R_{dm}: Re-conceptualizing Distribution Matching as a Reward for Diffusion Distillation

Linqian Fan[1,2†], Peiqin Sun[1‡], Tiancheng Wen[1], Shun Lu[1], and Chengru Song[1]

[1]KlingAI Research, China [2]Tsinghua University, China

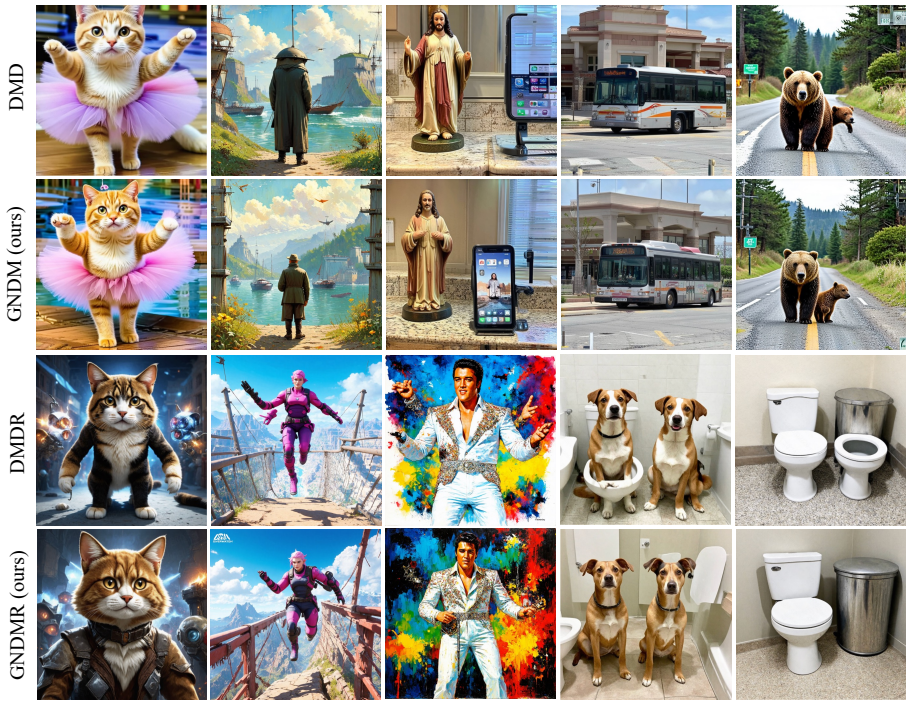


Fig. 1: (Top) Samples from 4-step vanilla DMD and our GNDM, and (Bottom) samples from 4-step DMDR and our GNDMR, based on SD3. Our method achieves better perceptual fidelity with fewer artifacts and better details.

Abstract. Diffusion models achieve state-of-the-art generative performance but are fundamentally bottlenecked by their slow, iterative sampling process. While diffusion distillation techniques enable high-fidelity, few-step generation, traditional objectives often restrict the student's performance by anchoring it solely to the teacher. Recent approaches

† Work done during an internship at KlingAI Research.
‡ Corresponding author.

have attempted to break this ceiling by integrating Reinforcement Learning (RL), typically through a simple summation of distillation and RL objectives. In this work, we propose a novel paradigm by **re-conceptualizing distribution matching as a reward**, denoted as R_{dm}. This unified perspective bridges the algorithmic gap between Diffusion Matching Distillation (DMD) and RL, providing several primary benefits: **(1) Enhanced Optimization Stability:** We introduce Group Normalized Distribution Matching (GNDM), which adapts standard RL group normalization to stabilize R_{dm} estimation. By leveraging group-mean statistics, GNDM establishes a more robust and effective optimization direction. **(2) Seamless Reward Integration:** Our reward-centric formulation inherently supports adaptive weighting mechanisms, allowing for the fluid combination of DMD with external reward models. **(3) Improved Sampling Efficiency:** By aligning with RL principles, the framework readily incorporates Importance Sampling (IS), leading to a significant boost in sampling efficiency. Extensive experiments demonstrate that GNDM outperforms vanilla DMD, reducing the FID by 1.87. Furthermore, our multi-reward variant, GNDMR, surpasses existing baselines by striking an optimal balance between aesthetic quality and fidelity, achieving a peak HPS of 30.37 and a low FID-SD of 12.21. Ultimately, R_{dm} provides a flexible, stable, and efficient framework for real-time, high-fidelity synthesis. Code is available at: https://github.com/Flq2002/Rdm.

Keywords: Diffusion Models · Distribution Matching Distillation · Reinforcement Learning

1 Introduction

Diffusion models [5, 10, 30, 32, 37] have established a new state-of-the-art in generative modeling, but they are fundamentally limited by their iterative sampling process, which incurs significant computational overhead. To achieve high-fidelity, real-time synthesis, researchers have explored various distillation strategies [25, 27, 33, 36, 44, 45, 47]. Among these, Distribution Matching Distillation (DMD) [44, 45] has been widely adopted due to its exceptional ability to enable high-fidelity generation in just one or a few steps.

However, traditional distillation objectives inherently bottleneck the performance of student models, as the optimization target is derived exclusively from a pretrained teacher [12, 22]. To break this performance ceiling, recent studies [11, 12, 24] have integrated Reinforcement Learning (RL) with diffusion distillation. However, these approaches typically rely on **a naive linear combination** of RL objectives and distillation losses. Departing from these paradigms, we adopt a fundamentally different perspective: **re-conceptualizing distribution matching as a reward**. This formulation allows the distillation objective to be seamlessly integrated and jointly optimized with task-specific rewards within a unified reward framework, as illustrated in Fig. 2.

Building upon this conceptual shift, we formally define distribution matching as R_{dm}. This transition reveals a critical challenge: as image noise increases, the

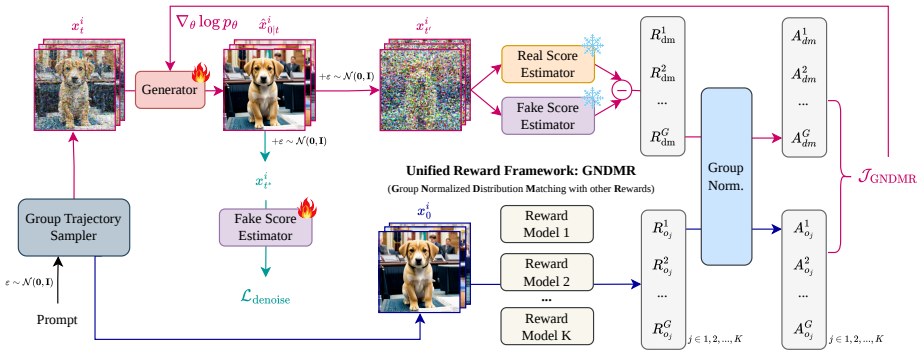


Fig. 2: Our unified reward framework GNDMR. After re-conceptualizing distribution matching as a reward, R_{dm} and other rewards perform GRPO simultaneously.

variance of R_{dm} amplifies significantly. This escalating variance destabilizes the estimated optimization direction and leads to inefficient training, a phenomenon we analyze in Sec. 4.2. To mitigate this, we draw inspiration from established RL practices and introduce Group Normalization (GN) to provide a stabilized, superior optimization gradient, resulting in Group Normalized Distribution Matching (GNDM). Moreover, framing DMD strictly as a reward maximization problem unlocks several critical advantages when integrating with other rewards: First, R_{dm} naturally informs the design of an effective adaptive weighting function to balance multiple objectives. Second, it seamlessly incorporates Importance Sampling (IS), which significantly improves sampling efficiency during training. Most importantly, this paradigm shift effectively constructs a algorithmic bridge between diffusion distillation and the broader RL ecosystem. We can now effortlessly leverage established RL techniques [6, 19, 20, 35] to further refine the distillation process under a single, unified mathematical umbrella.

Our contributions are summarized as follows:

- We propose R_{dm}, which natively incorporates powerful RL techniques into the diffusion distillation pipeline. This unification resolves the optimization conflicts inherent in prior joint-training methods and enables more intuitive control over training dynamics.
- We introduce Group Normalized Distribution Matching (GNDM) to provide high-fidelity directional guidance and propose a unified reward framework (GNDMR) to holisticlly optimize the distillation process.
- We conduct extensive experiments demonstrating that GNDM achieves superior distillation performance over vanilla DMD. Furthermore, our unified GNDMR framework surpasses existing baselines and yields a highly optimal balance between visual aesthetics and distillation efficiency.

2 Related Work

Distribution Matching Distillation. Distribution Matching Distillation (DMD) [45] is a foundation work that applies score-based distillation to large-scale diffusion models. A lot of follow-up work emerged to enhance its stability, theoretical grounding, and generation quality. DMD2 [44] integrates a GAN loss to eliminate the reliance on costly paired regression data. Flash-DMD [4] designs a timestep-aware strategy and incorporates pixel-GAN to achieve faster convergence and stable distillation. While numerous works need GAN to achieve better performance, TDM [26] combines trajectory distillation and distribution matching for better alignment and eliminates GAN. Decoupled DMD [18] mathematically decomposes the DMD objective, revealing that Classifier-Free Guidance (CFG) augmentation acts as the primary generative engine while distribution matching serves as a regularizer, enabling optimized decoupled noise schedules. DMDR [12] combines DMD with RL, utilizing the distribution matching loss as a regularization mechanism to safely allow the student generator to explore and ultimately outperform the teacher.

Reinforcement Learning for Diffusion Models. Reinforcement learning (RL) has been widely adopted to align diffusion models with human preferences. Various algorithms have been developed for this purpose: ReFL [43] achieves strong performance but inherently relies on differentiable reward models, whereas Direct Preference Optimization (DPO) [38] optimizes the policy using pairwise data. Alternatively, Denoising Diffusion Policy Optimization (DDPO) [2] requires only a scalar reward, a process that Group Relative Policy Optimization (GRPO) [19] further simplifies by eliminating the critic model via group normalization. Applying RL to distilled diffusion models is typically treated as an independent post-training phase: Pairwise Sample Optimization (PSO) [28] first adapted policy optimization for distilled models, and Hyper-SD [31] incorporated human feedback to further boost accelerated generation. Breaking away from decoupled training, DMDR [12] introduced the first framework to simultaneously optimize both DMD and RL objectives. Our proposed method is also built upon the DMDR[1] framework.

3 Preliminaries

3.1 Distribution Matching Distillation

The goal of Distribution Matching Distillation (DMD) [45] is to distill a multi-step diffusion model (teacher) into a high-fidelity, few-step generator (student) G_θ. The primary objective of DMD is Distribution Matching Loss (DML), which minimizes the reverse-KL divergence between the teacher's distribution p_{real} and student's distribution p_{fake}. The gradient of DML is:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{\varepsilon,t'} \left(s_{\text{real}}(x_{t'}) - s_{\text{fake}}(x_{t'}) \right) \nabla_\theta G_\theta(\varepsilon), \tag{1}$$

[1] We discuss only the GRPO-based variant of DMDR.

where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t' \sim \mathcal{U}(T_{\min}, T_{\max})$ and x'_t is the diffused sample obtained by injecting noise into $x_0 = G_\theta(\varepsilon)$ at diffused time step t'. s_{real} and s_{fake} are score functions given by score estimator μ_{real} and μ_{fake}. During training, μ_{fake} with parameter ψ is initialized with μ_{real}, and updating to track the distribution of G_θ through denoising diffusion objective:

$$\mathcal{L}_{\text{denoise}} = ||\mu^{\psi}_{\text{fake}}(x_{t'}, t') - x_0||_2^2. \tag{2}$$

In few-step generation training, $G_\theta(\varepsilon)$ is revised by backward simulation as introduced from DMD2 [44]. We follow SDE-based inference methods, starting from Standard Gaussian noise, it iteratively perform denoising $\hat{x}_{0|t} = G_\theta(x_t)$ and noising $x_{t-1} = \alpha_{t-1}\hat{x}_{0|t} + \sigma_{t-1}\varepsilon$. Thus, we have:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\alpha_{t-1}G_\theta(x_t), \sigma^2_{t-1}\mathbf{I}), \tag{3}$$

and we can redefine the gradient of one-step DML in Eq. (1) to multi-step DML:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{t', x_t \sim G_\theta} \left(s_{\text{real}}(x_{t'}) - s_{\text{fake}}(x_{t'})\right) \nabla_\theta G_\theta(x_t) \tag{4}$$

3.2 Denoising Diffusion Policy Optimization

Following Denoising Diffusion Policy Optimization (DDPO) [2], we map the denoising process to the following multi-step Markov decision process (MDP):

$$s_t \triangleq (t, x_t, c), \quad \pi(a_t \mid s_t) = p_\theta(x_{t-1} \mid x_t, c), \quad P(s_{t+1} \mid s_t, a_t) \triangleq \left(\delta_c, \delta_{t-1}, \delta_{x_{t-1}}\right),$$
$$a_t \triangleq x_{t-1}, \quad \rho_0(s_0) \triangleq \left(c, \delta_T, \mathcal{N}(\mathbf{0}, \mathbf{I})\right), \quad R(s_t, a_t) \triangleq \begin{cases} r(x_0, c), & t = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Where δ_x is Dirac delta distribution and T denoted the length of sampling trajectories.

After collecting denoising trajectories $\{x_T, x_{T-1}, ..., x_0\}$ and likelihoods $\log p_\theta$, we can use policy gradient estimator depicted in REINFORCE [29,41] to update parameter θ via gradient descent:

$$\nabla_\theta \mathcal{J}_{\text{DDPO}} = \mathbb{E}\left[r(x_0) \sum_{t=1}^{T} \nabla_\theta \log p_\theta(x_{t-1} \mid x_t)\right], \tag{5}$$

To overcome the limitation where optimization is confined to one step per sampling round due to the on-policy requirement of the gradient, we utilize an importance sampling estimator [13]. This formulation facilitates multi-step optimization by reweighting gradients from trajectories produced by θ_{old}:

$$\nabla_\theta \mathcal{J}_{\text{DDPO}_{\text{IS}}} = \mathbb{E}\left[r(x_0) \sum_{t=1}^{T} \frac{p_\theta(x_{t-1} \mid x_t)}{p_{\theta_{\text{old}}}(x_{t-1} \mid x_t)} \nabla_\theta \log p_\theta(x_{t-1} \mid x_t)\right] \tag{6}$$

In this context, the expectation is taken with respect to the denoising sequences generated under the previous parameter set θ_{old}.

Note that for ease of observation we omit the condition c in all formulas. Eq. (5) is similar in form to Eq. (4), which inspires us to regard term $(s_{\text{real}} - s_{\text{fake}})$ as a reward.

4 Methodology

4.1 R_{dm}: Distribution Matching as a Reward

The key to establishing the connection between Eq. (4) and Eq. (5) is uncovering the relationship between $\nabla_\theta G_\theta(x_t)$ and $\nabla_\theta \log p_\theta(x_{t-1}|x_t)$, which are intrinsically linked through Eq. (3). By taking the log-derivative of Eq. (3), we obtain the following identity:

$$\nabla_\theta \log p(x_{t-1} \mid x_t) = \frac{x_{t-1} - \mu_\theta(x_t)}{\sigma_{t-1}^2} \cdot \nabla_\theta \mu_\theta(x_t) \tag{7}$$

where $\mu_\theta(x_t) = \alpha_{t-1} G_\theta(x_t)$. Consequently, if we set

$$R_{\text{dm}}(x_t, x_{t-1}, t') = \frac{s_{\text{real}}(x_{t'}) - s_{\text{fake}}(x_{t'})}{x_{t-1} - \mu_\theta(x_t)} \cdot \frac{\sigma_{t-1}^2}{\alpha_{t-1}}, \tag{8}$$

the relationship between the $\nabla_\theta G_\theta(x_t)$ and $\nabla_\theta \log p_\theta(x_{t-1}|x_t)$ can be formulated as:

$$R_{\text{dm}}(x_t, x_{t-1}, t') \nabla_\theta \log p_\theta(x_{t-1}|x_t) = (s_{\text{real}}(x_{t'}) - s_{\text{fake}}(x_{t'})) \nabla_\theta G_\theta(x_t).$$

We define the policy gradient objective function for the distillation-based policy as:

$$\nabla_\theta \mathcal{J}_{\text{DDPO}_{\text{DM}}} = \mathbb{E}_{t', x_t \sim G_\theta} \left[R_{\text{dm}}(x_t, x_{t-1}, t') \nabla_\theta \log p(x_{t-1}|x_t) \right], \tag{9}$$

Comparing to Eq. (5), we use only one sample from the trajectory instead of all samples, ensuring consistency with the traditional DML as in Eq. (4). Finally, We have $\nabla_\theta \mathcal{J}_{\text{DDPO}_{\text{DM}}} = -\nabla_\theta \mathcal{L}_{DMD}$ strictly established.

4.2 Revisiting the Distribution Matching Reward

Next, we further explore the meaning of our new-defined distribution matching reward R_{dm}. By substituting $x_{t-1} - \mu_\theta(x_t) = \sigma_{t-1} \varepsilon^x$ (ε^x is the standard gaussian noise which sampled to obtain x_{t-1}), we can rewrite Eq. (8) as

$$R_{\text{dm}}(t, t') = R_s(t') \cdot \frac{\sigma_{t-1}}{\alpha_{t-1} \varepsilon^x}, \tag{10}$$

where $R_s(t') := s_{\text{real}}(x_{t'}) - s_{\text{fake}}(x_{t'})$.

The vector $R_s(t')$ functions as the critical guidance term that drives the generation of G_θ toward the manifold of p_{real} and away from p_{fake}.

However, $R_s(t')$ fluctuates across diffused samples $x_{t'}$ at timesteps t'. Specifically, as t' increases, the diffused sample $x_{t'}$ becomes increasingly blurred, resulting in more ambiguous directional information provided by $R_s(t')$, as shown in Fig. 3. Although recent works leverage the sample weighting mechanisms introduced by DMD [45] to enhance stability and fidelity, the variance of $R_s(t')$ is higher at larger values of t', which can mislead the optimization trajectory, thereby increasing the difficulty of effective model distillation. Further analysis in Sec. 5.3.

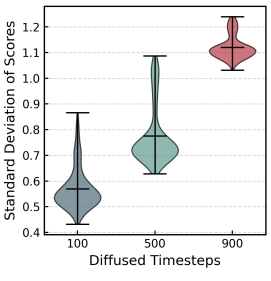


Fig. 3: Larger diffused timesteps, higher score variance.

Beyond the optimization challenges posed by the high variance at larger timesteps, the fundamental objective of the distribution matching reward R_{dm} diverges significantly from standard RL paradigms. Unlike conventional reward metrics such as HPS [42] or CLIP Score [8], where unbounded maximization often exacerbates reward hacking, R_{dm} essentially acts as a divergence measure that is optimized to approach zero. This stabilization dynamic is conceptually analogous to PREF-GRPO [40], where win-rate converges to 0.5. This fundamental shift from absolute maximization to distribution alignment introduces a critical advantage: R_{dm} provides an inherent safeguard against over-optimization, and it is a well-defined reward for regularization.

4.3 Group Normalized Distribution Matching

After re-concepting the distribution matching as a reward, we aim to address the inaccurate estimation issue in the calculation of $R_s(t')$. As Group Normalization (GN) on R_{dm} allows for a more stable estimation of the reward direction by the mean-subtraction mechanism, we naturally extend R_{dm} to a GRPO setting.

Recall the MDP formulation defined in Sec. 3.2. The generator G_θ samples a groups of G individual images $\{x_0^i\}_{i=1}^G$ and the corresponding trajectories $\{(x_T^i, x_{T-1}^i, ..., x_0^i)\}_{i=1}^G$. The advantage of i-th sample at trajectory step t and diffused timestep t' is

$$A_{\text{dm},t}^{i,t'} = \frac{R_{\text{dm}}(x_t^i, x_{t-1}^i, t') - \text{mean}(\{R_{\text{dm}}(x_t^i, x_{t-1}^i, t')\}_{i=1}^G)}{\text{std}(\{R_{\text{dm}}(x_t^i, x_{t-1}^i, t')\}_{i=1}^G)}. \tag{11}$$

The policy is updated by maximizing the Group Normalized Distribution Matching (GNDM) objective:

$$\mathcal{J}_{\text{GNDM}}(\theta) = \mathbb{E}_{t,t',\{x^i\}_{i=1}^G \sim G_{\theta_{\text{old}}}} \left[f(r, A_{\text{dm}}, \theta, \eta) \right], \tag{12}$$

where

$$f(r, A_{\text{dm}}, \theta, \eta) = \frac{1}{G}\sum_{i=1}^{G} \min\left(r_t^i(\theta)A_{\text{dm},t}^{i,t'}, \text{clip}\big(r_t^i(\theta), 1-\eta, 1+\eta\big)A_{\text{dm},t}^{i,t'}\right), \quad (13)$$

$$r_t^i(\theta) = \frac{p_\theta(x_t^i \mid x_{t-1}^i)}{p_{\theta_{\text{old}}}(x_t^i \mid x_{t-1}^i)}.$$

In each group, we share the generator timestep t and diffused timestep t', separately. It maximizes component diversity while ensuring consistency within the group. We prove it is effective in ablation study Sec. 5.3. Furthermore, we introduce importance sampling in Eq. (12). Since the denoising transition is Gaussian, the likelihood ratio $r_t^i(\theta)$ is available in closed form, allowing the generator to be updated multiple times from the same sampled trajectories. This reuse of trajectories improves sampling efficiency while maintaining performance, as discussed in Sec. 5.2.

4.4 GNDM with Other Rewards

To further improve the generative quality and alleviate mode seeking, following DMDR [12], we introduce other reward R_{o} in addition to the R_{dm}. As the same to Eq. (11), We employ GN to R_{o}:

$$A_{\text{o},t}^i = \frac{R_{\text{o}}(x_0^i) - \text{mean}(\{R_{\text{o}}(x_0^i)\}_{i=1}^G)}{\text{std}(\{R_{\text{o}}(x_0^i)\}_{i=1}^G)}. \quad (14)$$

Unlike $A_{\text{dm},t}^{i,t'}$, which is calculated as a dense reward on latent trajectory, $A_{\text{o},t}^i$ is calculated as a sparse reward on final image.

The total advantage of i-th sample $A_{\text{sum},t}^i$ at generation time t can be defined as the combination of the DM-derived advantage $A_{\text{dm},t}^{i,t'}$ and K auxiliary advantages $A_{\text{o}_j,t}^i$:

$$A_{\text{sum},t}^i = A_{\text{dm},t}^{i,t'} + \sum_{j=1}^{K} w_j A_{\text{o}_j,t}^i, \quad (15)$$

where w_j is the weighting function for j-th auxiliary reward. The final GRPO objective with multi-rewards is:

$$\mathcal{J}_{\text{GNDMR}}(\theta) = \mathbb{E}_{t,t',\{x^i\}_{i=1}^G \sim G_{\theta_{\text{old}}}}\left[f(r, A_{\text{sum}}, \theta, \eta)\right]. \quad (16)$$

4.5 Adaptive Weight Design in Practical Implementation

In practice, the term $x_{t-1} - \mu_\theta(x_t)$ introduces significant stochasticity and potential numerical instability, as x_{t-1} is obtained by adding gaussian noise from $\mu_\theta(x_t)$. To address this, we define a stabilized reward term R_{dm} by applying a sign-based normalization to the denominator to guarantee positive correlation. We also consider the weighting function proposed by DMD [45], as we found it

can improve distillation efficiency. The final distillation matching reward can be defined as

$$R_{\text{dm}}(x_t, x_{t-1}, t') = \frac{s_{\text{real}}(x_{t'}) - s_{\text{fake}}(x_{t'})}{\text{sign}(x_{t-1} - \mu_\theta(x_t))} \frac{CS}{||\mu_{\text{real}}(x_{t'}, t') - \hat{x}_{0|t}||_1}, \tag{17}$$

where $\text{sign}(x) = 1$ if $x > 0$, and -1 otherwise. C and S is the number of channels and spatial locations, respectively. After applying GN towards R_{dm}, we multiply the scalar w_{dm} with A_{dm} to maintain the original amplitude:

$$w_{\text{dm},t} = \frac{1}{|x_{t-1} - \mu_\theta(x_t)| + \epsilon} \frac{\sigma_{t-1}^2}{\alpha_{t-1}}. \tag{18}$$

This formulation ensures a stable optimization signal by reducing the variance inherent in the raw sampling residuals.

When incorporating other rewards, we found that multiplying the adaptive weight $\beta_{\text{dm},t}$ based on $w_{\text{dm},t}$ can effectively improve the target score without collapsing by keeping their amplitudes consistent:

$$\beta_{\text{dm},t} = \frac{||w_{\text{dm},t}||_1}{CS}. \tag{19}$$

As $w_{\text{dm},t}$ is pixel-wise, we keep it consistent with the other rewards' dimensions by averaging them out in the sample dimension. Intuitively, this treats the non-score terms in Eq. (10) as a dynamic magnitude-matching coefficient, so the auxiliary reward gradient stays on the same scale as the distribution-matching update during training. As discussed in Sec. 5.3, $\beta_{\text{dm},t}$ shows significance on both DMDR and GNDMR. Finally, we rewrite Eq. (15) with designed weight as

$$A^i_{\text{sum},t} = w_{\text{dm},t} A^{i,t'}_{\text{dm},t} + \beta_{\text{dm},t} \sum_j w_j A^i_{\text{o}_j,t}. \tag{20}$$

Algorithm 1 outlines the final training procedure. Additional details are provided in the supplementary materials.

5 Experiments

Experiment Setting. The distillation is conducted on the LAION-AeS-6.5+ [34] solely with its prompts. We use DFN-CLIP [7] and HPSv2.1 [42] as reward models by default, where HPS represents aesthetic quality and CLIP captures image-text alignment, jointly guiding DMD toward more diverse and semantically aligned modes. Meanwhile, to reduce sampling overhead, we propose a variant termed GNDMR-IS, which updates the generator twice for each sampling by importance sampling estimator. The trained distilled models are all flow-based [17, 21] and support inference on stochastic sampling [23]. We also consider cold start strategy for faster and stable convergence [12], more experiment details can be found in the supplementary materials.

Algorithm 1: GNDMR training procedure

Pretrained real diffusion model μ_{real}, number of prompts N, number of generated samples per prompt M (group size), number inference timesteps T. Distilled generator G_θ.

// Initialize generator and fake score estimators from pretrained model

$G_\theta \leftarrow \text{copyWeights}(\mu_{\text{real}})$, $\mu_{\text{fake}} \leftarrow \text{copyWeights}(\mu_{\text{real}})$

while *train* **do**

  // Sample trajectories

  Sample a batch of prompts $\mathcal{Q} = \{q_1, q_2, \ldots, q_N\}$

  for *each* $q \in \mathcal{Q}$ **do**

    Sample a group of M trajectories $\{(x_T^i, x_{T-1}^i, ..., x_0^i)\}_{i=1}^M$

  // Update fake score estimation model

  Sample generated timesteps t and diffused timesteps t'

  $x_{t'} = \text{forwardDiffusion}(\text{stopgrad}(G_\theta(x_t)), t')$

  $\mathcal{L}_{\text{denoise}} = \text{denoisingLoss}(\mu_{\text{fake}}(x_{t'}, t'), \text{stopgrad}(G_\theta(x_t)))$ // Eq. (2)

  $\mu_{\text{fake}} = \text{update}(\mu_{\text{fake}}, \mathcal{L}_{\text{denoise}})$

  // Compute rewards and advantages

  for *each* $q \in \mathcal{Q}$ **do**

    Sample generated timesteps t and diffused timesteps t'

    for $i = 1$ **to** M **do**

      $R_{\text{dm}}^i = R_{\text{dm}}(x_t^i, x_{t-1}^i, t')$, $R_o^i = RM(x_0^i)$ // Eq. (17)

      $A_{\text{dm},t}^{i,t'} = \text{GN}(R_{\text{dm}}^i)$, $A_{\text{o},t}^i = \text{GN}(R_o^i)$ // Eq. (11), Eq. (14)

      $A_{\text{sum},t}^i = \text{weightedAdd}(A_{\text{dm},t}^{i,t'}, A_{\text{o},t}^i)$ // Eq. (20)

  // Update generator

  for *train generator loop* **do**

    Use the same generated timesteps t when compute rewards

    $\mathcal{J}_{\text{GNDMR}}(\theta) = \text{GRPOLoss}(A_{\text{sum}}, r_t(\theta))$ // Eq. (16)

    $G_\theta = \text{update}(G_\theta, \mathcal{J}_{\text{GNDMR}}(\theta))$

Evaluation Metrics. To comprehensively evaluate our approach, we adopt a diverse set of metrics. Fine-grained image-text semantic alignment is measured via the Human Preference Score (HPS) v2.1 [42], while PickScore (PS) [15] gauges overall aesthetic quality and perceptual appeal. To capture broader nuances of human judgment such as object accuracy, spatial relations, and attribute binding, we incorporate the recently proposed Multi-dimensional Preference Score (MPS) [46]. Beyond perceptual assessments, CLIP Score (CS) [7] and FID [9] serve to quantitatively verify distillation effectiveness. We also compute FID-SD [39] by comparing the outputs of all baselines against images generated by the original pre-trained diffusion models.

5.1 Comparison with State-Of-The-Art (SOTA)

We validate the text-to-image generation performance in both aesthetics alignment and distillation efficiency on 10K prompts from COCO2014 [16] following the 30K split of Karpathy [14]. We compare our 4-step generative models GNDMR and its variant GNDMR-IS (update 2 times generator per sampling by importance sampling strategy) against SD3-Medium [5] and SD3.5-Medium [1], as well as other open-sourced SOTA distillation models. We reproduced DMD2 [44], as it serves as the foundational distribution-based model in this domain. Additionally, we implemented DMDR (with GRPO) [12], a pioneer method within the same category. To further evaluate our model's performance against RL techniques directly applied to the teacher model, we extended the application of Flow-GRPO [19] to the base model, positioning this as the ceiling of RL performance in this context.

Quantitative Comparison. As shown in Tab. 1, our quantitative analysis highlights the superiority of GNDMR across three key dimensions. Regarding **Aesthetics Alignment**, on both SD3 and SD3.5, GNDMR consistently outperforms existing 4-step models. It surpasses the teacher and achieves comparable aesthetic performance to this Flow-GRPO optimized model, underscoring the efficacy of our alignment formulation. Furthermore, it enables **Highly Efficient, Data-Free Distillation**. Operating without external image datasets, GNDMR achieves a strictly lower FID than the data-free DMDR baseline on both SD3 and SD3.5. For SD3, its remarkably low FID-SD indicates accurate teacher distribution matching without the aesthetic degradation seen in Hyper-SD, striking a superior balance between visual quality and distribution fidelity. Finally, our framework delivers **Reduced Training Cost**, the GNDMR-IS variant halves training expenses while maintaining highly competitive alignment, demonstrating the accelerated convergence and resource efficiency of our paradigm.

Qualitative Comparison. As shown in Fig. 4, GNDMR consistently achieves superior aesthetic quality, exhibiting richer details, enhanced color vibrancy, and stronger prompt alignment across diverse styles. Furthermore, our method significantly mitigates the visual artifacts prevalent in DMDR outputs. Our GNDMR demonstrates an optimal balance between high aesthetic appeal and clean image synthesis.

5.2 Importance Sampling Correction

One significant advantage of treating distribution matching as a reward is the ability to leverage the Importance Sampling (IS) estimator [13]. This enables multi-step updates for the student model from a single sampling iteration, effectively addressing the high sample demand typically associated with Reinforcement Learning (RL). Our experiments, conducted on SD3-Medium (512×512), employ HPSv2.1 as an auxiliary reward and report HPSv2.1 on HPDv2 test prompts [42].

As illustrated in Fig. 5a, while increasing the batch size (e.g., from 16×8 to 32×16) enhances reward optimization per training step, it substantially raises

Table 1: Comparison against state-of-the-art methods. * denotes our reproduced results. **Img-Free** represents whether the training requires external image data. **Cost** refers to the product of sample size and sampling iterations. The best results are marked in red, second-best in orange.

Method	NFE	Res.	Img-Free	HPS↑	PS ↑	MPS↑	CS↑	FID↓	FID-SD↓	Cost↓
Stable Diffusion 3 Medium Comparison										
Base Model (CFG=7)	50	1024	-	29.00	22.72	12.10	38.86	24.48	-	-
Flow-GRPO [19]* (CFG=7)	50	1024	-	30.35	22.90	12.56	38.52	26.63	12.39	-
Hyper-SD [31] (CFG=5)	8	1024	✗	27.20	21.90	11.22	37.76	26.94	10.09	-
LCM [23]	4	1024	✗	27.76	22.31	11.61	36.97	27.71	15.90	-
DMD2 [44]*	4	1024	✗	26.64	22.36	11.37	38.00	27.28	16.51	-
Flash-SD3 [3]	4	1024	✗	27.47	22.65	11.98	38.07	26.01	12.21	-
DMDR [12]*	4	1024	✓	29.50	22.77	11.98	38.10	29.10	14.11	-
GNDMR-IS	4	1024	✓	30.00	22.89	12.48	38.15	28.84	12.47	128*4k
GNDMR	4	1024	✓	30.37	22.88	12.53	38.20	28.02	12.21	128*8k
Stable Diffusion 3.5 Medium Comparison										
Base Model (CFG=3.5)	50	512	-	27.78	22.59	11.91	38.46	20.69	-	-
Flow-GRPO [19]* (CFG=3.5)	50	512	-	31.81	23.24	12.93	39.21	29.27	9.39	-
DMD2 [44]*	4	512	✗	30.44	22.92	12.73	38.59	26.64	14.63	-
DMDR [12]*	4	512	✓	30.83	23.07	12.80	38.22	26.05	16.73	-
GNDMR-IS	4	512	✓	30.88	22.94	12.86	38.39	25.60	16.68	128*3k
GNDMR	4	512	✓	31.25	23.15	12.93	38.59	24.44	13.93	128*6k

the sampling cost. By implementing 5 training iterations per sample with a clip range of $\eta = 0.5$, our 32×16 (w/ IS) configuration achieves a convergence rate and final performance comparable to the standard 32×16 setup, while requiring significantly fewer total samples even less than the baseline 16×8 configuration (see Fig. 5b).

Table 2: Ablation on Group Normalization (GN) on R_{dm}. We first only distill model for 500 iteration then continue training with different rewards.

Method	only R_{dm}	+HPS		+PS	
	FID↓	FID↓	HPS↑	FID↓	PS↑
GNDMR	23.07	24.47	30.22	22.32	22.76
w/o GN	24.94	25.40	30.39	24.61	22.71

Table 3: Ablation on sampling strategy (left) and timestep intervals of group normalization (right).

t'	t	FID ↓
		24.94
✓		24.34
✓	✓	23.07

Interval	FID ↓
[0, 300]	23.71
[300, 600]	23.46
[600, 1000]	23.43

5.3 Ablation Study

We explore the effectiveness of R_{dm} from multiple perspectives through ablation studies below. By default, we use SD3-Medium with 512×512 and first perform

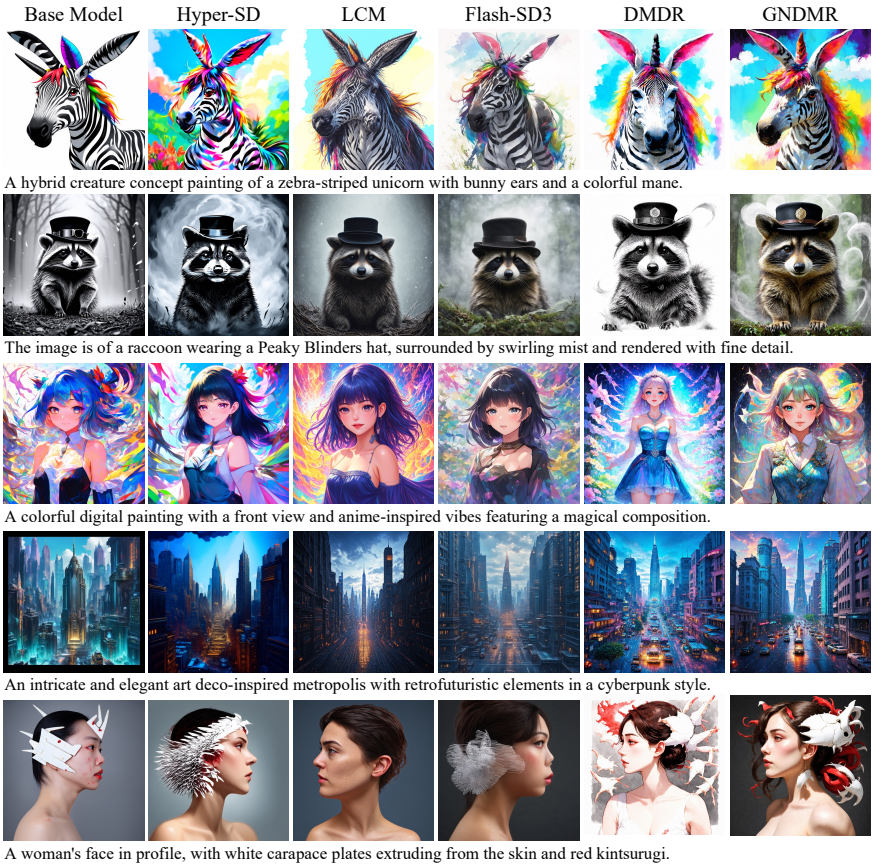
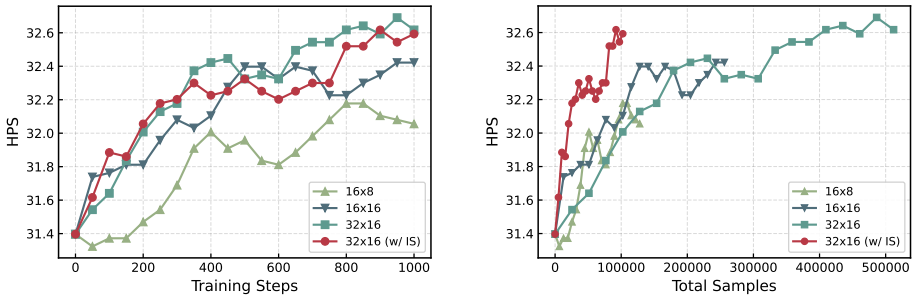


Fig. 4: Qualitative Results based on SD3. Our GNDMR has better aesthetics than other models and fewer artifacts than DMDR.

500 iterations vanilla DMD for fast training and observation, the results were reported on COCO30K.

Effect of group normalization on R_{dm}. The primary advantage of our GNDMR lies in applying Group Normalization (GN) to R_{dm}. To investigate this effect, we conduct the experiments shown in Tab. 2. During the first 500 distillation training iterations, GNDMR results in a lower FID, providing a better initialization for subsequent optimization. In the follow-up training, where HPS and PS are separately introduced as additional rewards, GNDMR continues to achieve lower FID while maintaining comparable performance on the target rewards. **Sampling strategy of generation timestep t and diffused timestep t'.** Since Group Normalization (GN) is applied to R_{dm}, the same t' is shared within each group, as t' determines the noise level of the diffused samples. We further examine whether sharing the same t within a group affects performance. The first row in Tab. 3 (left) corresponds to the vanilla DMD baseline.



(a) Equal Updates Comparison **(b)** Equal Sampling Budget Comparison

Fig. 5: Importance Sampling (IS) improves sampling efficiency. (a) With the same number of training steps, larger batch sizes lead to better reward optimization, but (b) they also require more samples. By introducing IS, the 32×16 (w/ IS) setting achieves comparable performance under a reduced sampling budget.

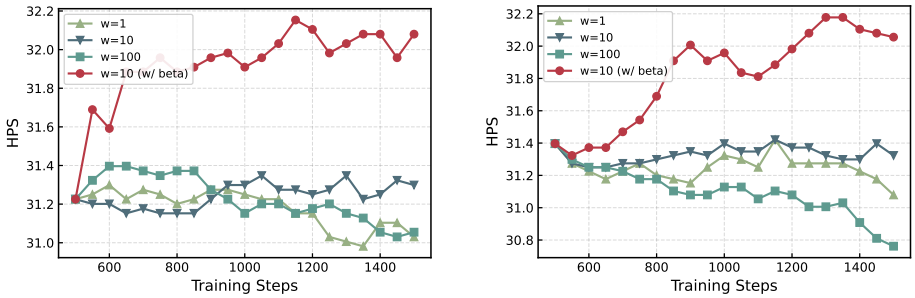
Sharing both t' and t within a group yields the best performance, consistent with Eq. (10), since R_{dm} depends on both t' and t.

Effect of diffused timestep intervals of group normalization on R_{dm}. As shown in Tab. 3 (right), larger timestep intervals correspond to higher noise levels, which increase the variance of R_{dm} estimation as shown in Fig. 3. Applying GN to regions with larger interval values effectively stabilizes the estimation and leads to lower FID.

Effect of $\beta_{\text{dm},t}$. As illustrated in Fig. 6, it is challenging to consistently improve the target reward through simple static weighting w. This difficulty arises because the distillation matching incorporates a dynamic coefficient $w_{\text{dm},t}$ that evolves during training. Without scaling the reward by the same level coefficient $\beta_{\text{dm},t}$, it is nearly impossible to maintain a proper balance between the two terms using a fixed weight w, often leading to unstable optimization or suboptimal reward gains. By contrast, applying $\beta_{\text{dm},t}$ to the reward term ensures a synchronized weighting scheme, allowing the reward to improve steadily without collapse.

6 Conclusion

In this work, we present a novel framework for improving diffusion distillation by re-conceptualizing distribution matching as a reward. Our approach bridges the gap between diffusion distillation and RL, resulting in more efficient and stable training. Through extensive experiments, we demonstrate that GNDM outperforms vanilla DMD, achieving a notable reduction in FID scores. Furthermore, the GNDMR framework, which integrates additional rewards, achieves an optimal balance between aesthetic quality and fidelity. Our method offers a flexible, efficient, and stable framework for real-time, high-fidelity image synthesis, while also providing a novel direction for the application of the latest RL techniques



(a) Effect of $\beta_{\text{dm},t}$ on DMDR **(b)** Effect of $\beta_{\text{dm},t}$ on GNDMR

Fig. 6: Ablation study on the effect of $\beta_{\text{dm},t}$ and the weighting factor w. "w/ beta" denotes the configuration using $\beta_{\text{dm},t}$ as defined in Eq. (19), whereas other variants set $\beta_{\text{dm},t} = 1$. The HPS is evaluated on the HPDv2 test prompts. The results demonstrate that incorporating $\beta_{\text{dm},t}$ leads to more stable and superior reward optimization compared to static weighting under both DMDR and GNDMR setting.

in diffusion model distillation, paving the way for future advancements in the field.

References

1. AI, S.: Sd3.5. https://github.com/Stability-AI/sd3.5 (2024)
2. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
3. Chadebec, C., Tasar, O., Benaroche, E., Aubin, B.: Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 15686–15695 (2025)
4. Chen, G., Huang, S., Liu, K., Zhu, J., Qu, X., Chen, P., Cheng, Y., Sun, Y.: Flash-dmd: Towards high-fidelity few-step image generation with efficient distillation and joint reinforcement learning. arXiv preprint arXiv:2511.20549 (2025)
5. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
6. Fan, J., Wei, T., Cheng, C., Chen, Y., Liu, G.: Adaptive divergence regularized policy optimization for fine-tuning generative models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
7. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
8. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Jia, Z., Nan, Y., Zhao, H., Liu, G.: Reward fine-tuning two-step diffusion models via learning differentiable latent-space surrogate reward. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12912–12922 (2025)
12. Jiang, D., Liu, D., Wang, Z., Wu, Q., Li, L., Li, H., Jin, X., Liu, D., Li, Z., Zhang, B., et al.: Distribution matching distillation meets reinforcement learning. arXiv preprint arXiv:2511.13649 (2025)
13. Kakade, S., Langford, J.: Approximately optimal approximate reinforcement learning. In: Proceedings of the nineteenth international conference on machine learning. pp. 267–274 (2002)
14. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
15. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in neural information processing systems **36**, 36652–36663 (2023)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
17. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
18. Liu, D., Gao, P., Liu, D., Du, R., Li, Z., Wu, Q., Jin, X., Cao, S., Zhang, S., HOI, S., Li, H.: Decoupled DMD: CFG augmentation as the spear, distribution matching as the shield. In: The Fourteenth International Conference on Learning Representations (2026)
19. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. Advances in neural information processing systems **38**, 40783–40818 (2026)
20. Liu, S.Y., Dong, X., Lu, X., Diao, S., Belcak, P., Liu, M., Chen, M.H., Yin, H., Wang, Y.C.F., Cheng, K.T., Choi, Y., Kautz, J., Molchanov, P.: GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. In: Forty-third International Conference on Machine Learning (2026)
21. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
22. Lu, Y., Ren, Y., Xia, X., Lin, S., Wang, X., Xiao, X., Ma, A.J., Xie, X., Lai, J.H.: Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16818–16829 (2025)
23. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
24. Luo, W.: Diff-instruct++: Training one-step text-to-image generator model to align with human preferences. arXiv preprint arXiv:2410.18881 (2024)
25. Luo, W., Huang, Z., Geng, Z., Kolter, J.Z., Qi, G.j.: One-step diffusion distillation through score implicit matching. Advances in Neural Information Processing Systems **37**, 115377–115408 (2024)
26. Luo, Y., Hu, T., Sun, J., Cai, Y., Tang, J.: Learning few-step diffusion models by trajectory distribution matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17719–17728 (2025)

27. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14297–14306 (2023)
28. Miao, Z., Yang, Z., Lin, K., Wang, Z., Liu, Z., Wang, L., Qiu, Q.: Tuning timestep-distilled diffusion model using pairwise sample optimization. In: The Thirteenth International Conference on Learning Representations (2025)
29. Mohamed, S., Rosca, M., Figurnov, M., Mnih, A.: Monte carlo gradient estimation in machine learning. Journal of Machine Learning Research **21**(132), 1–62 (2020)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
31. Ren, Y., Xia, X., Lu, Y., Zhang, J., Wu, J., Xie, P., Wang, X., Xiao, X.: Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. Advances in neural information processing systems **37**, 117340–117362 (2024)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
33. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
34. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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)
35. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
36. Song, Y., Dhariwal, P.: Improved techniques for training consistency models. In: The Twelfth International Conference on Learning Representations (2024)
37. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
38. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
39. Wang, F.Y., Huang, Z., Bergman, A., Shen, D., Gao, P., Lingelbach, M., Sun, K., Bian, W., Song, G., Liu, Y., et al.: Phased consistency models. Advances in neural information processing systems **37**, 83951–84009 (2024)
40. Wang, Y., Li, Z., Zang, Y., Zhou, Y., Bu, J., Wang, C., Lu, Q., Jin, C., Wang, J.: Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. arXiv preprint arXiv:2508.20751 (2025)
41. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning **8**(3), 229–256 (1992)
42. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
43. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
44. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)

45. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
46. Zhang, S., Wang, B., Wu, J., Li, Y., Gao, T., Zhang, D., Wang, Z.: 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)
47. Zhou, M., Wang, Z., Zheng, H., Huang, H.: Guided score identity distillation for data-free one-step text-to-image generation. In: The Thirteenth International Conference on Learning Representations (2025)