# ENERGY-GUIDED DISTRIBUTION MATCHING DISTILLATION FOR DIVERSE FEW-STEP VISUAL GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Diffusion distillation enables high-quality generation in only a few sampling steps, but aggressive acceleration often substantially reduces sample diversity. In particular, Distribution Matching Distillation (DMD) optimizes a reverse-KL objective whose mode-seeking behavior can collapse the teacher's rich multimodal distribution. Existing diversity-preserving methods mitigate this issue through trajectory-level regularization, forcing individual student trajectories to reproduce teacher trajectories under a prescribed noise correspondence. We argue that such pairing is unnecessarily restrictive: diversity should instead be preserved at the distribution level. We introduce Energy-Guided Distribution Matching Distillation (EG-DMD), which complements DMD with population-level teacher supervision for improved mode coverage. At each training step, EG-DMD jointly rolls out multiple student and teacher trajectories to an intermediate diffusion time and matches their empirical distributions through a Sinkhorn-divergence energy. Unlike paired regression, this objective optimizes the coupling between student and teacher populations, preserving teacher modes without enforcing a particular noise-to-sample correspondence. Across diffusion models with diverse architectures, EG-DMD consistently improves the diversity–quality frontier over DMD and existing diversity-preserving baselines while retaining four-step inference. On SDXL, EG-DMD improves ImageReward over the strongest baseline by 8.9%, while increasing DINO and CLIP diversity by 13.3% and 23.6%, respectively. These results establish population-level energy guidance as a simple, data-free, and architecture-agnostic principle for high-fidelity, diverse few-step generation.

## 1 INTRODUCTION

Diffusion and flow-based models (Ho et al., 2020; Song et al., 2021b; Lipman et al., 2022; Sohl-Dickstein et al., 2015) have become the prevailing tool for visual content generation with complex conditioning, such as text-to-image (Rombach et al., 2021; Esser et al., 2024) and text-to-video (Ho et al., 2022; Wan et al., 2025) synthesis. Despite their success, they typically require a sequential and iterative sampling procedure, which requires a large number of calls of the denoiser, leading to prohibitively slow inference speed that hinders latency-sensitive applications (Han et al., 2026b). To address this bottleneck, extensive efforts have been paid to *distill* the multi-step diffusion teacher model into a few-step student model. Amongst these approaches, distribution matching distillation (DMD) has become an outstanding means that demonstrates exceptional few-step sample quality on text-to-image and text-to-video diffusion models (Yin et al., 2024b;a; 2025; Huang et al., 2025).

However, the success of DMD comes with an undesirable side-effect. Its reverse Kullback–Leibler (KL) divergence-based distillation objective collapses the rich modes exhibited in the teacher model into the locally dominant ones (Zheng et al., 2025; Wu et al., 2026), incurring a severe sacrifice in diversity in the few-step samples. Such an effect is strongly against the goal of generative modeling, where the preservation of the diverse and multi-modal target distribution is inherently essential.

Earlier attempts approach this problem by incorporating an auxiliary perceptual or adversarial loss that measures between the student generated distribution and a reference data distribution (Yin et al., 2024b;a; Sauer et al., 2024). Yet, they require additional efforts in dataset curation, and the poten-



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.




Figure 1: Sampled images from the 4-step model distilled by EG-DMD. Our method produces high quality images across a diverse set of latent-space and pixel-space text-to-image diffusion models.

tial distribution shift in the pre-collected dataset can be detrimental to the distillation performance. Recently, Wu et al. (2026) propose to regularize the distillation by compressing several sampling steps of the teacher model into the initial student step. Despite enhanced mode coverage, it only enforces a trajectory-level regularization and ignores the distributional constraint, leading to inferior quality-diversity tradeoff in the distilled distribution.

In this work, we present *energy-guided* distribution matching distillation (EG-DMD) that preserves both the rich modality in the teacher distribution and the superior sample fidelity. Our core innovation is the energy guidance, a loss function that complements DMD by minimizing a distributional discrepancy measured at a certain intermediate diffusion timestep. In particular, we instantiate the energy function as the Sinkhorn divergence, which permits an efficient sample-based estimator and has shown effectiveness in distributional preservation (Han et al., 2026a; He et al., 2026). Practically, at each student training step, we first rollout the teacher and student in parallel to explore different modes up to an intermediate diffusion time and then compute the estimator using the sampled states, which is then combined with the DMD optimization as an auxiliary guidance term. Crucially, our approach shifts the paradigm from trajectory compression in Wu et al. (2026) to distribution-level regularization, avoiding suboptimal commitment to the teacher distribution's leading modes and remarkably enhancing the sample quality-diversity frontier.

We perform extensive experimental evaluations of EG-DMD on an array of text-to-image diffusion models with diverse architectures, including UNet and DiT-based latent-space models and pixel-space models. The results strongly verify the generality of EG-DMD across different models, showing that EG-DMD situates at the very frontier in sample quality and diversity and achieves state-of-the-art distillation performance. Specifically, when distilling SDXL, EG-DMD achieves 8.9% higher ImageReward quality metric compared with Wu et al. (2026), while also lifting the DINO and CLIP diversity score by 13.3% and 23.6%, respectively. Overall, EG-DMD opens up new opportunity in high-fidelity yet distributionally diverse diffusion distillation for fast visual synthesis.

## 2 Related Work

**Diffusion distillation.** Diffusion distillation has emerged as an essential technique that aims to distill a few-step student from the multi-step teacher for fast sampling. A line of work resorts to the ODE sampling trajectory of the teacher model and distills the student by matching the trajectory-level endpoint, including progressive distillation (Salimans & Ho, 2022; Meng et al., 2023), consistency models (Song et al., 2023; Song & Dhariwal, 2024; Wang et al., 2024; Lu & Song, 2025; Kim et al., 2024), MeanFlows (Geng et al., 2025; 2026), and their variants (Boffi et al., 2025; Sabour et al., 2025). Yet, their trajectory-level objective incurs insufficient supervision on the student distribution, leading to off-manifold and over-smoothed samples with degraded visual quality. Distribution matching distillation (Yin et al., 2024b;a) and variational score distillation (Wang et al., 2023a) take a different path that follow the gradient direction of the reverse KL divergence between the student and teacher distribution. Despite the high sample quality, it inherits the mode-seeking property of the reverse KL and incur mode concentration and diversity collapse (Zheng et al., 2025). Auxiliary losses

have been proposed in the form of perceptual (Yin et al., 2024b) and adversarial (Yin et al., 2024a; Chadebec et al., 2025), yet they require data curation and are vulnerable to the distribution shift. Recently, DP-DMD (Wu et al., 2026) introduces a diversity-preserving loss that compresses several teacher sampling steps into the student's first step. However, the approach remains trajectory-based and lacks distributional regularization, yielding unsatisfactory quality-diversity tradeoff. Differently, our method enforces energy guidance, a distributional metric that guides the distillation to preserve the teacher distribution's low-level statistics, situating at the frontier of sample quality and diversity.

**Sample-based distribution matching.** Our energy guidance approach is related to the line of work that optimizes a generated distribution towards a target distribution using only samples. The idea dates back to kernelized two-sample test (Gretton et al., 2012) and moment matching (Li et al., 2015), and has its early application in generative modeling as a type of generative adversarial networks (Li et al., 2017; Genevay et al., 2018; Unterthiner et al., 2017). The idea also revives recently as a core component for inductive moment matching (Zhou et al., 2025), drifting models (Deng et al., 2026; He et al., 2026), and Wasserstein gradient flows (Han et al., 2026a), where the design varies in the choice of the divergence that induces different geometries and the curriculum of the learning algorithm. Notably, these approaches are typically applied in a feature space instead of the high-dimensional raw ambient space, since the estimation of the distributional metrics in high-dimensional space can be prohibitively noisy with limited number of samples. Zhang et al. (2026) shows that the intermediate layers of the diffusion teacher model itself can serve as such a feature space that produces meaningful optimization signal. While our energy guidance is also implemented via sample-based distribution matching, it is measured between two proxy distributions from the student and teacher to enforce mode coverage, instead of being a standalone distillation objective which requires to directly draw samples from the target distribution by *e.g.* constructing a dataset.

## 3 Preliminaries

**Diffusion Models.** Diffusion and flow models (Ho et al., 2020; Song et al., 2021b; Lipman et al., 2022; Liu et al., 2023) define a forward process that gradually perturbs data samples with Gaussian noise. Given a clean sample $\mathbf{x}_0 \sim p_{\text{data}}$, its noisy counterpart at time $t \in [0, 1]$ can be expressed as $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\alpha_t$ and $\sigma_t$ are determined by the noise schedule. We denote the marginal distribution of $\mathbf{x}_t$ by $p_t$ and its corresponding score function as $\mathbf{s}(\mathbf{x}_t, t) := \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$. A diffusion model parameterized by $\mathbf{s}_\phi(\mathbf{x}_t, t)$ is trained to approximate this time-dependent score through *e.g.* denoising score matching (Song et al., 2021b). Generation proceeds by reversing the process from Gaussian noise towards the data distribution, typically realized by calling certain diffusion ODE solver (Song et al., 2021a; Lu et al., 2022). Formally, we specify

$$\mathbf{x}_t = \text{Solve}\left(\mathbf{x}_s, \mathbf{s}_\phi, s, t, N\right), \qquad 0 \le t < s \le 1 \tag{1}$$

as the empirical solver that solves the trajectory from initial timestep $s$ to a target timestep $t$ using $N$ steps. A standard diffusion solver starts from an initial latent $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and runs the inference via $\mathbf{x}_0 = \text{Solve}\left(\mathbf{x}_1, \mathbf{s}_\phi, 1, 0, N\right)$ to obtain the clean sample $\mathbf{x}_0$, where $N$ is typically large to maintain the sample quality, making the inference procedure inherently costly and time-consuming.

**Distribution Matching Distillation.** Distribution Matching Distillation (DMD) (Yin et al., 2024b;a; Wang et al., 2023a) accelerates diffusion sampling by directly matching the distribution induced by a few-step student generator to that represented by a pre-trained diffusion teacher $\mathbf{s}_\phi$. Denote $G_\theta$ as the one-step student generator. Starting from $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the student produces a clean sample $\hat{\mathbf{x}}_0 = G_\theta(\mathbf{x}_1, 1)$ whose distribution is denoted by $q_\theta$. Define $\hat{\mathbf{x}}_t = \alpha_t \hat{\mathbf{x}}_0 + \sigma_t \boldsymbol{\epsilon}$ and its marginal as $q_{\theta,t}$. DMD performs distribution matching over these noisy marginals by minimizing the reverse KL

$$\mathcal{L}_{\text{DMD}}(\theta) = \mathbb{E}_t \left[ D_{\text{KL}} \left( q_{\theta,t} \,\|\, p_{\phi,t} \right) \right], \tag{2}$$

where $p_{\phi,t}$ is the corresponding noisy marginal of the teacher distribution. Importantly, optimizing Eq. 2 does not require explicit evaluation of either distribution density. Its gradient can instead be expressed through the difference between their score functions:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{t, \mathbf{x}_1, \boldsymbol{\epsilon}} \left[ w(t) \left( \frac{\partial \hat{\mathbf{x}}_0}{\partial \theta} \right)^\top \left( \mathbf{s}_\phi(\hat{\mathbf{x}}_t, t) - \mathbf{s}_\psi(\hat{\mathbf{x}}_t, t) \right) \right], \tag{3}$$

where $w(t)$ absorbs the timestep-dependent scaling induced by the diffusion parameterization. Here, $\mathbf{s}_\phi$ is the frozen teacher score model while $\mathbf{s}_\psi$ is an auxiliary *fake score* model trained online to approximate $\nabla_{\mathbf{x}_t} \log q_{\theta,t}(\mathbf{x}_t)$ using samples produced by the current student.

**Multi-step DMD.** Yin et al. (2024a) further extends the framework to multi-step generation using *backward simulation*. In particular, the one-step direct prediction is replaced by an $N$-step solver

$$\hat{x}_0 = G_\theta^{(N)}(\mathbf{x}_1, 1) \coloneqq \mathrm{Solve}(\mathbf{x}_1, g_\theta, 1, 0, N), \tag{4}$$

where $g_\theta(\mathbf{x}_t, t)$ is the student and the function $\mathrm{Solve}(\cdot)$ carries gradient during training, distinct from inference. We adopt the standard choice of $N = 4$, following Wu et al. (2026); Yin et al. (2024a).

# 4 METHOD

We introduce EG-DMD, a distributional regularization framework for few-step distribution matching distillation. Our key observation is that standard DMD provides a strong *quality-seeking* signal but does not explicitly constrain the student to retain the *population-level coverage* of the multi-step teacher (Sec. 4.1). We address this limitation by constructing intermediate proxy distributions for the teacher and student and matching them with a geometry-aware Sinkhorn divergence (Sec. 4.2 and 4.3). We then integrate the framework into DMD with dedicated practical designs in Sec. 4.4.

## 4.1 FROM MODE-SEEKING DMD TO POPULATION-LEVEL GUIDANCE

**Mode-seeking behavior of DMD.** Despite its high sample quality, DMD often reduces the diversity of a few-step generator relative to its multi-step teacher. In particular, the reverse-KL in Eq. 2 is evaluated under the current student distribution. Consequently, a mode of the teacher that receives negligible probability mass under $q_{\theta,t}$ contributes little optimization signal. This makes reverse KL inherently mode-seeking and allows a student to optimize towards only a subset of the modes represented by the teacher. This issue becomes particularly pronounced in few-step sampling, where early errors in global semantic structure cannot be easily recovered by subsequent denoising steps.

A natural remedy is to complement DMD with a *population-level coverage objective*. One possibility is to match student samples against a dataset or a large collection of fully generated teacher samples. Such approaches, however, introduce additional data curation costs. More importantly, for conditional generation, a finite offline dataset may provide only sparse coverage of the multiple valid outputs associated with each condition and may deviate from the teacher distribution.

**Energy-guided distribution matching.** Instead, we obtain the required population-level supervision directly from the pretrained teacher during training. For a prompt $\mathbf{c} \sim \mathcal{D}$, we construct a student proxy distribution $\tilde{q}_\theta^\tau(\cdot \mid \mathbf{c})$ and a teacher proxy distribution $\tilde{p}_\phi^\tau(\cdot \mid \mathbf{c})$ at an intermediate diffusion time $\tau \in [0, 1)$, where $\mathcal{D}$ denotes the prompt dataset. We define the *energy-guidance objective*

$$\mathcal{L}_{\mathrm{EG}}(\theta) = \mathbb{E}_{\mathbf{c}\sim\mathcal{D}}\left[\mathcal{E}\left(\tilde{q}_\theta^\tau(\cdot \mid \mathbf{c}), \tilde{p}_\phi^\tau(\cdot \mid \mathbf{c})\right)\right], \tag{5}$$

where the energy function $\mathcal{E}$ is a sample-based discrepancy between two probability measures. For Eq. 5 to operate effectively, both proxies should be inexpensive to sample and should retain the semantics that determine the diversity of the final generated samples. Moreover, $\mathcal{E}$ should be a metric that compares distributions and provides effective gradient signal. These requirements motivate our two central design choices: *intermediate-state proxy distributions* and a *Sinkhorn-divergence energy*.

## 4.2 INTERMEDIATE PROXY DISTRIBUTIONS

We design the student and teacher proxy distributions as the distributions at a particular intermediate diffusion timestep that are efficiently approximated by student and teacher solvers. Specifically, let

$$F_\theta^\tau(\mathbf{z}, \mathbf{c}) \coloneqq \mathrm{Solve}\left(\mathbf{z}, g_\theta, 1, \tau, 1\right) \tag{6}$$

denote the first student sampling step from initial noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the *proxy timestep* $\tau$. The corresponding student proxy distribution is therefore

$$\tilde{q}_\theta^\tau(\cdot \mid \mathbf{c}) = F_\theta^\tau(\cdot, \mathbf{c})_\# \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

where "$\#$" denotes the pushforward of a probability measure. We construct the teacher proxy at the *same* diffusion time $\tau$. Let $K_\tau$ denote the number of teacher solver steps required by the adopted teacher schedule to integrate from $t = 1$ to $t = \tau$. We define

$$T_\phi^\tau(\mathbf{z}, \mathbf{c}) \coloneqq \mathrm{Solve}\left(\mathbf{z}, \mathbf{s}_\phi, 1, \tau, K_\tau\right), \qquad \tilde{p}_\phi^\tau(\cdot \mid \mathbf{c}) = T_\phi^\tau(\cdot, \mathbf{c})_\# \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{8}$$

In practice, we parameterize $K_\tau = \lfloor (1-\tau) K_{\max} \rfloor$ where $K_{\max} = 30$ upper-bounds the solver steps.

**Interpretation of $\tau$.** The choice of $\tau$ controls the information exposed by the proxy. At larger $\tau$, the teacher has undergone only a short denoising trajectory, so the proxy emphasizes the earliest, coarse structural information in the denoised image. Reducing $\tau$ reveals progressively finer semantics and appearances, at the cost of additional teacher solves. In practice, we find that an intermediate anchor, *e.g.* $\tau = 0.5$, already provides sufficient semantic informativeness (see Sec. 5.3).

### 4.3 Sinkhorn Divergence as the Energy Function

We instantiate the energy functional in Eq. 5 using the Sinkhorn divergence, which measures the discrepancy between the student and teacher proxy distributions without imposing a prescribed correspondence. In contrast to trajectory-level regression, the transport coupling is optimized jointly over two populations, making the objective naturally suited for matching multimodal distributions.

**Population-level objective.** For distributions with finite second moments $q, p \in \mathcal{P}_2(\mathbb{R}^d)$, let $\Pi(q,p)$ denote the set of joint probability measures with distributions $q$ and $p$. Under the quadratic cost $c(\mathbf{x},\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$, the entropically regularized optimal transport cost is defined as

$$\mathrm{OT}_\varepsilon(q,p) := \inf_{\pi \in \Pi(q,p)} \left\{ \int \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \, \mathrm{d}\pi(\mathbf{x},\mathbf{y}) + \varepsilon D_{\mathrm{KL}}\big(\pi \,\|\, q \otimes p\big) \right\}, \tag{9}$$

where $\varepsilon > 0$ controls the regularization strength. The corresponding Sinkhorn divergence is

$$S_\varepsilon(q,p) = \mathrm{OT}_\varepsilon(q,p) - \frac{1}{2}\mathrm{OT}_\varepsilon(q,q) - \frac{1}{2}\mathrm{OT}_\varepsilon(p,p). \tag{10}$$

We therefore instantiate the population-level energy in Eq. 5 as

$$\mathcal{L}_{\mathrm{EG}}(\theta) = \mathbb{E}_{\mathbf{c}\sim\mathcal{D}}\left[ S_\varepsilon\left(\tilde{q}_\theta^\tau(\cdot \mid \mathbf{c}), \tilde{p}_\phi^\tau(\cdot \mid \mathbf{c})\right)\right]. \tag{11}$$

Sinkhorn divergence is particularly attractive for several reasons. First, it retains the geometry-aware matching principle in OT while remaining efficiently computable via Sinkhorn iterations (Sinkhorn & Knopp, 1967). Second, unlike trajectory-wise compression which prescribes a fixed correspondence between noise and sample, the transport coupling is determined jointly from two populations. Finally, the marginal constraints of OT discourage mass from being matched only to a small subset of samples, thereby mitigating mode dropping and promoting coverage (Han et al., 2026a).

**Sample-based estimator.** The proxy distributions in Eq. 7 and Eq. 8 are implicit and can only be accessed through samples. We therefore optimize a discrete approximation of Eq. 11. For each conditioning input $\mathbf{c}$, we independently sample two sets of initial latents $\mathbf{z}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0},\mathbf{I}), \mathbf{z}'_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0},\mathbf{I})$, where $i,j = 1,\dots,R$, and obtain the corresponding student and teacher proxy samples

$$\tilde{\mathbf{x}}_q^{(i)} = F_\theta^\tau(\mathbf{z}_i,\mathbf{c}), \qquad \tilde{\mathbf{x}}_p^{(j)} = T_\phi^\tau(\mathbf{z}'_j,\mathbf{c}). \tag{12}$$

These samples induce the empirical measures

$$\hat{q}_R = \frac{1}{R}\sum_{i=1}^R \delta_{\tilde{\mathbf{x}}_q^{(i)}}, \qquad \hat{p}_R = \frac{1}{R}\sum_{j=1}^R \delta_{\tilde{\mathbf{x}}_p^{(j)}}. \tag{13}$$

Substituting these empirical measures into Eq. 10 gives the sample-based energy

$$\widehat{\mathcal{L}}_{\mathrm{EG}} = S_\varepsilon(\hat{q}_R, \hat{p}_R), \tag{14}$$

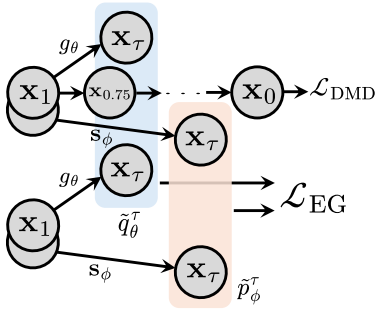


Figure 2: Flowchart of EG-DMD.

which can be efficiently evaluated using differentiable Sinkhorn iterations (see Alg. 2).

Concretely, Eq. 14 constructs an $R \times R$ pairwise cost matrix between student and teacher samples and optimizes a soft transport plan subject to uniform marginal constraints. As $R$ increases, the empirical measures provide increasingly rich approximations of the underlying proxy distributions. In practice, however, using a large $R$ is computationally expensive because each teacher sample requires a partial multi-step rollout. We therefore operate in the small-$R$ regime, where the choice of representation becomes particularly important, as we will detail next.

**Feature projection.** Directly estimating the transport geometry in the high-dimensional latent space can be statistically inefficient when $R$ is small. Inspired by Zhang et al. (2026), we use the pretrained teacher itself to provide a more informative representation for the finite-sample estimator. Specifically, let $\Phi(\mathbf{x}, \tau, \mathbf{c})$ denote an intermediate feature representation extracted from a frozen middle block of the teacher network. We spatially pool the feature map and flatten the resulting representation (see Table 6). We then project the samples via

$$\mathbf{a}_i = \Phi\left(\tilde{\mathbf{x}}_q^{(i)}, \tau, \mathbf{c}\right), \qquad \mathbf{b}_j = \mathrm{sg}\left[\Phi\left(\tilde{\mathbf{x}}_p^{(j)}, \tau, \mathbf{c}\right)\right], \tag{15}$$

and define the empirical feature distributions $\hat{q}_R^\Phi = \frac{1}{R}\sum_{i=1}^R \delta_{\mathbf{a}_i}$, $\hat{p}_R^\Phi = \frac{1}{R}\sum_{j=1}^R \delta_{\mathbf{b}_j}$. Our practical energy-guidance loss is consequently $\widehat{\mathcal{L}}_{\mathrm{EG}}^\Phi = S_\varepsilon\left(\hat{q}_R^\Phi, \hat{p}_R^\Phi\right)$. This feature transformation is a practical choice for constructing a more informative finite-sample estimate for the underlying distribution-matching principle. During training, the parameters of $\Phi$ remain frozen, while the gradient with respect to its student-side input is retained and backpropagates through the student proxy $F_\theta^\tau$.

### 4.4 Integration with Multi-Step DMD

Our energy guidance is designed to complement DMD where the two objectives act on different parts of the few-step trajectory. Specifically, we first compute the student and teacher proxy samples following Eq. 12, which are used to evaluate the guidance loss $\widehat{\mathcal{L}}_{\mathrm{EG}}$. Separately, for the DMD branch, we perform the $N$-step student sampling

$$\hat{\mathbf{x}}_{0,i} = \mathrm{Solve}\left(\mathbf{z}_i, g_\theta, \tau, 0, N\right), \tag{16}$$

and the standard DMD gradient is evaluated on $\hat{\mathbf{x}}_{0,i}$ using the frozen teacher $\mathbf{s}_\phi$ and online fake score $\mathbf{s}_\psi$. The resulting student update is

$$\nabla_\theta \mathcal{L} = \frac{1}{R}\sum_{i=1}^R \nabla_\theta \mathcal{L}_{\mathrm{DMD}}^{(i)} + \lambda \nabla_\theta \widehat{\mathcal{L}}_{\mathrm{EG}}^\Phi, \tag{17}$$

where $\lambda > 0$ controls the strength of the energy guidance. Notably, when performing Eq. 16, we also detach the first student step from the computation graph. By this means, the energy branch directly supervises the first, structure-forming transition while the DMD branch optimizes the subsequent refinement trajectory. Empirically, we find this separation beneficial: allowing both objectives to backpropagate through the entire trajectory can conflict their joint signal at the early denoising stage and leads to artifacts (see Fig. 5). The complete student update step is summarized in Alg. 1 with the conceptual flowchart depicted in Fig. 2. We leave the fake score update step untouched.

**Algorithm 1** EG-DMD Student Update Step

**Input:** student $g_\theta$, teacher $\mathbf{s}_\phi$, sampled prompt $\mathbf{c}$, proxy time $\tau$, number of proxy samples $R$, regularization $\varepsilon$, and energy weight $\lambda$.

1: Sample latents $\{\mathbf{z}_i\}_{i=1}^R, \{\mathbf{z}'_j\}_{j=1}^R \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
2: Compute student proxy samples
$\tilde{\mathbf{x}}_q^{(i)} \leftarrow \mathrm{Solve}(\mathbf{z}_i, g_\theta, 1, \tau, 1), \quad 1 \le i \le R.$
3: Compute teacher proxy samples w/o gradient
$\tilde{\mathbf{x}}_p^{(j)} \leftarrow \mathrm{Solve}(\mathbf{z}'_j, \mathbf{s}_\phi, 1, \tau, K_\tau), \quad 1 \le j \le R.$
4: Project features
$\mathbf{a}_i \leftarrow \Phi(\tilde{\mathbf{x}}_q^{(i)}, \tau, \mathbf{c}), \quad \mathbf{b}_j \leftarrow \mathrm{sg}[\Phi(\tilde{\mathbf{x}}_p^{(j)}, \tau, \mathbf{c})].$
5: Compute energy guidance
$$\widehat{\mathcal{L}}_{\mathrm{EG}}^\Phi \leftarrow S_\varepsilon\left(\frac{1}{R}\sum_i \delta_{\mathbf{a}_i}, \frac{1}{R}\sum_j \delta_{\mathbf{b}_j}\right).$$
6: Obtain student sample (first step detached):
$\hat{\mathbf{x}}_{0,i} \leftarrow \mathrm{Solve}\left(\mathbf{z}_i, g_\theta, \tau, 0, N\right).$
7: Estimate $\nabla_\theta \mathcal{L}_{\mathrm{DMD}}^{(i)}$ with $\hat{\mathbf{x}}_{0,i}$, $\mathbf{s}_\phi$, and $\mathbf{s}_\psi$ (Eq. 3).
8: Update $\theta$ with $\nabla_\theta \mathcal{L}$ defined in Eq. 17.

**Discussions.** DP-DMD supervises the first student step with a paired teacher target generated from the same initial noise. Up to time-dependent scaling, its auxiliary objective can be rewritten as

$$\mathcal{L}_{\mathrm{pair}} = \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|F_\theta^\tau(\mathbf{z}, \mathbf{c}) - T_\phi^\tau(\mathbf{z}, \mathbf{c})\right\|_2^2\right], \tag{18}$$

with proof provided in Appendix B. Although paired regression indeed upper-bounds marginal transport discrepancy (see Proposition A.1 in Appendix A.1), it additionally constrains the noise-to-sample correspondence through the shared initial latent. This constraint can favor contraction even when the student already matches the teacher distribution. We formalize this artifact in the following proposition. Here, we fix the prompt and proxy time and write $F = F_\theta^\tau(\cdot, \mathbf{c})$ and $T = T_\phi^\tau(\cdot, \mathbf{c})$.

**Proposition 4.1** (Contraction despite matched marginals). *Let $X = F(\mathbf{z})$ and $Y = T(\mathbf{z})$ have finite second moments and the same marginal distribution, but $\Delta = \mathbb{E}\|X - Y\|_2^2 > 0$. Define $F_s(\mathbf{z}) = m + s(F(\mathbf{z}) - m)$, where $m = \mathbb{E}X = \mathbb{E}Y$, and suppose that $F_s$ is a feasible student map with $s < 1$ sufficiently close to 1. Then*

$$\frac{d}{ds}\mathcal{L}_{\mathrm{pair}}(F_s)\Big|_{s=1} = \Delta > 0, \qquad \mathrm{Cov}(F_s(\mathbf{z})) = s^2\,\mathrm{Cov}(Y). \tag{19}$$

Table 1: **Controlled comparison of the DMD variants** on Pick-a-Pic and COCO2014-10K prompt datasets with **latent-space diffusion models**. The best and second-best results are highlighted in **bold** and underline, respectively. Numbers of the baselines are taken from Wu et al. (2026).

| Method | Image-Free | NFEs | Pick-a-Pic | | | | | | COCO2014-10K | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Diversity | | Quality | | Preference | | Diversity | | Quality | | Preference | |
| | | | DINO↑ | CLIP↑ | VQ-R1↑ | MIQA↑ | ImgR↑ | PicS↑ | DINO↑ | CLIP↑ | VQ-R1↑ | MIQA↑ | ImgR↑ | PicS↑ |
| *SD3.5-M* (CFG=3.5) | | | | | | | | | | | | | | |
| Base Model | - | 60 | 0.240 | 0.221 | 4.657 | 1.020 | 1.007 | 21.80 | 0.288 | 0.204 | 4.636 | 1.043 | 0.910 | 22.31 |
| DMD | ✓ | 4 | 0.137 | 0.133 | 4.649 | 1.016 | **1.189** | 21.75 | 0.210 | 0.154 | 4.690 | 1.060 | **1.053** | 22.40 |
| DMD-LPIPS | × | 4 | 0.169 | 0.169 | 4.598 | 1.005 | 1.063 | 21.62 | 0.204 | 0.168 | 4.599 | 1.012 | 0.949 | 22.29 |
| DMD-GAN | × | 4 | 0.183 | 0.162 | 4.525 | 0.984 | 1.033 | 21.63 | 0.214 | 0.174 | 4.584 | 0.983 | 0.751 | 22.02 |
| DP-DMD | ✓ | 4 | 0.179 | 0.182 | 4.646 | 1.017 | 1.142 | 21.76 | 0.250 | 0.182 | 4.689 | 1.032 | 0.988 | **22.41** |
| **EG-DMD** | ✓ | 4 | **0.191** | **0.188** | **4.688** | **1.045** | 1.156 | **21.78** | **0.261** | **0.190** | **4.757** | **1.071** | 1.015 | **22.41** |
| *SDXL* (CFG=8.0) | | | | | | | | | | | | | | |
| Base Model | - | 100 | 0.219 | 0.204 | 4.675 | 1.033 | 1.016 | 21.96 | 0.269 | 0.219 | 4.637 | 1.056 | 0.820 | 22.54 |
| DMD | ✓ | 4 | 0.109 | 0.133 | 4.667 | 0.971 | 0.951 | 21.68 | 0.139 | 0.143 | 4.643 | 0.982 | 0.712 | 22.21 |
| DMD-LPIPS | × | 4 | 0.136 | 0.139 | 4.610 | 0.976 | 0.883 | 21.74 | 0.181 | 0.137 | 4.723 | 0.984 | 0.729 | 22.38 |
| DMD-GAN | × | 4 | 0.126 | 0.124 | 4.624 | **1.019** | 1.036 | 21.80 | 0.157 | 0.117 | **4.789** | 1.030 | 0.801 | 22.63 |
| DP-DMD | ✓ | 4 | 0.173 | 0.161 | 4.591 | 0.954 | 1.011 | 21.75 | 0.204 | 0.157 | 4.765 | **1.041** | 0.835 | 22.45 |
| **EG-DMD** | ✓ | 4 | **0.196** | **0.199** | **4.671** | 0.980 | **1.101** | **21.94** | **0.233** | **0.199** | 4.744 | 1.035 | **0.849** | **22.65** |

*Consequently, a sufficiently small contraction decreases the paired loss while reducing the student's total variance.*

Proposition 4.1 shows that a distributionally correct student need not be stationary for paired regression: disagreement in the prescribed correspondence alone can create a loss-reducing contraction direction. This motivates our EG-DMD's population level supervision, which optimizes the transport coupling instead of prescribing a noise-to-sample correspondence. Note that the result targets on the paired auxiliary objective, rather than asserting contraction under every joint DMD update.

## 5 Experiments

In this section, we conduct comprehensive evaluation of our approach. We describe the detailed setup in Sec. 5.1 and present the main results in Sec. 5.2, with concrete ablation study in Sec. 5.3.

### 5.1 Experiment Setup

**Training setup.** We verify the efficacy of our approach on an array of text-to-image diffusion models with different architectures, including the MMDiT-based Stable Diffusion 3.5 Medium (Esser et al., 2024), the UNet-based SDXL (Podell et al., 2023), and the pixel-space diffusion model MiniT2I-B/16 (Wang et al., 2026). For the training configurations, we strictly follow Wu et al. (2026) and train on the prompt dataset DiffusionDB (Wang et al., 2023b) for 6,000 iterations, with one generator update every 5 fake score updates. We use the default classifier-free guidance scale for each model. More details are deferred to Appendix B.

Table 2: Controlled comparison of DMD variants with pixel-space diffusion model MiniT2I-B/16.

| Method | NFEs | Pick-a-Pic | | | | | | GenEval↑ |
|---|---|---|---|---|---|---|---|---|
| | | Diversity | | Quality | | Preference | | |
| | | DINO↑ | CLIP↑ | VQ-R1↑ | MIQA↑ | ImgR↑ | PicS↑ | |
| *MiniT2I-B/16* (CFG=5.0) | | | | | | | | |
| Base Model | 100 | 0.136 | 0.146 | 4.305 | 1.052 | 1.112 | 21.08 | 0.872 |
| DMD | 4 | 0.106 | 0.139 | **4.376** | **1.014** | 0.979 | 20.74 | 0.824 |
| DP-DMD | 4 | 0.113 | 0.136 | 4.303 | 0.992 | 0.965 | 20.75 | 0.827 |
| **EG-DMD** | 4 | **0.118** | **0.143** | 4.324 | 0.997 | **1.002** | **20.83** | **0.835** |

**Evaluation protocol.** We employ a comprehensive set of evaluation metrics that span across sample diversity, perceptual quality and preference, and prompt following, grounded on the protocol in Wu et al. (2026). *(I) Diversity.* We generate $M$ samples per test set prompt $\{\mathbf{x}^{(i)}\}_{i=1}^{M}$ and employ the diversity score (Wu et al., 2026) computed as $\text{DivScore} = 1 - \frac{2}{M(M-1)} \sum_{i<j} \cos\left(\mathcal{F}(\mathbf{x}^{(i)}), \mathcal{F}(\mathbf{x}^{(j)})\right)$ where $\mathcal{F}(\cdot)$ is a pretrained feature encoder. We use $M = 9$ to ensure a fair comparison against Wu et al. (2026). We employ two feature encoders: DINOv3-ViT-Large (Siméoni et al., 2025) for image structural information and CLIP-ViT-Large (Radford et al., 2021) for text-conditioned image semantics. *(II) Quality.* We leverage VisualQuality-R1 (Wu et al., 2025) and MANIQA (Yang et al., 2022) as the visual quality evaluators. *(III) Preference.* We adopt the widely used ImageReward (Xu et al., 2023) and PickScore (Kirstain et al., 2023) to reflect the human preference captured by dedicated reward models. *(IV) Prompt following.* We use GenEval (Ghosh et al., 2023) that measures the prompt following capability across different dimensions. For (I)-(III), we benchmark on two prompt datasets, namely Pick-a-Pic (Kirstain et al., 2023) and COCO2014-10K (Lin et al., 2014).

Table 3: **System-level comparison** of open-source few-step methods. The best and second-best results are highlighted in **bold** and underline. Numbers of baselines are taken from Wu et al. (2026).

| Method | Image-Free | NFEs | Pick-a-Pic | | | | | | COCO2014-10K | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Diversity | | Quality | | Preference | | Diversity | | Quality | | Preference | |
| | | | DINO↑ | CLIP↑ | VQ-R1↑ | MIQA↑ | ImgR↑ | PicS↑ | DINO↑ | CLIP↑ | VQ-R1↑ | MIQA↑ | ImgR↑ | PicS↑ |
| Base Model | - | 60 | 0.230 | 0.205 | 4.665 | 1.042 | 1.067 | 21.46 | 0.278 | 0.202 | 4.757 | 1.065 | 1.014 | 22.50 |
| Hyper-SD | × | 4 | 0.202 | **0.198** | 4.302 | 0.922 | 0.830 | 20.75 | **0.295** | **0.231** | 3.920 | 0.913 | 0.653 | 21.52 |
| Flash | × | 4 | 0.184 | 0.172 | 4.625 | 1.023 | 1.014 | **21.62** | 0.229 | 0.184 | 4.517 | 0.998 | 0.878 | 22.38 |
| TDM | ✓ | 4 | 0.148 | 0.167 | 4.675 | 1.013 | 1.134 | 21.21 | 0.196 | 0.172 | 4.617 | 1.046 | 0.998 | 21.49 |
| DP-DMD | ✓ | 4 | 0.162 | 0.181 | 4.673 | 1.001 | 1.128 | 21.15 | 0.197 | 0.174 | 4.672 | 1.048 | **1.034** | 22.29 |
| **EG-DMD** | ✓ | 4 | **0.215** | 0.192 | **4.679** | **1.027** | **1.144** | 21.29 | 0.255 | 0.189 | **4.759** | **1.055** | 1.021 | **22.40** |

**Baselines.** Besides the vanilla DMD, we also compare against several representative DMD variants. DMD-LPIPS (Yin et al., 2024b) adds a perceptual loss that enforces perceptual alignment in LPIPS feature space (Zhang et al., 2018) with samples from a pre-collected image dataset. DMD-GAN (Yin et al., 2024a) trains an adversarial discriminator, which complicates the distillation pipeline and increases the risk of introducing visual shortcuts. DP-DMD (Wu et al., 2026) introduces a trajectory-level loss that compresses multiple teacher steps into the initial step of the student. We keep the same teacher model, CFG, dataset, and training step, ensuring a controlled comparison. We further include a system-level comparison against several open-source few-step models. Hyper-SD (Ren et al., 2024) performs trajectory-level compression, Flash Diffusion (Chadebec et al., 2025) leverages adversarial distillation, and TDM (Luo et al., 2025) matches the student and teacher distribution along the diffusion trajectory. We strictly follow the setting in Wu et al. (2026) with Stable Diffusion 3 Medium (Esser et al., 2024) as the reference model.

### 5.2 Main Results

The main results on latent-space diffusion models are presented in Table 1 and 4, while the pixel-space MiniT2I results are in Table 2 and 7. We have the following findings. **1.** The vanilla DMD significantly degrades diversity metrics across all models. Despite the high quality scores, DMD sacrifices diversity and suffers from severe mode collapse. **2.** DMD-LPIPS and DMD-GAN offer limited signal in restoring the diversity lost in DMD and often at a notable cost in quality metrics. On Pick-a-Pic dataset with SD3.5-M, DMD-GAN degrades VQ-R1 from teacher's 4.657 to 4.525 and MIQA from 1.020 to 0.984. **3.** Our EG-DMD achieves superior performance at the frontier of diversity-quality tradeoff. Specifically, it achieves the highest diversity nearly at no cost of quality degradation. Notably, compared with the base model that uses more than 10× inference budget, EG-DMD is able to obtain even higher quality scores at a minor cost in diversity. For example, on COCO2014 prompt dataset with SD3.5-M, EG-DMD achieves a VQ-R1 score of 4.757 that outperforms the teacher model with a score of 4.636, while only mildly reducing CLIP diversity from 0.204 to 0.190. Under the same setting, DP-DMD only achieves VQ-R1 score of 4.689 while suffering a sharp diversity drop to 0.182. **4.** The strong performance of EG-DMD is consistently verified across an array of models, including latent/pixel-space DiT and UNet-based models. The results strongly favor our EG-DMD over DP-DMD in terms of achieving remarkably better tradeoff in diversity and quality.

Table 4: GenEval (Ghosh et al., 2023) results. Higher scores indicates better prompt following.

| Method | Overall | Single Obj. | Two Obj. | Count. | Colors | Pos. | Color Attr. |
|---|---|---|---|---|---|---|---|
| SD3.5-M | 0.657 | 0.98 | 0.78 | 0.64 | 0.83 | 0.18 | 0.53 |
| DP-DMD | 0.650 | 0.99 | 0.81 | **0.60** | 0.81 | 0.21 | 0.48 |
| **EG-DMD** | **0.665** | 0.99 | **0.86** | 0.59 | **0.82** | 0.21 | **0.52** |

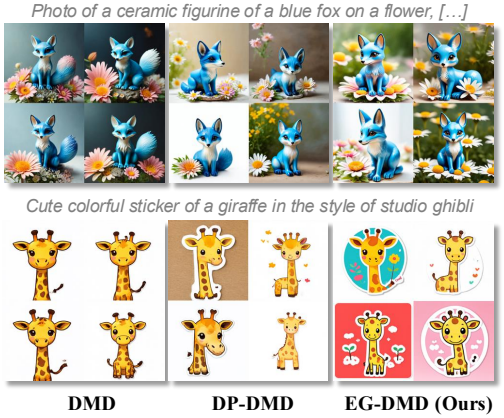



Figure 3: Qualitative comparisons (Pick-a-Pic). More samples are presented in Appendix D.

**System-level comparison.** The practical comparison in Table 3 against open-source few-step models again verifies the superiority of our approach, which convincingly prevails over trajectory-level consistency-based approaches (*e.g.* Hyper-SD) and adversarial distillation (*e.g.* Flash Diffusion).

**Training dynamics.** We plot the quality and diversity evaluation results during training in Fig. 4 *(right)*. It is clearly demonstrated that our EG-DMD pushes the quality enhancement much earlier and stronger than DMD and DP-DMD, while maintaining the diversity consistently at a high level.

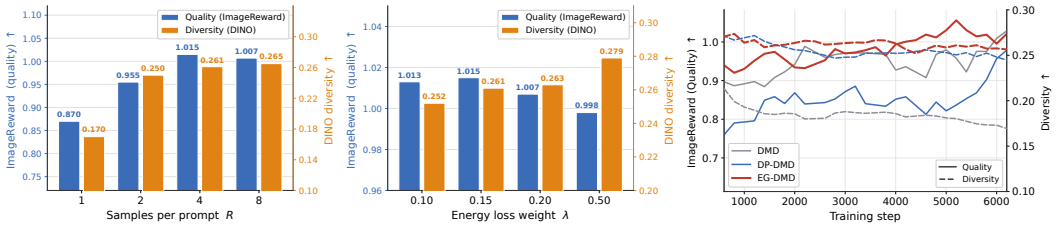



Figure 4: *Left:* Ablation on number of samples per prompt $R$. *Middle:* Ablation on the energy loss weight $\lambda$. *Right:* The training dynamics. The metrics are evaluated on COCO2014-10K.

**Qualitative comparison.** As demonstrated in Fig. 3, EG-DMD leads to much more diverse image samples compared with DMD. DP-DMD only performs trajectory-level compression and exhibits notable artifacts when trading for higher diversity, while our EG-DMD produces more visually high-quality images while preserving the diversity, due to the proposed population-level energy guidance.

### 5.3 Ablation Study

**The effect of samples per prompt $R$.** We investigate the role of the sample size $R$ per prompt in Fig. 4 *(left)*. We observe that trajectory-level distillation with $R = 1$ leads to inferior quality and diversity, while increasing $R$ to 4 peaks the overall performance. Scaling $R$ beyond it exhibits no clear performance gain while leading to higher training overhead due to more rollouts.

**The effect of the energy guidance weight $\lambda$.** We illustrate the effect of the energy loss weight $\lambda$ in Fig. 4 *(middle)*. As the weight increases, the optimization puts more efforts in maintaining the diversity by enforcing more strict alignment with the teacher's modes at intermediate timesteps. Notably, the quality remains fairly robust when $\lambda$ is not over-emphasized.

**The effect of proxy timestep $\tau$.** We ablate the choice of the proxy timestep in Table 5, which shows that a mid-range timestep $\tau = 0.5$ yields the best performance, while a larger $\tau$ exposes insufficient semantics for distribution matching and a smaller $\tau$ brings no benefits yet requiring more teacher steps.

**The entropic regularization $\varepsilon$.** As shown in Table 5, EG-DMD remains effective across a range of $\varepsilon$ values from 0.01 to 0.1. In general, a larger value of $\varepsilon = 1.0$ maintains diversity while yielding a slightly lower ImageReward score due to over-smoothed cost in the OT problem.

Table 5: Ablation study on the energy loss designs on SD3.5-M.

| | ImageReward | | DINO Div | |
|---|---|---|---|---|
| | Pic | COCO | Pic | COCO |
| ref. ($\tau = 0.5, \varepsilon = 0.1$) | 1.156 | 1.015 | 0.191 | 0.261 |
| $\tau = 0.8$ | 1.143 | 0.940 | 0.187 | 0.257 |
| $\tau = 0.2$ | 1.155 | 1.011 | 0.188 | 0.264 |
| $\varepsilon = 0.01$ | 1.146 | 0.956 | 0.192 | 0.263 |
| $\varepsilon = 1.00$ | 1.142 | 1.002 | 0.184 | 0.252 |
| w/o detach | 1.150 | 0.994 | 0.173 | 0.250 |
| share latent | 1.140 | 0.978 | 0.147 | 0.219 |
| w/o feat. enc. | 1.138 | 0.955 | 0.129 | 0.208 |



**Ref.** **w/o detach** **w/o feat. enc.**

Figure 5: Ablation visualizations.

**Other designs in the energy loss.** We observe in Table 5 that without detaching the student's first step for the DMD gradient, both the image quality and diversity degrade, since the two losses may yield competing optimization directions for the first student step and enlarge variance. Fig. 5 demonstrates the effect where the non-detached variant exhibits conflicted semantics. In addition, sharing the same initial latent when deriving the energy loss degrades diversity remarkably.

**Feature map.** It is observed in Table 5 that the feature encoding is essential for the energy loss to produce informative gradient direction. The distribution-level energy guidance becomes noisy and degrades the performance remarkably in the raw latent space with a very limited sample size $R = 4$. We highlight that we inherit an easy and generalizable feature encoding design by using the teacher model's mid-block, which does not rely on any external feature encoders.

## 6 Conclusion

We present EG-DMD for few-step visual generation that lies in the frontier of sample quality and diversity. Our core idea is to enforce an auxiliary energy guidance loss on top of distribution matching distillation, where the distribution of the first-step student prediction is measured against the teacher networks' partial rollout at a certain intermediate diffusion step. Such a design helps preserve the teacher distribution's diverse mode while maintaining high sample quality. We hope EG-DMD can establish new opportunities for high-fidelity, diverse, and multi-modal few-step visual generation.

### AI USE STATEMENT

In this work, we used generative AI tools to support qualitative and thematic data analysis. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, and design or provide feedback on research methodology or experiments. Additionally, we used generative AI tools to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

We propose an algorithm to enhance the diversity-quality frontier of few-step generators. Our work aims to contribute to machine learning methods and we do not identify particular ethical concerns that are worth emphasizing.

### REPRODUCIBILITY STATEMENT

We commit to ensure reproducibility of this work. We have included the algorithm in Alg. 1, the experiment setup in Sec. 5.1, and the detailed hyperparameter selection across all models in Appendix B. We also provide the code in the supplementary material, and the detailed theoretical proof in Appendix A.

## REFERENCES

Nicholas Matthew Boffi, Michael Samuel Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

Clement Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 3, 8

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026. 3

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow Transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024. 1, 7, 8

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 2681–2690. PMLR, 2019. 16

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International conference on artificial intelligence and statistics*, pp. 1608–1617. PMLR, 2018. 3

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025. 2

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026. 2

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, pp. 52132–52152, 2023. 7, 8

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012. 3, 19

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J Candès. One-step generative modeling via wasserstein gradient flows. *arXiv preprint arXiv:2605.11755*, 2026a. 2, 3, 5

Jiaqi Han, Juntong Shi, Puheng Li, Haotian Ye, Qiushan Guo, and Stefano Ermon. Adaptive spectral feature forecasting for diffusion sampling acceleration. *arXiv preprint arXiv:2603.01623*, 2026b. 1

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026. 2, 3

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020. 1, 3

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 1

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

Dongjun Kim, Chieh-Hsin Lai, WeiHsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *International Conference on Learning Representations*, volume 2024, pp. 44493–44525, 2024. 2

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. In *Advances in Neural Information Processing Systems*, pp. 36652–36663, 2023. 7

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017. 3

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International conference on machine learning*, pp. 1718–1727. PMLR, 2015. 3

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, 2014. 7

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1, 3

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. 3

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025. 2

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 *Advances in Neural Information Processing Systems*, 2022. 3

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. *arXiv preprint arXiv:2503.06674*, 2025. 8

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306. IEEE, 2023. 2

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 7

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-SD: Trajectory segmented consistency model for efficient image synthesis. In *Advances in Neural Information Processing Systems*, pp. 117340–117362, 2024. 8

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103, 2024. 1

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. 7

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 5

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015. 1

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. 3

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024. 2

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*, 2021b. 1, 3

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023. 2

Thomas Unterthiner, Bernhard Nessler, Calvin Seward, Günter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Coulomb gans: Provably optimal nash equilibria via potential fields. *arXiv preprint arXiv:1708.08819*, 2017. 3

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. WAN: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

Fu-Yun Wang, Zhaoyang Huang, Alexander Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. In *Advances in Neural Information Processing Systems*, pp. 83951–84009, 2024. 2

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. Minit2i: A minimalist baseline for text-to-image generation, 2026. URL https://peppaking8.github.io/#/post/minit2i. 7

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*, pp. 8406–8441, 2023a. 2, 3

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Association for Computational Linguistics*, pp. 893–911, 2023b. 7

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. In *Advances in Neural Information Processing Systems*, pp. 88167–88190, 2025. 7

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. In *Forty-third International Conference on Machine Learning*, 2026. 1, 2, 3, 4, 7, 8, 17

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. In *Advances in Neural Information Processing Systems*, pp. 15903–15935, 2023. 7

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022. 7

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, pp. 47455–47487, 2024a. 1, 2, 3, 4, 8

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b. 1, 2, 3, 8

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025. 1

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. 8

Yuan Zhang, Chenyi Li, Haodong Yu, Guoqing Ma, Jiajun Zha, Yuanming Yang, Bo Wang, Wei Tang, Wenbo Li, Haoyang Huang, et al. Teacher-feature drifting: One-step diffusion distillation with pretrained diffusion representations. *arXiv preprint arXiv:2605.07327*, 2026. 3, 6

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. *arXiv preprint arXiv:2510.08431*, 2025. 1, 2

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. In *Forty-second International Conference on Machine Learning*, 2025. 3

# A Theoretical Analysis and Proof

## A.1 Fixed-coupling upper bound

**Proposition A.1** (Paired supervision as a fixed-coupling upper bound). *Let $\mu = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and let $F, T : \mathbb{R}^d \to \mathbb{R}^m$ be measurable maps with finite second moments. Define*

$$q = F_{\#}\mu, \qquad p = T_{\#}\mu.$$

*Then*

$$W_2^2(q,p) \le \mathbb{E}_{\mathbf{z}\sim\mu}\left[\|F(\mathbf{z}) - T(\mathbf{z})\|_2^2\right]. \tag{20}$$

*Equality holds only when the coupling induced by the shared latent,*

$$\pi_{\text{shared}} = (F,T)_{\#}\mu,$$

*is itself an optimal transport coupling between $q$ and $p$.*

*Moreover, consider $R$ paired empirical samples $\mathbf{a}_i = F(\mathbf{z}_i)$ and $\mathbf{b}_i = T(\mathbf{z}_i)$ with uniform weights, and define $\hat{q}_R = R^{-1}\sum_i \delta_{\mathbf{a}_i}$ and $\hat{p}_R = R^{-1}\sum_i \delta_{\mathbf{b}_i}$. For the Sinkhorn divergence in Eq. 10,*

$$S_\varepsilon(\hat{q}_R, \hat{p}_R) \le \frac{1}{2R}\sum_{i=1}^{R}\|\mathbf{a}_i - \mathbf{b}_i\|_2^2 + \varepsilon \log R. \tag{21}$$

*Thus, paired regression optimizes a restricted, fixed-coupling surrogate, whereas Sinkhorn matching relaxes the prescribed correspondence and selects a transport plan according to the geometry of the two empirical populations.*

*Proof.* By construction, $\pi_{\text{shared}} = (F,T)_{\#}\mu$ has marginals $q$ and $p$ and therefore belongs to $\Pi(q,p)$. Since the squared 2-Wasserstein distance minimizes the transportation cost over *all* admissible couplings,

$$W_2^2(q,p) = \inf_{\pi\in\Pi(q,p)} \int \|\mathbf{x}-\mathbf{y}\|_2^2 \,\mathrm{d}\pi(\mathbf{x},\mathbf{y}) \tag{22}$$

$$\le \int \|\mathbf{x}-\mathbf{y}\|_2^2 \,\mathrm{d}\pi_{\text{shared}}(\mathbf{x},\mathbf{y}) \tag{23}$$

$$= \mathbb{E}_{\mathbf{z}\sim\mu}\left[\|F(\mathbf{z}) - T(\mathbf{z})\|_2^2\right]. \tag{24}$$

The inequality becomes equality precisely when $\pi_{\text{shared}}$ is an optimal coupling.

For the empirical case, consider the diagonal transport plan $P^{\text{diag}}$ assigning mass $1/R$ from $\mathbf{a}_i$ to $\mathbf{b}_i$ and zero elsewhere. It is a feasible coupling between the two uniform empirical measures. Under the quadratic cost in Eq. 9, its transportation cost is

$$\frac{1}{2R}\sum_{i=1}^{R}\|\mathbf{a}_i - \mathbf{b}_i\|_2^2.$$

Relative to the independent uniform coupling, whose mass is $1/R^2$ for each pair of sample indices,

$$D_{\text{KL}}\left(P^{\text{diag}} \,\|\, \hat{q}_R \otimes \hat{p}_R\right) = \sum_{i=1}^{R}\frac{1}{R}\log\frac{1/R}{1/R^2} = \log R.$$

Since entropic optimal transport minimizes over all feasible couplings,

$$\text{OT}_\varepsilon(\hat{q}_R, \hat{p}_R) \le \frac{1}{2R}\sum_{i=1}^{R}\|\mathbf{a}_i - \mathbf{b}_i\|_2^2 + \varepsilon\log R. \tag{25}$$

Finally, $\text{OT}_\varepsilon(\hat{q}_R, \hat{q}_R) \ge 0$ and $\text{OT}_\varepsilon(\hat{p}_R, \hat{p}_R) \ge 0$, so Eq. 10 gives

$$S_\varepsilon(\hat{q}_R, \hat{p}_R) \le \text{OT}_\varepsilon(\hat{q}_R, \hat{p}_R),$$

which establishes Eq. 21. □

### A.2 VARIANCE DEFICIT AND MATCHED-MARGINAL CONTRACTION

**Proposition A.2.** *Let $X = F(\mathbf{z})$ and $Y = T(\mathbf{z})$ have finite second moments. Suppose the paired loss is stationary under local affine output recalibration $F \mapsto AF + b$ at $(A, b) = (I, 0)$. Then*

$$\mathbb{E}X = \mathbb{E}Y, \qquad \mathrm{Cov}(Y) - \mathrm{Cov}(X) = \mathbb{E}[(Y - X)(Y - X)^\top] \succeq 0. \tag{26}$$

*Consequently, with $V(q) = \mathrm{tr}\,\mathrm{Cov}_q$,*

$$V(T_\# \mu) - V(F_\# \mu) = \mathcal{L}_{\mathrm{pair}}(F). \tag{27}$$

Proposition A.2 identifies a shrinkage mechanism: at an affine-calibrated paired solution, residual trajectory-fitting error is exactly the total variance missing from the student proxy. Its stationarity assumption holds, for example, at a paired-loss optimum in an affine-output-closed student class, but is not guaranteed at every checkpoint of joint DMD training (proof in Appendix A.2).

*Proof of Proposition A.2.* Set $E = Y - X$ and consider $\ell(A, b) = \mathbb{E}\|AX + b - Y\|^2$. Differentiating with respect to $b$ and $A$ at $(I, 0)$ gives

$$\mathbb{E}E = 0, \qquad \mathbb{E}[EX^\top] = 0.$$

Hence $X$ and $Y$ have the same mean and $\mathrm{Cov}(X, E) = 0$. Expanding the covariance of $Y = X + E$ gives

$$\mathrm{Cov}(Y) = \mathrm{Cov}(X) + \mathbb{E}[EE^\top],$$

which proves the matrix identity and its positive-semidefinite sign. Taking traces yields Eq. 27. □

*Proof of Proposition 4.1.* Set $X_c = X - m$, $Y_c = Y - m$, and $v = \mathbb{E}\|X_c\|_2^2 = \mathbb{E}\|Y_c\|_2^2$. Equality of the marginals gives

$$\Delta = 2v - 2\mathbb{E}\langle X_c, Y_c\rangle.$$

Consequently, the paired loss along the centered scaling satisfies the exact identity

$$\begin{aligned} \mathcal{L}_{\mathrm{pair}}(F_s) &= s^2 v + v - 2s\mathbb{E}\langle X_c, Y_c\rangle \\ &= \Delta + \Delta(s - 1) + v(s - 1)^2. \end{aligned} \tag{28}$$

Differentiating at $s = 1$ proves the first identity in Eq. 19. Since $\Delta > 0$, the common variance $v$ is positive. Whenever $F_{1-\eta}$ is feasible and $0 < \eta < \min\{1, \Delta/v\}$,

$$\mathcal{L}_{\mathrm{pair}}(F_{1-\eta}) - \mathcal{L}_{\mathrm{pair}}(F) = -\eta\Delta + \eta^2 v < 0.$$

Furthermore, $F_s(\mathbf{z}) - m = sX_c$, so

$$\mathrm{Cov}(F_s(\mathbf{z})) = s^2\,\mathrm{Cov}(X) = s^2\,\mathrm{Cov}(Y),$$

and its trace is strictly smaller for $0 < s < 1$. The assumptions are nonvacuous: for $z \sim \mathcal{N}(0, 1)$, the maps $F(z) = z$ and $T(z) = -z$ have identical marginals but $\Delta = 4$. No stationarity assumption is used in this proof. □

### A.3 FINITE-BATCH QUALITY–DIVERSITY SEPARATION

**Theorem A.3.** *Let $U, V \in \{-1, 1\}$ be balanced deterministic functions of a common latent, with $\mathbb{E}[UV] = \rho \in [0, 1)$. For $a > 0$, consider the teacher $Y = aV$ and the same restricted student family $X_r = rU$, $r \geq 0$, for both objectives. Let $\Psi_{R,\varepsilon}(r)$ be the expected Sinkhorn divergence between independent size-$R$ empirical samples from their marginals, using the fixed cost $\frac{1}{2}|x - y|^2$. Write $S_0 = \frac{1}{2}W_2^2$. Then*

$$r_{\mathrm{pair}} = \rho a, \qquad r_{R,0} = \kappa_R a, \qquad \kappa_R = 1 - 2\binom{2R}{R}4^{-R}. \tag{29}$$

*If $\rho < \kappa_R$, the unpaired minimizer has both smaller off-mode distortion $\mathcal{Q}(r) = (a - r)^2$ and larger cosine diversity $\mathcal{D}_b(r) = r^2/(b^2 + r^2)$, for every $b > 0$. Moreover, every global minimizer $r_{R,\varepsilon}$ satisfies*

$$|r_{R,\varepsilon} - \kappa_R a| \leq 2\sqrt{\varepsilon \log 2}. \tag{30}$$

*The strict improvement in both criteria therefore persists whenever $\rho + 2\sqrt{\varepsilon \log 2}/a < \kappa_R$.*

Theorem A.3 isolates a setting in which the student can match the teacher marginal at $r = a$, yet paired regression shrinks its modes; distribution matching improves both criteria when the minibatch and entropy biases are smaller than the pairing-induced shrinkage. Here $\mathcal{D}_b$ is cosine diversity under $e_b(x) = (x, b)/\sqrt{x^2 + b^2}$, so the result concerns explicit surrogate criteria rather than directly certifying image-level metrics.

Before formally proving Theorem A.3, we first introduce the following Lemma A.4.

For an atomic probability measure $q$, write $H(q) = -\sum_{x\in\mathrm{supp}(q)} q(\{x\}) \log q(\{x\})$, aggregating repeated atoms and using $0 \log 0 = 0$; as in the theorem, $S_0 = \frac{1}{2}W_2^2$.

**Lemma A.4** (Two-sided approximation on finite supports)**.** *For finitely supported $q, p$ and $\varepsilon > 0$,*

$$S_0(q,p) - \frac{\varepsilon}{2}\{H(q) + H(p)\} \le S_\varepsilon(q,p) \le S_0(q,p) + \varepsilon \min\{H(q), H(p)\}. \tag{31}$$

*Consequently, for uniform empirical measures of size $R$,*

$$\left|S_\varepsilon(\hat{q}_R, \hat{p}_R) - \tfrac{1}{2}W_2^2(\hat{q}_R, \hat{p}_R)\right| \le \varepsilon \log R. \tag{32}$$

*If each measure has at most two distinct atoms, $\log R$ can be replaced by $\log 2$, regardless of the number of repeated samples.*

*Proof.* For any finite-support coupling, its relative entropy against $q \otimes p$ is the mutual information and lies between zero and $\min\{H(q), H(p)\}$. Evaluating the regularized objective at an unregularized optimal plan gives

$$\mathrm{OT}_0(q,p) \le \mathrm{OT}_\varepsilon(q,p) \le \mathrm{OT}_0(q,p) + \varepsilon \min\{H(q), H(p)\}.$$

The diagonal self-coupling has zero transport cost and relative entropy $H(q)$, so $0 \le \mathrm{OT}_\varepsilon(q,q) \le \varepsilon H(q)$, and similarly for $p$. Subtract the self-costs to obtain Eq. 31. Since $\min(H(q), H(p)) \le \frac{1}{2}(H(q) + H(p))$, the absolute error is bounded by $\frac{\varepsilon}{2}(H(q) + H(p))$. Finally, $H(\hat{q}_R), H(\hat{p}_R) \le \log R$, or at most $\log 2$ on two-atom supports. □

Now we are ready to prove Theorem A.3.

*Proof of Theorem A.3.* The paired loss is

$$\mathcal{L}_{\mathrm{pair}}(r) = \mathbb{E}(rU - aV)^2 = r^2 + a^2 - 2\rho ar = (r - \rho a)^2 + a^2(1 - \rho^2), \tag{33}$$

so its unique minimizer is $r_{\mathrm{pair}} = \rho a$. At the population level, $\frac{1}{2}W_2^2(q_r, p) = \frac{1}{2}(r - a)^2$, since matching the signs is optimal. For every finite positive $\varepsilon$, the positivity and definiteness of the quadratic Sinkhorn divergence on compact supports (Feydy et al., 2019) likewise imply that $r = a$ is the unique population minimizer.

Now take independent size-$R$ batches. Let $K, L$ be the respective numbers of positive student and teacher samples. Then $K, L$ are independent $\mathrm{Bin}(R, 1/2)$ variables. The optimal plan matches signs except for mass $|K - L|/R$. Replacing a same-sign match by an opposite-sign match increases the half-quadratic cost by $2ar$. Therefore

$$S_0(\hat{q}_{r,R}, \hat{p}_R) = \tfrac{1}{2}(r - a)^2 + 2ar\frac{|K - L|}{R}. \tag{34}$$

Taking expectations gives

$$\Psi_{R,0}(r) = \tfrac{1}{2}(r^2 + a^2) - \kappa_R ar = \tfrac{1}{2}(r - \kappa_R a)^2 + \tfrac{a^2}{2}(1 - \kappa_R^2), \tag{35}$$

where $\kappa_R = 1 - 2\mathbb{E}|K - L|/R$. To evaluate this coefficient, note that $K - L$ has the same law as $J - R$ for $J \sim \mathrm{Bin}(2R, 1/2)$. Symmetry and

$$(j - R)\binom{2R}{j} = R\left\{\binom{2R-1}{j-1} - \binom{2R-1}{j}\right\}$$

yield the telescoping sum

$$\mathbb{E}|J - R| = 2\,4^{-R} \sum_{j=R+1}^{2R} (j - R)\binom{2R}{j} = R\binom{2R}{R}4^{-R}.$$

Table 6: Detailed hyperparameters used in the main experiments.

| | SD3.5-Medium | SDXL | MiniT2I-B/16 |
|---|---|---|---|
| *Architecture* | | | |
| backbone | MMDiT (latent) | UNet (latent) | MMJiT (pixel) |
| resolution | 1024×1024 | 1024×1024 | 512×512 |
| parameterization | velocity $\mathbf{v}$ | noise $\boldsymbol{\epsilon}$ | clean data $\mathbf{x}_0$ |
| *Training* | | | |
| student lr | 1e-5 | 5e-7 | 5e-5 |
| fake score lr | 1e-5 | 5e-7 | 1e-5 |
| teacher CFG | 3.5 | 8.0 | 5.0 |
| EG weight $\lambda$ | 0.15 | 0.005 | 0.02 |
| Sinkhorn $\varepsilon$ | 0.1 | 0.01 | 0.1 |
| *Feature Space* | | | |
| layer | Transformer block 12 (of 24) | UNet "mid_block" | double-stream block 8 (of 17) |
| channel dim | 1536 | 1280 | 768 |
| token grid | 64×64 | 32×32 | 32×32 |
| pool size | 8 | 0 (global) | 4 |

This proves Eq. 29.

For $X_r \in \{-r, r\}$, squared distance to the nearest valid teacher mode is $(a-r)^2$ for every $r \geq 0$. For the normalized embedding

$$e_b(x) = \frac{(x, b)}{\sqrt{x^2 + b^2}}, \qquad b > 0,$$

independent samples satisfy

$$\mathbb{E}[1 - \langle e_b(X_r), e_b(X_r')\rangle] = 1 - \|\mathbb{E}e_b(X_r)\|^2 = \frac{r^2}{b^2 + r^2}.$$

When $\rho < \kappa_R \leq 1$, $r_{\text{pair}} < r_{R,0} \leq a$, which strictly improves both criteria.

For positive $\varepsilon$, Lemma A.4 applies to the two distinct atoms in each batch and gives the uniform-in-$r$ bound

$$|\Psi_{R,\varepsilon}(r) - \Psi_{R,0}(r)| \leq \varepsilon \log 2.$$

The objective is continuous and coercive by this bound, so a global minimizer exists. If $r_{R,\varepsilon}$ is any such minimizer, comparison with $r_{R,0}$ gives

$$\tfrac{1}{2}|r_{R,\varepsilon} - \kappa_R a|^2 = \Psi_{R,0}(r_{R,\varepsilon}) - \Psi_{R,0}(\kappa_R a) \leq 2\varepsilon \log 2.$$

This proves Eq. 30. Write $d = 2\sqrt{\varepsilon \log 2}$. If $\rho + d/a < \kappa_R$, then

$$r_{R,\varepsilon} \geq \kappa_R a - d > \rho a, \qquad |a - r_{R,\varepsilon}| \leq a(1-\kappa_R) + d < a(1-\rho).$$

These two strict inequalities prove the positive-$\varepsilon$ diversity and quality claims, even if a perturbed minimizer is slightly larger than $a$. □

## B Additional Experiment Details

**Detailed hyperparameters.** We first specify the shared hyperparameters across all diffusion models while summarizing the detailed model-specific hyperparameters in Table 6. We use 4 student steps with $K_{\max} = 30$, following Wu et al. (2026). We use $\tau = 0.5$, $R = 4$, and 4 prompts per GPU per student training step. We keep generator and fake score update step at a frequency of 1:5. We train the model on 8 NVIDIA H200 GPUs with FSDP and bf16 precision. Gradient checkpointing is enabled. For Sinkhorn energy guidance, we employ a maximum of 30 Sinkhorn iteration steps with an early stop of tolerance 1e-3. Note that computing Sinkhorn iteration is cheap since it only requires tensor operations on the cost matrix with shape $B \times R \times R$ which is empirically very small. The cost matrix uses the median normalization to disentangle the scale of $\varepsilon$ from the scale of the feature. When computing the feature from the teacher middle block, we employ the same diffusion timestep at $\tau = 0.5$ to ensure consistency with the noise-level of the input. For the system-level comparison

**Algorithm 2** Empirical Sinkhorn Divergence Estimator

**Input:** Student features $\{\mathbf{a}_i\}_{i=1}^R$, teacher features $\{\mathbf{b}_j\}_{j=1}^R$, entropic regularization $\varepsilon$, number of Sinkhorn iterations $L$.

**function** SINKHORNOT($\{\mathbf{x}_i\}_{i=1}^R, \{\mathbf{y}_j\}_{j=1}^R, \varepsilon, L$)

Set uniform marginals $\mathbf{r} = \mathbf{s} = \frac{1}{R}\mathbf{1}_R$.

Construct the pairwise quadratic cost matrix

$$C_{ij} \leftarrow \frac{1}{2}\|\mathbf{x}_i - \mathbf{y}_j\|_2^2, \qquad 1 \le i, j \le R.$$

Construct the Gibbs kernel

$$K_{ij} \leftarrow \exp\left(-\frac{C_{ij}}{\varepsilon}\right).$$

Initialize $\mathbf{u} \leftarrow \mathbf{1}_R$ and $\mathbf{v} \leftarrow \mathbf{1}_R$.

**for** $\ell = 1, \dots, L$ **do**

$\mathbf{u} \leftarrow \mathbf{r} \oslash (K\mathbf{v})$.

$\mathbf{v} \leftarrow \mathbf{s} \oslash (K^\top \mathbf{u})$.

Recover the transport plan

$$P \leftarrow \mathrm{diag}(\mathbf{u}) K \mathrm{diag}(\mathbf{v}).$$

Compute the entropic OT objective

$$\widehat{\mathrm{OT}}_\varepsilon \leftarrow \sum_{i,j} P_{ij} C_{ij} + \varepsilon \sum_{i,j} P_{ij} \log \frac{P_{ij}}{r_i s_j}.$$

**return** $\widehat{\mathrm{OT}}_\varepsilon$.

Compute the cross-distribution transport cost

$$\widehat{\mathrm{OT}}_{qp} \leftarrow \text{SINKHORNOT}\left(\{\mathbf{a}_i\}_{i=1}^R, \{\mathbf{b}_j\}_{j=1}^R, \varepsilon, L\right).$$

Compute the student self-transport cost

$$\widehat{\mathrm{OT}}_{qq} \leftarrow \text{SINKHORNOT}\left(\{\mathbf{a}_i\}_{i=1}^R, \{\mathbf{a}_i\}_{i=1}^R, \varepsilon, L\right).$$

Compute the teacher self-transport cost

$$\widehat{\mathrm{OT}}_{pp} \leftarrow \text{SINKHORNOT}\left(\{\mathbf{b}_j\}_{j=1}^R, \{\mathbf{b}_j\}_{j=1}^R, \varepsilon, L\right).$$

Compute the empirical Sinkhorn divergence

$$\widehat{\mathcal{L}}_{\mathrm{EG}} \leftarrow \widehat{\mathrm{OT}}_{qp} - \frac{1}{2}\widehat{\mathrm{OT}}_{qq} - \frac{1}{2}\widehat{\mathrm{OT}}_{pp}.$$

**return** $\widehat{\mathcal{L}}_{\mathrm{EG}}$.

on Stable Diffusion 3 Medium in Table 3, we reuse exactly the same choice of hyperparameters as the Stable Diffusion 3.5 Medium experiments.

**DP-DMD derivation.** For the linear flow parameterization used by DP-DMD, we derive an equivalent form from their velocity regression objective. In particular, if

$$\mathbf{v}^{\mathrm{target}}(\mathbf{z}) = \frac{\mathbf{z} - T_\phi^\tau(\mathbf{z})}{1 - \tau} \tag{36}$$

and the one-step student state is

$$F_\theta^\tau(\mathbf{z}) = \mathbf{z} - (1 - \tau)\mathbf{v}_\theta(\mathbf{z}, 1), \tag{37}$$

then

$$\left\|F_\theta^\tau(\mathbf{z}) - T_\phi^\tau(\mathbf{z})\right\|_2^2 = (1 - \tau)^2 \left\|\mathbf{v}_\theta(\mathbf{z}, 1) - \mathbf{v}^{\mathrm{target}}(\mathbf{z})\right\|_2^2. \tag{38}$$

Hence, up to the constant $(1 - \tau)^2$, DP-DMD's diversity objective is a same-noise paired state-matching loss specified in Eq. 18.

## C Additional Experiment Results

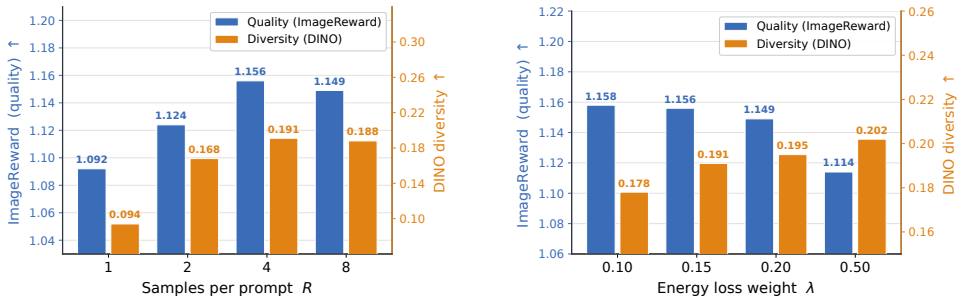



Figure 6: Ablation study on samples per prompt $R$ and the energy loss weight $\lambda$ evaluated on Pick-a-Pic dataset.

Table 7: **Controlled comparison of DMD variants with pixel-space diffusion model MiniT2I-B/16**. All methods are evaluated with the same backbone and inference budget. The best and second-best results are highlighted in **bold** and underline, respectively. This table is a complete version of Table 2 which additionally includes the evaluation results on COCO2014-10K.

| Method | NFEs | Pick-a-Pic | | | | | | COCO2014-10K | | | | | | GenEval↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Diversity | | Quality | | Preference | | Diversity | | Quality | | Preference | | |
| | | DINO↑ | CLIP↑ | VQ-R1↑ | MIQA↑ | ImgR↑ | PicS↑ | DINO↑ | CLIP↑ | VQ-R1↑ | MIQA↑ | ImgR↑ | PicS↑ | |
| *MiniT2I-B/16* (CFG=5.0) | | | | | | | | | | | | | | |
| Base Model | 100 | 0.136 | 0.146 | 4.305 | 1.052 | 1.112 | 21.08 | 0.125 | 0.112 | 4.566 | 1.124 | 1.228 | 22.54 | 0.872 |
| DMD | 4 | 0.106 | 0.139 | **4.376** | **1.014** | 0.979 | 20.74 | 0.104 | 0.096 | **4.549** | **1.076** | 1.048 | 22.23 | 0.824 |
| DP-DMD | 4 | 0.113 | 0.136 | 4.303 | 0.992 | 0.965 | 20.75 | 0.107 | 0.098 | 4.503 | 1.059 | 1.041 | 22.20 | 0.827 |
| **EG-DMD** | 4 | **0.118** | **0.143** | 4.324 | 0.997 | **1.002** | **20.83** | **0.114** | **0.105** | 4.524 | 1.064 | **1.072** | **22.26** | **0.835** |

**More ablation results.** We provide more ablation results that inspect the role of samples per prompt $R$ and the energy loss weight $\lambda$ evaluated on Pick-a-Pic dataset in Fig. 6. The results remain consistent with our findings in Fig. 4, supporting the efficacy of the proposed energy guidance loss.

Table 8: The choice of the energy function.

| | ImageReward | | DINO Div | |
|---|---|---|---|---|
| | Pic | COCO | Pic | COCO |
| *SD3.5-M* | | | | |
| Sinkhorn | 1.156 | 1.015 | 0.191 | 0.261 |
| MMD | 1.132 | 0.995 | 0.177 | 0.239 |
| *SDXL* | | | | |
| Sinkhorn | 1.101 | 0.849 | 0.196 | 0.233 |
| MMD | 1.018 | 0.825 | 0.182 | 0.212 |

**The choice of energy guidance function.** We further compare the Sinkhorn divergence with Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) as an alternative population-level energy function. As shown in Table 8, MMD also alleviates the diversity collapse of vanilla DMD, indicating that the benefit of population-level energy guidance is not specific to Sinkhorn. However, across both SD3.5-M and SDXL, MMD consistently yields lower ImageReward and DINO diversity than Sinkhorn. These results support our choice of Sinkhorn divergence, whose optimized transport coupling provides more effective distributional supervision.

**Inference speed comparison.** We provide a detailed inference speed benchmark on three diffusion model architectures in Table 9. We analyze both the denoise-only speedup and the full generation pipeline speedup, where the latter includes the time cost in the VAE decoding (if applicable) and the final image output. The student model distilled by EG-DMD achieves remarkable inference speedup compared with the multi-step teacher model.

Table 9: Inference speed comparison between the multi-step teacher model and our EG-DMD distilled student model.

| | teacher step | student step | Denoise-only | | | Full Generation | | |
|---|---|---|---|---|---|---|---|---|
| | | | teacher (sec/img) | student (sec/img) | speedup | teacher (sec/img) | student (sec/img) | speedup |
| SD3.5-M | 60 (w/ CFG) | 4 | 6.084 | 0.215 | 28.2× | 6.245 | 0.350 | 17.8× |
| SDXL | 100 (w/ CFG) | 4 | 5.771 | 0.205 | 28.1× | 5.901 | 0.321 | 18.4× |
| MiniT2I-B/16 | 100 (w/ CFG) | 4 | 3.063 | 0.121 | 25.3× | 3.098 | 0.135 | 23.0× |

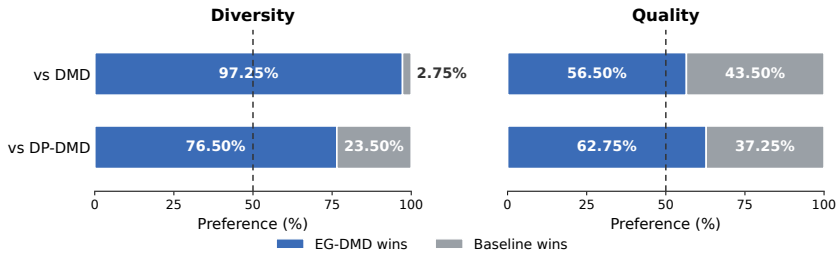



Figure 7: User study with the vote on EG-DMD vs DMD and EG-DMD vs DP-DMD.

**Training complexity analysis.** Denote the teacher rollout time cost as $T^{(t)}$ (in seconds per step), the student rollout cost as $T^{(s)}$, and the student gradient update step as $T_g^{(s)}$, which includes the feature encoding and backpropagation. Further denote the fake score update step cost as $T^{(f)}$ and $K$ fake steps are updated per one generator step. $K$ is kept as 5 throughout the paper, regardless of the choice of $R$. The naive implementation scales the teacher and student step by $R$, which gives the training complexity of $T(R) = T^{(g)} + K \cdot T^{(f)}$ where the generator step cost $T^{(g)} = R \cdot (T^{(t)} + T^{(s)} + T_g^{(s)})$. In practice, we profiled each quantity during training and we have $T^{(t)} = 5.6$, $T^{(s)} + T_g^{(s)} = 5.0$, and $T^{(f)} = 2.1$. Under our experiment setting, this gives the scaling of $T(R = 4)/T(R = 1) \approx 2.5\times$, which is much smaller than the $R/1 = 4$ scaling. Empirically, we can further enhance the efficiency by only updating one student rollout while using the rest of $R - 1$ rollouts with gradient detached, and compensate the energy loss scale by $R$ for unbiased gradient estimate. This removes the $R$-scaling in the student gradient step and further pushes the $R$-scaling on $T$ to $1.9\times$. Note that for the ablation study, we keep the same setting among all variants to ensure that comparison is performed under the same budget. The ablation study in Fig. 4 adopts the same setting for DMD and DP-DMD that scales the number of rollouts to $R = 4$ per prompt. Compared with the $R = 1$ case and the baseline DMD, the primary additional cost comes from multiple teacher rollouts performed for each prompt, which is essential to explore different modes in the teacher distribution. The significantly improved metrics in quality and diversity verify the necessity of this proposal while remaining to be an image-free approach.

**User study.** We further conduct a controlled user study on SD3.5-M using 40 randomly selected prompts with 20 participants experienced in evaluating text-to-image models. For each prompt, outputs from two methods are presented side by side with randomized ordering under matched prompts and random seed. Participants compare the methods in terms of diversity and image quality. 10 participants are responsible for the vote between EG-DMD and DMD and the other 10 perform the comparison between EG-DMD and DP-DMD. The voting results are then summarized. We report pairwise preference rates averaged over all prompts and participants with results displayed in Fig. 7. The user study consistently supports the superiority of EG-DMD in terms of achieving much higher diversity while preserving exceptional image quality.

## D Additional Visualizations

We present additional visualizations of the 4-step samples produced by DMD, DP-DMD, and our EG-DMD in Fig. 8, 9, and 10, for SD3.5-M and SDXL models. Our EG-DMD clearly exhibits higher sample diversity and exceptional visual quality. Samples are produced from the same set of initial latents across the three methods for head-to-head comparison.

## E Detailed Prompts

**Prompts used in Fig. 1.** We use the following prompts for the samples in Fig. 1 from top left to bottom right:

- *a white ceramic cup of latte with intricate latte art spelling "EGDMD" in the foam, top-down cafe shot, warm tones*

- *a vintage cafe racer motorcycle parked on a coastal road at golden hour, polished chrome, cinematic*
- *a golden retriever puppy running through a sunlit meadow of daisies, gentle motion blur, joyful, bokeh*
- *an astronaut standing on an alien desert planet with two moons, vibrant nebula sky, sci-fi concept art*
- *a glass of iced coffee with swirling cream on a sunny windowsill, condensation droplets, food photography*
- *a glowing neon sign on a rainy Tokyo alley at night that reads "EGDMD", vivid reflections on wet pavement, cinematic bokeh*
- *a tall stack of fluffy pancakes with maple syrup and fresh berries, bright morning light, food photography*
- *a crystal-clear alpine lake mirroring snow-capped peaks and pine forest, vivid blue sky, landscape photography*
- *a red fox curled up asleep in autumn leaves, soft morning light, golden bokeh, nature photography*
- *a steaming bowl of tonkotsu ramen with a soft-boiled egg and scallions, close-up food photography, shallow depth of field*
- *a fantasy castle floating among clouds with waterfalls cascading off the edges, epic volumetric light*
- *a humpback whale breaching the ocean surface at sunset, spray and golden backlight, dramatic wildlife photo*

**Prompts used in Fig. 5.** The prompts used in Fig. 5 are from COCO2014-10k:

- *A plethora of stop signs sitting in a storage area.*
- *a person riding a horse with trees in the back ground*

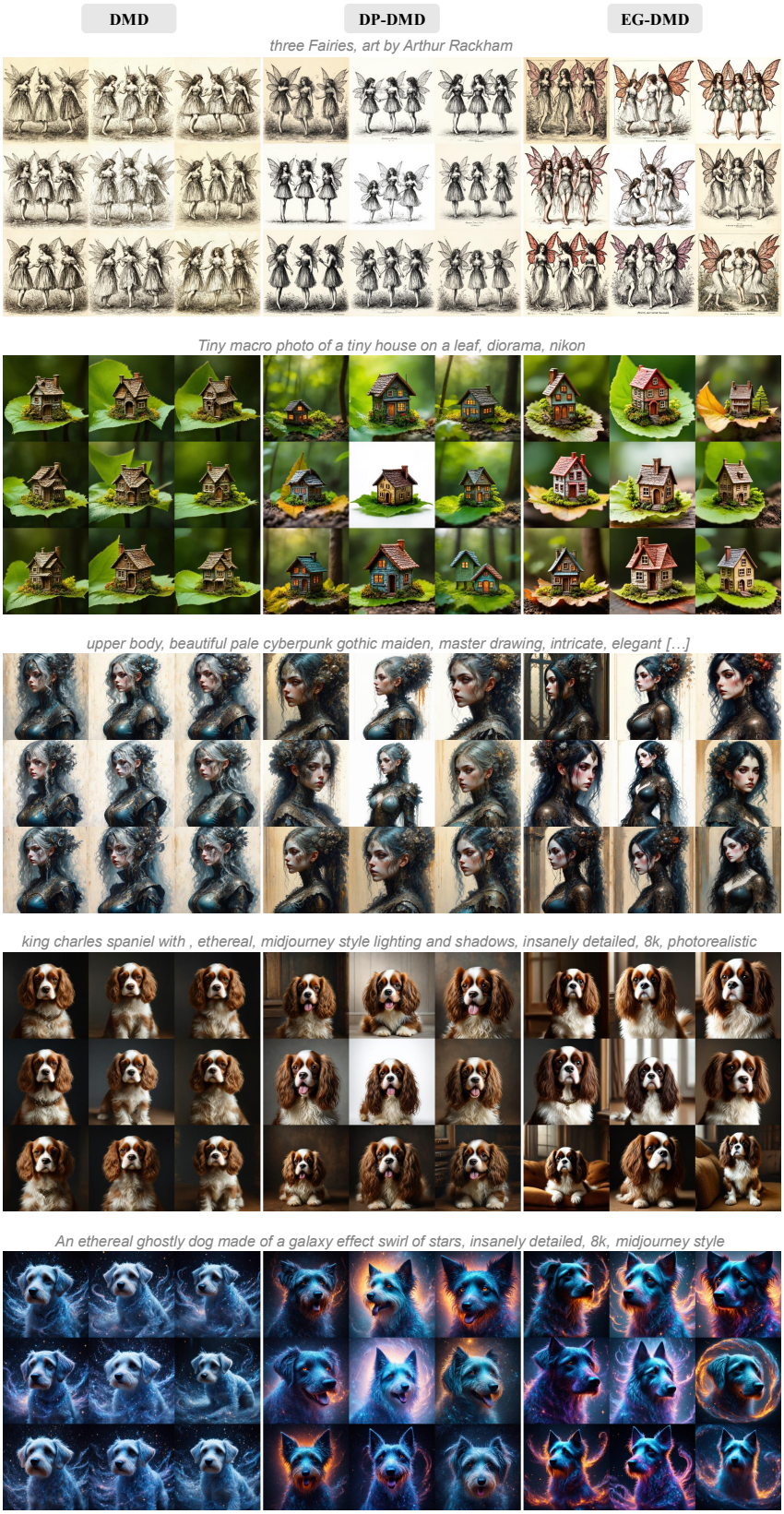



Figure 8: Qualitative comparison on SD3.5-M between DMD *(left)*, DP-DMD *(middle)*, and our EG-DMD *(right)*. EG-DMD demonstrates exceptional visual quality while preserving high diversity in the generated samples.




Figure 9: Qualitative comparison on SD3.5-M between DMD *(left)*, DP-DMD *(middle)*, and our EG-DMD *(right)*. EG-DMD demonstrates exceptional visual quality while preserving high diversity in the generated samples.

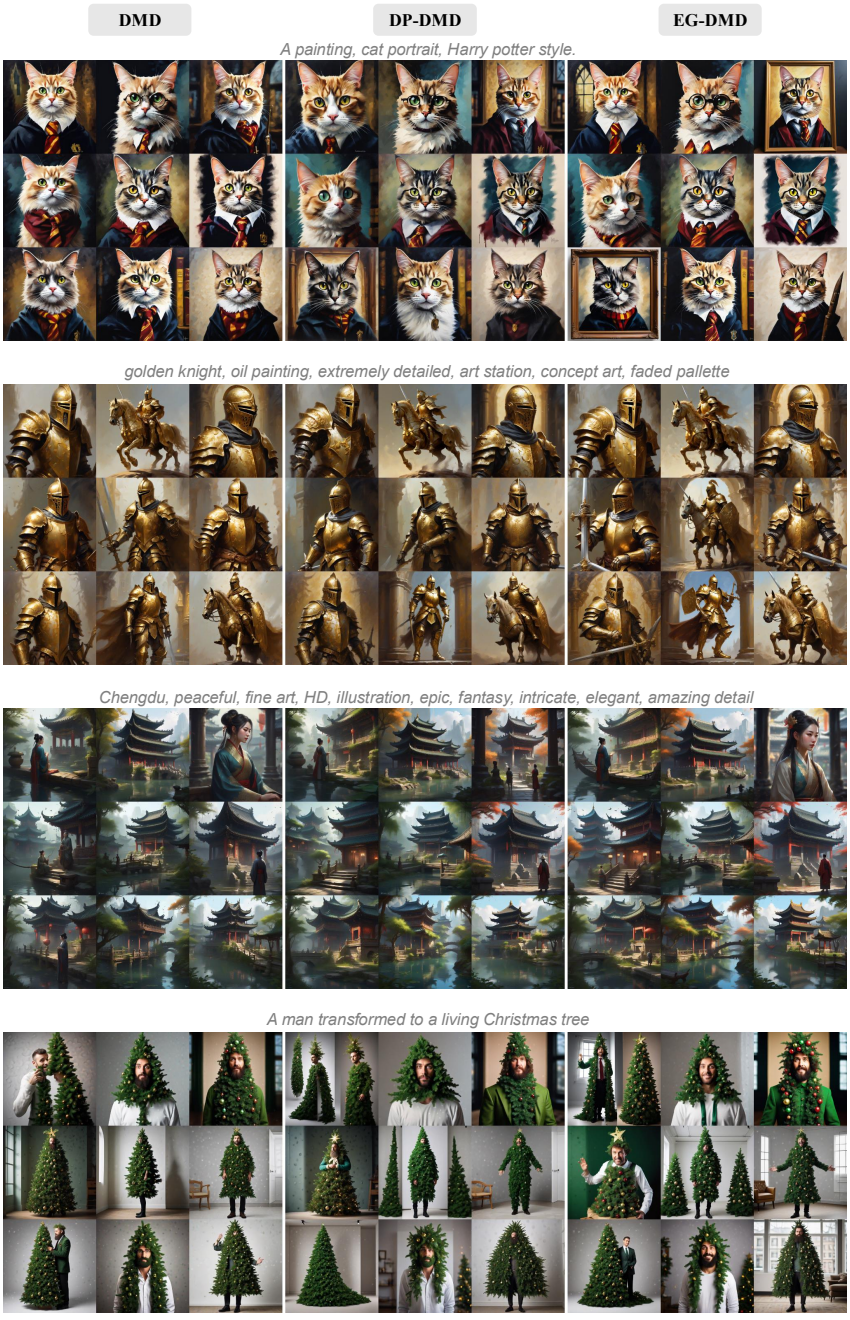



Figure 10: Qualitative comparison on SDXL between DMD *(left)*, DP-DMD *(middle)*, and our EG-DMD *(right)*. EG-DMD demonstrates exceptional visual quality while preserving high diversity in the generated samples.