Diff-Instruct with Diffused Reward: Towards Principled One-step Generator RL

Junyi Wu[1] **Weijian Luo**[2] **Haoyang Zheng**[1] **Ruizhe Zhang**[1] **Guang Lin**[1]
[1]Purdue University [2]Independent Researcher
{wu2393,zheng528,rzzhang,guanglin}@purdue.edu, pkulwj1994@icloud.com

Abstract

Recent advances in one-step text-to-image generation have enabled real-time synthesis with remarkable efficiency and quality. Previous reinforcement learning methods for one-step generators combine reward optimization in image space with distribution matching in the diffusion noise space. This paradigm brings challenges due to a mismatch between terminal reward optimization and the underlying generative dynamics. As a result, optimization tends to exploit stochastic degrees of freedom, often improving reward at the expense of image fidelity. To address this issue, we propose **Diff-Instruct with Diffused Reward** (DIDR), a data-free trajectory-level alignment framework derived from Integral KL minimization. DIDR propagates the RLHF-optimal reward-tilted clean-image distribution across all noise levels along the diffusion trajectory. We show that this objective admits the same minimizer as clean-image RLHF, while naturally inducing **the Diffused Reward Score** (DRS), which acts as a reward-driven correction to the reference score function. To make this practical, we further introduce **the Diffused Reward Proxy** (DRP), an efficient estimator of DRS based on differentiable short-step denoising. Extensive experiments demonstrate that DIDR consistently Pareto-dominates existing one-step SDXL baselines. Moreover, when transferred to a 6B DiT backbone (Z-IMAGE), DIDR surpasses its 50-step teacher in preference alignment with a single generation step and outperforms the 8-step Z-Image-Turbo with only 4 steps. Code is available at `https://github.com/Junyi-W/DIDR`.

1 Introduction

Deep generative models have achieved remarkable progress in text-to-image (T2I) synthesis [Rombach et al., 2022, Saharia et al., 2022, Ramesh et al., 2022, Esser et al., 2024, Nichol and Dhariwal, 2021, Karras et al., 2020], driving the development of one-step generators [Luo et al., 2023, Yin et al., 2024a, Zhou et al., 2024a, Sauer et al., 2023, Kang et al., 2023, Zheng and Yang, 2024, Liu et al., 2024, Xu et al., 2024] that map latent noise directly to images in a single forward pass via diffusion distillation [Luo, 2023] and GAN-based techniques [Goodfellow et al., 2014, Sauer et al., 2024, Zheng et al., 2026a]. While these models achieve real-time synthesis, they frequently exhibit suboptimal aesthetic quality and poor alignment with human preferences—a central requirement for real-world deployment.

Diff-Instruct[Luo et al., 2023] opens the door to train one-step image generators by distribution matching distillation[Yin et al., 2024b] by minimizing an Integral Kullback-Leibler divergence. Later works based on concepts of distribution matching explored the reinforcement learning post-training of one-step text-to-image generators by combining image-space rewards with latent space distribution matching [Luo, 2024, Luo et al., 2025]. However, these attempts lack a trajectory-level target directly induced by the KL-regularized RLHF optimum. These methods apply preference signals exclusively at the clean image output x_0 while imposing KL regularization over the full noisy trajectory. This creates a structural mismatch: the reward acts at the clean-image level while the regularizer spans all

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: 1024×1024 images from one-step SDXL aligned via DIDR. Prompts in Appendix I.

noise levels, so the combined objective does not correspond to any proper RLHF-induced target on the generator distribution. We refer to the resulting tendency to over-tilt toward high-reward outputs as **terminal reward domination**, and make it precise in a tractable example (§3.1). In contrast, DIDR derives a principled trajectory-level target directly from the KL-regularized RLHF optimum, ensuring that reward and regularization are balanced at every noise level.

To close this gap, we propose **Diff-Instruct with Diffused Reward** (DIDR), a principled reinforcement learning framework for one-step image generators without using image data. Our core insight is to systematically propagate reward signals defined on clean images to every noise level of the diffusion trajectory. We realize this by starting from the KL-regularized RLHF objective, identifying its optimal clean-image target as the reward-tilted density $q^*(x_0|c) \propto q_0(x_0|c)\exp(r(x_0, c)/\tau)$, and diffusing this target through the reference process to obtain trajectory-level reward-tilted marginals $q_t^*(x_t|c)$. Minimizing the Integral KL (IKL) divergence against these marginals yields a target score that decomposes into the reference score plus a reward-induced correction, the **Diffused Reward Score (DRS)**, defined at every noise level through reference posterior.

Since the DRS involves an intractable posterior expectation, we further derive the **Diffused Reward Proxy (DRP)**: a differentiable estimator that runs short-step denoising chains from x_t using the frozen reference model, propagating reward gradients stably back through the trajectory to yield a practical, data-free algorithm. Empirically, compared with Diff-Instruct* [Luo et al., 2025] (the strongest prior one-step SDXL baseline), $\text{DIDR}_{\text{longer}}$ improves PickScore from 23.1 to 23.9 and ImageReward from 1.01 to 1.10 (+8.9%); the standard DIDR achieves the best PickScore–FID trade-off among one-step methods. We also scale DIDR to the 6B Z-Image backbone, where it surpasses the 50-step base in a

Table 1: **Comparison of DIDR with representative distillation and reward-alignment methods.** *Noise-level reward*: reward defined on x_t.

Method	Data-free	Few-step	Reward alignment	Noise-level reward
Diff-Instruct [Luo et al., 2023]	✓	✓	×	×
Diff-Instruct++ [Luo, 2024]	✓	✓	✓	×
Diff-Instruct* [Luo et al., 2025]	✓	✓	✓	×
DMD2 [Yin et al., 2024a]	×	✓	×	×
Flow-GRPO [Liu et al., 2025]	✓	×	✓	×
AWM [Xue et al., 2025]	✓	×	✓	×
DiffusionNFT [Zheng et al., 2026b]	✓	×	✓	×
DIDR (Ours)	✓	✓	✓	✓

single step and outperforms the 8-step Z-Image-Turbo with 4 steps. Table 1 summarizes how DIDR differs from prior approaches. We summarize our contributions as follows:

- We identify *terminal reward domination* (a structural mismatch in prior endpoint-only methods where optimizers exploit noise to collapse onto high-reward regions) and propose DIDR, a principled trajectory-level alignment framework. Grounded in our proof of Integral KL equivalence to KL-regularized RLHF, DIDR rigorously propagates the RLHF-optimal, reward-tilted clean-image density forward to define a principled alignment target at *every* noise level.
- We analytically derive DRS, which exactly decomposes the trajectory target score into a reference score plus a reward-induced correction. To make this tractable, we introduce DRP, a zero-data estimator that computes this correction via differentiable short-step posterior denoising through a frozen reference model, enabling stable reward-gradient propagation across the entire trajectory.
- Extensive experiments demonstrate that DIDR sets a new state-of-the-art for one-step alignment without requiring any image training data. On one-step SDXL at 1024×1024, standard DIDR Pareto-dominates all existing one-step baselines in the PickScore–FID trade-off, while $\text{DIDR}_{\text{longer}}$ pushes PickScore to 23.9 and ImageReward from 1.01 to 1.10 ($+8.9\%$). Furthermore, DIDR successfully transfers to a 6B DiT-based Z-Image backbone, surpassing its 50-step teacher on preference metrics in a single step and outperforming the 8-step Z-Image-Turbo with 4 steps.

2 Preliminaries

Diffusion Trajectories and One-step Generators. Let x_0 denote a clean image drawn from the reference data distribution $q_0(x_0|c)$, conditioned on a text prompt c. A diffusion model defines a forward noising process that gradually corrupts x_0 into pure noise over a continuous time variable $t \in [0, T]$. This is characterized by a transition kernel $q_t(x_t|x_0)$, which induces the noisy marginal distributions $q_t(x_t|c) = \int q_t(x_t|x_0)q_0(x_0|c)\mathrm{d}x_0$. To reverse this process for image generation, one needs the Stein score of the marginals, $\nabla_{x_t} \log q_t(x_t|c)$, a vector field pointing toward higher data density. In practice, a neural network $s_{\text{ref}}(x_t, t, c)$ is trained to approximate this score via Denoising Score Matching (DSM) [Vincent, 2011, Song et al., 2021]. In our framework, this trained multi-step diffusion model serves as the *reference model* and is kept frozen.

While the reference model generates high-quality images, its iterative sampling process is computationally expensive. To achieve real-time synthesis, a one-step generator [Goodfellow et al., 2014, Song et al., 2023] bypasses the iterative ODE/SDE integration entirely. It maps a standard Gaussian noise vector $z \sim p_z$ directly to a clean image in a single forward pass: $x_0 = g_\theta(z, c)$. This mapping induces an implicit clean-image distribution, denoted as $p_{\theta,0}(x_0|c)$. By hypothetically injecting the same forward noise into these generated samples, we can define the generator's noisy trajectory $p_{\theta,t}(x_t|c) = \int q_t(x_t|x_0)p_{\theta,0}(x_0|c)\mathrm{d}x_0$. This distinction between the reference trajectory q_t and the generator trajectory $p_{\theta,t}$ is crucial, as our objective relies on matching their reward-tilted dynamics.

Trajectory Distillation and Terminal-reward Alignment. Standard one-step distillation, such as Diff-Instruct [Luo et al., 2023], trains one-step generators by minimizing an Integral KL (IKL) divergence across the entire diffusion trajectory, anchoring the generator's noisy marginals to the bare reference marginals $\{q_t\}_{t\in[0,T]}$. To incorporate human preferences, recent methods like Diff-Instruct* (DI*, [Luo et al., 2025]) and Diff-Instruct++ (DI++, [Luo, 2024]) extend this objective by naively

appending a terminal reward. For a generic divergence $\mathcal{D}$, the objective becomes:

$$\mathcal{L}_{\text{term}}(\theta) = -\mathbb{E}_{c,\, x_0 \sim p_{\theta,0}(x_0|c)}[r(x_0, c)] + \tau \int_0^T w(t)\, \mathcal{D}\big(p_{\theta,t}(x_t|c) \| q_t(x_t|c)\big) \mathrm{d}t. \tag{1}$$

Critically, this formulation evaluates the reward exclusively at the clean-image endpoint (x_0), while applying KL regularization against the unrewarded reference marginals ($\{q_t\}$) across all noise levels. This structural mismatch creates a fundamental optimization loophole: because forward noise naturally masks image details, the regularization penalty against mode divergence drastically weakens at high noise levels. Optimizers readily exploit this trajectory vulnerability, abandoning the reference distribution to collapse onto high-reward regions without incurring sufficient KL penalty. We formally identify this failure mode as *terminal reward domination*. Resolving this inherent misalignment necessitates a principled trajectory-level target, which directly motivates our construction in Section 3.

The Principled KL-Regularized Target. To construct a mathematically sound trajectory objective, we must first establish the correct alignment target at the clean-image level. Standard Reinforcement Learning from Human Feedback (RLHF) [Christiano et al., 2017, Ouyang et al., 2022] formulates preference optimization by balancing reward maximization against a KL-divergence penalty to explicitly prevent fidelity degradation:

$$\mathcal{L}(\theta) = \mathbb{E}_{c,\, x_0 \sim p_\theta(x_0|c)}\big[-r(x_0, c)\big] + \tau\, \mathcal{D}_{\text{KL}}\big(p_\theta(x_0|c) \| q_0(x_0|c)\big), \tag{2}$$

where the temperature $\tau > 0$ governs the trade-off between semantic preference and reference fidelity. Crucially, this objective admits a unique, closed-form global optimum—the reward-tilted density:

$$q^*(x_0|c) = \frac{1}{Z(c)}\, q_0(x_0|c) \exp\left(\frac{r(x_0, c)}{\tau}\right), \tag{3}$$

where $Z(c)$ is the normalizing partition function. As detailed in Appendix C.1, minimizing Eq. (2) is mathematically equivalent to directly minimizing $\tau \mathcal{D}_{\text{KL}}(p_\theta \| q^*)$. Thus, q^* represents the fundamentally correct, perfectly balanced alignment target for clean images. Our core insight, developed next, is that this optimal target can be rigorously propagated forward through the diffusion process to guide every intermediate noise level without structural mismatch.

3 The Proposed Diff-Instruct with Diffused Reward

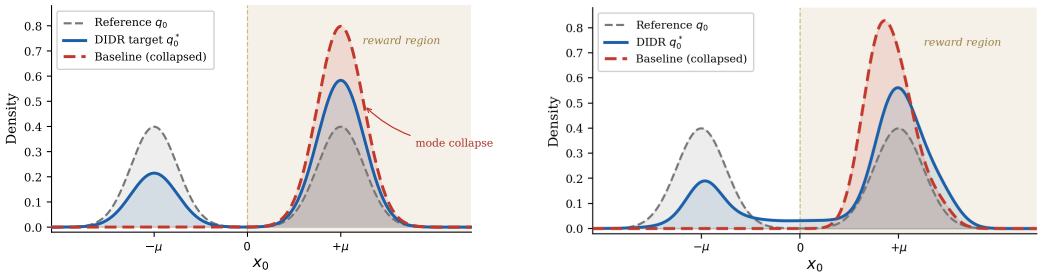
DIDR resolves the target mismatch of prior methods through four steps. (i) We illustrate how the structural mismatch in $\mathcal{L}_{\text{term}}$ can lead to over-tilting via a tractable example (§3.1). (ii) We diffuse the RLHF-optimal clean-image target q^* through the reference process to obtain principled trajectory-level marginals $\{q_t^*\}$, and show that minimizing the IKL against $\{q_t^*\}$ is equivalent to KL-regularized RLHF (§3.2). (iii) We derive the target score as the reference score plus a reward-induced correction (DRS) defined at every noise level (§3.2). (iv) We approximate DRS with DRP via differentiable short-step denoising, and train the generator using a Teaching Assistant (TA) score model (§3.3–§3.4).

3.1 Terminal Reward Domination

As noise increases, the forward diffusion process smooths modal structure, weakening the KL cost of suppressing an unrewarded mode. Because the reward acts only at the clean level while this weakening occurs across the full trajectory, the reward signal may dominate: the optimizer can over-tilt toward high-reward outputs relative to the RLHF target q^*. We refer to this behavior as *terminal reward domination* and prove it analytically in the tractable example below.

Example 1 (Terminal reward domination) *Let $p_0 = \frac{1}{2}\mathcal{N}(-\mu, \sigma^2) + \frac{1}{2}\mathcal{N}(\mu, \sigma^2)$ with reward $r(x_0) = \mathbf{1}[x_0 > 0]$. While q^* remains a proper mixture for all finite τ, $\mathcal{L}_{\text{term}}$ fully collapses onto the rewarded mode whenever $\tau \le \gamma(1-\sigma^2)/[2\mu^2(-\log\sigma^2)]$, a threshold that need not be small (Appendix C.5, Figure 2).*

This theoretical result shows that q^* (Eq. (3)) is a strictly better alignment target: it stays calibrated for all τ precisely because reward and KL regularization are balanced at the same level. Section 3.2 formalizes how to lift q^* to a full trajectory objective.



(a) **Theoretical prediction** (§3.1, $\tau = 1$): $\mathcal{L}_{\text{term}}$ collapses onto the rewarded mode (red) while q^* remains a soft mixture (blue).

(b) **1-D empirical validation** (setup in Appendix D): DIDR vs. DI++ as the baseline.

Figure 2: **Terminal reward domination.** The reward signal in $\mathcal{L}_{\text{term}}$ tends to dominate the KL regularizer, causing the optimizer to over-tilt toward the high-reward mode relative to the balanced RLHF target q^*. **(a)** plots the theoretical prediction from §3.1; **(b)** shows the 1-D empirical result (Appendix D), confirming the collapse in practice even at $\tau = 1$.

3.2 Reward-Tilted Trajectory Objective

Our core construction is the reward-tilted trajectory $\{q_t^*\}$, a principled optimization target defined at every noise level. We obtain it by lifting the RLHF optimum $q^*(x_0|c)$ (Eq. (3)) to all noise levels via the reference forward process:

$$q_t^*(x_t|c) = \int q_t(x_t|x_0)\, q^*(x_0|c)\, \mathrm{d}x_0. \tag{4}$$

The family $\{q_t^*\}_{t\in[0,T]}$ is the diffusion trajectory of the RLHF minimizer q^*, providing a well-defined target at every noise level. Proposition 2 establishes the formal equivalence.

Objective. We train the generator by minimizing the IKL divergence between its forward-noised marginals $p_{\theta,t}$ and the reward-tilted trajectory $\{q_t^*\}$:

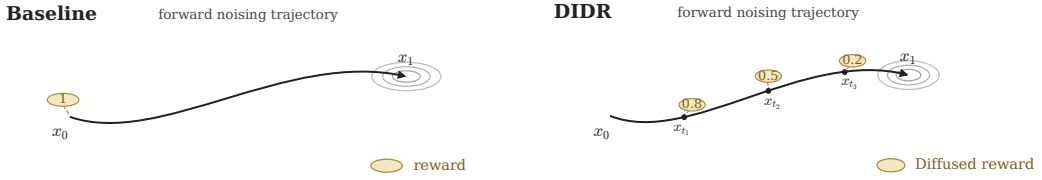
$$\mathcal{L}_{\text{DIDR}}(\theta) = \mathbb{E}_{c\sim\mathcal{C}}\left[\int_0^T w(t)\, \mathcal{D}_{\text{KL}}\big(p_{\theta,t}(\cdot|c)\|q_t^*(\cdot|c)\big)\mathrm{d}t\right]. \tag{5}$$

By minimizing IKL against q_t^* rather than the bare reference marginals, DIDR encodes the full RLHF signal inside the trajectory without a separate terminal reward term, inheriting the broad distribution support and stable gradient flow of the IKL formulation while correctly injecting reward signals at every noise level.

Proposition 2 (IKL shares the RLHF minimizer) *For any clean distribution $p_0(\cdot|c)$ and its forward marginals $p_t(\cdot|c)$, if $Z(c) < \infty$ and $w(t) > 0$ a.e., then*

$$\arg\min_{p_0} \mathcal{L}_{\text{RLHF}}(p_0|c) = \arg\min_{p_0} \mathcal{L}_{\text{DIDR}}(p_0|c) = \{q^*(\cdot|c)\}.$$

(See Appendix C.2 for proof.) The shared minimizer shows that the ideal IKL objective targets the RLHF optimum q^*. Note that this equivalence holds at the distribution level; the practical algorithm introduces score, posterior, and finite-sample approximations. Nevertheless, $\{q_t^*\}$ provides a well-defined optimization target throughout training.



(a) Terminal-reward alignment: reward only at x_0.

(b) DIDR: reward at every x_t via DRS.

Figure 3: **Reward propagation comparison.** Prior methods inject reward only at the clean endpoint x_0 while regularizing against bare reference marginals. DIDR propagates preference through the full trajectory via the Diffused Reward Score.

Theorem 3 (Score-based DIDR gradient) *Differentiating the Integral KL through the generator and the forward noising process gives the ideal score-based gradient*

$$\nabla_\theta \mathcal{L}_{\mathrm{DIDR}} = \mathbb{E}_{\substack{c\sim\mathcal{C},\, z\sim p_z,\, t\sim\pi(t) \\ x_0=g_\theta(z,c) \\ x_t\sim q_t(\cdot|x_0)}} \left[w(t)\big(s_\theta(x_t,t,c) - s_{\mathrm{ref}}(x_t,t,c) - s_r(x_t,t,c)\big) \frac{\partial x_t}{\partial \theta} \right]. \tag{6}$$

where $s_\theta(x_t,t,c) = \nabla_{x_t} \log p_{\theta,t}(x_t|c)$. *Also,* $s_{\mathrm{ref}}(x_t,t,c) = \nabla_{x_t} \log q_t(x_t|c)$. *The Diffused Reward Score (DRS) is*

$$s_r(x_t,t,c) = \nabla_{x_t} \log \mathbb{E}_{x_0\sim q(x_0|x_t,c)} \left[\exp\left(\frac{r(x_0,c)}{\tau} \right) \right]. \tag{7}$$

(See Appendix C.3 for proof.) The DRS term s_r adds a reward-driven correction to the reference score, steering the generator toward the RLHF optimum at every noise level. Because s_r is defined through the posterior $q(x_0|x_t,c)$, it propagates preference information only where the clean image is identifiable from the noisy observation, naturally attenuating reward guidance at high noise. The generator update is therefore driven by the mismatch between its own score and the reward-tilted target score $s_{\mathrm{ref}} + s_r$.

Remark 4 (Connection to classifier guidance) *Classifier guidance [Dhariwal and Nichol, 2021] is the first-order approximation of the DRS; the DRS keeps the full log-exponential tilt.*

3.3 Diffused Reward Proxy

Although the DRS provides a principled reward signal at every noise level, computing Eq. (7) requires the posterior $q(x_0|x_t,c)$, which is unavailable in closed form for a learned diffusion model. Assuming posterior samples admit a reparameterization $x_0 = \mathcal{G}(x_t,\boldsymbol{\epsilon},c)$ differentiable in x_t, differentiating the log-normalizer yields the pathwise form

$$s_r(x_t,t,c) = \mathbb{E}_{x_0\sim q(x_0|x_t,c)} \left[\frac{\exp\left(\frac{r(x_0,c)}{\tau}\right)}{\mathbb{E}_{\tilde{x}_0\sim q(\tilde{x}_0|x_t,c)}\left[\exp\left(\frac{r(\tilde{x}_0,c)}{\tau}\right)\right]} \cdot \frac{1}{\tau} \nabla_{x_t} r(x_0,c) \right], \tag{8}$$

where $\nabla_{x_t} r(x_0,c)$ is the gradient through a posterior sample x_0 treated as a function of x_t: the DRS is a reward-softmax-weighted average of pathwise reward gradients. We make this tractable by approximating the unavailable posterior with differentiable short-step denoising chains from x_t under the frozen reference model, $\hat{x}_0 = \mathcal{G}_{\mathrm{ref}}(x_t,\boldsymbol{\epsilon},c)$, giving DRP formalized below.

Proposition 5 (Diffused Reward Proxy) *Let* $\hat{x}_0^{(k)}$ *be* K *independent* S*-step denoising samples from the frozen reference model,*

$$\hat{x}_0^{(k)} = \mathcal{G}_{\mathrm{ref}}(x_t,\boldsymbol{\epsilon}^{(k)},c), \qquad k=1,\dots,K, \qquad \boldsymbol{\epsilon}^{(k)} \sim p(\boldsymbol{\epsilon}). \tag{9}$$

For fixed $\boldsymbol{\epsilon}$*, each chain is differentiable with respect to* x_t*. The DRP approximates the unavailable posterior expectation in Eq.* (8) *by the Monte Carlo estimator*

$$\widehat{s}_r(x_t,t,c) = \frac{1}{\tau} \sum_{k=1}^{K} \omega^{(k)} \nabla_{x_t} r(\hat{x}_0^{(k)},c), \qquad \omega^{(k)} = \frac{\exp(r(\hat{x}_0^{(k)},c)/\tau)}{\sum_j \exp(r(\hat{x}_0^{(j)},c)/\tau)}. \tag{10}$$

Here the pathwise reward gradient is

$$\nabla_{x_t} r(\hat{x}_0^{(k)},c) = \left(\frac{\partial \mathcal{G}_{\mathrm{ref}}(x_t,\boldsymbol{\epsilon}^{(k)},c)}{\partial x_t} \right)^{\top} \nabla_{\hat{x}_0} r(\hat{x}_0^{(k)},c). \tag{11}$$

(See Appendix C.4 for derivation.) The softmax weights $\omega^{(k)}$ concentrate gradient mass on high-reward denoised images, while the pathwise Jacobian $\partial \mathcal{G}_{\mathrm{ref}}/\partial x_t$ propagates clean-image reward information back to x_t through the frozen denoising chain without storing reference model activations.

Remark 6 (Why the proxy works) *At low noise, the posterior* $q(x_0|x_t,c)$ *is concentrated, so short denoising chains yield accurate proxies. At high noise,* x_t *carries little information about* x_0*, so the DRS correction is less sensitive to the accuracy of individual posterior samples; a short denoising proxy captures the leading-order correction.*

3.4 Practical Training with a Teaching Assistant

Two substitutions make Eq. (6) implementable. First, the generator score $s_\theta = \nabla_{x_t} \log p_{\theta,t}$ is unavailable for an implicit one-step generator; following Luo et al. [2023], we approximate it with a Teaching Assistant (TA) score model s_ψ trained concurrently via denoising score matching (DSM) [Vincent, 2011, Song et al., 2021] on noisy generator samples, so that $s_\psi(x_t, t, c) \approx s_\theta(x_t, t, c)$. Second, for text-to-image generation we replace the reference score s_{ref} with the CFG-corrected score

$$\tilde{s}_{\text{ref}}(x_t, t, c) = s_{\text{ref}}(x_t, t, \emptyset) + \alpha_{\text{cfg}}\big[s_{\text{ref}}(x_t, t, c) - s_{\text{ref}}(x_t, t, \emptyset)\big], \tag{12}$$

where $\alpha_{\text{cfg}} \geq 1$ amplifies the prompt-conditional component.

Remark 7 (Implicit and explicit reward signals) *The CFG-corrected reference score $\tilde{s}_{\text{ref}}$ can be viewed as an implicit reward-tilted score induced by text guidance, while the DRS s_r is an explicit correction from the chosen reward model. These two signals are complementary and can be combined, giving* DIDR *flexibility in the reward design.*

With these substitutions, the practical generator update is

$$\text{Grad}(\theta) = \mathbb{E}_{\substack{c\sim\mathcal{C},\, z\sim p_z,\, t \\ x_0 = g_\theta(z,c),\, x_t\sim q_t(\cdot|x_0)}} \left[w(t)\left(s_\psi(x_t, t, c) - \tilde{s}_{\text{ref}}(x_t, t, c) - s_r(x_t, t, c)\right) \frac{\partial x_t}{\partial \theta} \right]. \tag{13}$$

DIDR alternates two stages (Algorithm 1, Figure 6). **Stage I** keeps s_ψ synchronized with the evolving g_θ: sample $x_0 = g_\theta(z, c)$, diffuse to x_t, and update s_ψ via DSM. **Stage II** updates g_θ: draw K differentiable S-step denoising chains from x_t, compute the DRP weights, and backpropagate Eq. (13) to θ. Both the reference diffusion model and the reward model remain frozen throughout.

4 Empirical Results

Models. We evaluate DIDR on two backbones. In our primary experiments, we apply DIDR to one-step SDXL [Podell et al., 2024] at 1024×1024 resolution, initializing the generator from the DMD2-SDXL-1step checkpoint [Yin et al., 2024a] and using the pretrained SDXL diffusion model as both the frozen reference s_{ref} and the TA initialization. We report a standard variant (DIDR) and an extended-training variant ($\text{DIDR}_{\text{longer}}$). To assess generalizability, we additionally apply DIDR to the Z-Image backbone [Tongyi-MAI, 2025] (Z-Image DIDR), initializing from Z-Image-Turbo. We train 1-step and 4-step Z-Image DIDR variants and compare each against Z-Image-Turbo with twice as many steps. Full hyperparameters are in Appendix H.

Training Data and Reward. All models are trained on text prompts from LAION-Aesthetic-6.25+ [Zhou et al., 2024a]; no image data is required. As the reward model $r(x_0, c)$, we use PickScore [Kirstain et al., 2023], a widely accepted human preference model trained on large-scale pairwise human annotations that scores how well a generated image matches human preference given a text prompt. The training reward uses the raw unscaled PickScore, which is $1/98.86$ of the official reported values in Table 2, making τ=0.01 a reasonable temperature for the DRP softmax.

Evaluation. We report three categories of metrics: *Preference* — PickScore [Kirstain et al., 2023], ImageReward [Xu et al., 2023], HPSv2.1 [Wu et al., 2023], and Aesthetic Score [Schuhmann, 2022]; *Text alignment* — CLIPScore [Hessel et al., 2021, Radford et al., 2021], DPG-Bench [Hu et al., 2024], and GenEval [Ghosh et al., 2023]; and *Fidelity* — FID [Heusel et al., 2017]. PickScore, ImageReward, Aesthetic Score, and CLIPScore are evaluated on 1k prompts from the MSCOCO-2017 validation set [Lin et al., 2014]; FID is evaluated on 30k samples; HPSv2.1, DPG-Bench, and GenEval each use their own standard benchmark prompts. Since PickScore is also our training reward, the remaining preference metrics use reward models not used for training. Further details are provided in Appendix E.

4.1 Quantitative Results

$\text{DIDR}_{\text{longer}}$ achieves the highest PickScore among evaluated one-step text-to-image models. As shown in Table 2, $\text{DIDR}_{\text{longer}}$ achieves the highest PickScore (23.9) and surpasses all multi-step reference models on all four preference metrics under our evaluation protocol, using only 2.6B parameters and a single inference step. Gains span both the training reward and metrics not used for training (ImageReward $+8.9\%$, HPSv2.1 $+0.6$, Aesthetic Score $+0.15$ over DI*), suggesting that alignment does not simply overfit to PickScore. Per-category GenEval and HPSv2.1 results are provided in Tables 5 and 6 (Appendix F).

Table 2: Comparison of DIDR against baselines at 1024×1024 resolution. Multi-step reference models (top block) are shown for reference only (no bold). **Bold**: best result within each group. ↑/↓: higher/lower is better.

				Preference ↑				Text Alignment ↑			
Model	**Steps**	**Arch.**	**Params**	**PickS.**	**ImgR.**	**HPSv2.1**	**Aesth.**	**CLIP**	**DPG**	**GenEval**	**FID ↓**
Multi-step reference											
SDXL [Podell et al., 2024]	50	UNet	2.6B	22.8	0.82	28.98	5.45	33.86	74.42	0.551	15.9
SDXL-DPO [Wallace et al., 2024]	50	UNet	2.6B	22.8	0.92	30.45	5.61	33.97	74.24	0.549	21.5
SD3.5-large [AI, 2024]	28	DiT	8B	23.0	1.01	30.06	5.42	33.85	84.88	0.717	17.7
FLUX-dev [Labs, 2024]	50	DiT	12B	23.2	1.05	30.65	5.54	32.88	83.86	0.672	24.8
Z-Image [Tongyi-MAI, 2025]	50	DiT	6B	22.5	0.99	30.57	5.43	33.15	88.14	0.840	14.3
One-step SDXL											
SDXL DMD2 [Yin et al., 2024a]	1	UNet	2.6B	22.3	0.83	29.99	5.47	33.31	74.01	0.557	**13.7**
SDXL Diff-Instruct [Luo et al., 2023]	1	UNet	2.6B	22.4	0.92	31.51	5.53	33.35	74.98	0.560	19.3
SDXL Diff-Instruct++ [Luo, 2024]	1	UNet	2.6B	22.4	0.92	31.44	5.58	**33.37**	75.01	0.560	19.0
SDXL Diff-Instruct* [Luo et al., 2025]	1	UNet	2.6B	23.1	1.01	33.29	5.68	33.09	74.97	0.555	18.9
SDXL DIDR (Ours)	1	UNet	2.6B	23.5	1.04	33.77	5.82	33.03	**75.06**	**0.579**	18.8
SDXL $\text{DIDR}_{\text{longer}}$ (Ours)	1	UNet	2.6B	**23.9**	**1.10**	**33.89**	**5.83**	32.97	74.99	0.566	20.6
Z-Image backbone, 4–8 steps											
Z-Image-Turbo [Tongyi-MAI, 2025]	8	DiT	6B	23.0	1.01	31.86	5.39	33.08	84.86	**0.820**	25.2
Z-Image DIDR (Ours)	4	DiT	6B	**23.1**	**1.09**	**33.62**	**5.59**	**33.11**	**86.80**	0.730	**24.8**
Z-Image backbone, 1–2 steps											
Z-Image-Turbo [Tongyi-MAI, 2025]	1	DiT	6B	21.0	0.39	24.54	4.68	33.28	77.05	0.572	36.7
Z-Image-Turbo [Tongyi-MAI, 2025]	2	DiT	6B	22.5	0.96	31.16	5.40	33.11	85.10	**0.763**	23.1
Z-Image DIDR (Ours)	1	DiT	6B	**22.6**	**1.08**	**31.40**	**5.46**	**33.35**	**86.63**	0.692	**22.1**

Standard DIDR Pareto-dominates all one-step alignment baselines on preference vs. fidelity. Since the target distribution is the reward-tilted q^* (Eq. (3)), some FID increase is theoretically expected; the key question is which method best navigates this trade-off. As shown in Figure 4a, DIDR (PickScore 23.5, FID 18.8) Pareto-dominates all one-step alignment baselines and simultaneously surpasses FLUX-dev and SDXL-DPO on both axes. DIDR also maintains text alignment, achieving the highest DPG-Bench (75.06) and GenEval (0.579) among one-step SDXL methods, with only a marginal CLIPScore dip, which is expected since text-alignment metrics are not directly optimized by the reward model.

DIDR generalizes across architectures and model scales. On the 6B Z-Image backbone, both Z-Image DIDR variants outperform Z-Image-Turbo with twice as many steps on every metric except GenEval: the 4-step model beats the 8-step Turbo (HPSv2.1 33.62 vs. 31.86), and the 1-step model beats the 2-step Turbo while also surpassing the 50-step Z-Image base on all four preference metrics. These gains are driven by training, not initialization—the 1-step Z-Image-Turbo initially scores ImageReward 0.39 and FID 36.7, vs. 1.08 and 22.1 for Z-Image DIDR.

4.2 Qualitative Comparison

DIDR consistently improves visual quality over prior one-step methods while preserving perceptual naturalness (Figures 1 and 5). Compared with DI++ and DI*, DIDR produces sharper fine detail and more natural color—skin tones and textures are faithfully rendered rather than over-saturated or painterly, reflecting a more favorable reward–fidelity trade-off. On the Z-Image backbone, Z-Image DIDR recovers dramatically from the severely degraded 1-step Turbo initialization and matches 8-step Z-Image-Turbo in perceptual sharpness and composition with a single step. A qualitative comparison against multi-step baselines (SDXL, Z-Image, FLUX-dev, SD3.5-Large) is provided in Figure 8.

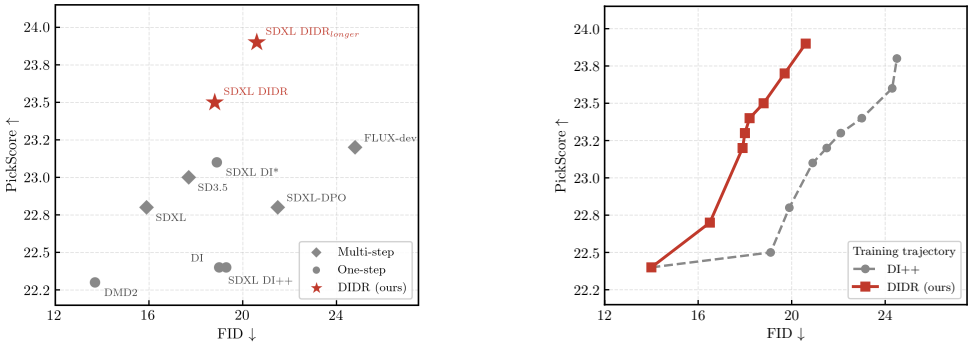


Figure 4: **PickScore–FID analysis**. *Left*: final model comparison on MSCOCO-2017; DIDR Pareto-dominates all one-step alignment baselines. *Right*: training trajectories, DIDR vs. Diff-Instruct++.

4.3 Ablation Studies

Effect of DRS. Figure 4b shows that DIDR Pareto-dominates DI++ at every training checkpoint, confirming that trajectory-level reward propagation provides a structural advantage over terminal reward alone.

Effect of (K, S) and temperature τ. Both K and S contribute independently: raising either from 1 to 4 improves all metrics, and $(K, S)=(4, 4)$ achieves the best result on all four metrics (Table 3). Decreasing τ monotonically improves PickScore (22.6→23.8) at the cost of FID (16.3→22.2); we select $\tau=0.01$ as the operating point balancing both (Table 4, Figure 7).

Table 3: **Effect of (K, S).** Both axes contribute independently; $(K, S)=(4, 4)$ is best. †: default; **bold**: best in column.

K	S	PickS. ↑	ImgR. ↑	Aesth. ↑	FID ↓
1	1	22.6	0.80	5.32	24.1
1	2	22.8	0.86	5.41	22.4
1	4	23.1	0.92	5.66	20.8
2	1	22.7	0.84	5.44	23.5
4	1	22.8	0.90	5.50	22.0
4†	4†	**23.5**	**1.04**	**5.82**	**18.8**

Table 4: **Effect of τ.** Lower τ improves preference at the cost of FID; $\tau=0.01$ is the operating point. $(K, S)=(4, 4)$ fixed. †: default; **bold**: best in column.

τ	PickS. ↑	ImgR. ↑	Aesth. ↑	FID ↓
1.00	22.6	0.88	5.56	**16.3**
0.50	22.9	0.93	5.41	17.1
0.10	22.9	1.02	5.68	17.5
0.05	23.2	1.01	5.74	17.8
0.01†	23.5	1.04	5.82	18.8
0.001	**23.8**	**1.17**	**5.83**	22.2

5 Related Work

RLHF and Preference Alignment for Diffusion Models. Preference alignment for multi-step diffusion models uses trajectory-accessible signals via supervised fine-tuning [Dai et al., 2023, Podell et al., 2024], reward backpropagation [Prabhudesai et al., 2023, Clark et al., 2024, Black et al., 2024, Fan et al., 2023], offline preference optimization [Wallace et al., 2024, Yang et al., 2024, Hong et al., 2026], or online RL [Liu et al., 2025, Xue et al., 2025, Zheng et al., 2026b] (Table 1). These methods rely on explicit denoising trajectories and do not directly address one-step generators with implicit distributions; DIDR derives a compatible trajectory objective that applies to implicit generators without requiring trajectory access during inference.

Classifier Guidance and Reward-Guided Sampling. Classifier guidance [Dhariwal and Nichol, 2021] adds $\nabla_{x_t} \log p(c|x_t)$ to the score, which is a first-order approximation of the DRS (Remark 4), and related inference-time methods [Chung et al., 2023, Prabhudesai et al., 2023, Clark et al., 2024, Kong et al., 2026] apply differentiable rewards along the trajectory. However, all such methods modify only the sampling procedure and leave the generator weights unchanged; DIDR instead trains an aligned one-step generator.



Figure 5: **Qualitative comparison at 1024×1024.** *Left*: SDXL backbone (SDXL DIDR, DI++, DI*, DMD2). *Right*: Z-Image backbone (Z-Image DIDR 1 NFE, Turbo 1 NFE, Turbo 8 NFE). Backbone-specific prompts are listed in Appendix I.

6 Conclusion and Limitations

Conclusion. We identified terminal reward domination as a fundamental failure mode of trajectory distillation with terminal rewards, and proposed DIDR to resolve it. The core idea is to lift the RLHF-optimal clean-image target q^* to a principled trajectory objective via the Diffused Reward Score, and approximate it tractably with the Diffused Reward Proxy through differentiable short-step denoising—requiring no image training data. DIDR achieves the best PickScore–FID trade-off among one-step alignment methods, improves all four preference metrics over the strongest prior baseline, and transfers to the 6B Z-Image backbone, where it surpasses its 50-step teacher on preference metrics in a single step and outperforms the 8-step Z-Image-Turbo with 4 steps. These results indicate that principled trajectory-level reward propagation is an effective alignment strategy across architectures and scales.

Limitations. Since DRP propagates gradients through the reward model, it may amplify reward-model biases or trigger reward hacking when the reward model is imperfect. Each generator update also requires $K \times S$ differentiable reference-denoising steps, increasing training cost relative to endpoint-only methods. Finally, counting, fine-grained structure, and complex multi-object scenes remain open challenges for single-step generation (Figure 9). Extending DIDR to multi-step generators and developing reward-ensemble strategies to mitigate hacking are natural directions for future work.

Acknowledgments and Disclosure of Funding

G.L. would like to acknowledge the support from the National Science Foundation (DMS-2533878, DMS-2053746, DMS-2134209, ECCS-2328241, CBET-2347401 and OAC-2311848), and U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research program under the "Uncertainty Quantification for Multifidelity Operator Learning (MOLUcQ)" project (Project No. 81739), DE-SC0023161, the SciDAC LEADS Institute, and DOE–Fusion Energy Science, under grant number: DE-SC0024583. We thank the anonymous reviewers and area chairs for their constructive feedback.

References

S. AI. Stable Diffusion 3.5. `https://stability.ai/news/stable-diffusion-3-5`, 2024.

K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training Diffusion Models with Reinforcement Learning. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2024.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

H. Chung, J. Kim, M. T. McCann, M. L. Klasky, and J. C. Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2023.

K. Clark, P. Vicol, K. Swersky, and D. J. Fleet. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2024.

X. Dai, J. Hou, C.-Y. Ma, S. Tsai, J. Wang, R. Wang, P. Zhang, S. Vandenhende, X. Wang, A. Dubey, et al. Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack. In *arXiv preprint arXiv:2309.15807*, 2023.

P. Dhariwal and A. Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 8780–8794, 2021.

P. Esser, S. Kulal, A. Blattmann, R. Aichele, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Proc. of the International Conference on Machine Learning (ICML)*, 2024.

Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee. Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

D. Ghosh, H. Hajishirzi, and L. Schmidt. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. CLIPScore: A Reference-Free Evaluation Metric for Image Captioning. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, 2021.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

J. Hong, S. Paul, N. Lee, K. Rasul, J. Thorne, and J. Jeong. Margin-aware Preference Optimization for Aligning Diffusion Models without Reference. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2026.

X. Hu, R. Wang, Y. Fang, B. Fu, P. Cheng, and G. Yu. ELLA: Equip Diffusion Models with LLM for Enhanced Semantic Alignment. In *arXiv preprint arXiv:2403.05135*, 2024.

M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park. Scaling up GANs for Text-to-Image Synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and Improving the Image Quality of StyleGAN. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2015.

Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

C. X. Kong, Y. Wang, H. Zheng, W. Luo, and G. Lin. Ultra Fast PDE Solving via Physics Guided Few-step Diffusion. *arXiv preprint arXiv:2602.03627*, 2026.

B. F. Labs. FLUX.1 Models. `https://github.com/black-forest-labs/flux`, 2024.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2014.

J. Liu, G. Liu, J. Liang, Y. Li, J. Liu, X. Wang, P. Wan, D. Zhang, and W. Ouyang. Flow-GRPO: Training Flow Matching Models via Online RL. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

X. Liu, C. Gong, and Q. Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *arXiv preprint arXiv:2209.03003*, 2022.

X. Liu, X. Zhang, J. Ma, J. Peng, and Q. Liu. InstaFlow: One Step is Enough for High-Quality Diffusion-Based Text-to-Image Generation. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2024.

W. Luo. A Comprehensive Survey on Knowledge Distillation of Diffusion Models. *arXiv preprint arXiv:2304.04262*, 2023.

W. Luo. Diff-Instruct++: Training One-step Text-to-image Generator Model to Align with Human Preferences. *Transactions on Machine Learning Research (TMLR)*, 2024.

W. Luo, T. Hu, S. Zhang, J. Sun, Z. Li, and Z. Zhang. Diff-Instruct: A Universal Approach for Transferring Knowledge from Pre-trained Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

W. Luo, C. Zhang, D. Zhang, and Z. Geng. David and Goliath: Small One-step Model Beats Large Diffusion with Score Post-training. In *Proc. of the International Conference on Machine Learning (ICML)*, 2025.

A. Q. Nichol and P. Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *Proc. of the International Conference on Machine Learning (ICML)*, 2021.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2024.

M. Prabhudesai, A. Goyal, D. Pathak, and K. Fragkiadaki. Aligning Text-to-Image Diffusion Models with Reward Backpropagation. In *arXiv preprint arXiv:2310.03739*, 2023.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proc. of the International Conference on Machine Learning (ICML)*, 2021.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. Technical report, OpenAI, 2022.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. In *Proc. of the International Conference on Machine Learning (ICML)*, 2023.

A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach. Adversarial Diffusion Distillation. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2024.

C. Schuhmann. LAION-Aesthetics Predictor V2. `https://github.com/christophschuhmann/improved-aesthetic-predictor`, 2022.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2021.

Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency Models. In *Proc. of the International Conference on Machine Learning (ICML)*, 2023.

Tongyi-MAI. Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer. *arXiv preprint arXiv:2511.22699*, 2025.

P. Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion Model Alignment Using Direct Preference Optimization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. `https://github.com/tgxs002/HPSv2`, 2023.

J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Y. Xu, Y. Zhao, Z. Xiao, and T. Hou. UFOGen: You Forward Once Large Scale Text-to-Image Generation via Diffusion GANs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

S. Xue, C. Ge, S. Zhang, Y. Li, and Z.-M. Ma. Advantage Weighted Matching: Aligning RL with Pretraining in Diffusion Models. *arXiv preprint arXiv:2509.25050*, 2025.

K. Yang, J. Tao, J. Lyu, C. Ge, J. Chen, W. Shen, X. Zhu, and X. Li. Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

T. Yin, M. Gharbi, T. Park, R. Zhang, E. Shechtman, F. Durand, and W. T. Freeman. Improved Distribution Matching Distillation for Fast Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park. One-step Diffusion with Distribution Matching Distillation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

B. Zheng and T. Yang. Diffusion Models Are Innate One-Step Generators. *arXiv preprint arXiv:2405.20750*, 2024.

H. Zheng, X. Liu, C. X. Kong, N. Jiang, Z. Hu, W. Luo, W. Deng, and G. Lin. Ultra-Fast Language Generation via Discrete Diffusion Divergence Instruct. *Proc. of the International Conference on Learning Representation (ICLR)*, 2026a.

K. Zheng, H. Chen, H. Ye, H. Wang, Q. Zhang, K. Jiang, H. Su, S. Ermon, J. Zhu, and M.-Y. Liu. DiffusionNFT: Online Diffusion Reinforcement with Forward Process. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2026b.

M. Zhou, Z. Wang, H. Zheng, and H. Huang. Long and Short Guidance in Score identity Distillation for One-Step Text-to-Image Generation. *arXiv preprint arXiv:2406.01561*, 2024a.

M. Zhou, H. Zheng, Z. Wang, M. Yin, and H. Huang. Score Identity Distillation: Exponentially Fast Distillation of Pretrained Diffusion Models for One-step Generation. In *Proc. of the International Conference on Machine Learning (ICML)*, 2024b.

Contents

1 Introduction **1**

2 Preliminaries **3**

3 The Proposed Diff-Instruct with Diffused Reward **4**

3.1 Terminal Reward Domination 4

3.2 Reward-Tilted Trajectory Objective 5

3.3 Diffused Reward Proxy 6

3.4 Practical Training with a Teaching Assistant 7

4 Empirical Results **7**

4.1 Quantitative Results 7

4.2 Qualitative Comparison 8

4.3 Ablation Studies 9

5 Related Work **9**

6 Conclusion and Limitations **10**

A Method Supplements **15**

B Additional Qualitative Results **15**

C Theoretical Derivations **15**

C.1 Proof of the RLHF-Target Equivalence (Eq. (2)) 15

C.2 Proof of Proposition 2 (IKL Shares the RLHF Minimizer) 15

C.3 Proof of Theorem 3 (Score-based DIDR Gradient) 17

C.4 Proof of Proposition 5 (Diffused Reward Proxy) 18

C.5 A Bimodal Gaussian Example of Terminal Reward Domination 20

D 1-D Toy Experiment **22**

E Evaluation Metrics **22**

F Per-Category Results **23**

G Full Training Algorithm **23**

H Implementation Details **24**

H.1 SDXL Experiments 24

H.2 Z-Image Experiments 25

I Image Generation Prompts **26**

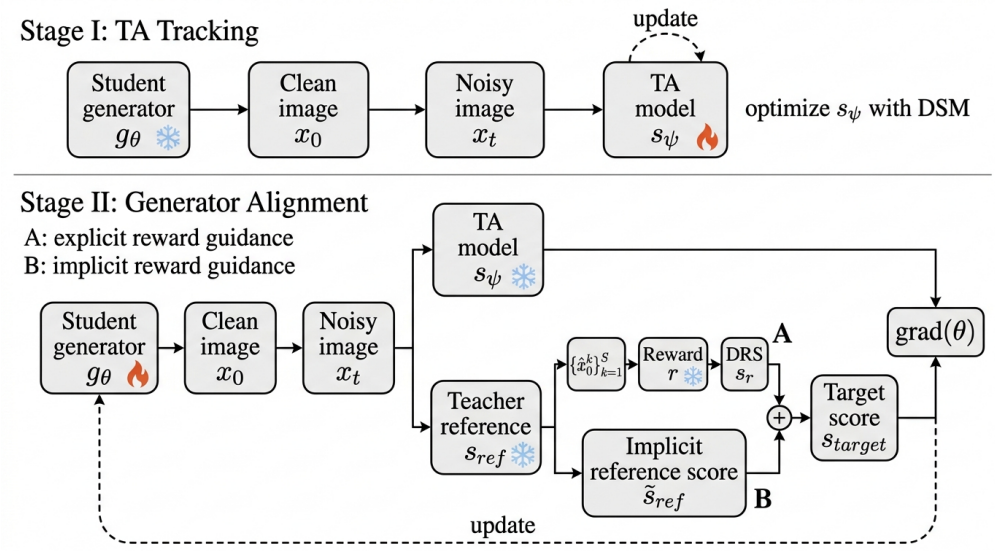


Figure 6: **The DIDR training framework. Stage I (TA):** s_ψ tracks the generator marginals via DSM. **Stage II (Generator):** g_θ is updated toward the reward-tilted target score $\tilde{s}_{\text{ref}} + s_r$.

Algorithm 1: DIDR Training (see Appendix G for full pseudocode)

while not converged **do**

  Stage I – TA update: sample $x_0 = g_\theta(z, c)$, diffuse to x_t, and update s_ψ by DSM to track $p_{\theta,t}$;

  Stage II – Generator update:;

    sample $x_0 = g_\theta(z, c)$ and diffuse to x_t;

    run K differentiable S-step reference denoising chains from x_t to obtain $\{\hat{x}_0^{(k)}\}_{k=1}^K$;

    compute $s_r = \frac{1}{\tau}\sum_k \omega^{(k)} \nabla_{x_t} r(\hat{x}_0^{(k)}, c)$, with $\omega^{(k)} \propto \exp(r(\hat{x}_0^{(k)}, c)/\tau)$;

    update θ using $w(t)(s_\psi - \tilde{s}_{\text{ref}} - s_r)\partial x_t/\partial\theta$;

end

A Method Supplements

B Additional Qualitative Results

This section provides supplementary visual comparisons. Figure 7 illustrates the qualitative effect of the temperature hyperparameter τ on generated images. Figure 8 compares one-step DIDR against 50-step multi-step baselines on matched prompts. Figure 9 documents representative failure cases.

C Theoretical Derivations

C.1 Proof of the RLHF-Target Equivalence (Eq. (2))

Proof 1 *Let* $Z(c) \triangleq \int q_0(x_0|c)\exp(r(x_0, c)/\tau)\mathrm{d}x_0 < \infty$ *and* $q^*(x_0|c) \triangleq q_0(x_0|c)\exp(r/\tau)/Z(c)$. *Expanding* $\mathcal{L}(\theta)$*:*

$$\mathcal{L}(\theta) = \tau \int p_\theta \left[\log p_\theta - \log q_0 - \frac{r}{\tau}\right] \mathrm{d}x_0 = \tau \int p_\theta \left[\log p_\theta - \log\left(q_0\, e^{r/\tau}\right)\right] \mathrm{d}x_0. \tag{14}$$

Since $q_0 e^{r/\tau} = Z(c)\, q^*$*, this becomes* $\tau\, \mathcal{D}_{\mathrm{KL}}(p_\theta \| q^*) - \tau \log Z(c)$*. As* $Z(c)$ *is independent of* θ*, the two objectives share the same minimizer.*

C.2 Proof of Proposition 2 (IKL Shares the RLHF Minimizer)

Proof 2 ***(i) Non-negativity.*** *Follows immediately from* $\mathcal{D}_{\mathrm{KL}} \geq 0$ *and* $w(t) > 0$.

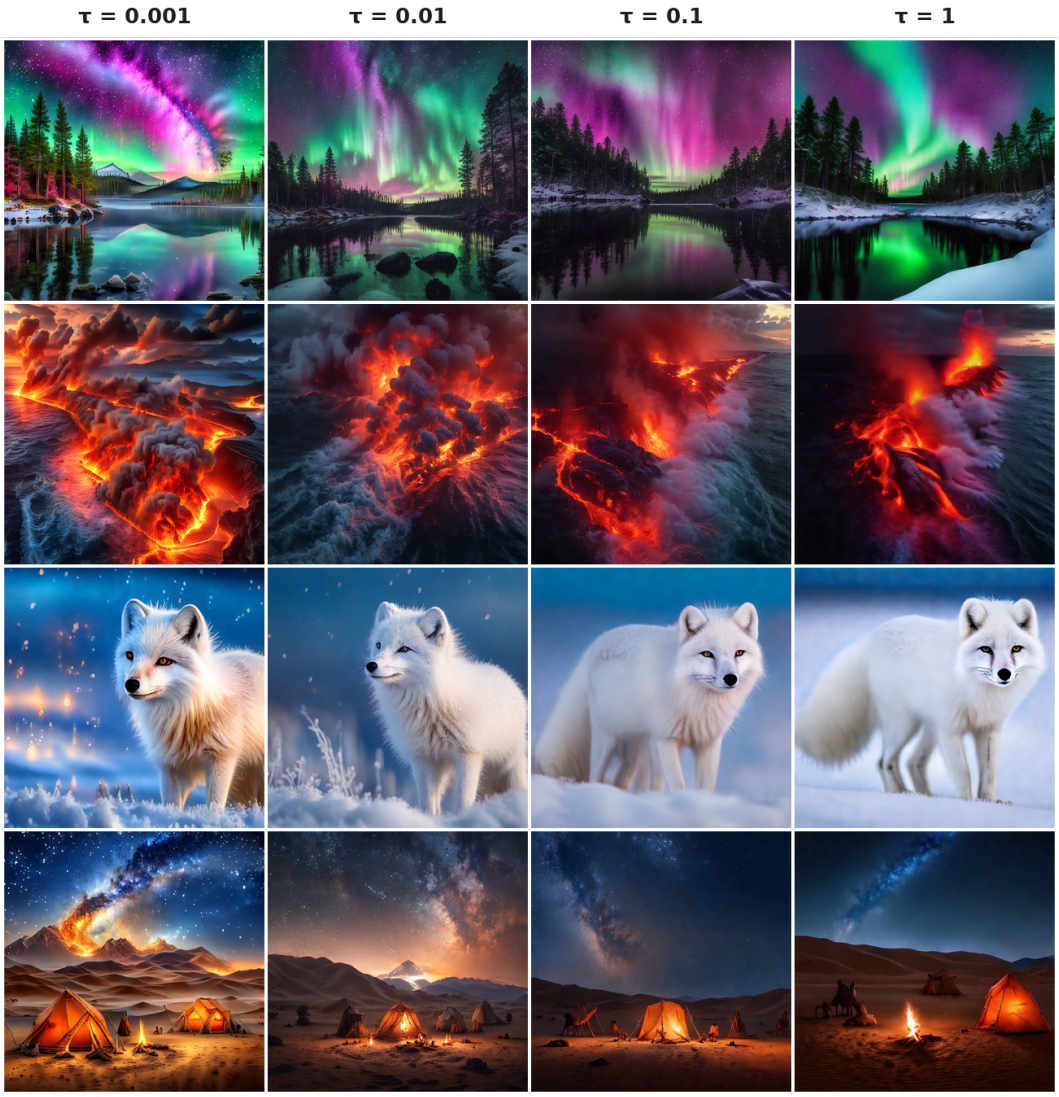


Figure 7: **Qualitative effect of temperature τ.** Images generated from the same prompt at decreasing τ (left to right). Smaller τ sharpens reward weighting and increases visual appeal, but introduces over-saturation and fine-detail artifacts at very low values, illustrating the preference–fidelity trade-off. Prompts in Appendix I.

(ii) Convexity. *$\mathcal{D}_{\mathrm{KL}}(p\|q^*)$ is convex in p for fixed q^*; a positively weighted integral of convex functionals is convex.*

(iii) Unique minimizer. *$\mathcal{L}_{\mathrm{IKL}} = 0$ iff $\mathcal{D}_{\mathrm{KL}}(p_t\|q_t^*) = 0$ for a.e. $t \in (0,T]$, since $w(t) > 0$ and $\mathcal{D}_{\mathrm{KL}} = 0$ iff the distributions coincide. This forces $p_t = q_t^*$ on a set of full measure in $(0,T]$, hence on a sequence $t_n \downarrow 0$. Since SDE marginals are weakly continuous in t (standard for Lipschitz drift and diffusion coefficients), taking $n \to \infty$ gives $p_0 = q_0^* = q^*$, which by Eq. (3) is the unique minimizer of $\mathcal{L}$.*

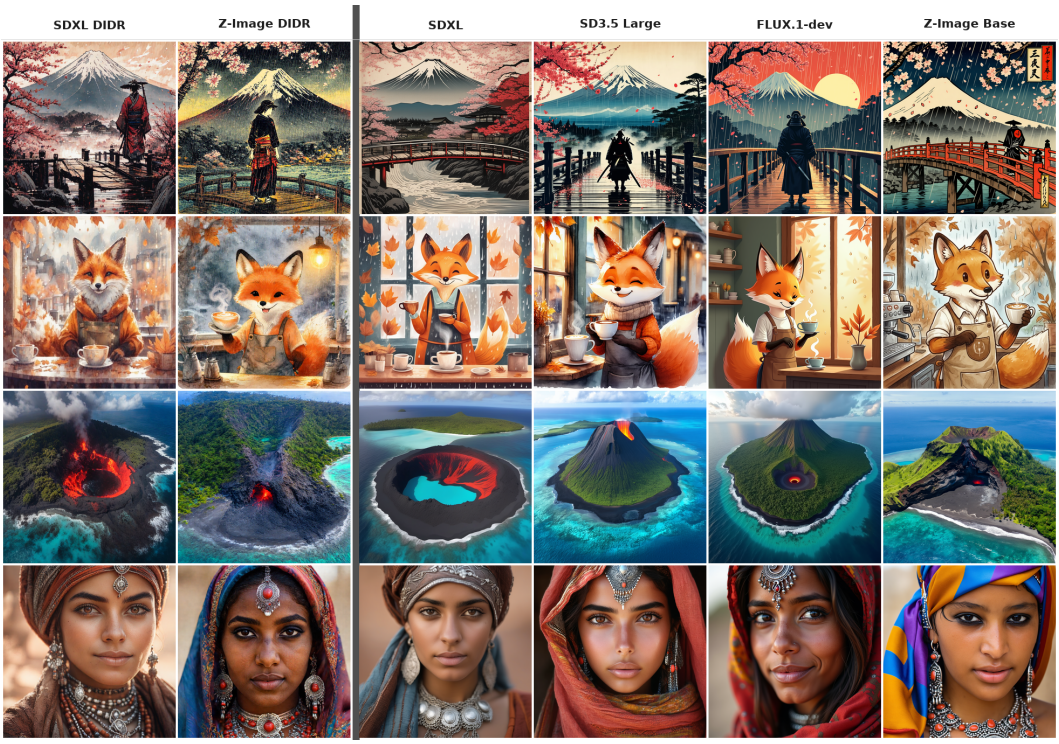


Figure 8: **Qualitative comparison against multi-step baselines.** One-step DIDR compared with 50-step SDXL, 50-step Z-Image, 50-step FLUX-dev, and 28-step SD3.5-Large on matched prompts. DIDR achieves comparable or superior visual quality and prompt fidelity in a single inference step. Prompts in Appendix I.



Figure 9: **Failure cases of DIDR:** anatomical errors, incorrect counts, structural distortions, and subject–background entanglement. Prompts in Appendix I.

C.3 Proof of Theorem 3 (Score-based DIDR Gradient)

Proof 3 *Let* $x_t = \mathcal{F}(g_\theta(z,c),\mathbf{w},t)$, $v_\theta \triangleq \partial x_t/\partial\theta$, *and* $s_\theta \triangleq \nabla_{x_t}\log p_{\theta,t}$. *Differentiating* $\mathcal{D}_{\mathrm{KL}}(p_{\theta,t}\|q_t^*) = \int p_{\theta,t}\log(p_{\theta,t}/q_t^*)\mathrm{d}x_t$ *in* θ*:*

$$\nabla_\theta\,\mathcal{D}_{\mathrm{KL}}(p_{\theta,t}\|q_t^*) = \underbrace{\int \frac{\partial p_{\theta,t}}{\partial\theta}\log\frac{p_{\theta,t}}{q_t^*}\,\mathrm{d}x_t}_{\text{(I)}} + \underbrace{\int p_{\theta,t}\,\frac{\partial\log p_{\theta,t}}{\partial\theta}\,\mathrm{d}x_t}_{\text{(II)}}. \tag{15}$$

Term (II) equals $\partial_\theta\int p_{\theta,t}\mathrm{d}x_t = 0$.

Continuity equation. *For any smooth test function* ϕ*, differentiating both sides of* $\mathbb{E}_{z,\mathbf{w}}[\phi(x_t)] = \int\phi\,p_{\theta,t}\mathrm{d}x_t$ *in* θ*:*

$$\int\phi\,\frac{\partial p_{\theta,t}}{\partial\theta}\mathrm{d}x_t = \mathbb{E}_{z,\mathbf{w}}\big[\nabla_{x_t}\phi(x_t)\cdot v_\theta\big] = \int p_{\theta,t}\,\nabla_{x_t}\phi\cdot v_\theta\,\mathrm{d}x_t = -\int\phi\,\nabla_{x_t}\cdot(p_{\theta,t}\,v_\theta)\,\mathrm{d}x_t, \tag{16}$$

where the last step is integration by parts. Since ϕ *is arbitrary,*

$$\frac{\partial p_{\theta,t}}{\partial\theta} = -\nabla_{x_t}\cdot(p_{\theta,t}\,v_\theta). \tag{17}$$

KL gradient. *Substituting* (17) *into term (I) and integrating by parts:*

$$\begin{aligned}(\mathrm{I}) &= -\int \nabla_{x_t} \cdot (p_{\theta,t}\, v_\theta)\, \log \frac{p_{\theta,t}}{q_t^*}\, \mathrm{d}x_t \\ &= \int p_{\theta,t}\, v_\theta \cdot \nabla_{x_t} \log \frac{p_{\theta,t}}{q_t^*}\, \mathrm{d}x_t \\ &= \int p_{\theta,t}\, v_\theta \cdot \big(s_\theta(x_t,t,c) - \nabla_{x_t} \log q_t^*(x_t|c)\big)\, \mathrm{d}x_t. \end{aligned} \tag{18}$$

Rewriting as an expectation over $(z, \mathbf{w})$ *via* $\int p_{\theta,t}\, \varphi\, \mathrm{d}x_t = \mathbb{E}_{z,\mathbf{w}}[\varphi(x_t)]$ *and substituting* $v_\theta = \partial x_t/\partial\theta$*:*

$$\nabla_\theta\, \mathcal{D}_{\mathrm{KL}}(p_{\theta,t} \| q_t^*) = \mathbb{E}_{z,\mathbf{w}}\left[\big(s_\theta(x_t,t,c) - \nabla_{x_t} \log q_t^*(x_t|c)\big) \frac{\partial x_t}{\partial\theta}\right]. \tag{19}$$

Multiplying by $w(t)$*, integrating over* $t \in (0,T]$*, and exchanging the* t*-integral with the expectation via Fubini gives Eq.* (6)*.*

Target score decomposition *(*$\nabla_{x_t} \log q_t^* = s_{\mathrm{ref}} + s_r$*). Substituting* $q^*(x_0|c) = q_0(x_0|c) \exp(r/\tau)/Z(c)$ *into* $q_t^*(x_t|c) = \int q_t(x_t|x_0)\, q^*(x_0|c)\mathrm{d}x_0$ *and using the Bayes factorization* $q_0(x_0|c)\, q_t(x_t|x_0) = q_t(x_t|c)\, q(x_0|x_t,c)$*:*

$$\begin{aligned} q_t^*(x_t|c) &= \frac{1}{Z(c)} \int q_t(x_t|x_0)\, q_0(x_0|c) \exp\!\left(\frac{r(x_0,c)}{\tau}\right) \mathrm{d}x_0 \\ &= \frac{q_t(x_t|c)}{Z(c)} \int q(x_0|x_t,c) \exp\!\left(\frac{r(x_0,c)}{\tau}\right) \mathrm{d}x_0 \\ &= \frac{q_t(x_t|c)}{Z(c)}\, \mathbb{E}_{x_0\sim q(x_0|x_t,c)}\left[\exp\!\left(\frac{r(x_0,c)}{\tau}\right)\right]. \end{aligned} \tag{20}$$

Taking the logarithm of the factored form:

$$\log q_t^*(x_t|c) = \log q_t(x_t|c) + \log \mathbb{E}_{x_0\sim q(x_0|x_t,c)}\left[\exp\!\left(\frac{r(x_0,c)}{\tau}\right)\right] - \log Z(c). \tag{21}$$

Applying ∇_{x_t}*: since* $Z(c)$ *does not depend on* x_t*, its gradient vanishes, giving*

$$\nabla_{x_t} \log q_t^*(x_t|c) = \underbrace{\nabla_{x_t} \log q_t(x_t|c)}_{=\, s_{\mathrm{ref}}(x_t,t,c)} + \underbrace{\nabla_{x_t} \log \mathbb{E}_{x_0\sim q(x_0|x_t,c)}\left[\exp\!\left(\frac{r(x_0,c)}{\tau}\right)\right]}_{\mathrm{DRS}(x_t,t,c)}, \tag{22}$$

which is exactly the decomposition in Eq. (6)*. The DRS term is intractable as written; Appendix C.4 derives a differentiable estimator for it.*

C.4 Proof of Proposition 5 (Diffused Reward Proxy)

Proof 4 *We derive a differentiable estimator for*

$$\mathrm{DRS}(x_t,t,c) = \nabla_{x_t} \log \mathbb{E}_{x_0\sim q(x_0|x_t,c)}\left[\exp\!\left(\frac{r(x_0,c)}{\tau}\right)\right]. \tag{23}$$

By the identity $\nabla \log f = \nabla f / f$*:*

$$\mathrm{DRS}(x_t,t,c) = \frac{\nabla_{x_t}\, \mathbb{E}_{x_0\sim q(x_0|x_t,c)}\left[\exp\!\left(\frac{r(x_0,c)}{\tau}\right)\right]}{\mathbb{E}_{\tilde{x}_0\sim q(\tilde{x}_0|x_t,c)}\left[\exp\!\left(\frac{r(\tilde{x}_0,c)}{\tau}\right)\right]}. \tag{24}$$

To handle the numerator, we obtain differentiable approximate samples from the reference posterior $q(x_0|x_t,c)$ *by running a denoising chain with the frozen reference model. Denote this map* $x_0 = \mathcal{G}_{\mathrm{ref}}(x_t, \boldsymbol{\epsilon})$*, initialized at* $\hat{x}_{t_S} = x_t$*, where* $\boldsymbol{\epsilon} = (\epsilon^{(S)}, \ldots, \epsilon^{(1)})$ *is the sequence of injected noise vectors (empty for deterministic chains). The chain differs by model type:*

VP diffusion (SDXL). *At each step $j = S, S-1, \ldots, 1$, the Tweedie formula gives a clean-image estimate,*

$$\hat{x}_0^{(j)} = \frac{\hat{x}_{t_j} + \sigma_{t_j}^2 \, s_{\mathrm{ref}}(\hat{x}_{t_j}, t_j, c)}{\alpha_{t_j}},$$

followed by a stochastic DDPM-style update

$$\hat{x}_{t_{j-1}} = \alpha_{t_{j-1}} \hat{x}_0^{(j)} + \sqrt{\sigma_{t_{j-1}}^2 - \tilde{\sigma}_{t_j}^2} \, \frac{\hat{x}_{t_j} - \alpha_{t_j} \hat{x}_0^{(j)}}{\sigma_{t_j}} + \tilde{\sigma}_{t_j} \, \epsilon^{(j)}, \qquad \epsilon^{(j)} \sim \mathcal{N}(0, I).$$

where $\tilde{\sigma}_{t_j} = \sigma_{t_{j-1}} \sqrt{1 - \dfrac{\alpha_{t_j}^2 \, \sigma_{t_{j-1}}^2}{\alpha_{t_{j-1}}^2 \, \sigma_{t_j}^2}}$ is the DDPM posterior standard deviation. Injecting independent noise $\epsilon^{(j)}$ for each of the K chains yields K diverse approximate posterior samples; the output of each chain is $\hat{x}_0^{(1)}$.

Flow matching (Z-Image). *A fixed Euler schedule yields identical $\hat{x}_0$ across all K chains. Instead, each chain draws S timesteps uniformly at random from $(0, t_S]$, runs the deterministic Euler ODE on its own schedule,*

$$\hat{x}_{t_{j-1}^{(k)}} = \hat{x}_{t_j^{(k)}} + \left(t_{j-1}^{(k)} - t_j^{(k)}\right) v_{\mathrm{ref}}\left(\hat{x}_{t_j^{(k)}}, t_j^{(k)}, c\right),$$

and returns $\hat{x}_0^{(1)} = \hat{x}_{t_1^{(k)}} - t_1^{(k)} \, v_{\mathrm{ref}}(\hat{x}_{t_1^{(k)}}, t_1^{(k)}, c)$, yielding K diverse approximate denoising endpoints induced by randomized discretization paths.

In both cases $\mathcal{G}_{\mathrm{ref}}$ is differentiable in x_t for fixed $\boldsymbol{\epsilon}$, since all operations are compositions of smooth network evaluations and linear maps. Under this reparameterization the expectation becomes $\mathbb{E}_{\boldsymbol{\epsilon}}[\exp(r(\mathcal{G}_{\mathrm{ref}}(x_t, \boldsymbol{\epsilon}), c)/\tau)]$, in which x_t enters explicitly through $\mathcal{G}_{\mathrm{ref}}$ rather than through the measure. Interchanging gradient and expectation via dominated convergence (justified when r is smooth with bounded gradient and $\mathcal{G}_{\mathrm{ref}}$ is uniformly Lipschitz in x_t):

$$\nabla_{x_t} \mathbb{E}_{x_0 \sim q(x_0|x_t,c)} \left[\exp\left(\frac{r(x_0, c)}{\tau} \right) \right] \approx \mathbb{E}_{\boldsymbol{\epsilon}} \left[\nabla_{x_t} \exp\left(\frac{r(\mathcal{G}_{\mathrm{ref}}(x_t, \boldsymbol{\epsilon}), c)}{\tau} \right) \right]. \tag{25}$$

Applying the chain rule:

$$\nabla_{x_t} \exp\left(\frac{r(\mathcal{G}_{\mathrm{ref}}(x_t, \boldsymbol{\epsilon}), c)}{\tau} \right) = \exp\left(\frac{r(x_0, c)}{\tau} \right) \cdot \frac{1}{\tau} \underbrace{\frac{\partial r(x_0, c)}{\partial x_0} \frac{\partial \mathcal{G}_{\mathrm{ref}}(x_t, \boldsymbol{\epsilon})}{\partial x_t}}_{\triangleq \nabla_{x_t} r(x_0, c)}, \tag{26}$$

where $\nabla_{x_t} r(x_0, c)$ denotes the pathwise gradient of r through the differentiable denoising chain. Converting back to the $q(x_0|x_t, c)$ expectation:

$$\nabla_{x_t} \mathbb{E}_{x_0 \sim q(x_0|x_t,c)} \left[\exp\left(\frac{r(x_0, c)}{\tau} \right) \right] = \mathbb{E}_{x_0 \sim q(x_0|x_t,c)} \left[\exp\left(\frac{r(x_0, c)}{\tau} \right) \cdot \frac{1}{\tau} \nabla_{x_t} r(x_0, c) \right]. \tag{27}$$

Dividing numerator and denominator:

$$\mathrm{DRS}(x_t, t, c) = \mathbb{E}_{x_0 \sim q(x_0|x_t,c)} \left[\frac{\exp\left(\frac{r(x_0,c)}{\tau} \right)}{\mathbb{E}_{\tilde{x}_0 \sim q(\tilde{x}_0|x_t,c)} \left[\exp\left(\frac{r(\tilde{x}_0,c)}{\tau} \right) \right]} \cdot \frac{1}{\tau} \nabla_{x_t} r(x_0, c) \right], \tag{28}$$

which is precisely the softmax-weighted gradient estimator in Eq. (10)*.*

Remark 8 (Regularity assumptions) *The interchange of gradient and expectation above requires: (i) r is smooth with bounded gradient, and (ii) $\mathcal{G}_{\mathrm{ref}}$ is uniformly Lipschitz in x_t. Condition (i) holds for standard differentiable reward models such as PickScore. Condition (ii) holds because $\mathcal{G}_{\mathrm{ref}}$ is a finite composition of smooth network evaluations and linear maps, hence globally Lipschitz on any compact domain.*

C.5 A Bimodal Gaussian Example of Terminal Reward Domination

All results below hold under the mild assumption that the two modes are well-separated ($\mu/\sigma \gg 1$), in which regime $\Phi(-\mu/\sigma) \approx 0$ and every approximation becomes exact as $\mu/\sigma \to \infty$. The $\approx$ symbol denotes equality up to corrections of order $\Phi(-\mu/\sigma)$.

Setup. Consider the bimodal reference

$$p_0 = \tfrac{1}{2}\mathcal{N}(-\mu, \sigma^2) + \tfrac{1}{2}\mathcal{N}(\mu, \sigma^2), \qquad \mu > \sigma > 0,$$

the generator family

$$q_{\alpha,0} = (1-\alpha)\mathcal{N}(-\mu, \sigma^2) + \alpha\mathcal{N}(\mu, \sigma^2), \qquad \alpha \in [0,1],$$

and the binary reward $r(x_0) = \mathbf{1}[x_0 > 0]$.

Exact reward expectation. By direct computation,

$$\begin{aligned}\mathbb{E}_{q_{\alpha,0}}[r] &= (1-\alpha)\, P\big(\mathcal{N}(-\mu,\sigma^2) > 0\big) + \alpha\, P\big(\mathcal{N}(\mu,\sigma^2) > 0\big) \\ &= (1-\alpha)\,\Phi\big(-\tfrac{\mu}{\sigma}\big) + \alpha\,\Phi\big(\tfrac{\mu}{\sigma}\big) = \alpha + (1-2\alpha)\,\Phi\big(-\tfrac{\mu}{\sigma}\big)\,, \end{aligned} \tag{29}$$

where Φ is the standard normal CDF. Since $\Phi(-\mu/\sigma) \to 0$ as $\mu/\sigma \to \infty$,

$$\mathbb{E}_{q_{\alpha,0}}[r] \approx \alpha, \qquad \frac{d}{d\alpha}\mathbb{E}_{q_{\alpha,0}}[r]\big|_{\alpha=1} = 1 - 2\Phi\big(-\tfrac{\mu}{\sigma}\big) \approx 1.$$

Forward marginals. Under the VP forward process with $\bar{\alpha}_t = e^{-\gamma t}$, $t \in [0,\infty)$, the forward kernel is $q_t(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)I)$, so the marginals are

$$q_{\alpha,t} = (1-\alpha)\phi_{-,t} + \alpha\phi_{+,t}, \qquad p_t = \tfrac{1}{2}\phi_{-,t} + \tfrac{1}{2}\phi_{+,t},$$

where $\phi_{\pm,t} = \mathcal{N}(\pm m_t, \Sigma_t)$ with

$$m_t = \sqrt{\bar{\alpha}_t}\,\mu, \qquad \Sigma_t = 1 - \bar{\alpha}_t(1-\sigma^2).$$

Objective and convexity. The endpoint-reward objective (to be *minimized*) is

$$\mathcal{L}_{\text{term}}(\alpha) = -\mathbb{E}_{q_{\alpha,0}}[r] + \tau \int_0^\infty D_t(\alpha)\,\mathrm{d}t, \qquad D_t(\alpha) := \mathrm{KL}(q_{\alpha,t}\|p_t).$$

Since $q_{\alpha,t}$ is affine in α and KL divergence is convex in its first argument, $D_t(\alpha)$ is convex in α, and $-\mathbb{E}_{q_{\alpha,0}}[r]$ is linear in α. Hence $\mathcal{L}_{\text{term}}$ is convex in α, so $\mathcal{L}'_{\text{term}}$ is non-decreasing on $[0,1]$. Consequently, $\alpha^* = 1$ (collapse to the positive mode) occurs whenever the left derivative satisfies

$$\mathcal{L}'_{\text{term}}(1^-) \le 0,$$

since convexity then forces $\mathcal{L}'_{\text{term}}(\alpha) \le \mathcal{L}'_{\text{term}}(1^-) \le 0$ for all $\alpha \in [0,1)$, making $\mathcal{L}_{\text{term}}$ non-increasing throughout and the minimum attained at $\alpha^* = 1$.

Computing $D'_t(1)$. Differentiating $D_t(\alpha) = \mathrm{KL}(q_{\alpha,t}\|p_t)$ with respect to α and using $\partial_\alpha q_{\alpha,t} = \phi_{+,t} - \phi_{-,t}$ (together with the fact that $\partial_\alpha \int q_{\alpha,t}\,\mathrm{d}x = 0$, so the term from differentiating the $\log q_{\alpha,t}$ factor vanishes by normalization), we obtain

$$D'_t(\alpha) = \int (\phi_{+,t} - \phi_{-,t})\,\log\frac{q_{\alpha,t}(x)}{p_t(x)}\,\mathrm{d}x.$$

At $\alpha = 1$, $q_{1,t} = \phi_{+,t}$, so

$$D'_t(1) = \int (\phi_{+,t}(x) - \phi_{-,t}(x))\,\log\frac{\phi_{+,t}(x)}{\frac{1}{2}(\phi_{+,t}(x) + \phi_{-,t}(x))}\,\mathrm{d}x.$$

Since $\phi_{\pm,t} = \mathcal{N}(\pm m_t, \Sigma_t)$, the Gaussian log-ratio is

$$\log\frac{\phi_{+,t}(x)}{\phi_{-,t}(x)} = -\frac{(x-m_t)^2 - (x+m_t)^2}{2\Sigma_t} = \frac{2m_t x}{\Sigma_t},$$

so $\phi_{-,t}(x)/\phi_{+,t}(x) = \exp(-2m_t x/\Sigma_t)$. Substituting:

$$\log \frac{\phi_{+,t}(x)}{\frac{1}{2}(\phi_{+,t} + \phi_{-,t})} = \log \frac{2}{1 + e^{-2m_t x/\Sigma_t}}.$$

We now simplify by symmetry. Split the integral into contributions from $\phi_{+,t}$ and $\phi_{-,t}$:

$$D_t'(1) = \int \phi_{+,t}(x) \log \frac{2}{1 + e^{-2m_t x/\Sigma_t}} \, \mathrm{d}x - \int \phi_{-,t}(x) \log \frac{2}{1 + e^{-2m_t x/\Sigma_t}} \, \mathrm{d}x.$$

In the second integral, substitute $x \mapsto -x$; since $\phi_{-,t}(-x) = \phi_{+,t}(x)$:

$$\int \phi_{-,t}(x) \log \frac{2}{1 + e^{-2m_t x/\Sigma_t}} \, \mathrm{d}x = \int \phi_{+,t}(x) \log \frac{2}{1 + e^{2m_t x/\Sigma_t}} \, \mathrm{d}x.$$

Combining both integrals:

$$D_t'(1) = \int \phi_{+,t}(x) \log \frac{1 + e^{2m_t x/\Sigma_t}}{1 + e^{-2m_t x/\Sigma_t}} \, \mathrm{d}x = \mathbb{E}_{x \sim \phi_{+,t}} \left[\log \frac{1 + e^{2m_t x/\Sigma_t}}{1 + e^{-2m_t x/\Sigma_t}} \right]. \tag{30}$$

Finally, we invoke the well-separated assumption. Define $a_t \triangleq m_t/\sqrt{\Sigma_t}$. Under $\phi_{+,t} = \mathcal{N}(m_t, \Sigma_t)$, a typical sample satisfies $x \approx m_t$, so $2m_t x/\Sigma_t \approx 2m_t^2/\Sigma_t = 2a_t^2$. In the regime $\mu/\sigma \gg 1$ (so $a_t \gg 1$ for all t where the diffusion has not yet erased the modes), $e^{2m_t x/\Sigma_t} \gg 1$ and $e^{-2m_t x/\Sigma_t} \approx 0$ throughout the support of $\phi_{+,t}$. Therefore,

$$\log \frac{1 + e^{2m_t x/\Sigma_t}}{1 + e^{-2m_t x/\Sigma_t}} \approx \log e^{2m_t x/\Sigma_t} = \frac{2m_t x}{\Sigma_t}.$$

Substituting into Eq. (30) and using $\mathbb{E}_{x \sim \phi_{+,t}}[x] = m_t$:

$$D_t'(1) \approx \mathbb{E}_{x \sim \phi_{+,t}} \left[\frac{2m_t x}{\Sigma_t} \right] = \frac{2m_t}{\Sigma_t} \, \mathbb{E}_{x \sim \phi_{+,t}}[x] = \frac{2m_t^2}{\Sigma_t}.$$

Collapse condition. Combining the derivative of the reward term (≈ -1 after negation) and the regularizer:

$$\mathcal{L}_{\text{term}}'(1) \approx -1 + \tau \int_0^\infty \frac{2m_t^2}{\Sigma_t} \, \mathrm{d}t.$$

The condition $\mathcal{L}_{\text{term}}'(1) \leq 0$ gives the collapse threshold

$$\tau \leq \frac{1}{B_{\text{crit}}}, \qquad B_{\text{crit}} \triangleq \int_0^\infty \frac{2m_t^2}{\Sigma_t} \, \mathrm{d}t.$$

Closed-form evaluation of B_{crit}. Substituting $m_t^2 = \bar{\alpha}_t \mu^2$ and $\Sigma_t = 1 - \bar{\alpha}_t(1 - \sigma^2)$:

$$B_{\text{crit}} = \int_0^\infty \frac{2\mu^2 \, e^{-\gamma t}}{1 - e^{-\gamma t}(1 - \sigma^2)} \, \mathrm{d}t.$$

Change variables $u = e^{-\gamma t}$, so $\mathrm{d}u = -\gamma u \, \mathrm{d}t$ and the limits $t : 0 \to \infty$ become $u : 1 \to 0$:

$$B_{\text{crit}} = \int_1^0 \frac{2\mu^2 \, u}{1 - (1 - \sigma^2)u} \cdot \frac{-\mathrm{d}u}{\gamma u} = \frac{2\mu^2}{\gamma} \int_0^1 \frac{\mathrm{d}u}{1 - (1 - \sigma^2)u}.$$

For $\sigma^2 \neq 1$, the integral evaluates via $\int_0^1 \frac{\mathrm{d}u}{1 - cu} = \frac{-\log(1-c)}{c}$ with $c = 1 - \sigma^2$:

$$B_{\text{crit}} = \frac{2\mu^2}{\gamma} \cdot \frac{-\log \sigma^2}{1 - \sigma^2}.$$

Therefore, the collapse condition is

$$\boxed{\tau \leq \frac{\gamma(1 - \sigma^2)}{2\mu^2(-\log \sigma^2)} \quad \Longrightarrow \quad \alpha^* = 1.}$$

The case $\sigma^2 = 1$ follows by L'Hôpital's rule (or a direct integral): $\int_0^1 \mathrm{d}u/(1 - 0) = 1$, giving

$$\boxed{\tau \leq \frac{\gamma}{2\mu^2} \quad \Longrightarrow \quad \alpha^* = 1.}$$

This shows that terminal reward domination occurs under the integral forward KL penalty for any finite τ below a closed-form threshold. Larger γ (faster diffusion) raises the critical τ, making collapse to the rewarded mode easier to trigger at any fixed regularization strength.

D 1-D Toy Experiment

This section describes the setup for the 1-D empirical validation of terminal reward domination (Figure 2(b)).

Reference distribution and reward. The reference is a symmetric bimodal Gaussian

$$q_0 = \tfrac{1}{2}\mathcal{N}(-\mu, \sigma^2) + \tfrac{1}{2}\mathcal{N}(\mu, \sigma^2), \qquad \mu = 2,\ \sigma = 0.5,$$

with ideal binary reward $r_{\mathrm{hard}}(x) = \mathbf{1}[x > 0]$. For gradient-based training, we use the smooth approximation

$$r(x) = \mathrm{sigmoid}(\beta x), \qquad \beta = 20,$$

so that $r(x)$ approximates $r_{\mathrm{hard}}(x)$ except in a narrow neighborhood of the decision boundary while remaining differentiable.

Forward process. VP diffusion with rate $\gamma = 20$: $\bar{\alpha}_t = e^{-\gamma t}$, $t \in [0, T]$, $T = 0.25$, giving $\bar{\alpha}_T = e^{-5} \approx 0.007$. This places the experiment firmly in the collapse regime: the closed-form domination threshold derived in Appendix C.5 evaluates to $\tau_{\mathrm{crit}} \approx 1.35 > \tau = 1$.

Networks. Both s_{ref} and s_ψ are MLPs with input $(x, t) \in \mathbb{R}^2$, three hidden layers of width 128, SiLU activations, and a scalar output (noise-prediction parameterization; converted to score via $s = -\hat{\varepsilon}/\sqrt{1 - \bar{\alpha}_t}$). The generator $g_\theta : \mathcal{N}(0, 1) \to \mathbb{R}$ uses the same architecture with a scalar input z.

Training pipeline.

1. **Reference model.** s_{ref} is trained for 10,000 steps via DSM on samples from q_0 (Adam, lr=3×10^{-4}, batch 2,048). Weights are frozen afterwards.
2. **Generator distillation.** g_θ is initialized by regressing against 30-step DDIM samples from s_{ref} using matched noise (3,000 steps, Adam, lr=10^{-3}, batch 2,048).
3. **Teaching Assistant initialization.** s_ψ is copied from s_{ref} and fine-tuned during alignment.
4. **Alignment.** 6,000 outer steps; each step runs 5 TA DSM updates (Adam, lr=3×10^{-4}) followed by one generator update (Adam, lr=10^{-4}, batch 2,048). The two methods differ only in s_{target}:

Method	Reward location
DI++	endpoint x_0 only
DIDR	every noise level via DRS

Evaluation. Final generators are evaluated on $n = 10{,}000$ samples using the hard decision boundary $x > 0$. For the ideal binary reward r_{hard}, the theoretical optimal weight under q^* is $P_{q^*}(x > 0) = \mathrm{sigmoid}(1/\tau) \approx 0.731$. DI++ collapses to $P(x > 0) \approx 1$ while DIDR converges to $P(x > 0) \approx 0.73$, confirming the theory.

E Evaluation Metrics

We use automatic metrics that measure complementary aspects of text-to-image generation. Preference metrics estimate human visual preference or aesthetics; text-alignment metrics measure whether the image satisfies the prompt; FID measures distributional fidelity against real images. For all metrics except FID, higher is better.

Preference metrics. PickScore [Kirstain et al., 2023] is a CLIP-based human preference model trained on Pick-a-Pic pairwise user preferences, and is also the reward used for DIDR training. ImageReward [Xu et al., 2023] is a learned image-text reward model trained from human preference annotations for text-to-image outputs. HPSv2.1 [Wu et al., 2023] evaluates human preference using the Human Preference Score benchmark and reports results over its standard prompt categories. Aesthetic Score [Schuhmann, 2022] uses the LAION aesthetic predictor, which estimates visual appeal from CLIP image embeddings and does not directly measure prompt faithfulness.

Text-alignment metrics. CLIPScore [Hessel et al., 2021, Radford et al., 2021] measures image-text compatibility using CLIP embedding similarity. DPG-Bench [Hu et al., 2024] evaluates semantic adherence on dense prompts containing multiple attributes and relations. GenEval [Ghosh et al., 2023] is an object-focused benchmark that checks compositional correctness, including object presence, counting, color binding, position, and attribute relations.

Fidelity metric. FID [Heusel et al., 2017] compares generated and real image distributions in Inception feature space:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\Big),$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are Gaussian approximations of real and generated feature distributions. Lower FID indicates closer distributional match to the reference images, but it does not directly measure prompt alignment or human preference.

F Per-Category Results

Tables 5 and 6 provide per-category breakdowns of GenEval and HPSv2.1 scores respectively. On GenEval, DIDR achieves the highest overall score among one-step SDXL methods, with particularly strong gains on counting (0.51 vs. 0.43 for DI*) and color attribute binding (0.25 vs. 0.22). On HPSv2.1, $\text{DIDR}_{\text{longer}}$ leads across all four style categories. For the Z-Image backbone, Z-Image DIDR improves substantially over the 1-step Z-Image-Turbo initialization on all categories, suggesting that the gains come from the alignment procedure rather than initialization quality. Against Z-Image-Turbo with twice as many steps, Z-Image DIDR leads on all four HPSv2.1 styles at 4 steps and on three of four at 1 step, but trails on GenEval at both budgets, most notably on position (0.65 vs. 0.34); we attribute this to the PickScore reward, which favors aesthetics over compositional correctness.

Table 5: GenEval per-category breakdown. **Bold**: best result within each group. −: not reported.

				GenEval Category ↑						
Model	**Steps**	**Arch.**	**Params**	**Single Obj.**	**Two Obj.**	**Counting**	**Colors**	**Color Attr.**	**Position**	**Overall ↑**
Multi-step reference										
SDXL [Podell et al., 2024]	50	UNet	2.6B	0.97	0.75	0.38	0.85	0.22	0.14	0.55
SDXL-DPO [Wallace et al., 2024]	50	UNet	2.6B	0.97	0.75	0.38	0.84	0.22	0.14	0.55
SD3.5-large [AI, 2024]	28	DiT	8B	1.00	0.92	0.67	0.85	0.59	0.27	0.72
FLUX-dev [Labs, 2024]	50	DiT	12B	0.99	0.86	0.68	0.77	0.49	0.24	0.67
Z-Image [Tongyi-MAI, 2025]	50	DiT	6B	1.00	0.94	0.78	0.93	0.77	0.62	0.84
One-step SDXL										
SDXL DMD2 [Yin et al., 2024a]	1	UNet	2.6B	0.98	0.68	0.47	0.88	0.22	**0.11**	0.56
SDXL Diff-Instruct [Luo et al., 2023]	1	UNet	2.6B	0.99	0.75	0.46	0.88	0.19	0.09	0.56
SDXL Diff-Instruct++ [Luo, 2024]	1	UNet	2.6B	0.99	0.74	0.48	0.87	0.20	0.09	0.56
SDXL Diff-Instruct* [Luo et al., 2025]	1	UNet	2.6B	0.99	0.74	0.43	0.85	0.22	**0.11**	0.56
SDXL DIDR (Ours)	1	UNet	2.6B	0.99	0.73	**0.51**	**0.89**	**0.25**	**0.11**	**0.58**
SDXL $\text{DIDR}_{\text{longer}}$ (Ours)	1	UNet	2.6B	**1.00**	**0.76**	0.47	0.88	0.19	0.10	0.57
Z-Image backbone, 4–8 steps										
Z-Image-Turbo [Tongyi-MAI, 2025]	8	DiT	6B	**1.00**	**0.95**	**0.77**	**0.89**	**0.68**	**0.65**	**0.82**
Z-Image DIDR (Ours)	4	DiT	6B	**1.00**	0.93	0.68	0.82	0.61	0.34	0.73
Z-Image backbone, 1–2 steps										
Z-Image-Turbo [Tongyi-MAI, 2025]	1	DiT	6B	0.97	0.51	0.56	0.80	0.30	0.29	0.57
Z-Image-Turbo [Tongyi-MAI, 2025]	2	DiT	6B	**1.00**	**0.90**	**0.76**	**0.86**	**0.60**	**0.47**	**0.76**
Z-Image DIDR (Ours)	1	DiT	6B	**1.00**	0.85	0.63	0.81	0.53	0.34	0.69

HPSv2.1 Per-Category Results

G Full Training Algorithm

Algorithm 2 provides the complete pseudocode for DIDR, expanding the concise version in Algorithm 1. The outer loop alternates between two stages. Stage I updates the Teaching Assistant s_ψ via denoising score matching to track the generator's current marginals. Stage II updates the generator g_θ: for each training step it samples a prompt and noise, forward-diffuses the generated image to a random time t, draws K posterior samples via S-step reference denoising, computes the DRP estimate s_r as a softmax-weighted gradient, and applies the IKL gradient. The CFG-corrected reference score $\tilde{s}_{\text{ref}}$ is used throughout to maintain text alignment.

Table 6: HPSv2.1 per-category breakdown. Same conventions as Table 2.

Model	Steps	Arch.	Params	HPSv2.1 ↑				
				Animation	Concept-Art	Painting	Photo	Average
Multi-step reference								
SDXL [Podell et al., 2024]	50	UNet	2.6B	30.89	29.16	28.91	26.97	28.98
SDXL-DPO [Wallace et al., 2024]	50	UNet	2.6B	32.08	30.80	30.82	28.12	30.45
SD3.5-large [AI, 2024]	28	DiT	8B	31.87	30.05	30.31	28.01	30.06
FLUX-dev [Labs, 2024]	50	DiT	12B	31.95	30.35	31.21	29.10	30.65
Z-Image [Tongyi-MAI, 2025]	50	DiT	6B	32.53	30.42	29.58	29.76	30.57
One-step SDXL								
SDXL DMD2 [Yin et al., 2024a]	1	UNet	2.6B	31.53	29.86	29.43	29.14	29.99
SDXL Diff-Instruct [Luo et al., 2023]	1	UNet	2.6B	32.78	31.39	31.29	30.59	31.51
SDXL Diff-Instruct++ [Luo, 2024]	1	UNet	2.6B	32.74	31.42	31.20	30.41	31.44
SDXL Diff-Instruct* [Luo et al., 2025]	1	UNet	2.6B	35.10	33.63	33.14	31.28	33.29
SDXL DIDR (Ours)	1	UNet	2.6B	35.64	34.21	33.81	31.41	33.77
SDXL $\text{DIDR}_{\text{longer}}$ (Ours)	1	UNet	2.6B	**35.76**	**34.43**	**33.83**	**31.53**	**33.89**
Z-Image backbone, 4–8 steps								
Z-Image-Turbo [Tongyi-MAI, 2025]	8	DiT	6B	33.44	31.99	31.26	30.76	31.86
Z-Image DIDR (Ours)	4	DiT	6B	**35.24**	**33.95**	**33.27**	**32.00**	**33.62**
Z-Image backbone, 1–2 steps								
Z-Image-Turbo [Tongyi-MAI, 2025]	1	DiT	6B	26.50	24.22	23.63	23.81	24.54
Z-Image-Turbo [Tongyi-MAI, 2025]	2	DiT	6B	32.87	31.30	30.26	**30.22**	31.16
Z-Image DIDR (Ours)	1	DiT	6B	**33.39**	**31.62**	**30.38**	30.21	**31.40**

Algorithm 2: DIDR: Diff-Instruct with Diffused Reward (full pseudocode)

Input : prompt dataset $\mathcal{C}$; generator g_θ; prior p_z; reward model $r(\cdot, c)$; CFG scale α_{cfg}; reference diffusion s_{ref}; Teaching Assistant (TA) s_ψ; posterior samples K; temperature τ; denoising steps S; TA update rounds K_{TA}; weights $w(t), \lambda(t)$.

while not converged **do**

  Stage I: Update Teaching Assistant (TA);

  for $k = 1$ **to** K_{TA} **do**

    Sample $c \sim \mathcal{C}, z \sim p_z, t \sim \pi(t)$;

    Generate $x_0 = g_\theta(z, c)$, diffuse $x_t \sim p_t(x_t|x_0)$;

    Update ψ by minimizing DSM loss: $\mathcal{L}(\psi) = \lambda(t)\, \|s_\psi(x_t, t, c) - \nabla_{x_t} \log p_t(x_t|x_0)\|_2^2$;

  end

  Stage II: Update Generator;

  Sample $c \sim \mathcal{C}, z \sim p_z, t \sim \pi(t)$;

  Generate $x_0 = g_\theta(z, c)$, diffuse $x_t \sim p_t(x_t|x_0)$;

  Compute CFG score: $\tilde{s}_{\text{ref}} = s_{\text{ref}}(x_t, t, \emptyset) + \alpha_{\text{cfg}}\, [s_{\text{ref}}(x_t, t, c) - s_{\text{ref}}(x_t, t, \emptyset)]$;

  Compute DRP estimate s_r via posterior sampling:;

  for $k = 1$ **to** K **do**

    Run S-step differentiable denoising from x_t: $\hat{x}_0^{(k)} \leftarrow \text{Denoise}_S(x_t, s_{\text{ref}}, c)$;

    Compute $r^{(k)} = r(\hat{x}_0^{(k)}, c)$ and pathwise gradient $\nabla_{x_t} r^{(k)}$;

  end

  Compute softmax weights: $\omega^{(k)} = \dfrac{\exp(r^{(k)}/\tau)}{\sum_{k'=1}^{K} \exp(r^{(k')}/\tau)}$;

  Compute $s_r = \dfrac{1}{\tau} \displaystyle\sum_{k=1}^{K} \omega^{(k)}\, \nabla_{x_t} r^{(k)}$;

  Update θ via Eq. (13): $\text{Grad}(\theta) = w(t)\, \big(s_\psi(x_t, t, c) - (\tilde{s}_{\text{ref}} + s_r)\big)\, \frac{\partial x_t}{\partial \theta}$;

end

return θ, ψ.

H Implementation Details

H.1 SDXL Experiments

Generator Architecture. SDXL [Podell et al., 2024] is a latent diffusion model built on a UNet backbone with approximately 2.6B parameters. It employs a dual text-encoder design: a CLIP ViT-L encoder producing $d_1 = 768$-dimensional embeddings $c_{\text{CLIP}} \in \mathbb{R}^{768}$ and an OpenCLIP ViT-bigG encoder producing $d_2 = 1280$-dimensional embeddings $c_{\text{bigG}} \in \mathbb{R}^{1280}$; these are concatenated to

form the conditioning vector

$$c = [c_{\text{CLIP}};\, c_{\text{bigG}}] \in \mathbb{R}^{2048}. \tag{31}$$

Images $x \in \mathbb{R}^{3\times1024\times1024}$ are encoded by the SDXL VAE with $8\times$ spatial downsampling into a 4-channel latent $z = \mathcal{E}(x) \in \mathbb{R}^{4\times128\times128}$, and decoded back via $\hat{x} = \mathcal{D}(z)$. The base UNet is trained natively at 1024×1024 with micro-conditioning on original and target image sizes to suppress training-resolution artifacts. SDXL supports an optional two-stage pipeline (base UNet for high-noise denoising, refiner UNet for low-noise enhancement); we use only the base UNet as the reference model s_{ref}, which is queried with classifier-free guidance:

$$\tilde{s}_{\text{ref}}(x_t, t, c) = s_{\text{ref}}(x_t, t, \emptyset) + \alpha_{\text{cfg}}\big(s_{\text{ref}}(x_t, t, c) - s_{\text{ref}}(x_t, t, \emptyset)\big). \tag{32}$$

The forward VP-SDE corrupts data as

$$q_t(x_t \mid x_0) = \mathcal{N}\big(\alpha_t\, x_0,\, \sigma_t^2 I\big), \quad \alpha_t^2 + \sigma_t^2 = 1, \tag{33}$$

where (α_t, σ_t) follow the noise schedule of the reference SDXL checkpoint. The one-step generator g_θ shares the same UNet architecture and latent space, mapping a noise vector $z \sim \mathcal{N}(0, I)$ directly to a clean latent,

$$\hat{z}_0 = g_\theta(z, c) \in \mathbb{R}^{4\times128\times128}, \tag{34}$$

which is decoded to a 1024×1024 image by the frozen SDXL VAE. We initialize g_θ from the DMD2-SDXL-1step checkpoint [Yin et al., 2024a]. Both s_{ref} and the Teaching Assistant (TA) s_ψ are initialized from the pretrained 50-step SDXL base model. We set $\sigma_{\text{init}} = 2.5$, the noise magnitude at which the generator input z is injected into the forward diffusion during training, following Diff-Instruct [Luo et al., 2023] and SiD [Zhou et al., 2024b].

Training Setup. All models are trained on text prompts from LAION-Aesthetic-6.25+ [Zhou et al., 2024a] with no image data. We use PickScore [Kirstain et al., 2023] as the reward model. Both g_θ and s_ψ are optimized with Adam [Kingma and Ba, 2015] ($\beta_1 = 0.0$, $\beta_2 = 0.999$), learning rate 5×10^{-6}, per-GPU batch size 1, and effective batch size 512 via gradient accumulation. The CFG scale for all reference model evaluations is $\alpha_{\text{cfg}} = 7.5$. The generator loss time weighting is $w(t) = 1$ for all t [Luo et al., 2023], and the TA DSM weighting $\lambda(t)$ follows the default SDXL training schedule. DRP hyperparameters are $K = 4$, $S = 4$, $\tau = 0.01$. Training is conducted on 8 H100 GPUs (DIDR: 48 hours; $\text{DIDR}_{\text{longer}}$: 72 hours).

Ablation Experiments. All ablation variants (Tables 3 and 4) share the same setup as standard DIDR: same initialization, optimizer, and training duration (48 hours on 8 H100 GPUs). Only the hyperparameter under study is varied; all others are fixed at the default values ($K = 4$, $S = 4$, $\tau = 0.01$).

H.2 Z-Image Experiments

Generator Architecture. Z-Image [Tongyi-MAI, 2025] is a 6.15B-parameter text-to-image model built on a Scalable Single-Stream Diffusion Transformer (S3-DiT) with 30 transformer layers (hidden dimension 3840, 30 attention heads, FFN intermediate dimension 10240). S3-DiT concatenates text tokens h_{text} (from a Qwen3-4B encoder) and VAE image tokens h_{img} (encoded by the Flux VAE) into a single unified sequence

$$h = [h_{\text{text}};\, h_{\text{img}}], \tag{35}$$

which is processed jointly by all transformer layers, with 3D Unified RoPE for positional encoding and RMSNorm / QK-Norm / Sandwich-Norm for stability. In contrast to SDXL's VP-SDE, Z-Image adopts a rectified flow (flow matching) parameterization [Liu et al., 2022]. The forward process interpolates linearly between data x_0 and noise $\epsilon \sim \mathcal{N}(0, I)$:

$$x_t = (1 - t)\, x_0 + t\, \epsilon, \quad t \in [0, 1], \tag{36}$$

and the network is trained to predict the velocity field $v^*(x_t, t, c) = \epsilon - x_0$. Inference integrates the ODE

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_\phi(x_t, t, c) \tag{37}$$

via Euler steps $x_{t+\Delta t} = x_t + \Delta t\, v_\phi(x_t, t, c)$, with the 50-step base model serving as s_{ref} (expressed in score form as $s_{\text{ref}} = -[x_t + (1 - t)v_\phi]/t$ for $t > 0$).

The one-step generator g_θ maps noise $z \sim \mathcal{N}(0, I)$ directly to a clean image,

$$\hat{x}_0 = g_\theta(z, c), \tag{38}$$

and is initialized from Z-Image-Turbo [Tongyi-MAI, 2025], a consistency-distilled variant capable of high-quality generation in 8 steps. Initializing from Z-Image-Turbo rather than the raw 50-step base model provides a stronger single-step starting point, as it has already undergone consistency distillation. The frozen reference s_{ref} (equivalently v_{ref}) and TA initialization are taken from the full 50-step Z-Image base model.

Training Setup. We follow the same data and reward setup as the SDXL experiments (LAION-Aesthetic-6.25+ prompts, PickScore reward). The optimizer is Adam ($\beta_1 = 0.0$, $\beta_2 = 0.999$) with a reduced learning rate of 1×10^{-6} to accommodate the larger model scale. We use per-GPU batch size 1, effective batch size 256, and CFG scale $\alpha_{\text{cfg}} = 7.5$. DRP hyperparameters are $K = 2$, $S = 4$, $\tau = 0.01$. Training is conducted on 8 H100 GPUs (96 hours). The 4-step variant uses identical hyperparameters and is trained for 192 hours.

DRP Posterior Sampling for Flow Matching. As derived in Appendix C.4, diversity across the K posterior chains is induced by randomized discretization: each chain draws S timesteps uniformly at random from $(0, t_S]$ and runs the deterministic Euler ODE on its own schedule, yielding K diverse approximate denoising endpoints.

Table 7: Hyperparameter summary for all DIDR model variants.

Hyperparameter	**SDXL DIDR**	**SDXL DIDR$_{\text{longer}}$**	**Z-Image DIDR (1-step)**	**Z-Image DIDR (4-step)**
Generator init	DMD2-SDXL-1step	DMD2-SDXL-1step	Z-Image-Turbo	Z-Image-Turbo
Reference diffusion	SDXL (VP)	SDXL (VP)	Z-Image (flow)	Z-Image (flow)
Reward model	PickScore	PickScore	PickScore	PickScore
CFG scale α_{cfg}	7.5	7.5	7.5	7.5
Posterior samples K	4	4	2	2
Denoising steps S	4	4	4	4
Temperature τ	0.01	0.01	0.01	0.01
Learning rate	5×10^{-6}	5×10^{-6}	1×10^{-6}	1×10^{-6}
Effective batch size	512	512	256	256
Training time	48 hours	72 hours	96 hours	192 hours

I Image Generation Prompts

The following lists the text prompts used to generate all qualitative figures in the paper.

Figure 1 — Teaser

1. A tranquil dormant mountain dusted with fresh snow under a clear pale blue daytime sky, calm still air, soft diffuse light, muted cool whites and blues, serene and peaceful alpine landscape
2. Close-up macro of a human eye with a vivid blue iris, glittering sparkles and cosmic galaxy-like reflections, extreme detail, macro photography
3. Portrait of a young African woman wearing a blue and yellow headwrap and a pearl earring, classical painting style, warm studio lighting, elegant and dignified
4. A majestic tall sailing ship navigating through a glowing cosmic nebula, teal and cyan hues, dramatic fantasy scene, epic scale
5. A vivid explosion of colorful paint splashing in mid-air, rainbow hues of orange, yellow, magenta, blue and purple, high-speed photography, dark background
6. A tabby cat leaping through the air in a frozen action shot, photorealistic, natural daylight background
7. An assorted sushi platter with rolls and nigiri, glowing neon blue ring lighting, futuristic and vibrant food photography

8. A lush bouquet of deep red roses in a round green ceramic vase, soft natural window light, elegant floral still life
9. A woman's portrait blending photorealism with blue watercolor ink splashes, artistic double-exposure style, ethereal and dreamy

Figure 5 — Qualitative Comparison

SDXL comparison:

1. Dramatic close-up portrait of a rugged elderly man with a weathered face, white beard, dark flat cap, overcast sea in the background
2. A baby elephant in rain boots jumping in puddles under a rainbow, cheerful children's book illustration.
3. An oil painting of a polar bear mother and cub drifting on an ice floe at Arctic dusk, warm amber sky.

Z-Image comparison:

1. A beautiful young woman in a white linen dress, Mediterranean coastal town background, warm golden hour light, photorealistic portrait
2. A majestic white stallion galloping through crashing ocean waves at sunrise, cinematic photography.
3. A young wizard girl casting a glowing spell in a floating sky library, anime key visual style.

Figure 7 — Temperature Ablation

1. Aurora borealis over a calm reflective lake surrounded by snow-covered pine trees, winter night
2. Volcanic eruption with lava flows meeting snow and ice, dramatic fire and steam, epic landscape
3. White arctic fox standing in snow, winter landscape, photorealistic wildlife photography
4. Desert campsite at night with glowing orange tents and campfire, Milky Way visible overhead

Figure 8 — Multi-step Comparison

1. Japanese ukiyo-e woodblock print of Mount Fuji with cherry blossom trees and a lone figure on a wooden bridge
2. Cute anthropomorphic fox in a cozy autumn café, illustrated anime art style, fallen leaves, warm tones
3. Aerial drone view of a tropical volcanic island with a turquoise crater lake, surrounded by ocean
4. Portrait of a Rajasthani woman with traditional Indian jewelry and a colorful headscarf, photorealistic

Figure 9 — Failure Cases

1. A woman placing a tray of food into a kitchen oven
2. A large crowd of people standing in a queue outdoors
3. A round wooden dining table in a bright room
4. A person cutting raw meat on a kitchen counter surrounded by fresh vegetables

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction explicitly state the three contributions (DIDR framework, DRS/DRP derivation, empirical results on SDXL and Z-Image), and the experimental section validates each claim.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 contains a dedicated Limitations paragraph covering reward-model bias and reward hacking risks, increased training cost from $K \times S$ denoising steps, and persistent failure modes in single-step generation (counting, fine-grained structure, complex scenes).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results (Example 1, Propositions 1–2, Theorem 1, and associated Remarks) are numbered, cross-referenced in the main text, and proved in Appendix C, with the target-score decomposition folded into the proof of Theorem 1.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4 describes models, training data, reward model, and evaluation benchmarks. Full hyperparameters (optimizer, learning rate, batch size, K, S, τ, GPU setup) are provided in Appendix H.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The training and evaluation code is released at `https://github.com/Junyi-W/DIDR`. The method uses only publicly available prompts, models, and reward models.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 describes experimental setup; Appendix H provides implementation details including the optimizer, learning rates, batch size, DRP hyperparameters, training duration, and hardware.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported due to the high compute cost of rerunning full training; each model variant requires hundreds of H100 GPU hours. Results are reported as single-run evaluations on standard benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix H reports GPU type (H100), number of GPUs, and approximate training time for each experiment (SDXL and Z-Image runs).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves no human subjects, no scraped personal data, and no harmful applications. The work aligns generative models with human preference signals using publicly available reward models and text prompts.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 6 (Limitations) notes positive impacts (improved preference alignment and visual quality for one-step generators) and negative risks (reward-model bias amplification and reward hacking). The failure-case appendix (Figure 9) further illustrates current scope limitations.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: No model weights or datasets are released at this time. Upon release, standard safety filters (as used in the base SDXL and Z-Image models) will apply.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All pretrained models (SDXL, DMD2, Z-Image, PickScore), datasets (LAION-Aesthetic, MSCOCO), and evaluation benchmarks (HPSv2.1, GenEval, DPG-Bench) are properly cited throughout the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new datasets or publicly released model checkpoints are introduced in this submission.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects; all evaluations use automatic metrics.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects are involved in this research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as part of the core method; any use was limited to writing assistance.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.