

FLASH-DMD: UNIFYING DISTILLATION AND REFINEMENT FOR HIGH-FIDELITY FEW-STEP IMAGE GENERATION

Anonymous authors

Paper under double-blind review



Figure 1: Samples from Flash-DMD, taking less than 3% training cost of DMD2.

ABSTRACT

Diffusion Models have emerged as a leading class of generative models, yet their iterative sampling process remains computationally expensive. Timestep distillation is a promising technique to accelerate generation, but it often requires extensive training and leads to a degradation in image quality. Furthermore, fine-tuning these distilled models to optimize for specific objectives, such as aesthetic appeal or user preference, using Reinforcement Learning (RL) is notoriously unstable, and easily falls into reward hacking. In this work, we introduce Flash-DMD, a novel framework that enables fast convergence with distillation and stable RL-based refinement for stable optimization. Specifically, we first propose an efficient timestep-aware distillation strategy that significantly reduce training cost with enhanced human preference and realism. Second, and most critically, we introduce a joint training scheme where the model is fine-tuned with an RL objective while the timestep distillation training continues simultaneously. We demonstrate that the stable, well-defined loss from the ongoing distillation acts as a powerful regularizer, effectively stabilizing the RL training process and preventing policy collapse. Our experiments show that our proposed Flash-DMD not only converges significantly faster but also achieves state-of-the-art generation quality in the 4-step sampling regime, outperforming existing methods in human preference evaluations. Our work presents an effective paradigm for training efficient, high-fidelity, and stable generative models. Codes are attached in the supplementary.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024; Labs, 2024) have demonstrated remarkable success in text-to-image generation in recent years. However,

generating high-quality images necessitates numerous iterative denoising steps, resulting in substantial computational overhead, especially as models grow in size and complexity. This overhead poses a significant obstacle to real-time or resource-constrained deployment. To address this issue, various diffusion distillation techniques have been developed to distill multi-step teacher diffusion models into efficient student models that can produce comparable image quality in just one or a few inference steps (Luo et al., 2023a; Wang et al., 2024; Yin et al., 2024c;b; Lin et al., 2024; Chadebec et al., 2025; Ge et al., 2025; Lu et al., 2025). However, existing distillation methods are often inefficient and resource-intensive, requiring thousands of GPU hours for training. This significantly limits its accessibility to research groups and institutions with limited resources, and hinders rapid iteration and deployment in practical applications.

Among existing distillation methods, Distribution Matching Distillation (DMD) methods (Yin et al., 2024c;b; Lu et al., 2025; Ge et al., 2025) stand out for their superior generative quality, leveraging variational score distillation objectives (Yu et al., 2023) to align the output distributions of student and teacher models. However, this objective function suffers from unstable training and a tendency to mode seeking. Some approaches have employed adversarial methods to mitigate these problems. DMD2 (Yin et al., 2024b) proposes latent adversarial regulations with real images and designs a Two-Time scale Update Rule (TTUR) to stabilize training, but it combines the GAN (Sauer et al., 2024b;a) framework with DMD in a naive manner, neglecting the timestep aware feature of timestep distilled diffusion models, and the fake score μ_{fake} is trained both to discriminate real and generated images, but also to track the distribution changes of student models. These design compromises its efficiency in matching the distribution of teacher models. Furthermore, DM loss is inefficient in the latter part of distillation as it is hard to guide detailed learning, preventing it from effectively guiding the student diffusion model. These observations motivate our core research questions:

- Q1 In the early phase, how can we more effectively coordinate distribution matching with perceptual realism enhancement to accelerate convergence and stabilize training?*
- Q2 In the later phase, how can we refine the student model for better visual details and perceptual fidelity when gradient signals from distribution matching become less informative?*

To address the inefficiencies of the distribution matching methods, we proposed Flash-DMD, a twofold method in the four-step distillation task that outperforms DMD2 in the human preference benchmark with a much smaller training cost, while preserving superior perceptual realism. Specifically, our Flash-DMD follows different principles in the early and later generation phases.

In the early phase, the denoising performance at different stages varied, so the distillation target should also differ. We accordingly decoupled the GAN and distribution matching frameworks. At high-noise timesteps, the denoising model’s primary objective is to learn global composition and structure from the teacher, and DM loss is effective to process noisy latents. Therefore, we use a pure DM loss to align the student model with the teacher model’s output distribution. At low-noise timesteps, the model focuses on refining fine-grained details and enhancing perceptual realism. Thus, we use an adversarial loss to match the distribution of real images and improve the photorealism of the final outputs.

In the later phase, as gradient signals from distribution matching diminish, we directly optimize for visual quality and human preference. Our approach combines the DMD framework with latent reinforcement learning to efficiently refine the model’s handling of fine-grained details. Simultaneously, the distribution matching (DM) loss constrains domain shift, preventing the mode collapse and visual degeneration common in few-step reinforcement learning. This ensures the final results maintain high realism, avoiding reward hacking and the emergence of “oil painting” artifacts.

By combining faster convergence in the early phase with finer optimisation in the latter phase, we demonstrate the efficiency and superior performance of our method of distilling from SDXL to produce high-quality, realistic images. Notably, our method achieves the highest human preference scores while requiring the lowest training cost to date.

To summarize, our main contributions are threefold:

- We introduce Flash-DMD, a highly efficient framework for Distribution Matching Distillation. we decouple training objectives via a timestep-aware strategy to efficiently distill the fundamental distribution of the teacher model in low-SNR timesteps and refine perceptual quality and texture in high-SNR timesteps, and we counteract the mode-seeking of the

DM loss with a SAM-based Pixel-GAN that robustly enhances realism. The combination of these strategies and the stabilized score estimator allows for fastest convergence and stabilized distillation.

- Extensive experiments demonstrate that our method achieves superior performance compared to both the teacher model and baseline in human preference metrics, using only 2.1% of the training cost of DMD2. Further improvements are attained with an additional 6.2%–35.4% of DMD2’s training cost with even enhanced realism and textural detail.
- We successfully integrate reinforcement learning into the distillation process, creating a unified framework that seamlessly combines post-training refinement with distillation. This innovation eliminates the need for separate reinforcement and distillation phases, significantly reducing computational training costs while enabling efficient optimization of fine-grained details and perceptual fidelity in few-step image generation.

2 RELATED WORK

Diffusion Distillation. *Progressive Distillation* (Salimans & Ho, 2022; Meng et al., 2023; Lin et al., 2024) reduce inference steps in diffusion models by iteratively halving them, ultimately producing a one-step generator. Although effective, this iterative process is computationally expensive and is constrained by the preceding teacher model’s quality, leading to compounding errors. *Consistency Distillation* (Luo et al., 2023a;b) enforces a consistency constraint for the diffusion models, stipulating that any point on a given trajectory will revert to its starting point. However, it leads to performance degradation in the few-step inference. To migrate the issue, recent works (Wang et al., 2024; Ren et al., 2024; Zheng et al., 2024) have segmented the trajectory and progressively perform distillation on timestep segments, thereby achieving enhanced stability and fidelity. *Adversarial Distillation* introduces a discriminator to align the few-step student’s output with the multi-step teacher’s, either at the pixel level (Sauer et al., 2024b) or latent level (Sauer et al., 2024a), DMD2(Yin et al., 2024b). DMD2(Yin et al., 2024b) also uses latent adversarial training to match real-world data distribution, but its straightforward combination of adversarial loss and distribution matching may introduce conflicting objectives that can hinder overall distillation efficiency. *Score Distillation* was subsequently adapted for the distillation of diffusion models themselves (Yin et al., 2024c; Nguyen & Tran, 2024; Franceschi et al., 2023). An early approach, Distribution Matching Distillation (DMD) (Yin et al., 2024c), aims to minimize the KL-divergence between the teacher and student distributions. DMD2 (Yin et al., 2024b) replaced regression loss with adversarial loss for better realism. Building on this, Adversarial Distribution Matching (ADM) (Lu et al., 2025) introduced a GAN framework with Hinge loss, while SenseFlow (Ge et al., 2025) optimized scorers and discriminators for efficient distillation of larger models.

Reinforcement Learning in T2I Generation. Reinforcement learning is rapidly migrating to image generation tasks to align large-scale diffusion models with human feedback. Direct Preference Optimization (DPO) (Wallace et al., 2024; Liang et al., 2025; Lee et al., 2025; Li et al., 2025b; Miao et al., 2025; Zhang et al., 2025a) and Group Relative Policy Optimization (GRPO)(Liu et al., 2025; Xue et al., 2025; Li et al., 2025a; Wang et al., 2025; He et al., 2025) are two popular paradigms. The former methods construct offline or online win-lose pairs and back-propagate the preference order by the Bradley-Terry formed objectives. The latter methods sample a group of images on the SDE/mixed ODE-SDE trajectory, calculate the normalized advantage within the group, and constrain the policy generation direction. However, current research on performing RL on few-step models remains quite limited. Miao et al. (2025) proposed pairwise sample optimization (PSO), strengthening the relative likelihood margin between the training and reference sets. In contrast to PSO, we leverage a reward model operating in the latent space without relying on extrinsic positive targets(e.g., human-preferred labels, stylized images), which exclusively utilizes self-generated image pairs produced by the model itself and reduces computational overhead during training.

3 METHODOLOGY

3.1 PRELIMINARY OF DISTRIBUTION MATCHING DISTILLATION

Given pretrained diffusion model $\mathcal{T}_\phi(x_t, t)$ as teacher model, where x_t is noisy sample at timestep $t \sim \mathcal{U}(1, T)$, DMD(Yin et al., 2024c) and DMD2(Yin et al., 2024b) distill it into few-step effi-

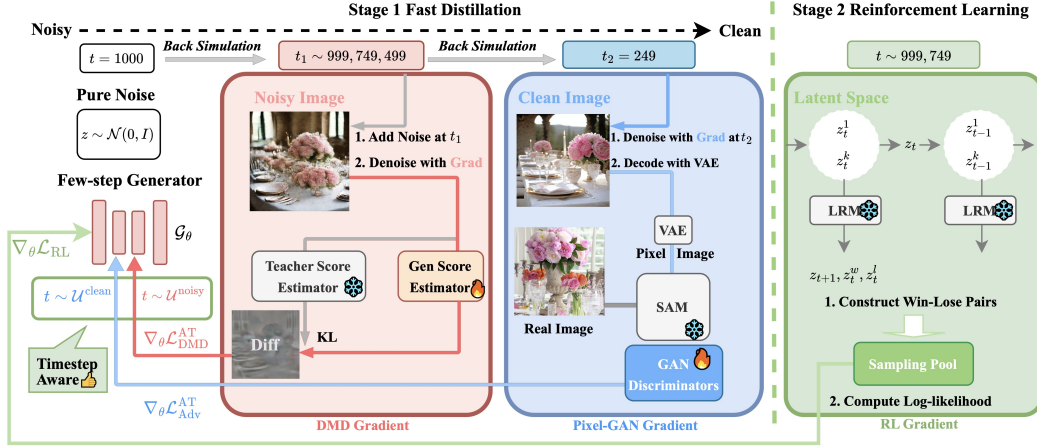


Figure 2: Framework of our proposed method Flash-DMD. There are two training stages: fast distillation and reinforcement learning. In stage 1, we take timestep-aware strategy and PixelGAN to efficiently create a realistic version of the teacher. In stage 2, latent preference optimization on high noise timesteps is utilized to enhance generation fidelity to surpass the teacher.

cient generator $\mathcal{G}_\theta(x_t, t)$ by minimizing the reverse KL divergence between the teacher model's distribution p_{tea} and the few-step generator's distribution p_{gen} . DMD series methods estimate p_{tea} through score estimator $\mu_{\text{tea}}(x_t, t)$, and p_{gen} is tracked with estimator $\mu_{\text{gen}}(x_t, t)$. Score function of the diffused distribution is:

$$s_{\text{gen}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{gen}}(x_t, t)}{\sigma_t^2}; s_{\text{tea}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{tea}}(x_t, t)}{\sigma_t^2} \quad (1)$$

where $\alpha_t, \sigma_t > 0$ are scalars determined by the noise schedule. s_{gen} and s_{tea} are vector fields that point towards higher density of distribution. Gradient of Distribution Matching objective w.r.t. θ is,

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathbb{E}_{z, t} \left[- (s_{\text{tea}}(\mathcal{G}_\theta(z, t)) - s_{\text{gen}}(\mathcal{G}_\theta(z, t))) \frac{d\mathcal{G}_\theta(z, t)}{d\theta} \right], \quad (2)$$

where $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t \sim \mathcal{U}(\mathbf{0}, \mathbf{T})$. In addition to the Distribution Matching objective, DMD2 introduces the combination with adversarial training with real images. Gradient of generator's adversarial objective w.r.t. θ is,

$$\nabla_\theta \mathcal{L}_{\text{AdvGen}} = \mathbb{E}_{z, t} \left[\log \mathcal{D}(\mathcal{G}_\theta(z, t)) \frac{d\mathcal{G}_\theta(z, t)}{d\theta} \right], \quad (3)$$

where $\mathcal{D}$ is the discriminator forward process. The score estimator of teacher $\mathcal{T}_\phi(\cdot)$ is itself, the generator's score estimator $\mu_{\text{gen}}^\psi(\cdot)$ is initialized with $\mathcal{T}_\phi(\cdot)$, and is dynamically updated to track p_{gen} with diffusion loss:

$$\mathcal{L}_{\text{Diffusion}} = \mathbb{E}_{x_{t-1}, t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\mu_{\text{gen}}^\psi(x_t, t) - \epsilon\|_2^2], \quad (4)$$

DMD2 reuses the parameter ψ of $\mu_{\text{gen}}^\psi(x_t, t)$ and extra trainable heads to distinguish p_{gen} and real image distribution p_{real} . Gradient of estimators's adversarial objective w.r.t. ψ is,

$$\nabla_\psi \mathcal{L}_{\text{AdvDisc}} = \mathbb{E}_{z, t, x_{\text{real}} \sim p_{\text{real}}} \left[\log \mathcal{D}(x_{\text{real}}) \frac{d\mathcal{D}(x_{\text{real}})}{d\psi} - \log \mathcal{D}(\mathcal{G}_\theta(z, t)) \frac{d\mathcal{D}(\mathcal{G}_\theta(z, t))}{d\psi} \right], \quad (5)$$

3.2 TRAINING INEFFICIENCY OF DMD SERIES

Despite their impressive performance, methods in the DMD series are characterized by significant computational overhead during distillation. This is evident in the extensive training schedules required by prominent models. For instance, the original DMD(Yin et al., 2024c) required 20,000 iterations with a batch size of 2,304 to distill Stable Diffusion v1.5 (Rombach et al., 2022) for single-step generation. Similarly, DMD2 (Yin et al., 2024b) used 24,000 iterations to distill SDXL (Podell et al., 2023) for four-step generation, and ADM (Lu et al., 2025) used 16,000 iterations

for single-step SDXL distillation. Given the strong empirical results and open-source implementation of DMD2, we select it as the foundation for our investigation into these inefficiencies. One primary source of inefficiency in DMD2 stems from its optimization strategy. As noted by Cheng et al. (2025), DMD2 simultaneously optimizes the model using two distinct gradients: a distribution-matching gradient (Eq. equation 2) and an adversarial gradient (Eq. equation 3). A direct summation of these gradients can introduce conflicting objectives, potentially steering the model toward a suboptimal state. This conflict can degrade both the accuracy of the distribution matching and the perceptual quality of the generated images, thereby hindering efficient convergence. A second challenge lies in the dual role assigned to the generator’s score estimator. It is tasked with two demanding objectives: tracking the output distribution of $\mathcal{G}_\theta(\cdot)$ (Eq. equation 4) and discriminating between real and generated samples (Eq. equation 5). To stabilize this complex dynamic, a two-time scale update rule (TTUR) is employed in DMD2, where score estimator is updated five times for every single update of the generator $\mathcal{G}_\theta(\cdot)$. This significantly contributes to the model’s overall training inefficiency.

3.3 FASTER CONVERGENCE IN EARLY PHASE

Adversarial Training is Necessary. DMD2 framework optimizes the generator by naively summing the Distribution-Matching (DM) loss from the teacher and an adversarial loss against real images at every timestep. This superposition of gradients can result in suboptimal and inefficient optimization. When we remove the adversarial teacher entirely, we observe that under pure DM loss supervision, the generator rapidly converges to a suboptimal domain, producing outputs with unnaturally high contrast and lacking fine-grained textures. We attribute this behavior to the mode-seeking nature of the reverse KL divergence, an observation also discussed in ADM (Lu et al., 2025). This finding underscores the necessity of the adversarial loss with real images for perceptual fidelity.

Decoupling Losses with a Timestep-Aware Strategy. We observe that the generator’s objective changes throughout the denoising process. For a 4-step distilled model, the initial, high-noise timesteps (low Signal-to-Noise Ratio, or SNR) primarily establish global composition and structure, and the low-noise timesteps (high SNR) focus on refining details, textures, and color tones to enhance realism. This observation is corroborated by findings of Cheng et al. (2025) in video generation tasks, which noted that the adversarial training in DMD2 is most active at high SNRs, whereas DM loss excels at guiding the model through high-noise regimes. Based on these insights, we assign DM loss and adversarial loss to distinct timesteps: *during the high-noise regime* (steps 1-3 in a 4-step model), we optimize the generator exclusively with the DM loss (Eq. equation 2). This allows the model to efficiently learn the teacher’s fundamental distribution and ODE trajectory in the early phases of generation. *For the low-noise step*, we apply the adversarial loss against real images, enabling the model to refine perceptual quality and texture in the final denoising step.

During each generator update, we sample one timestep $t \sim 999, 749, 499, 249$ and x_t from DMD2’s back-simulation forward process $\mathcal{B}$ to compute the DM loss:

$$x_t = \text{Detach}(\mathcal{B}(z, t)); \nabla_{\theta} \mathcal{L}_{\text{DMD}}^{\text{AT}} = \mathbb{E}_{z,t} \left[- (s_{\text{tea}}(\mathcal{G}_\theta(x_t, t)) - s_{\text{gen}}(\mathcal{G}_\theta(x_t, t))) \frac{d\mathcal{G}_\theta(x_t, t)}{d\theta} \right], \quad (6)$$

then employ $\mathcal{B}$ to propagate the denoised output x_{t-1} to a final clean image x_0 :

$$x_{t_1} = \mathcal{G}_\theta(x_t, t); x_0 = \text{Detach}(\mathcal{B}(x_{t_1}, 0)) \quad (7)$$

where Detach denotes stop gradient, x_0 is then used for the adversarial loss computation. We perform diffusion forward on x_0 to obtain the noisy sample x_{249} . Gradient for adversarial loss is:

$$\nabla_{\theta} \mathcal{L}_{\text{AdvGen}}^{\text{TA}} = \left[\mathbb{E}_{t=249} \log \mathcal{D}(\text{Decode}(\mathcal{G}_\theta(x_{249}, t))) \frac{d\mathcal{G}_\theta(x_{249}, t)}{d\theta} \right] \quad (8)$$

where the $\mathcal{D}(\cdot)$ is the pixel-level discriminator, and Decode is the decode process of SDXL-VAE (Podell et al., 2023). This timestep-aware strategy reduces interference between these two optimization objectives. Our experiments demonstrate that this approach significantly improves training efficiency while generating high-quality images with enhanced realism and textural detail.

Pixel-GAN alleviates mode-seeking. In order to enforce realism and structural coherence, the present study performs adversarial learning directly in the pixel space using a discriminator based on features that is new and innovative. In contrast to a conventional latent-space GAN, the discriminator in our model is constructed upon the frozen vision encoder of the Segment Anything Model

(SAM) to extract hierarchical features with multiple trainable discriminator heads attached. The discriminator’s trainable parameters ω are updated via the Hinge loss:

$$\mathcal{L}_{\text{AdvDisc}}^{\text{PG}} = \mathbb{E}_{x_{\text{real}}} [-\log \mathcal{D}_{\omega}(x_{\text{real}})] + \mathbb{E}_{z, t} [\log \mathcal{D}_{\omega}(\text{Decode}(\mathcal{G}_{\theta}(x_t, t)))] \quad (9)$$

The discriminator is characterized by its exceptional sensitivity to local geometric structures and fine-grained textures, a capability that is facilitated by SAM’s powerful, general-purpose representations, as noted by Lu et al. (2025). This pixel-level supervision exerts a stringent realism constraint from the training’s earliest stages, compelling the generator to expeditiously discern and anchor to diverse, high-fidelity modes within the data distribution. This approach effectively prevents premature convergence to simplistic or blurry solutions (mode-seeking).

Stabilize Score Estimator. In contrast to DMD2 Yin et al. (2024b), where the score estimator is required to serve as a discriminator (Eq. equation 5) and thus faces conflicting optimization objectives, our approach trains μ_{gen}^{ψ} solely via the diffusion loss (Eq. equation 4), eliminating the tasking burden and training complexity for $\mathcal{G}_{\theta}(\cdot)$. Our experiments show that updating the score estimator only once or twice per generator update (TTUR=1,2) is sufficient for stable and accurate distribution tracking. This lightweight coupling leads to more stable training dynamics and superior sample fidelity compared to DMD2 with TTUR=5, while reducing computational overhead.

Similar to implicit distribution alignment proposed by Ge et al. (2025), we also adopt an Exponential Moving Average (EMA) update strategy to ensure that the score estimator μ_{gen}^{ψ} accurately tracks the evolving distribution of the generator $\mathcal{G}_{\theta}(\cdot)$. Specifically, after each generator update, we inject the latest generator parameters into the score estimator using an EMA coefficient λ_{ema} , i.e.,

$$\psi \leftarrow \lambda_{\text{ema}}\psi + (1 - \lambda_{\text{ema}})\theta, \quad (10)$$

which enables μ_{gen}^{ψ} to closely follow the generator’s trajectory with minimal additional updates.

Putting everything together. We introduce Flash-DMD, a highly efficient framework for Distribution Matching Distillation. In summary, Flash-DMD’s training objectives via a timestep-aware strategy efficiently distill the fundamental distribution of the teacher model in low-SNR timesteps and refine perceptual quality and texture in the final high-SNR timestep. To counteract the mode-seeking tendency of the DM loss, we introduce a SAM-based Pixel-GAN that robustly enhances realism. The combination of these strategies and the stabilized score estimator enables a more effective and balanced optimization of the generator $\mathcal{G}_{\theta}(\cdot)$.

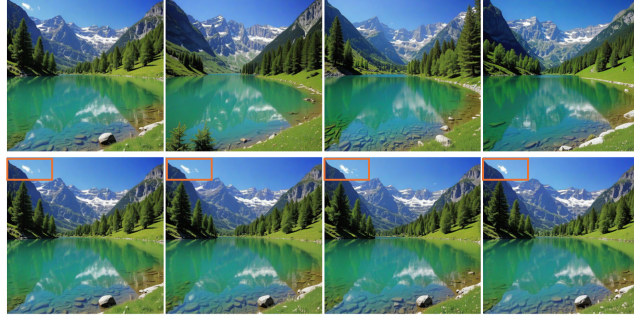
3.4 REINFORCEMENT LEARNING FOR FINER PERFORMANCE IN THE LATER PHASE

Reinforcement learning boosts model performance. Leveraging the above training paradigm, we have developed a student generator capable of rivaling the teacher model. Subsequently, our focus shifts to enhancing its performance beyond that of the teacher and deploying it in practical scenarios. Preference optimization provides a direct shortcut to significantly improve image fidelity, aesthetics, and detail richness with limited resources.

Choosing a suitable Reward Model. Selecting an appropriate reward model is a critical step. We begin with an empirical investigation of the prevalent models and systematically summarize their characteristics in the appendix. The empirical findings lead to the following conclusions: 1) Most models operate in pixel space, with only the Latent Reward Model (LRM) exploring in latent space. 2) Although the timestep constitutes a critical factor for text-to-image models, it is neglected by most reward models, with a few exceptions. 3) In general, these reward models score on dimensions such as image fidelity, aesthetics, and text-image alignment. In light of the analysis above, we incorporate LRM into our distillation training pipeline.

The best of both worlds. Born from stable diffusion models, LRM can score noisy latent representations at any timestep without the need for complex VAE decoder transformations. This capability inherently meets our demands. A subsequent key challenge is how to integrate LRM into our framework to achieve greater gains effectively. To adapt the approach to our method, we introduce several key changes: 1) In the sampling process for win-lose pair construction and log-likelihood computation stage, we replace the original noise scheduler, Denoising Diffusion Implicit Model (DDIM), with Latent Consistency Model (LCM) since the distilled model is compatible with the latter. 2) As shown in Fig. 3, with the same initialized noise, it is evident that images sampled at high-noise steps exhibit better diversity in layout and fine-grained details compared to those from

low-noise steps. As a result, we only perform stochastic sampling in the high-noise phase, altering latent representations that are deemed optimal/suboptimal by the reward model. 3) The dynamic threshold is removed because the original dynamic boundary range is insufficient to adapt to the distribution characteristics of distilled models adequately. 4) We combine the logarithmic likelihood loss with the vanilla loss of Flash-DMD during the training process rather than applying preference optimization separately. We will explore a reasonable time scale update rule in the appendix file.



Prompt: Designs Similar to Lake Verde In The Alps

Figure 3: Sampling variance analysis at different time steps. The first row displays samples obtained at the 999th denoising step, while the second row corresponds to the 499th step.

Win-lose Pair Construction and Final RL Loss. Given a generator $\mathcal{G}_\theta(\cdot)$ distilled from a pre-trained diffusion model $\mathcal{T}_\phi(\cdot)$, it can sample clean images from pure noise $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, conditioned on text prompt c , within $T = 4$ steps. At high-noise timesteps, we sample a set of k noisy latent images $\{z_{t-1}^1, \dots, z_{t-1}^k\}$ from the same initial latent image z_t . LRM predicts preference scores. The samples corresponding to the highest and lowest normalized scores are selected as win-lose pairs, thereby constructing paired training data $(z_t, z_{t-1}^w, z_{t-1}^l)$ to the sampling pool. These pairs are subsequently used to minimize the loss function:

$$\mathcal{L}_{rl} = -\mathbb{E}_{z_{t-1}^w, z_{t-1}^l \sim p_\theta(z_{t-1}|z_t, c)} [\log \sigma(\beta \log(\frac{p_\theta(z_{t-1}^w|z_t, c)}{p_{ref}(z_{t-1}^w|z_t, c)} - \beta \log(\frac{p_\theta(z_{t-1}^l|z_t, c)}{p_{ref}(z_{t-1}^l|z_t, c)}))] \quad (11)$$

where $p_\theta(z_{t-1}|z_t, c)$ denotes the backward process to denoise z_t and can be formulated as follows:

$$p(z_{t-1}|z_t) = \mathcal{N}(\mu_t, \sigma_t^2 \epsilon_t), \epsilon_t \sim \mathcal{N}(0, I) \quad (12)$$

$$\mu_t = \sqrt{\bar{\alpha}_{t-1}} \cdot \hat{z}_{0,t} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(z_t, t) \quad (13)$$

$$\hat{z}_{0,t} = c_{out}(t) \cdot \left(\frac{z_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_\theta(z_t|c)}{\sqrt{\bar{\alpha}_t}} \right) + c_{skip}(t) \cdot z_t \quad (14)$$

where $\hat{z}_{0,t}$ denotes clean latent images predicted by noise predictor $\epsilon_\theta(\cdot)$ at timestep t . $c_{skip}(\cdot)$ and $c_{out}(\cdot)$ are differentiable functions predefined in the LCM scheduler.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

Experiment Setup For the first phase, we utilize a filtered set from the LAION 5B (Schuhmann et al., 2022) dataset to provide high-quality image-text pairs for training, following the setting of (Yin et al., 2024a). For discriminator conditioning, we adopt the vision encoder from Kirillov et al. (2023) as the backbone to extract image representations. The structure of trainable discriminator heads follows the 2D architecture of (Lu et al., 2025). For the second phase, we adopt the training dataset from the first phase and utilize the Latent Reward Model from (Zhang et al., 2025a). We sample a set of noisy latent images at the high-noise timesteps $t = 749, 999$ and set $k = 4$ following (Zhang et al., 2025a). We conduct all of our experiments on NVIDIA H20 GPUs.

Table 1: Comparison with other distillation methods on COCO-10k dataset.

Method	#NFE	ImageReward $\uparrow$	CLIP $\uparrow$	Pick $\uparrow$	HPSv2 $\uparrow$	MPS $\uparrow$	Cost $\downarrow$
SDXL	100	0.7143	0.3295	0.2265	0.2865	11.87	-
LCM-SDXL	4	0.5562	0.3250	0.2236	0.2818	11.11	-
SDXL-Lightning	4	0.6952	0.3268	0.2285	0.2888	12.15	-
SDXL-Turbo	4	0.8338	0.3302	0.2286	0.2899	12.25	-
NitroSD-Realism	4	0.9112	0.3274	0.2291	0.2975	12.43	-
NitroSD-Vibrant	4	0.8419	0.3201	0.2205	0.2865	11.13	-
DMD2-SDXL	4	0.8748	0.3302	0.2309	0.2937	12.41	128*24k
Flash-DMD under Phase 1							
TTUR1-1k	4	0.9509	0.3292	<u>0.2322</u>	0.2968	<u>12.67</u>	64*1k (2.1%)
TTUR2-4k	4	0.9450	0.3291	<u>0.2322</u>	0.2969	12.65	64*4k (8.3%)
TTUR2-8k	4	0.9740	<u>0.3298</u>	0.2327	<u>0.2981</u>	12.71	64*8k (16.7%)
TTUR5-18k	4	0.9426	0.3302	0.2319	0.2982	12.63	64*18k (37.5%)

Table 2: Comparison with other methods with reinforcement learning on COCO-10k dataset.

Method	#NFE	ImageReward $\uparrow$	CLIP $\uparrow$	Pick $\uparrow$	HPSv2 $\uparrow$	MPS $\uparrow$	GPU Hours $\downarrow$
Hyper-SDXL	4	1.085	0.3300	0.2324	0.3030	12.45	400 A100
PSO-DMD2	4	0.9157	0.3285	0.2338	0.2897	12.53	160 A100
LPO-SDXL	40	1.042	0.3324	0.2342	0.2965	12.58	92 A100
Flash-DMD	4	1.004	0.3285	0.2346	0.2930	12.84	12 H20

Evaluation Tasks and Baseline The evaluation of image generators is conducted on 10K prompts from COCO 2014 (Lin et al., 2014), adhering to the DMD2 (Yin et al., 2024a) framework, containing 10,000 images. We present the result of CLIP score (Radford et al., 2021) (ViT-B/32) to evaluate text-image similarity, and we adopt a set of advanced preference-based metrics to thoroughly evaluate the quality of generated images from multiple human-aligned perspectives. We use HPSv2 (Wu et al., 2023) to measure fine-grained image-text semantic alignment, focusing on how well the generated content adheres to the input prompt. ImageReward (Xu et al., 2023) and PickScore (Kirstain et al., 2023) are employed to assess overall aesthetic quality and perceptual appeal, reflecting general human preferences in visual coherence and composition. Furthermore, we evaluate multidimensional human preferences using MPS (Zhang et al., 2024), a recently proposed metric that captures diverse aspects of human judgment, such as object accuracy, spatial relation, and attribute binding-beyond global similarity. Together, these metrics provide a comprehensive and human-centric evaluation of both fidelity and preference in text-to-image generation. To demonstrate the effectiveness of phase 1 distillation, we compare our 4-step generative models against SDXL (Podell et al., 2023), as well as other open-sourced timestep distillation methods, including LCM-SDXL (Luo et al., 2023a), SDXL-Lighting (Lin et al., 2024), SDXL-Turbo (Sauer et al., 2024b), Realism and Vibrant version of NitroSD (Chen et al., 2025), and DMD2 (Yin et al., 2024b). For phase 2, we further evaluate our approach by comparing it with three reinforcement learning-finetuned models, Hyper-SDXL (Ren et al., 2024), PSO-DMD2 (Miao et al., 2025), and LPO-SDXL (Zhang et al., 2025a).

4.2 EXPERIMENT ANALYSIS

Faster and Better Distillation under Phase 1. Our method achieves highly efficient distillation from the teacher model, leading to state-of-the-art performance across all benchmarks, by decoupling the distribution matching (DM) and adversarial losses with a timestep-aware strategy, employing PixGAN to alleviate mode-seeking behavior, and stabilizing the generator’s score estimator. As shown in Tab. 1, we distill SDXL using various two-time update rules (TTUR), which corresponds to the update frequency ratio between the score estimator and generator. In contrast to DMD2, which uses a TTUR of 5 and thus hinders training efficiency, we experiment with TTUR values of 1, 2, and 5. Our results demonstrate significant improvements in both efficiency and performance. With a TTUR of 5, our model surpasses DMD2 on all benchmarks while requiring only 37.5% of the training cost (batch size $\times$ training steps). Reducing the TTUR to 2 allows us to maintain superior human preference scores and comparable text-image consistency with only 8.3% of DMD2’s training cost. In the most extreme case, training for only 1,000 steps with a TTUR of 1, with merely 2.1% of DMD2’s training cost, we still yield a higher human preference score. Notably, under all tested settings, our model consistently outperforms the original teacher model.

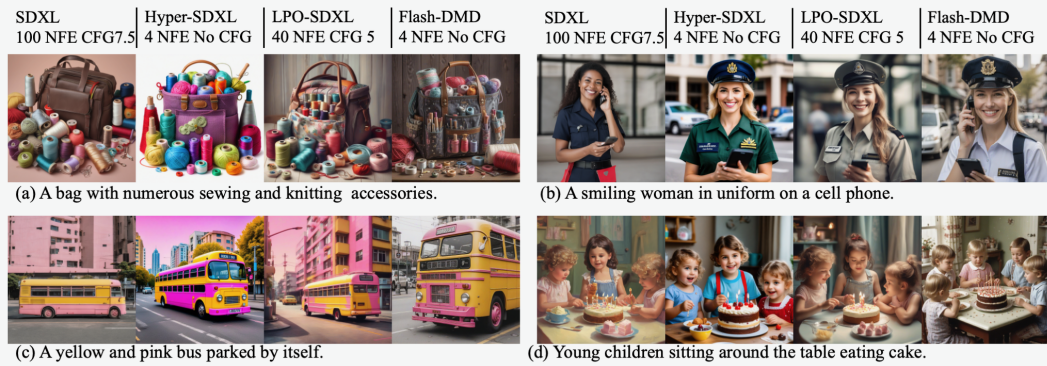


Figure 4: Qualitative comparisons with other reinforcement approaches on SDXL.

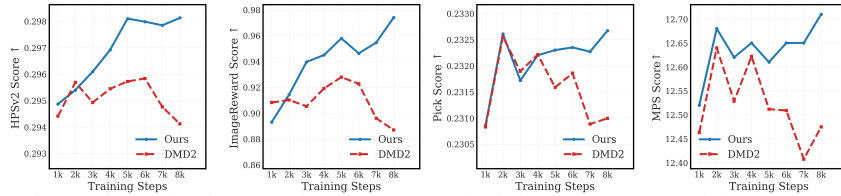


Figure 5: Evaluation results of DMD2 and Flash-DMD (ours) with TTUR=2.

Boost Performance under Phase 2. By incorporating reinforcement learning, Flash-DMD achieves a performance comparable to other reinforcement approaches with fewer computational resources, as shown in the Tab. 2 and Fig. 4. Flash-DMD scores the highest on PickScore and MPS, and image fidelity surpasses SDXL and other competitors. Although Hyper-SDXL has the highest ImageReward score and HPSv2 score, it generates overexposed colors and unnatural images. LPO-SDXL gets the highest CLIP score, but it produces oversmoothed images. We speculate that the reason may be that these models use the trained models directly for reinforcement training. This is prone to reward hacking and only rewards the results preferred by the reward model. Meanwhile, Flash-DMD introduces preference optimization during the training process. With the constraints from distribution matching and PixelGAN, the problem of reward hacking can be alleviated.

4.3 ABLATION STUDIES

Phase1: Comparison with DMD2 under smaller TTUR. We set TTUR=2 for ablate the performance of DMD2 and Flash-DMD under phase 1. The results, presented in Fig. 5, show that our method exhibits stable and continuous improvement throughout the training process. In contrast, DMD2 shows slight initial gains but quickly degrades as training progresses. This comparison validates that our approach offers much better training stability and efficiency under these conditions. To keep the main text focused on core experiments and results, we have moved some detailed ablation studies to the appendix. These additional experiments further validate the design choices of our method and their impact on model performance. Readers can refer to the appendix for more comprehensive data and analysis that support the conclusions presented in the main text.

5 CONCLUSION

We present Flash-DMD, a twofold approach that addresses the inefficiencies of existing diffusion distillation methods by leveraging timestep-aware objectives and optimizing the distillation process. In the early phase, Flash-DMD accelerates convergence by coordinating distribution matching and perceptual realism enhancement. In the later phase, it refines visual details using latent reinforcement learning while preventing mode collapse and artifacts. Experiments show Flash-DMD achieves superior generation quality and the highest human preference scores with significantly reduced training costs. Our method makes diffusion distillation more efficient and accessible, paving the way for advancements in low-step generative modeling. We hope our findings can open new avenues for research and contribute to advancing the field of visual generation.

ETHICS STATEMENT

All authors affirm that this work adheres to the ICLR Code of Ethics (<https://iclr.cc/public/CodeOfEthics>). We are committed to responsible stewardship of machine learning research and have considered the broader impacts of our work in alignment with the principles of contributing to societal well-being, upholding scientific excellence, avoiding harm, and ensuring fairness. Our method builds upon existing diffusion model frameworks and does not involve direct experimentation with human subjects. All datasets used in this work are established public benchmarks commonly adopted in the generative modeling community, and no new data involving personal or sensitive information has been collected or released. We have taken care to ensure that our approach minimizes potential for misuse and does not introduce new risks related to bias, discrimination, or privacy. The focus of this work is on improving training efficiency and generation quality within a controlled research setting. While advanced generative models may have downstream applications with ethical implications, we emphasize transparency in methodology and encourage future deployment only in socially responsible contexts. We support open and reproducible science while complying with all relevant academic and institutional standards.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of the results presented in this paper. To this end, we provide detailed descriptions of our method, training procedure, and evaluation protocols in the main text and appendix. We will publicly release both the source code for training and inference, as well as the trained models, to facilitate replication and further research. This open-sourcing plan is intended to support full reproducibility and community engagement.

REFERENCES

- Clement Chadebec, Onur Tasar, Eyal Benaroch, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15686–15695, 2025.
- Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Nitrofusion: High-fidelity single-step diffusion through dynamic adversarial training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 7654–7663. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00717. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_NitroFusion_High-Fidelity_Single-Step_Diffusion_through_Dynamic_Adversarial_Training_CVPR_2025_paper.html.
- Jiaxiang Cheng, Bing Ma, Xuhua Ren, Hongyi Jin, Kai Yu, Peng Zhang, Wenyue Li, Yuan Zhou, Tianxiang Zheng, and Qinglin Lu. Pose: Phased one-step adversarial equilibrium for video diffusion models. *arXiv preprint arXiv:2508.21019*, 2025.
- 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 *Forty-first international conference on machine learning*, 2024.
- Jean-Yves Franceschi, Mike Gartrell, Ludovic Dos Santos, Thibaut Issenhuth, Emmanuel de Bézenac, Mickaël Chen, and Alain Rakotomamonjy. Unifying gans and score-based diffusion as generative particle models. *Advances in Neural Information Processing Systems*, 36: 59729–59760, 2023.
- Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. Senseflow: Scaling distribution matching for flow-based text-to-image distillation. *CoRR*, abs/2506.00523, 2025. doi: 10.48550/ARXIV.2506.00523. URL <https://doi.org/10.48550/arXiv.2506.00523>.
- Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025.

- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.
- 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.
- Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18465–18475. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01721. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lee_Calibrated_Multi-Preference_Optimization_for_Aligning_Diffusion_Models_CVPR_2025_paper.html.
- Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025a.
- Zejian Li, Yize Li, Chenye Meng, Zhongni Liu, Yang Ling, Shengyuan Zhang, Guang Yang, Changyuan Yang, Zhiyuan Yang, and Lingyun Sun. Inversion-dpo: Precise and efficient post-training for diffusion models. *CoRR*, abs/2507.11554, 2025b. doi: 10.48550/ARXIV.2507.11554. URL <https://doi.org/10.48550/arXiv.2507.11554>.
- Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13199–13208. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01232. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liang_Aesthetic_Post-Training_Diffusion_Models_from_Generic_Preferences_with_Step-by-step_Preference_CVPR_2025_paper.html.
- Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 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.
- Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online RL. *CoRR*, abs/2505.05470, 2025. doi: 10.48550/ARXIV.2505.05470. URL <https://doi.org/10.48550/arXiv.2505.05470>.
- Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J. Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. *CoRR*, abs/2507.18569, 2025. doi: 10.48550/ARXIV.2507.18569. URL <https://doi.org/10.48550/arXiv.2507.18569>.
- Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023a.
- Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023b.

- Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 14297–14306. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01374. URL <https://doi.org/10.1109/CVPR52729.2023.01374>.
- Zichen Miao, Zhengyuan Yang, Kevin Lin, Ze Wang, Zicheng Liu, Lijuan Wang, and Qiang Qiu. Tuning timestep-distilled diffusion model using pairwise sample optimization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=fXnE4gB64o>.
- Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7807–7816, 2024.
- 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.
- 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*, pp. 8748–8763. PmlR, 2021.
- 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. *Advances in Neural Information Processing Systems*, 37:117340–117362, 2024.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.
- 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, 2024a.
- Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024b.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.
- Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 8228–8238. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00786. URL <https://doi.org/10.1109/CVPR52733.2024.00786>.
- 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. *Advances in neural information processing systems*, 37:83951–84009, 2024.
- Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025.

- Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.
- 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.
- Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing GRPO on visual generation. *CoRR*, abs/2505.07818, 2025. doi: 10.48550/ARXIV.2505.07818. URL <https://doi.org/10.48550/arXiv.2505.07818>.
- Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.
- Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024b.
- Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024c.
- Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023.
- Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8018–8027, 2024.
- Tao Zhang, Cheng Da, Kun Ding, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. *CoRR*, abs/2502.01051, 2025a. doi: 10.48550/ARXIV.2502.01051. URL <https://doi.org/10.48550/arXiv.2502.01051>.
- Tao Zhang, Cheng Da, Kun Ding, Huan Yang, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. *arXiv preprint arXiv:2502.01051*, 2025b.
- Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation. *CoRR*, 2024.

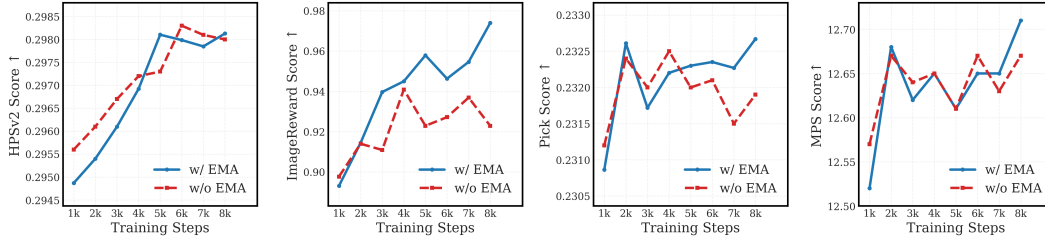


Figure 6: Evaluation results of Flash-DMD (ours) with or without EMA on ImageReward, PickScore, and HPSv2. The training steps range from 1,000 to 8,000. Both models are trained with a two-time scale update rule (TTUR). The generator and the score estimator are updated at a rate of 1:2, i.e., TTUR=2.

A APPENDIX

A.1 THE USE OF LARGE LANGUAGE MODELS (LLMs)

In this work, we utilized large language models (LLMs) as assistive tools to refine the writing. Specifically, LLMs were employed to enhance the clarity, coherence, and linguistic quality of these sections without altering the core content or scientific contributions of the paper. The use of LLMs was limited to language refinement and did not play a significant role in research ideation, methodology development, or experimental design. This statement is provided to ensure transparency in accordance with the guidelines for LLM usage.

A.2 ABLATION STUDIES

Phase 1: Significance of EMA in Score Estimator. As described in Sec. 3.3, we employ a Exponential Moving Average (EMA) strategy to help the score estimator more accurately track the generator’s distribution, especially under high-frequency updates. To validate the effectiveness of this approach, we conduct an ablation study comparing performance with and without the EMA strategy. As shown in Fig. 6, the model with EMA achieves higher ImageReward and PickScore scores in later training stages. The HPSv2 scores remain nearly identical. This confirms that the EMA strategy enhances visual quality and human preference without compromising text alignment.

Phase 2: Trade-Off in Time Scale Update Rule for RL. Rather than superimposing multiple loss functions via a weighted sum, we use an alternating update strategy to update the generator, applying different loss functions at different frequencies. At the beginning, we initialize the generator and the fake score estimator with weights from TTUR1-1k experiment, and the real score estimator is initialized with SDXL. We trained the model for 2,000 iterations on a single H20 GPU under different frequency ratios between reinforcement loss and distribution matching loss (1:1, 2:1, 5:1, 10:1). The metric comparisons are shown in Table 3. The 5:1 ratio achieves the highest score among these settings.

Phase 2: Other ablation experiments on reinforcement learning. *Online training VS Post-training.* We compare online training with post-training by LPO (Zhang et al., 2025b) alone. Our method demonstrates superior performance over Post-Train LPO, which validates the advantage of our proposed training paradigm. and full-noise timestep sampling. *High-noise VS all noise.* Training only on high-noise steps achieves better results than training on all-noise steps. *Including PixelGAN loss.* Furthermore, the incorporation of an additional Pixel-Gan objective yields a positive gain, resulting in a marginal improvement in the metrics. Fig. ?? supplements the results with and without GAN loss across 1,000 to 5,000 iterations. We selected the 5k-step model with GAN loss as our final enhanced version, as it delivered the best performance. The training cost for this model was 12 H20 GPU hours.

Table 3: Comparison of Different Variants in Reinforcement Learning Experiments.

	ImageReward	PickScore	MPS	CLIP	GPU Hours
Base-TTUR1-1k	0.9508	0.2322	12.672	0.3292	–
1:1	0.9135	0.2330	12.755	0.3284	5.72
2:1	0.9315	0.2329	12.770	0.3271	5.16
5:1	0.9808	0.2345	12.764	0.3275	4.83
10:1	0.9640	0.2344	12.685	0.3272	4.72
Post-Train LPO	0.9795	0.2345	12.689	0.3284	4.96
all noise	0.9421	0.2331	12.800	0.3294	7.3
+ pixelgan	0.9678	0.2345	12.812	0.3280	5.77
Flash-DMD	<u>1.0004</u>	<u>0.2346</u>	<u>12.813</u>	<u>0.3285</u>	12.0

Table 4: Mainstream Image Reward Models

Model	Space	Evaluation Dimension	VLM	Step-aware
PickScore	Pixel	Fidelity	CLIP	×
ImageReward	Pixel	Fidelity/Aes.	BLIP	×
MPS	Pixel	Align./Fidelity/Aes.	CLIP	×
HPSv2	Pixel	Align./Fidelity	CLIP	×
SPM	Pixel	Align./Aes.	CLIP	✓
LRM	Latent	Align./Aes.	SD1.5/SDXL	✓
VisionReward	Pixel	Align./Fidelity	Llama	×
UnifiedReward	Pixel	Align.	Qwen2.5VL	×

A.3 ANALYSIS OF IMAGE REWARD MODELS

In the Tab. 4, we summarize the scope space and evaluation dimensions of popular image reward models. We also include the Visual Large Model (VLM) used, and whether the timestep is considered for these models.

A.4 ADDITIONAL QUALITATIVE RESULTS

We show additional qualitative comparison as in Figure7, demonstrating that our model not only surpasses other distillation models but also outperforms the teacher model in refining image quality. Specifically, we compare our results with SDXL, SDXL-Lighting(Lin et al., 2024), SDXL-Turbo(Sauer et al., 2024b), Hyper-SDXL(Ren et al., 2024), DMD2(Yin et al., 2024b), LPO (Zhang et al., 2025b), Realism version of NitroSDChen et al. (2025), PSO(Miao et al., 2025). Additionally, we present qualitative examples of Phase 1-generated images in Figure 8, further highlighting the strengths of our approach.

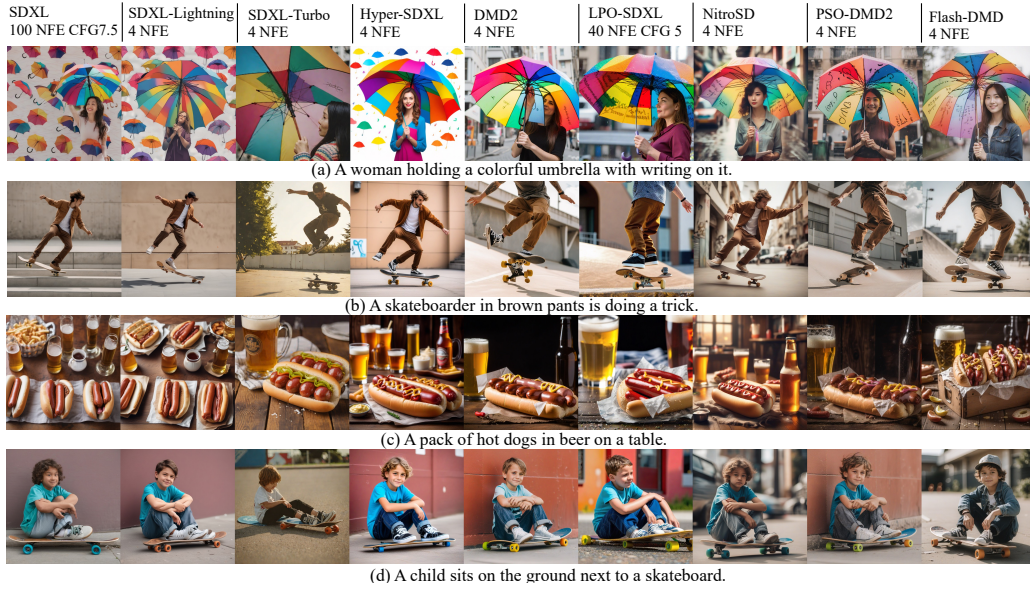


Figure 7: Qualitative comparisons with other models.



Figure 8: Qualitative results of our first phase model.