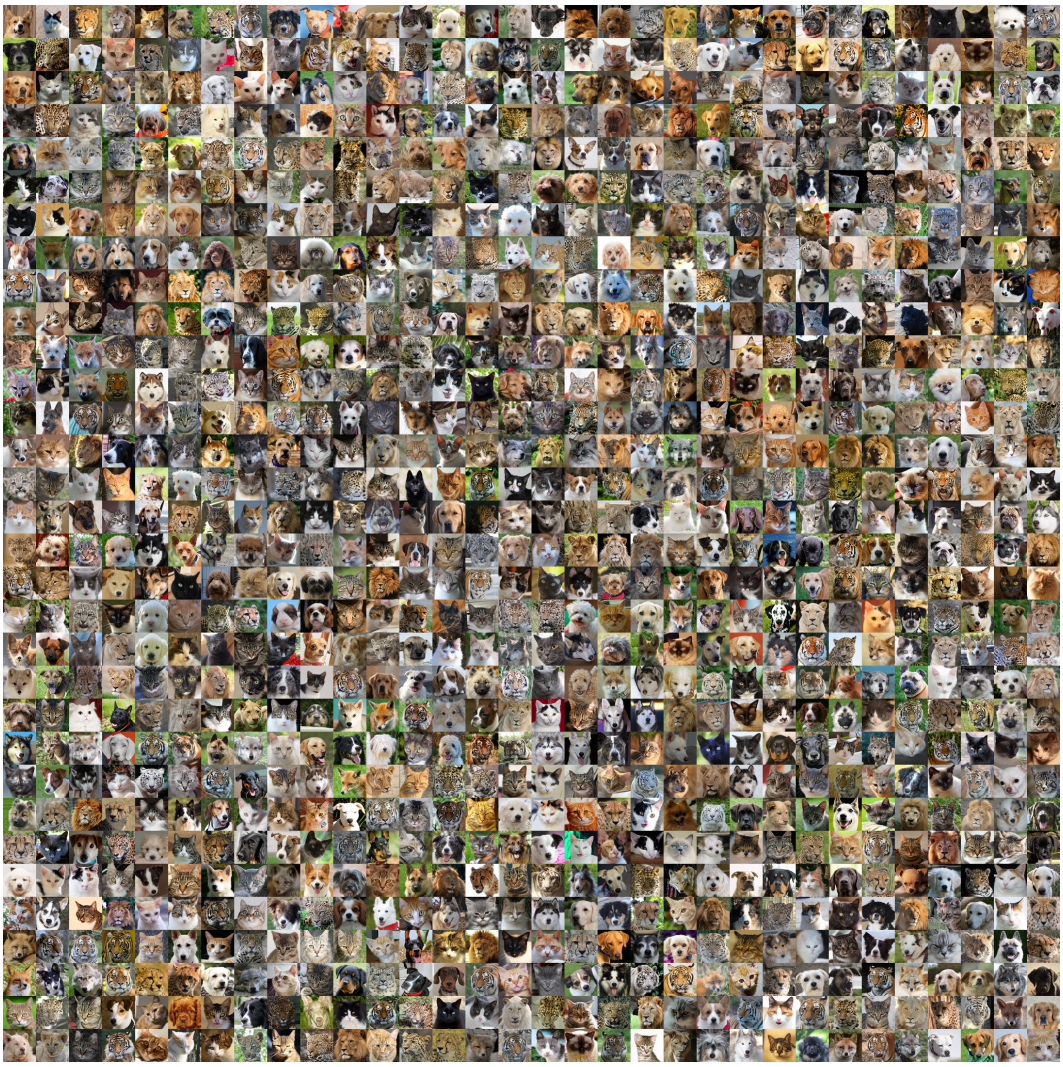
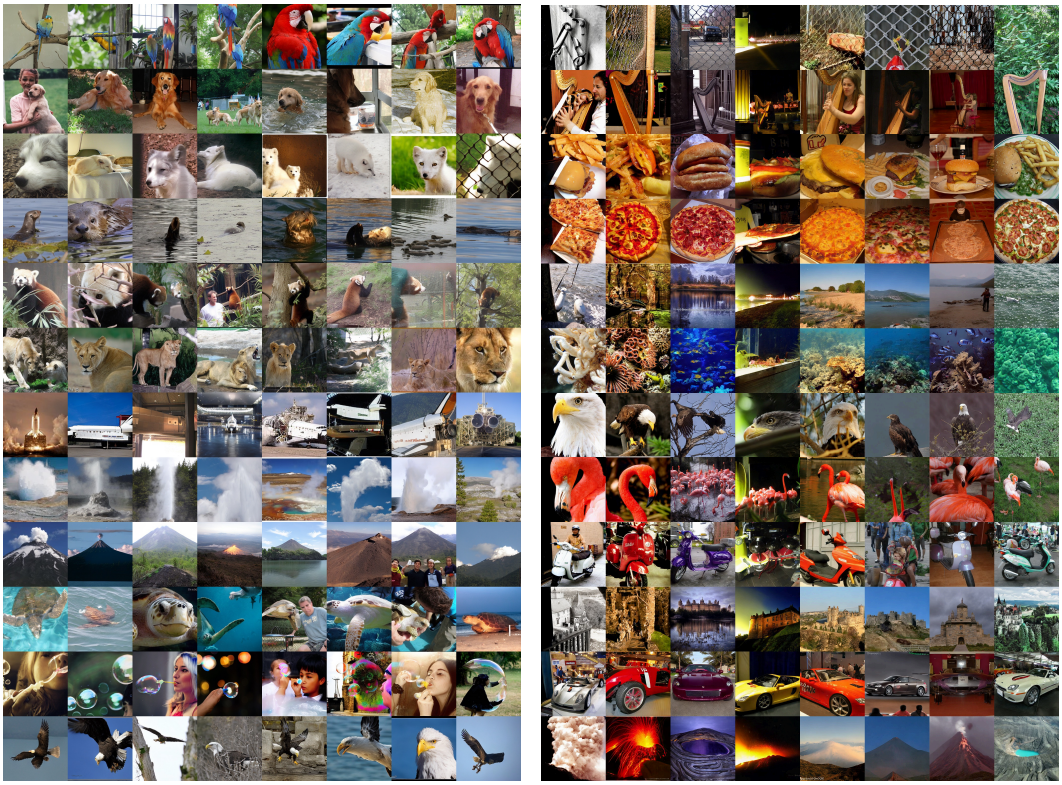


Figure 16: AFHQv2 64×64. Samples from MuSD (NFE $= 1$, FID $= 1.192$).



(a) DiT-XL/2, ImageNet 256×256 (b) EDM2-S, ImageNet 512×512

Figure 17: One-step samples from MuSD on ImageNet. (a) DiT-XL/2 teacher; each image uses a different random seed. Rows from top: macaw, golden retriever, arctic fox, otter, red panda, lion, space shuttle, geyser, volcano, loggerhead, bubble, and bald eagle. (b) EDM2-S teacher; each column uses the same initial noise for all classes. Rows from top: chain-link fence, harp, cheeseburger, pizza, lakeside, coral reef, bald eagle, flamingo, motor scooter, castle, sports car, and volcano.