Distribution Matching Distillation Meets Reinforcement Learning

Dengyang Jiang[1,2], Dongyang Liu[4,2], Zanyi Wang[4], Qilong Wu[2], Liuzhuozheng Li[1], Hengzhuang Li[1], Xin Jin[2], Changsheng Lu[1], Zhen Li[2], Mengmeng Wang[3], Steven Hoi[2], Peng Gao[2], and Harry Yang[1⋆]

[1]HKUST [2]Alibaba Group [3]ZJUT [4]CUHK

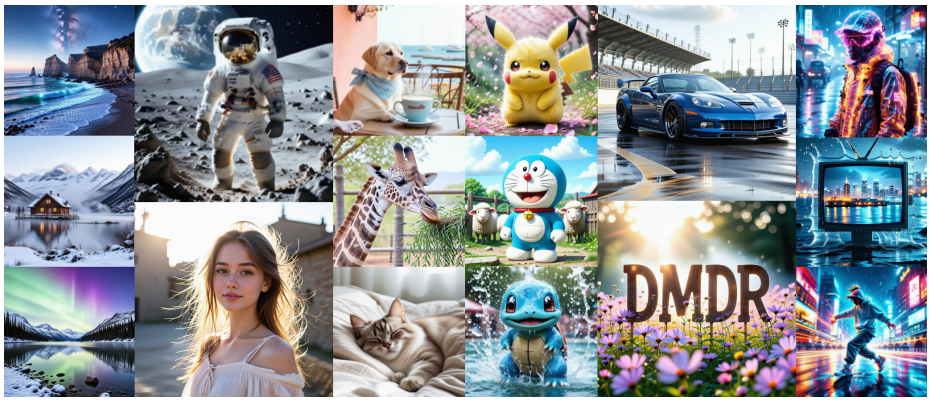


Fig. 1: Images generated by SD3.5 Large finetuned through our DMDR using only 4 NFE. See more in Appendix.

Abstract. Distribution Matching Distillation (DMD) facilitates efficient inference by distilling multi-step diffusion models into few-step variants. Concurrently, Reinforcement Learning (RL) has emerged as a vital tool for aligning generative models with human preferences. While both represent critical post-training stages for large-scale diffusion models, existing studies typically treat them as independent, sequential processes, leaving a systematic framework for their unification largely unexplored. In this work, we demonstrate that jointly optimizing these two objectives yields mutual benefits: RL enables more preference-aware and controllable distillation rather than uniformly compressing the full data distribution, while DMD serves as an effective regularizer to mitigate reward hacking during RL training. Building on these insights, we propose DMDR, a unified framework that incorporates Reward-Tilted Distribution Matching optimization alongside two dynamic distillation training strategies in the initial stage, followed by the joint DMD and RL optimization in the second stage. Extensive experiments demonstrate that DMDR achieves state-of-the-art visual quality and prompt adherence among few-step generation methods, even surpassing the performance of its multi-step teacher model.

⋆ Corresponding author

Keywords: Diffusion Model · Step Distillation with Distribution Matching Distillation· Reinforcement Learning

1 Introduction

Diffusion models pre-trained on large-scale datasets have achieved remarkable performance in visual generation tasks [8, 22, 39, 40, 52]. However, their sampling process typically requires numerous iterative denoising steps [16, 25, 50], each involving a full forward pass through a large neural network. This requirement makes high-resolution text-to-image synthesis computationally demanding. Furthermore, base models often fail to produce images that align closely with human preferences and aesthetic criteria [20, 36, 55].

To enhance practical utility, researchers typically employ two post-training stages. First, step-distillation [4, 31, 33, 42, 60, 61], which compresses the original model into a generator capable of few-step sampling (e.g., 4 steps) for efficient sampling. Among such methods, Distribution Matching Distillation (DMD) [61] is widely recognized for its effectiveness in large-scale scenarios [7, 11, 29, 60] and has been adopted in prominent industrial applications [23, 47, 52]. Second, reinforcement learning (RL) [3, 27, 28, 49, 53, 57, 59, 66] aligns the model with human preferences to improve aesthetic quality. Despite progress in both domains, most existing studies address these objectives in isolation. Early efforts to incorporate RL into the distillation process were largely confined to a two-stage paradigm [9, 32, 37, 42] (exploring how to conduct RL on an already distilled model). Consequently, there remains a critical gap in the field: *a systematic framework to explore the unification and potential synergy between these two "temporarily separated" stages.*

To bridge this gap, we propose DMDR, a unified framework that seamlessly integrates DMD and RL into a synergistic training pipeline. Our core insight is that concurrent optimization yields mutual benefits: (1) Preference-Aware Distillation: RL steers the distillation process toward high-reward regions, breaking the "performance ceiling" inherent in purely mimicking a teacher; (2) Distributional Regularization: The DMD objective serves as a robust regularizer for RL, mitigating reward hacking by maintaining proximity to the teacher's manifold. Specifically, DMDR operates in two stages. In the initial stage, we implement Reward-Tilted Distribution Matching (RT-DM), which distills toward a reward-tilted teacher distribution to prioritize human-preferred regions. Concurrently, we introduce Dynamic Distribution Guidance (DynaDG) and Dynamic Renoise Sampling (DynaRS) to stabilize the "cold-start" phase through annealed distributional overlap maximization. In the second stage, we transition to a joint optimization paradigm. In this phase, the RL objective drives aggressive preference alignment, while the DMD mechanism serves as a robust regularizer to prevent reward hacking.

Our experimental results show that our method leads to state-of-the-art few-step generative models. Moreover, our method is not only compatible with different models (e.g., flow-based, denoising-based) but also with various RL algo-

rithms (e.g., ReFL, DPO, GRPO). This ensures the long-term effectiveness of our method as the multi-step model and RL algorithm evolve.

In summary, our main contributions are as follows:

- We propose DMDR, which shows that DMD and RL can be trained simultaneously with mutual benefits.
- We design a Reward-Tilted Distribution Matching and two dynamic distillation training strategies for better integrating RL and distillation in the initial phase.
- We validate our method on various models and RL algorithms, all of which achieve excellent performance.

2 Related Work

We highlight key related studies here, and discuss others in Appendix.

Distribution matching distillation. Distribution Matching Distillation [61] (DMD) is the first work to successfully apply the principle score-based distillation [41, 51, 54] to large-scale text-to-image models. Intuitively, it strives to ensure that any sample realized by the student at a given noise level occurs with exactly the same probability as it would under the teacher's distribution, thereby preserving the multi-step generative priors in the few-step model. From then on, a lot of follow-up work emerged [2, 5, 7, 29, 34, 58, 60], for example, DMD2 [60] employs a discriminator to align the student model distribution with a specific target distribution like GAN [13] does; TDM [34] incorporates DMD loss in the sampling process of the student model for better alignment; f–distill [58] replaces the original reverse Kullback–Leibler (KL) divergence to the proposed f-divergence for covering different divergences with different properties. Our work also starts with DMD, but we don't focus on how to better "imitate" the teacher. Instead, we aim to incorporate reinforcement learning in the distillation process to enable a more controllable and preference-aligned distillation.

Reinforce learning for diffusion models. Reinforce learning helps to align diffusion models to human preferences by training a reward model and using it to guide generation [3,27,49,53,57,59,66]. There are many algorithms to conduct RL, for example, DDPO [3] adapts PPO via image log-likelihoods; ReFL [57] bypasses likelihoods by optimizing outputs with frozen-reward gradients; Diffusion-DPO [53] adapts DPO to diffusion for paired human preference data; and recent GRPO extensions [27,59] use GRPO to diffusion models by coupling the training loss with SDE samplers. Although significant progress has been made in RL for diffusion models, most of the work has focused on multi-step models. Here, we find that these algorithms are also adapted to the few-step model when carried out in conjunction with distribution matching distillation.

3 Method

3.1 Preliminary

Distribution matching distillation. DMD [61] compresses a multi-step diffusion model (teacher) into a few-step generator (student) G by minimizing the time-averaged approximate Kullback-Leibler (KL) divergence between the real distribution $p_{\text{real},t}$ and the synthetic distribution $p_{\text{fake},t}$. Note that G is optimized via gradient descent, and this gradient admits a compact expression as the difference of two score functions:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{dmd}} &= \mathbb{E}_t[\nabla_\theta \,\text{KL}(p_{\text{fake},t} \| p_{\text{real},t})] \\ &= -\mathbb{E}_t\left[\int\Big(s_{\text{real}}\big(F_t\big)\big) - s_{\text{fake}}\big(F_t\big)\Big)\frac{\mathrm{d}G_\theta(z)}{\mathrm{d}\theta}\,\mathrm{d}z\right],\end{aligned} \tag{1}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$, θ denotes the parameters of G, and F_t is the forward diffusion operator that injects noise at time t to $G_\theta(z)$. The quantities s_{real} and s_{fake} are the score functions estimated by diffusion models μ_{real} and μ_{fake}, respectively. During training, the μ_{fake} is updated via diffusion loss $\mathcal{L}_{\text{diff}}$ (for denoising-based models, the prediction target is noise, while for flow-based ones, it is velocity) on synthetic samples produced by the few-step generator. Moreover, as introduced in DMD2 [60], $G_\theta(z)$ is a noisy synthetic image produced by the current G running several steps, which is called backward simulation. The generator G then denoises these simulated images, and the outputs are supervised with the above loss functions.

Reinforce learning for diffusion models. Although there are various RL algorithms [27,53,57], All their goal is to optimize the diffusion model to generate the samples that can maximize a score given by reward models [14, 20, 55]. This score can be expressed as aesthetic, texture, prompt following, and so on. At the same time, a regularization term is often added to balance the trade-off between reward maximization and the deviation from the original model [67]. Thus, the loss can be formulated as follows:

$$\mathcal{L}_{\text{rl}} = \mathbb{E}\left[r(\boldsymbol{x}_0) - \text{KL}(p_{new} \| p_{\text{ref}})\right], \tag{2}$$

where r is the reward function, p_{new} and p_{ref} is the distribution of the trainable model and the reference (this can be a pre-train data distribution [57] or the original model's [27, 53, 66]), $\boldsymbol{x}_0$ is the clean image that is generated by the training generation model.

3.2 Problem Formulation and Pipeline Overview.

It is noted that DMD effectively reduces the sampling steps of the diffusion model, while RL enables the diffusion model to generate instruction-aligned and human-preferred images. Both of these are crucial for the practical application of the diffusion model. However, at present, the majority of the work [27, 29, 48, 59, 60, 66] only focuses on studying and improving one aspect, without exploring the

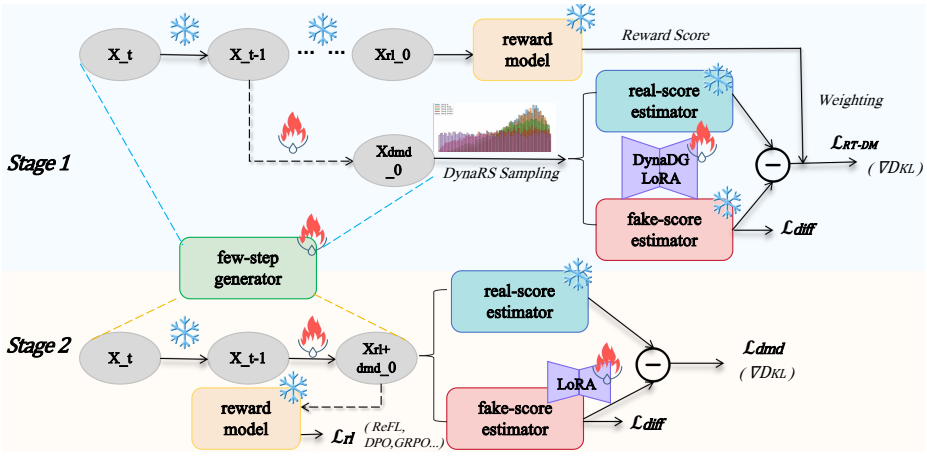


Fig. 2: Overview of DMDR. It follows a two-stage training paradigm: Stage 1 (Reward-Tilted Distribution Matching) employs reward-weighted distillation to integrate preference signals into the early distillation phase, stabilized by Dynamic Distribution Guidance (DynaDG) and Dynamic Renoise Sampling (DynaRS). Stage 2 (Joint RL + DMD) performs direct reward maximization regularized by the DMD loss.

synergy between the two. And only a small portion of the works [5, 9, 32, 37, 42] attempt to introduce RL into the distillation process, but they still focus on how to conduct RL on a distilled model. This raises a question: *Can these two be combined into one stage and benefit each other?*

We address this challenge question by proposing DMDR, as shown in Figure 2, the training process of DMDR is divided into two operational stages: Stage 1: RT-DM with Dynamic Distillation. (Section 3.4 and Section 3.5) Stage 2: Joint RL + DMD Optimization(Section 3.3). The following sections will provide the detailed analysis and illustration. And the detailed pseudo-code of DMDR is provided in Appendix.

3.3 Bring DMD and RL Together with Benefit

We now show formally and empirically that DMD operates as a superior regularization mechanism that subsumes the traditional KL penalty in RL for few-step model, yielding robust joint optimization. From the perspective of RL, standard preference optimization requires a penalty (e.g., $\mathrm{KL}(p_{new} \| p_{\text{ref}})$) to prevent reward hacking [27, 57]. In few-step models, sequentially applying RL after distillation uses the already-distilled distribution $P_{\text{fake-ref}}$ as the regularizer (Fig. 3). Because distillation inherently loses some modes, $P_{\text{fake-ref}}$ acts as an impoverished anchor, leading to rapid overfitting and mode collapse [42]. More fundamentally, instead of viewing DMD and RL as two disparate losses added heuristically, we reframe the joint objective as maximizing reward under a distribution-level structural constraint. As the joint gradient can be decomposed into two orthogonal objectives [46, 51, 67]: a "reward ascent" direction

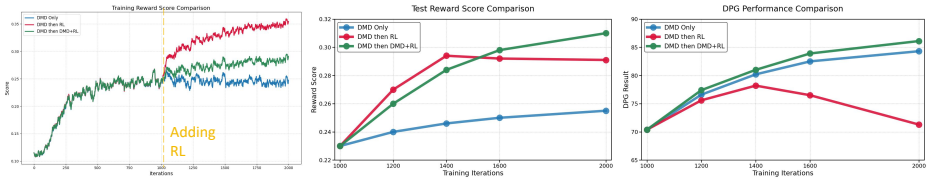


Fig. 4: Verification of the joint optimization synergy. Directly conduct RL (red) on the distilled model exhibits high training rewards but suffers from reward hacking and semantic collapse (low DPG score). In contrast, the joint DMD+RL paradigm (green) effectively boosts the performance (better than the DMD baseline (blue) in all metric) while preventing reward hacking. Better to zoom in to check the effect.

that increases human preference, and a "manifold projection" term that corrects off-manifold displacement by pulling the student back toward the teacher's regions.The DMD gradient $\nabla_\theta \mathcal{L}_{\text{dmd}}$ corresponds exactly to the maximum likelihood update projecting the student output distribution toward the learned data distribution of the multi-step teacher P_{real}. When combined linearly with the reward gradient $\nabla_\theta \mathcal{L}_{\text{rl}}$, any displacement induced by r that exits the teacher's high-density manifold generates a large discrepancy $s_{\text{real}} - s_{\text{fake}}$. The $\mathcal{L}_{\text{dmd}}$ gradient thus projects the reward ascent strictly along the teacher's support.

Simultaneously, from the perspective of distillation, pure DMD forces the student to uniformly mimic the teacher, implicitly establishing a "performance ceiling" capped by distilling P_{real} equally. Incorporating the RL objective addresses this fundamental limitation: RL steers the distillation toward high-reward regions, allowing the student to surpass the teacher while the DMD regularizer ensures semantic integrity is preserved.

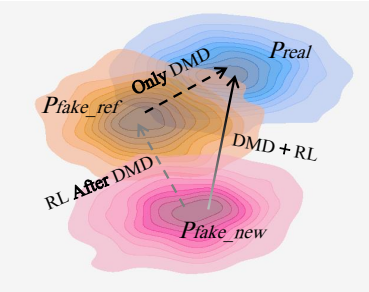


Fig. 3: Visualization of the difference of optimization directions.

To validate the analysis, we conduct a comparative study using SD3-M [8] as the base model, with rewards provided by HPS v2.1 [55] and ReFL-based RL [57] for training. We compare three paradigms: (i) DMD Only (Blue line); (ii) DMD then RL (Sequential: distillation followed by RL in red line); and (iii) DMD then DMD+RL (Joint approach: distillation followed by joint optimization in green line). We evaluate these models across training/test reward curves and the DPG_Bench score [18]. As shown in Fig. 4(a), when RL is introduced (after 1000 iterations), the sequential approach achieves the highest training reward. However, this gain is obtained through hacking score. Fig. 4(b) and (c) reveal that this it suffers from severe reward hacking and catastrophic forgetting. While its training reward climbs, its test reward plateaus, and its DPG score collapses. This confirms that without the strong distributional anchor of the teacher, the few-step model quickly drifts into narrow, high-reward modes that sacrifice semantic in-

tegrity [42]. In contrast, the joint optimization demonstrates a clear synergistic effect. First, it effectively breaks the "performance ceiling" of the teacher; the test reward significantly surpasses the DMD-only baseline. Second, by treating the DMD loss as a persistent structural regularizer, the model keeps improving DPG score while increasing aesthetic quality. This indicates that the DMD objective prevents the RL process from hacking, while the RL objective guides the distillation toward more human-preferred regions.

Based on the empirical synergy observed above, we formally define the optimization objective. The total loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dmd}} + \lambda_{\text{rl}}\mathcal{L}_{\text{rl}}, \tag{3}$$

where $\mathcal{L}_{\text{dmd}}$ is the original distribution matching loss with the gradient formulated in Eq. 1, and $\mathcal{L}_{\text{rl}}$ is the plug-and-play reward maximization loss from the RL branch (Eq. 1), the gradient can be differentiable type like ReFL, or policy type like GRPO, etc. The hyperparameter λ_{rl} is a balancing coefficient used to trade off the two losses.

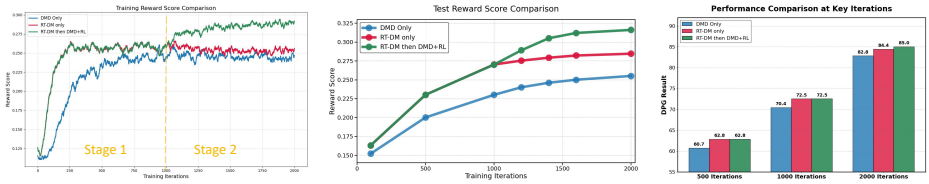
Table 1: Comparison of different training strategies. The joint approach avoids the "distillation gap" of "RL then distillation" pipeline, achieving the highest preference score with $\sim 10\times$ rollout speedup.

Method	HPS	DPG	Time
SD3-M, resolution 1024×1024, local batchsize 16 on H100			
Multi-Step RL	31.06	86.43	9.24s $\sim$ 10.57s
Multi-Step RL then Distillation	30.42	85.92	-
Few-Step Joint RL and DMD	31.42	86.08	0.43s $\sim$ 1.21s

3.4 Reward-Tilted Distribution Matching for Integrating RL into the Early Phase of DMD

While we have demonstrated that DMD and RL can be trained jointly with mutual benefit, we still rely on a two-stage paradigm that incorporates RL only after an initial "cold start" distillation phase [32, 42]. This requirement stems from the observation that reward models typically necessitate a baseline generation capability to produce reliable feedback [49, 57]. Specifically, in DMD frameworks with flow matching model, the image used for gradient computation is defined as $\boldsymbol{x}_t - (0-t)G_\theta(\boldsymbol{x}_t)$, where $\boldsymbol{x}_t$ is iteratively sampled. Notably, t can be any training time-step; thus, when training a 4-step model, the clean image can be generated in only one to four steps. In the early training phase, however, such few-step sampling often fails to produce coherent images, rendering reward signals noisy and unreliable for direct optimization.

To address this, we propose a Reward-Tilted Distribution Matching (RTDM) paradigm to "softly" integrate reward signals into the distillation process. Specifically, we utilize the reward score as a weighting factor for the DMD loss. This formulation enables a more flexible optimization objective compared to direct reward maximization. This is because the reward score serves as a static weighting factor, we avoid the necessity of backpropagating through the multi-step denoising process, which allows us to evaluate the reward on the complete trajectory of the few-step model (e.g., 4 steps) without the computational overhead of a large gradient graph, bypassing the need for single-step $\boldsymbol{x}_0$ approximations to make the reward score more precise and highly valuable as a reference



(a) Training Reward Score Comparison (b) Test-set Reward Score Comparison (c) DPG Score Comparison

Fig. 5: Effectiveness of RT-DM and two-stage training strategy. In Stage 1, the RT-DM paradigm (red/green) significantly accelerates preference alignment compared to the DMD baseline (blue). In Stage 2, while "RT-DM only" (red) eventually saturates due to the constraints of the teacher's manifold, the transition to joint DMD+RL (green) successfully breaks this performance plateau by directly optimizing the reward score. This complete two-stage approach achieves the highest test rewards and DPG scores. Better to zoom in to check the effect.

for optimization efforts. More profoundly, we are changing *which* distribution the student learns by smoothly integrating reward signals to make the student not learn toward the vanilla P_{real}, but toward a *reward-tilted teacher distribution* defined proportionally to $r(x_0)p_{\text{real}}$. The gradient is formulated as:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{RT-DM}} &= \mathbb{E}_t[\nabla_\theta \,\text{KL}(p_{\text{fake},t} \| (p_{\text{real},t}) r(\boldsymbol{x}_0))] \\ &= -\mathbb{E}_t\left[\int e^{\frac{r(\boldsymbol{x}_0)}{\|r(\boldsymbol{x}_0)\|}} \Big(s_{\text{real}}\big(F_t\big) - s_{\text{fake}}\big(F_t\big)\Big) \frac{\mathrm{d}G_\theta(z)}{\mathrm{d}\theta}\,\mathrm{d}z\right].\end{aligned} \tag{4}$$

To evaluate the effectiveness of the proposed soft integration, we compare three configurations in Fig. 5: (i) DMD Only (Blue); (ii) RT-DM for the entire duration (Red); and (iii) RT-DM followed by joint DMD+RL (Green). As shown in Fig. 5(a) and (b), the RT-DM paradigm (Red/Green) provides a significant "warm start" effect, accelerating preference alignment in the early phase (Stage 1) compared to the DMD-only baseline. This confirms that weighting the DMD loss with reward scores successfully incorporates preference signals even when few-step generations are still maturing, avoiding the noise instability of direct RL. However, the performance of the "RT-DM only" approach (Red) eventually saturates. We hypothesize that since RT-DM primarily acts by re-weighting the teacher's distribution gradients, its ability to drive the model toward high-reward regions is inherently constrained by the teacher's manifold. To further unlock the model's potential, we introduce a more direct reward maximization phase (Green) in Stage 2. By transitioning to joint DMD+RL optimization once the model has reached a stable baseline, we can leverage direct RL gradients to aggressively push the performance beyond the initial plateau. As a result, the joint approach achieves the highest test rewards and DPG scores, demonstrating that RT-DM serves as an ideal bridge to transition from pure distillation to direct preference alignment.

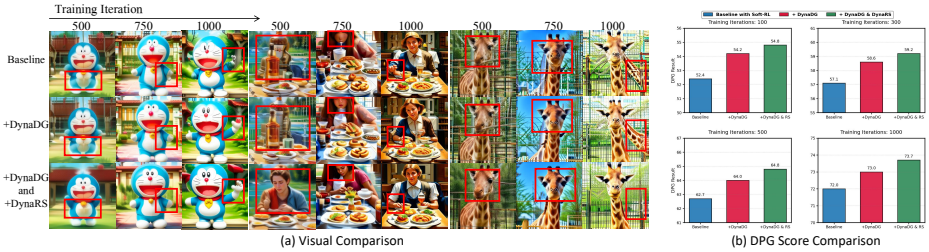


Fig. 7: Impact of dynamic training strategies. (a) Qualitative comparisons show that our dynamic strategies facilitate faster global structure building (highlighted in red boxes). (b) Quantitative DPG scores further confirm that DynaDG and DynaRS provide more reliable optimization signals, consistently outperforming the baseline throughout Stage 1. Better to zoom in to check the effect.

3.5 Annealed Distributional Overlap Maximization for Better Initial Distillation

During the initial phase of distillation, DMD encounters a classic "cold start" challenge: the synthetic samples generated by the few-step student exhibit substantial divergence from the real-world pre-training samples. This profound disjointness between P_{fake} and P_{real} impedes the teacher model's ability to estimate reliable scores, rendering the gradient $\nabla \log P_{\text{real}} - \nabla \log P_{\text{fake}}$ unreliable for optimization. To mitigate this, we introduce a unified principle of *annealed distributional overlap maximization*. We propose two complementary techniques to artificially increase the overlap between P_{fake} and P_{real} early in training, progressively relaxing them as the model converges.

Dynamic Distribution Guidance (DynaDG). DynaDG maximizes overlap primarily by actively shifting the target distribution. The poor initial Image quality causes P_{fake} to be disjoint from P_{real} (Figure 6, left). To solve this, we inject trainable LoRA modules [17] into the real score estimator to "pull" the perceived P_{real} toward the nascent P_{fake} (Figure 6, right). This ensures a reliable optimization signal by bridging the manifold gap. As training progresses, we anneal the LoRA scale in the teacher model toward zero, allowing the real score estimator to seamlessly revert to the true P_{real} once P_{fake} is safely anchored.

Dynamic Renoise Sampling (DynaRS). While DynaDG moves the distributions, DynaRS maximizes overlap by exploiting regions where distributions naturally intersect. By heavily biasing the sampling of renoise levels t toward higher noise values at the start of training, we ensure that both initial samples and reference samples are dominated by Gaussian noise, placing them in a similar regime.

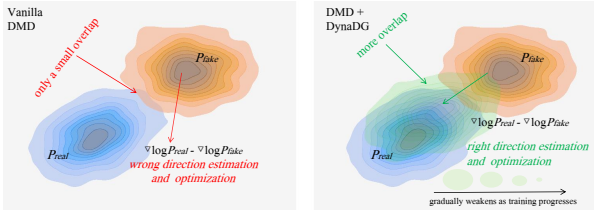


Fig. 6: Illustration for Dynamic Distribution Guidance.

At these high noise levels, the score estimator provides highly reliable gradients that emphasize global structure [64, 65]. As the student generator improves, the renoise bias is progressively annealed to uniform sampling, allowing the model to capture fine-grained textural details from a wider range of noise levels.

As shown in Figure 7 (b), our dynamic training strategies significantly enhance the initial training phase. Qualitative results presented in Figure 7 left further corroborate these findings, showing a faster global structure building when DynaDG and DynaRS are employed.

3.6 Distinction from Prior Few-step RL Works

Some prior works also try to introduce RL for the few-step diffusion model [9, 30, 42, 44], but they mainly perform it after the distillation, and the problem mentioned by Hyper-SD [42] of quickly deviating from the original domain still occurs. Flash-DMD [5] and Diff-Instruct++ [32] are the close-related works to ours. However, they both follow a "distill-then-distill+RL" paradigm (the former adds DPO-style RL after DMD2, while the latter adds ReFL-style RL after Diff-Instruct) because early few-step samples would provide unreliable reward signals for direct RL. We solve this by proposing RT-DM, which enables RL from the beginning of distillation by using reward scores as static weights for the DMD loss to achieve soft guidance and by evaluating rewards on samples generated with more denoising steps along the same trajectory, yielding more reliable reward signals. The effectiveness is verified in Figure 5.

4 Experiments

4.1 Experimental Setup

We provide a brief overview of our training and evaluation schemes here and give very detailed and specific implementation settings for each model in Appendix. For training, we adopt denoising-based model (SDXL-Base [40]) and flow-based models (SD3-Medium [8], SD3.5-Large [1]) for distillation, using prompts from t2i-2M. Additionally, we use ReFL [57] with DFN-CLIP [10] and HPSv2.1 [55] as reward models by default. As for evaluation, we validate the effect of DMDR through extensive experiments. First, we use prompts sampled from a recent public dataset ShareGPT-4o-Image [6] to generate images and follow previous work [29, 34] that report CLIP Score [14], Aesthetic Score [45], Pick Score [20], and Human Preference (HP) Score [55]. Next, in order to obtain a more comprehensive evaluation, we also compare our distilled models with their teachers on two popular benchmarks, DPG_Bench [18] and GenEval [12]. Noted that there are some differences from the settings we made in the Method section. We train 1.5K steps in stage1 and stage 2, respectively, to achieve better performance.

Method for comparison. We compare our method against several categories of existing work, including foundational multi-step base models [1, 8, 40], RL only approaches [57], distillation only methods [4, 24, 31, 34, 60], distillation then RL

Table 2: System-level comparison against state-of-the-art methods. ⋆ denotes our reproduced results based on the official code. Best performance is marked in **Bold**.

Method	NFE	Post-Train	CLIP Score↑	Aesthetic Score↑	Pick Score↑	HP Score↑
SDXL-Base						
Base-Model	50	-	34.7588	5.6480	22.1085	27.1477
ReFL [57]⋆	50	RL	35.2107	5.9045	22.8076	**33.1874**
LCM [31]	2	Distill	28.4664	5.1026	20.0603	17.6837
Lightning [24]	1	Distill	32.0283	5.6761	21.4868	26.3615
DMD2 [60]	1	Distill	34.3046	5.6238	21.7293	26.5986
DMDR (ours)	1	Distill & RL	**35.6241**	**6.1184**	**22.8498**	32.0065
DMD2 [60]	4	Distill	34.5169	5.7043	22.1546	28.5655
DMD2-PSO [37]⋆	4	Distill then RL	34.0128	5.8032	22.2644	29.8732
DMDR (ours)	4	Distill & RL	**35.4940**	**6.0324**	**22.7122**	32.9832
SD3-Medium						
Base-Model	50	-	34.9025	5.5942	22.1801	28.4021
ReFL [57]⋆	50	RL	**35.1851**	5.7821	22.0064	32.0745
DMD2 [60]⋆	4	Distill	33.9421	5.6137	21.7688	27.3675
Flash [4]	4	Distill	34.2634	5.6702	21.5921	26.6542
TDM [34]	4	Distill	34.0301	5.6250	22.0010	27.7522
Hyper-SD [42]	8	Distill then RL	32.0234	5.2489	20.2831	22.4544
DMDR (ours)	4	Distill & RL	35.0142	**5.8876**	**22.4892**	**32.1145**
SD3.5-Large						
Base-Model	50	-	35.5509	5.7014	22.4856	28.8135
LADD [44]	4	Distill then RL	35.0480	5.4514	22.2451	27.8470
DMDR (ours)	4	Distill & RL	**35.9757**	**6.1541**	**22.9072**	**32.7368**

frameworks [37, 42]. We do not compare with methods like Flash-DMD [5] and Diff-Instruct++ [32] that do not provide model weights or open source training code because we cannot guarantee a fair and reproducible evaluation.

4.2 Main Results

We begin by assessing DMDR's text-to-image generation performance using prompts sampled from ShareGPT-4o-Image [6], ensuring that there is no overlap with the prompts utilized during training. Table 2 presents a summary of the distillation results in comparison to other methods. Models distilled through our approach exhibit state-of-the-art (SOTA) capabilities in terms of prompt coherence and the generation of high-quality, aesthetically pleasing images. Notably, our method demonstrates effectiveness across various architectures (e.g., UNet, Transformer), model sizes (e.g., 2B, 8B), and paradigms (e.g., flow-based, denoising-based), highlighting the broad applicability and universality of the proposed approach. Additionally, we provide qualitative comparisons in Figure 8, where our method produces images characterized by superior quality and enhanced prompt coherence.

To further verify our method's effectiveness and provide a more comprehensive comparison between the few-step distillation model and the multi-step teacher model, we conduct the comparison on two commonly used benchmarks [12,

Table 3: GenEval Comparison: 4-step (ours) *vs*. its multi-step teacher.

Model	Overall	Single	Two	Count.	Colors	Pos.	Attr.
SDXL-Base							
Teacher	0.55	0.98	0.74	0.39	0.85	0.15	0.23
Ours	**0.57**	**0.99**	**0.76**	**0.44**	**0.85**	0.12	**0.25**
SD3-Medium							
Teacher	0.62	0.98	0.74	0.63	0.67	0.34	0.36
Ours	**0.65**	**0.99**	**0.84**	0.57	**0.81**	0.27	**0.45**
SD3.5-Large							
Teacher	0.71	0.98	0.89	0.73	0.83	0.34	0.47
Ours	**0.72**	**0.99**	**0.93**	0.69	0.80	0.31	**0.59**

Table 4: DPG_Bench Comparison: 4-step (ours) *vs*. its multi-step teacher.

Model	Overall	Global	Entity	Attribute	Relation	Other
SDXL-Base						
Teacher	74.65	83.27	82.43	80.91	86.76	80.41
Ours	**76.48**	**83.72**	**82.58**	**83.69**	84.76	**83.54**
SD3-Medium						
Teacher	84.08	87.90	91.01	88.83	80.70	88.68
Ours	**85.30**	**90.49**	90.07	**90.56**	**87.17**	**91.23**
SD3.5-Large						
Teacher	84.12	91.48	90.22	87.81	91.20	89.49
Ours	**85.33**	90.47	**90.53**	**90.68**	87.44	**90.20**

Fig. 8: Visual comparison between the teachers, selected competing methods [42, 44, 60], and ours. All images are generated using identical noise. Our model produces images with superior quality and prompt coherence.

18], Although we do not perform RL optimization on the metric of these two benchmarks, table 4 and 3 show that the overall score of our four-step model consistently outperforms its multi-step teacher on DPG_Bench and GenEvalacross all base models. These results indicate that DMDR has successfully freed the student model from the constraints of the multi-step teacher model and stimulated its capabilities during the distillation process.

We also included a user study in Appendix, which further demonstrates the superiority of our method in terms of human preferences when integrating RL into the distillation process.

4.3 Ablation Study

In this section, we conduct a systematic investigation into the key components of DMDR. Unless otherwise specified, all experiments use the SD3-M base model with a 4-step sampling configuration.

Table 5: Ablation on different components. Default settings are marked in purple.

(a) Ablation on different reward models.

Method	CLIP Score	Aesthetic Score	Pick Score	HP Score	DPG overall
Pick [20]	34.3403	5.9021	22.9674	31.7665	84.86
AE [45]	34.1748	6.1024	22.3858	31.6620	84.61
CLIP [14]	35.2087	5.5832	22.0644	28.0507	85.10
HPS [55]	34.4088	5.9021	22.5108	32.9065	85.04
CLIP + HPS	35.0142	5.8876	22.4892	32.1145	85.30

(b) Ablation on RL algorithm in stage 2.

Method	CLIP Score	Aesthetic Score	Pick Score	HP Score	DPG overall
stage 1 init	*33.3648*	*5.5744*	*21.0070*	*28.8371*	*82.4371*
w/ RL (ReFL)	35.0142	5.8876	22.4892	32.1145	85.30
w/ RL (DPO)	34.0234	5.8340	21.9844	30.8352	85.00
w/ RL (GRPO)	34.6675	5.9324	22.3248	31.4474	85.08

(c) Ablation on Value of λ_{rl}.

Value	CLIP Score	Aesthetic Score	Pick Score	HP Score	DPG overall
0.1	34.7422	5.8508	22.3064	31.6565	85.24
0.5	35.0142	5.8876	22.4892	32.1145	85.30
1.0	35.0087	5.84430	22.4709	32.8982	85.04
2.0	35.0322	5.7806	22.1098	32.7683	82.48

Versatility across reward models. A key advantage of ourusing a DINOv2-based reward model to maximize classification accuracyDMDR is the ability to steer the distillation process toward diverse human preferences. As shown in Table 5a, we evaluate the framework using different reward models (RMs) targeting specific attributes. The results provide a clear insight: optimizing with a specific RM leads to a significant performance gain in its corresponding metric, confirming that DMDR effectively injects the desired preference signal into the few-step manifold. Furthermore, by integrating multiple rewards (CLIP + HPS), the model achieves a more balanced and superior generative performance, similar to the findings in DanceGRPO [59]. This flexibility allows users to customize the distillation "flavor" according to specific downstream tasks.

Compatibility with RL paradigms. To verify the algorithmic universality of our unified framework, we compare three distinct RL strategies in the Stage 2 joint optimization phase: ReFL [57], DPO [37, 53], and GRPO [27, 59]. As shown in Table 5b, all three algorithms yield substantial improvements over the "Stage 1 init" baseline across all evaluation dimensions. Furthermore, an interesting observation is that using the ReFL in DMDR achieves overall better performance, and our analysis is that: traditional ReFL for multi-step models ignores earlier sampling steps and only trains the last few steps before the output image [57] because the large computational graph generated by the multi-step denoising process would result in a huge memory consumption. Subsequent solutions [49, 56] mainly focus on saving forward noise trajectory and sampling a subset of steps to optimize, although this can allow the gradient to return to the early denoising step, it is also prone to error accumulation. On the contrary, in our few-step model, the model is trained to predict a clean image at each step's state [60], which enables the gradient of ReFL to directly return to the initial state, thus obtaining an effective optimization.

Table 6: 1-step SiT DMDR Results on ImageNet 256×256. We report the performance with different CFG scales and RL optimization. Metrics include FID [15], Recall [21], Inception Score (IS) [43], and Precision [21].

CFG	RL	FID ↓	Recall ↑	Inception Score ↑	Precision ↑
No	No	2.13	0.64	232.15	0.77
1.5	No	5.20	0.48	387.24	0.88
4	No	15.70	0.11	453.65	0.86
No	Yes	6.95	0.40	416.01	0.90
1.5	Yes	13.43	0.24	494.43	0.93

The balancing Role of DMD Regularization. The synergy between distillation and RL is governed by the balancing coefficient λ_{rl} in our joint objective. Table 5c illustrates the trade-off between reward maximization and distributional stability. When λ_{rl} is set too low; the model remains constrained by the teacher's manifold, limiting the potential for preference alignment. Conversely, as it increases beyond an optimal threshold, we observe a noticeable decline in a general capacity (reflected by the DPG overall score), despite high reward scores. This trend provides a crucial insight: the DMD loss serves as a vital structural regularizer. It prevents the model from "reward hacking"(We also showed in Figure 4. The optimal balance ensures that the model pushes the boundaries of quality while staying anchored to the teacher's robust generative priors.

5 Limitation and Discussion

Quality vs. Diversity. To further analyze the behavior of DMDR, we conduct class-conditional experiments on ImageNet 256×256 using the SiT [19, 35, 62] backbone for 1-step generation and text-to-image settings used in the former paper. As shown in Table 6, a fundamental trade-off between generative quality and diversity emerges. Our results reveal that both Classifier-Free Guidance (CFG) (this also observed in Decoupled-DMD [26]) and Reinforcement Learning (here we use a DINOv2-based [38] reward model to maximize classification accuracy) act as "density sharpeners" during the distillation process.

We believe this reduction in diversity is not a limitation unique to our approach, but a characteristic of score-based distillation and preference alignment, which steer the model toward high-reward, high-confidence regions. We also provide additional experiments with LPIPS diversity [63] on SD3-medium in Table 7, which shows similar degradation as on ImageNet and for other methods, while DMDR provides a favorable quality/alignment trade-off. Although DMD regularization mitigates reward hacking compared with direct/sequential RL (shown in Figure 4), it cannot completely remove the inherent diversity cost of reward optimization. This trade-off is tunable: a smaller λ_{rl} improves diversity with reduced preference gains as shown in Table 8, allowing users to balance the trade-off. More explorations on adaptive reward weighting and other diversity-preserving RL objectives could be future work.

Table 7: Diversity evaluation on ShareGPT-4o-Image [6].

Metric	Base-Model	DMD	DMD2	Base + ReFL	DMD + PSO	DMDR
HPS↑	28.4021	28.0248	27.3675	32.0745	30.1180	32.1145
LPIPS↑	0.6840	0.5664	0.5832	0.5602	0.5540	0.5548

Table 8: Impact of the trade-off λ_{rl}.

Metric	Base-Model	DMD	DMDR $\lambda_{\text{rl}}0.1$	DMDR $\lambda_{\text{rl}}0.5$	DMDR $\lambda_{\text{rl}}1$
HPS↑	28.4021	28.0248	30.4024	32.1145	32.4786
LPIPS↑	0.6840	0.5664	0.5602	0.5548	0.5266

6 Conclusion

We introduces DMDR, a unified framework that transforms Distribution Matching Distillation (DMD) and Reinforcement Learning (RL) from independent, sequential stages into a synergistic training pipeline. We demonstrate that joint optimization is mutually beneficial: RL steers distillation toward high-reward regions to break the "teacher ceiling," while DMD regularizes RL to effectively mitigate reward hacking. By employing a two-stage strategy—incorporating RTDM and dynamic training followed by joint optimization—DMDR achieves state-of-the-art few-step performance across diverse architectures. Ultimately, our approach enables few-step models to not only match but consistently surpass the visual quality and prompt adherence of their original multi-step teacher models.

7 Acknowledgment

This work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under the Early Career Scheme (Project No. 26210225).

We also want to thank David Liu, Ruoyi Du, Aiming Hao, Xiangpeng Yang, Bin Fu, Bo Zhang, Jiajun Zha, and Sizhe Dang for the helpful discussions and suggestions during the progress of DMDR project.

References

1. AI., S.: Sd3.5. `https://github.com/Stability-AI/sd3.5` (2024)
2. Bandyopadhyay, H., Entezari, R., Scott, J., Adithyan, R., Song, Y.Z., Jampani, V.: Sd3. 5-flash: Distribution-guided distillation of generative flows. arXiv preprint arXiv:2509.21318 (2025)
3. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
4. 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)

5. Chen, G., Huang, S., Liu, K., Zhu, J.X., 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 **abs/2511.20549** (2025)
6. Chen, J., Cai, Z., Chen, P., Chen, S., Ji, K., Wang, X., Yang, Y., Wang, B.: Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. arXiv preprint arXiv:2506.18095 (2025)
7. Cheng, Z., Sun, P., Li, J., Lin, T.: Twinflow: Realizing one-step generation on large models with self-adversarial flows. arXiv preprint arXiv:2512.05150 (2025)
8. 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)
9. Eyring, L., Karthik, S., Dosovitskiy, A., Ruiz, N., Akata, Z.: Noise hypernetworks: Amortizing test-time compute in diffusion models. arXiv preprint arXiv:2508.09968 (2025)
10. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
11. Ge, X., Zhang, X., Xu, T., Zhang, Y., Zhang, X., Wang, Y., Zhang, J.: Seneflow: Scaling distribution matching for flow-based text-to-image distillation. arXiv preprint arXiv:2506.00523 (2025)
12. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
14. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
15. 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)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
18. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
19. Jiang, D., Wang, M., Li, L., Zhang, L., Wang, H., Wei, W., Dai, G., Zhang, Y., Wang, J.: No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. arXiv preprint arXiv:2505.02831 (2025)
20. 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)
21. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in neural information processing systems **32** (2019)
22. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
23. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025)

24. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)
25. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
26. Liu, D., Gao, P., Liu, D., Du, R., Li, Z., Wu, Q., Jin, X., Cao, S., Zhang, S., Li, H., et al.: Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. arXiv preprint arXiv:2511.22677 (2025)
27. 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. arXiv preprint arXiv:2505.05470 (2025)
28. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Qin, W., Xia, M., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
29. 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. arXiv preprint arXiv:2507.18569 (2025)
30. Lu, Y., Xia, X., Zhang, M., Kuang, H., Zheng, J., Ren, Y., Xiao, X.: Hyper-bagel: A unified acceleration framework for multimodal understanding and generation. arXiv preprint arXiv:2509.18824 (2025)
31. 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)
32. Luo, W.: Diff-instruct++: Training one-step text-to-image generator model to align with human preferences. arXiv preprint arXiv:2410.18881 (2024)
33. Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., Zhang, Z.: Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. Advances in Neural Information Processing Systems **36**, 76525–76546 (2023)
34. Luo, Y., Hu, T., Sun, J., Cai, Y., Tang, J.: Learning few-step diffusion models by trajectory distribution matching. arXiv preprint arXiv:2503.06674 (2025)
35. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)
36. Ma, Y., Wu, X., Sun, K., Li, H.: Hpsv3: Towards wide-spectrum human preference score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15086–15095 (2025)
37. Miao, Z., Yang, Z., Lin, K., Wang, Z., Liu, Z., Wang, L., Qiu, Q.: Tuning timestep-distilled diffusion model using pairwise sample optimization. arXiv preprint arXiv:2410.03190 (2024)
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
39. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
40. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

41. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
42. 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. arXiv preprint arXiv:2404.13686 (2024)
43. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
44. Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., Rombach, R.: Fast high-resolution image synthesis with latent adversarial diffusion distillation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
45. 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)
46. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
47. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
48. Shao, S., Yi, H., Guo, H., Ye, T., Zhou, D., Lingelbach, M., Xu, Z., Xie, Z.: Magicdistillation: Weak-to-strong video distillation for large-scale few-step synthesis. arXiv preprint arXiv:2503.13319 (2025)
49. Shen, X., Li, Z., Yang, Z., Zhang, S., Zhang, Y., Li, D., Wang, C., Lu, Q., Tang, Y.: Directly aligning the full diffusion trajectory with fine-grained human preference. arXiv preprint arXiv:2509.06942 (2025)
50. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
51. 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)
52. Team, Z.I.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)
53. 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)
54. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in neural information processing systems **36**, 8406–8441 (2023)
55. 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)
56. Wu, X., Hao, Y., Zhang, M., Sun, K., Huang, Z., Song, G., Liu, Y., Li, H.: Deep reward supervisions for tuning text-to-image diffusion models. In: European Conference on Computer Vision. pp. 108–124. Springer (2024)
57. 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)
58. Xu, Y., Nie, W., Vahdat, A.: One-step diffusion models with f-divergence distribution matching. arXiv preprint arXiv:2502.15681 (2025)

59. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
60. 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)
61. 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)
62. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: International Conference on Learning Representations (2025)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
64. Zhang, Y., Dong, W., Tang, F., Huang, N., Huang, H., Ma, C., Lee, T.Y., Deussen, O., Xu, C.: Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. ACM Transactions on Graphics (TOG) **42**(6), 1–14 (2023)
65. Zhang, Z., Zhang, Q., Lin, H., Xing, W., Mo, J., Huang, S., Xie, J., Li, G., Luan, J., Zhao, L., et al.: Towards highly realistic artistic style transfer via stable diffusion with step-aware and layer-aware prompt. arXiv preprint arXiv:2404.11474 (2024)
66. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. arXiv preprint arXiv:2509.16117 (2025)
67. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)