ADVFIDELITY: UNIFYING PREFERENCE ALIGNMENT AND FEW-STEP IMAGE GENERATION WITHIN REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Practical image generation demands both alignment with human preference and few-step efficiency, yet the corresponding post-training procedures, reinforcement learning (RL) and few-step distillation, remain difficult to unify. We observe that few-step RL training carries an intrinsic distillation effect. Although few-step sampling on its own collapses into incoherent structures, generations evolve toward structural coherence as RL training proceeds. However, this effect could not remove grid-pattern artifacts, and preference rewards are inaccurate on such low-level degradations. Therefore, we introduce a fidelity reward that makes the intrinsic distillation effect explicit and reliable by supervising the removal of these artifacts. In AdvFidelity, this reward is realized as a discriminator that is updated dynamically along with the generator. The discriminator is trained on freshly constructed pairs at each training iteration. The negatives in each pair are drawn from the current generator and differ from their clean counterparts primarily in artifacts, so the discriminator keeps tracking the generator's evolving failure modes. The preference and fidelity rewards are differentiable and are optimized along the sampling trajectory used at inference via leap-based gradient propagation. In this way, a single RL stage performs preference alignment and few-step distillation simultaneously. AdvFidelity generates preference-aligned images in 8 steps without classifier-free guidance, matching or exceeding the quality of full-step CFG-guided generation.

1 INTRODUCTION

Few-step distillation and reinforcement learning (RL) are two important procedures in the post-training of image generation models. In real-world production, they are usually performed sequentially: RL is first applied for human preference alignment (Liu et al., 2026; Xue et al., 2025a; Zheng et al., 2025; Xue et al., 2025b), and few-step distillation is subsequently performed for efficient inference (Song et al., 2023; Kim et al., 2024; Luo et al., 2023; Yin et al., 2024a). This sequential pipeline is cumbersome and often suboptimal, as distillation may wash out the reward gains obtained from RL. Recent works, such as DMDR (Jiang et al., 2025) and RTDMD (Huang et al., 2026), adopt a joint training approach: they first warm up with distribution matching distillation (DMD) (Yin et al., 2024b) and then optimize the RL and DMD objectives in parallel.

The core idea of these methods—training RL and DMD in parallel—is straightforward yet effective; however, they share a common limitation rooted in the externality of the distillation anchor. The teacher distribution must be fixed before RL begins and thus cannot reflect the reward signal, so the RL and DMD objectives pull the policy in different directions: one toward higher rewards, the other toward a reward-blind anchor. Their parallel optimization is therefore difficult to balance; and in practice the anchor is the frozen pre-trained base model itself (e.g., the original SD3.5-M), so the mismatch only grows as the student evolves under reward optimization.

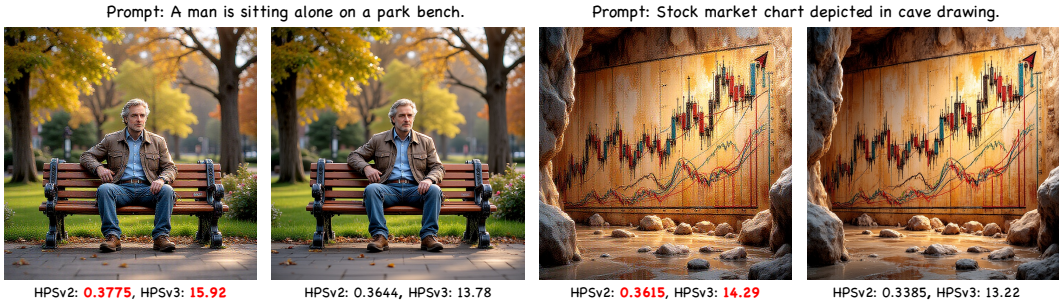
In this work, we show that the separate distillation stage—together with its frozen, RL-agnostic teacher—can be eliminated altogether: instead of introducing a heavy distillation framework such as DMD into RL, we achieve both goals within the RL framework by introducing a fidelity reward. We start from a key observation: existing RL pipelines, which directly optimize few-step rollouts

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for efficiency (Zheng et al., 2025; Liang et al., 2026), already exhibit an implicit few-step distillation effect—as training progresses, the quality of few-step generations steadily improves, even when the reward is highly specialized (e.g., an OCR reward in Figure 1a). This effect, however, is incomplete: the generated images still contain subtle noise artifacts that common reward functions fail to detect (Figure 1b), especially at later training iterations. This observation suggests that such a stage is unnecessary in principle—what RL lacks is explicit fidelity supervision. An intuitive idea is thus to introduce an artifact-sensitive fidelity reward to complement the specialized rewards.



(a) Few-step RL with a sparse text-rendering reward: collapsed generations regain structural coherence during training, while grid-pattern artifacts persist after the reward is satisfied and are removed only by the fidelity reward (rightmost).



(b) Artifact-laden generations (left of each pair) receive preference scores comparable to or even higher than those of clean images, providing insufficient supervision to recover the clean image distribution.

Figure 1: Motivation of AdvFidelity.

The core idea of the fidelity reward is simple: train an artifact-sensitive discriminator (Goodfellow et al., 2020) to serve as the reward. However, this idea raises two practical challenges, which we address with two corresponding insights. First, a frozen discriminator remains vulnerable to reward hacking, as the evolving RL policy may generate novel artifacts that the discriminator has never been trained on. We therefore argue that the discriminator should co-evolve adversarially with the generator, updating online throughout training to adapt to newly emerging artifacts. Second, if the positive and negative samples differ systematically in high-level attributes (e.g., real images and generations drawn from different style distributions), the discriminator can be hacked through such shortcuts—separating images by style rather than by artifacts. We therefore need to construct proper training data so that the discriminator focuses on image quality rather than image style.

Building on these two insights, we propose AdvFidelity, which introduces an adversarially trained discriminator as a general fidelity reward that co-evolves with the policy rather than anchoring it to a frozen base model—to eliminate artifacts in few-step generation, thereby unifying RL-based preference alignment and high-fidelity few-step generation. AdvFidelity has two key design choices corresponding to the two problems above. First, it alternates between updating the generator and the discriminator, with the latter trained on high-quality samples from the target distribution as positives and images produced by the current generator as negatives. Second, we propose a content-matched pair construction strategy that builds positive–negative pairs in which the two images share almost the same style and scene layout but differ in low-level details. Such pairs force the discriminator to rely on image quality rather than high-level semantics when distinguishing between them.

Built upon the LeapAlign RL framework (Liang et al., 2026), AdvFidelity matches or surpasses full-step RL with a few sampling steps, both on text-to-image (T2I) generation (FLUX.1-dev with the

HPSv2 reward) and on a production-level image editing task (virtual dressing with QwenImage-Edit and a garment-consistency reward). The two settings exercise both anchor constructions (Sec. 3.4): target-anchored pairs for the strongly stylized HPSv2 reward, and real-anchored pairs for the photorealistic dressing task, where real images are reused directly without any preparatory RL stage. Notably, on QwenImage-Edit, whose classifier-free guidance (CFG) is not distilled, our few-step generations without CFG match full-step RL with CFG. To the best of our knowledge, AdvFidelity is the first framework to unify preference alignment and few-step generation end-to-end in a single RL stage, without introducing DMD or trajectory distillation. Its few-step, CFG-free generations match or surpass full-step RL, whether or not the full-step baseline itself relies on CFG.

Our contributions are summarized as follows.

- We introduce an adversarial fidelity reward, instantiated as an online discriminator that co-evolves with the policy, to improve the fidelity of few-step rollouts during RL training.
- We propose an online strategy for constructing content-matched positive–negative pairs, enabling the discriminator to focus on low-level artifacts.
- Extensive experiments demonstrate that AdvFidelity generates preference-aligned images in just 4 or 8 sampling steps without classifier-free guidance (CFG), matching or surpassing full-step generation whether or not it uses CFG, across both T2I generation (FLUX.1-dev with HPSv2) and production-level virtual dressing (QwenImage-Edit with a garment-consistency reward).

2 Related Work

Reinforcement Learning Reinforcement learning has become a major approach for aligning generative models with human preferences. Policy-gradient methods formulate the denoising process as a Markov Decision Process and optimize terminal rewards through policy gradients. DDPO (Black et al., 2024) establishes this paradigm, and the group-relative objective of GRPO (Shao et al., 2024) is extended to rectified flow models by Flow-GRPO (Liu et al., 2026), DanceGRPO (Xue et al., 2025b), and MixGRPO (Li et al., 2025). DPO-based methods (Rafailov et al., 2023; Wallace et al., 2024; Liang et al., 2025) instead optimize preference pairs without explicit reward modeling. Direct-gradient methods exploit differentiable sampling to backpropagate reward signals through the generation process. ReFL (Xu et al., 2023b), DRaFT (Clark et al., 2024), and DRTune (Wu et al., 2024) update selected timesteps, and LeapAlign (Liang et al., 2026) enables reward gradients to reach arbitrary steps through two-step leap trajectories. These RL methods achieve outstanding preference alignment, yet they run with fewer sampling steps during training than at inference (Zheng et al., 2025; Liang et al., 2026) only to reduce the cost of trajectory rollouts. The role of this few-step training trajectory remains underexplored, and the distillation ability inherent in RL is overlooked.

Few-Step Distillation Few-step distillation compresses multi-step teacher models into generators requiring only 1–8 inference steps. Consistency Models (CM) (Song et al., 2023), Latent Consistency Models (LCM) (Luo et al., 2023), and Phased Consistency Models (PCM) (Wang et al., 2024) enforce trajectory-level consistency for fast generation. Score-based methods such as Distribution Matching Distillation (DMD) (Yin et al., 2024b) improve few-step quality by matching student and teacher distributions through an auxiliary fake score network, and DMD2 (Yin et al., 2024a) further extends this setup with an additional discriminator. These distillation methods rely on auxiliary networks such as fake score models or discriminators, which complicate the training pipeline. Applying distillation after RL, moreover, undermines the ability that RL has acquired. We instead introduce a dynamically updated reward that carries out distillation and RL at the same time.

Combining Reinforcement Learning and Distillation Recent work pursues joint distillation-RL by combining Distribution Matching Distillation with RL objectives, as in DMDR (Jiang et al., 2025) and RTDMD (Huang et al., 2026). They first train a few-step generator through warm-up DMD (Yin et al., 2024b) and then optimize it with parallel RL and DMD objectives, which adds a fake-score network and a multi-stage schedule on top of the RL loop. AdvFidelity instead performs distillation within the reward optimization itself, modeling fidelity as a reward from a discriminator trained on content-matched pairs. Through leap-based trajectory optimization, the fidelity and preference rewards are jointly optimized on the same few-step trajectory, enabling distillation and preference alignment in a single stage.

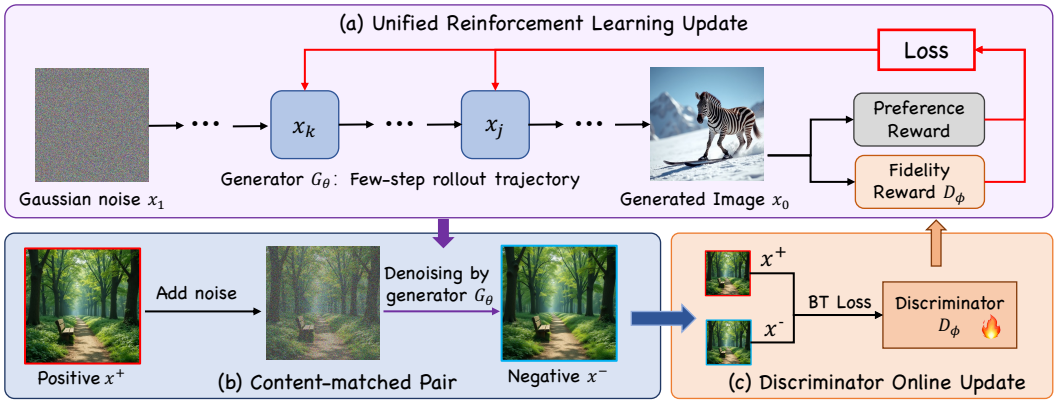


Figure 2: **Framework of AdvFidelity.** (a) Joint RL update: the few-step generator G_θ (π_θ) rolls out a trajectory from Gaussian noise x_1 to the generated image x_0. (b) Content-matched pair: each positive x^+ is perturbed with moderate noise and reconstructed by the current G_θ to form its matched negative x^-. (c) Online update: the discriminator D_ϕ, which serves as the fidelity reward, is updated on such pairs with a hinge loss.

Comparison with Adv-GRPO. The closest work to ours is Adv-GRPO (Mao et al., 2026), which similarly employs an iteratively updated discriminator as an adversarial reward against reward hacking; our discriminator architecture follows theirs. The differences therefore lie not in the discriminator itself, but in goal, reward semantics, and optimizer coupling. First, Adv-GRPO improves conventional text-to-image generation under hackable rewards and does not concern few-step generation or distillation, whereas AdvFidelity removes few-step artifacts to unify preference alignment and few-step generation in a single RL stage. Second, its positives are a fixed set of images generated offline by another model (Qwen-Image (Wu et al., 2025)), so the discriminative signal largely captures distribution-level discrepancies (e.g., style and model fingerprints) rather than artifacts—a property that Adv-GRPO indeed embraces for distribution transfer and style customization. Our content-matched pairs instead align style and layout by construction (Sec. 3.4), leaving low-level artifacts as the only reliable discriminative signal. Third, AdvFidelity is optimizer-agnostic (instantiated with a ReFL-family framework), whereas Adv-GRPO is built upon GRPO.

3 Method

3.1 Preliminaries: RL on Few-Step Rollouts

We consider text-conditioned flow-matching generators that transport a noise sample $\epsilon \sim \mathcal{N}(0, I)$ to an image conditioned on a prompt c. Efficient generators operate in the few-step regime, producing samples with only τ denoising steps; we write the resulting stochastic policy as $x \sim \pi_\theta(\cdot \mid c; \tau)$.

RL-based post-training optimizes the policy against a reward $r(x, c)$ encoding human preferences or task-specific criteria:

$$\max_\theta \ \mathbb{E}_{c\sim\mathcal{C}}\, \mathbb{E}_{x\sim\pi_\theta(\cdot|c;\tau)}\big[r(x,c)\big]. \tag{1}$$

Two representative families of optimizers are commonly used. Advantage-based methods (e.g., GRPO (Shao et al., 2024)) sample a group of rollouts per prompt and normalize rewards into advantages. Reward-feedback (ReFL) methods (Xu et al., 2023b; Liang et al., 2026) instead treat the differentiable reward as a loss and backpropagate it through the sampling chain. Unless otherwise specified, we instantiate our method with LeapAlign (Liang et al., 2026), a ReFL-family framework that directly optimizes few-step rollouts.

3.2 Overview of AdvFidelity

AdvFidelity eliminates the separate distillation stage by internalizing fidelity supervision as an additional reward within the RL objective itself. Concretely, we augment the specialized reward with a fidelity reward r_{fid} produced by a discriminator that co-evolves with the policy:

$$\max_\theta \ \mathbb{E}_{c\sim\mathcal{C}}\, \mathbb{E}_{x\sim\pi_\theta(\cdot|c;\tau)}\big[r_{\text{spec}}(x,c) + \lambda\, r_{\text{fid}}(x)\big], \tag{2}$$

where r_{spec} denotes the specialized reward (e.g., HPSv2, PickScore) and λ balances the two terms. Note that r_{fid} is non-stationary by design, as ϕ co-evolves with the policy throughout training (Sec. 3.3). Under ReFL, the policy is updated by minimizing a thresholded reward loss following

LeapAlign (Liang et al., 2026), $\mathcal{L}(\theta) = \mathbb{E}\big[\max(0,\, m - r(x))\big]$, which bounds the reward scale and stops optimization once the margin m is met. Figure 2 illustrates the resulting single-stage pipeline, in which policy optimization and discriminator training alternate. Two design choices make this simple objective work: an adversarially co-evolving discriminator (Sec. 3.3), and a content-matched pair construction that forces the discriminator to focus on low-level artifacts (Sec. 3.4).

3.3 Adversarial Fidelity Reward

Discriminator as reward in RL update. The fidelity reward is produced by a discriminator D_ϕ that scores how free an image is from generation artifacts. Following Mao et al. (2026), D_ϕ builds on a frozen DINOv2 (Oquab et al., 2023) backbone with a lightweight MLP head h_ϕ that operates on the [CLS] token only (no patch tokens), fed at the RL resolution; the frozen backbone provides a stable feature space, while adaptivity comes from the online-trained head. The reward is simply the discriminator score, $r_{\text{fid}}(x) = D_\phi(x)$.

Training the discriminator online. D_ϕ is trained with a hinge loss on high-quality positives x^+ and negatives x^- produced on the fly by the current policy (Sec. 3.4):

$$\mathcal{L}_D = \mathbb{E}_{x^+}\big[\max(0, 1 - D_\phi(x^+))\big] + \mathbb{E}_{x^-}\big[\max(0, 1 + D_\phi(x^-))\big]. \tag{3}$$

Since the [CLS] token is a global representation while few-step artifacts are often localized (e.g., grid patterns and texture noise), a whole-image score may dilute them. We therefore apply random-crop augmentation during both generator and discriminator training: each image is evaluated both as a whole and through randomly cropped sub-regions, so that local regions receive dedicated scrutiny and the discriminator cannot rely on global statistics (e.g., overall tone or border effects). We alternate between updating the policy θ with Eq. 2 and updating the head ϕ with Eq. 3. In this race the generator leads: each policy update may shift the artifact distribution, and the discriminator must chase it. We therefore perform multiple discriminator updates per policy update so that it catches up with the latest artifacts before its supervision degrades; the exact update ratio is deferred to the implementation details.

Benefits of adversarial co-evolution. This alternating process induces an adversarial dynamic. Under the pressure of the fidelity reward, the generator suppresses the artifacts currently penalized by D_ϕ; yet as its distribution shifts, new artifacts inevitably emerge, and the discriminator—trained on fresh negatives from the current policy (Sec. 3.4)—adapts to them in turn, much like the generator–discriminator dynamics in GANs. A frozen reward lacks this chase-and-adapt loop and is eventually hacked, whereas adversarial co-evolution drives artifacts to be progressively eliminated rather than merely displaced into forms the reward cannot see.

Potential for generalization to other RL methods. As a scalar reward, r_{fid} is agnostic to the downstream RL optimizer: under ReFL-style training it is backpropagated through the sampling chain together with r_{spec}; under advantage-based methods such as GRPO, it enters advantage computation as an additional reward term. This decoupling allows AdvFidelity to work with a wide range of RL frameworks and specialized rewards.

3.4 Content-Matched Pair Construction

The hinge loss in Eq. 3 teaches D_ϕ whatever separates positives from negatives—including style shortcuts, if the two populations differ systematically in high-level attributes. We therefore construct positive–negative pairs that share style and scene layout and differ only in low-level details, so that artifacts become the only reliable discriminative signal.

Noise-anchored negatives. Given a positive sample x^+, we corrupt it with partial noise at level t,

$$x_t = (1-t)\, x^+ + t\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{4}$$

and let the current policy complete the remaining denoising from x_t, yielding the negative x^-; as x_t is already partially denoised, this completion uses fewer steps than full inference (one or two steps in practice). Here c is a prompt paired with x^+, readily available from existing text–image datasets or a captioner. The construction is in fact agnostic to the conditioning signal—it operates purely on images, and thus extends to non-text tasks such as image super-resolution. With $t \leq 0.5$, the intermediate state x_t retains the structure of x^+: the policy inherits its style and layout, and the residual discrepancy between x^+ and x^- concentrates in fine details—precisely where few-step artifacts concentrate. The noise level t thus controls the difficulty of the discrimination task; we use a fixed value per task ($t \in \{0.3, 0.5\}$). Since negatives are generated on the fly, the pair distribution tracks the policy throughout training.

Algorithm 1 AdvFidelity Training

Require: few-step policy π_θ, discriminator D_ϕ, specialized reward r_{spec}, anchor set $\{x^+\}$, noise level t, reward weight λ, discriminator update ratio K

while not converged **do**
  Sample prompts c and few-step rollouts $x \sim \pi_\theta(\cdot \mid c; \tau)$
  $r \leftarrow r_{\text{spec}}(x, c) + \lambda\, D_\phi(x)$; update θ with r (Eq. 2)
  for $k = 1$ to K **do**
    $x_t \leftarrow (1-t)\, x^+ + t\, \epsilon; \quad x^- \sim \pi_\theta(\cdot \mid x_t, c; \tau)$
    Update ϕ with the hinge loss on (x^+, x^-) (Eq. 3)
  end for
end while

Choosing the anchor. Fidelity is relative to the target distribution: since content-matching already aligns all high-level attributes within each pair, the positives only need fine details that match the target distribution in style, leaving artifacts as the only within-pair discrepancy. This admits three anchor sources in ascending order of cost:

- *Real-anchored pairs.* For photorealistic tasks (e.g., e-commerce, film production), real images are ready-made positives; data construction is free.
- *Base-anchored pairs.* When prompts lack paired ground-truth images but the target distribution remains photorealistic (e.g., GenEval), positives are full-step samples from the base model itself—again free of any RL stage.
- *Target-anchored pairs.* When the reward drives the target distribution away from photorealism (e.g., HPSv2 on open-ended prompts), positives are full-step samples from a model after a vanilla RL stage—a one-off data preparation, while the deployable generator is still trained in a single RL stage.

3.5 Training Procedure

Algorithm 1 summarizes the alternating training loop of AdvFidelity. The reward weight λ, the margin m, the noise level t, and the discriminator update ratio K are fixed hyperparameters; their values are deferred to the implementation details.

Connection to adversarial training. AdvFidelity is genuinely adversarial: under ReFL-style optimization, gradients of the discriminator score flow into the generator through the reward term, and the generator–discriminator pair undergoes chase-and-adapt dynamics analogous to GAN training. The differences lie in role and construction: (i) the discriminator contributes one term within a multi-reward RL objective rather than serving as the sole training target; (ii) its negatives are content-matched, noise-anchored completions rather than full generations from scratch; and (iii) its goal is to supervise the fidelity of a preference-aligned few-step policy, not to define the generative distribution itself. Compared with the recent adversarial-reward design of Mao et al. (2026), our discriminator serves a distinct purpose—removing few-step artifacts so that preference alignment and few-step generation unify in a single RL stage—and is paired with content-matched pair construction (Sec. 3.4) to suppress style shortcuts.

4 Experiments

We evaluate AdvFidelity in three settings: the first two follow the experimental protocols of LeapAlign (Liang et al., 2026) (preference alignment on HPDv2 prompts, and compositional generation on GenEval prompts), and the third is a production-oriented virtual dressing task. In each setting, the full-step baseline is LeapAlign RL trained on the same base model; our models sample in 4 or 8 steps without classifier-free guidance (CFG), while the baseline uses CFG whenever the base model is not CFG-distilled. Anchor construction (Sec. 3.4) follows the target distribution of each setting: target-anchored pairs for the strongly stylized HPDv2 setting, base-anchored pairs for GenEval, and real-anchored pairs for virtual dressing; the latter two require no preparatory RL stage.

4.1 Preference Alignment on HPD

Implementation Details. We fine-tune FLUX.1-dev on the HPDv2 prompt set with HPSv2 (Wu et al., 2023) as the preference reward. Evaluation reports HPSv2, PickScore (Kirstain et al.,

Table 1: Quantitative comparison of preference alignment on FLUX.1-dev. **Bold** and underlined values denote the best and second best results. Higher is better. †HPSv3 is shaded and reported for reference only, as it assigns unreliable scores to artifact-laden images.

		In-Domain	Out-of-Domain				
Method	**NFE**	**HPSv2**	**PickScore**	**ImgRwd**	**UR-Align**	**UR-IQ**	**HPSv3†**
Flux.1-dev	50	0.3083	22.6970	1.0210	3.4428	3.4792	13.3406
DanceGRPO	50	0.3451	23.1186	1.2347	3.4660	3.6199	14.8336
MixGRPO	50	0.3692	23.5184	**1.6155**	3.4393	3.6241	14.7530
ReFL	50	0.3852	23.6299	1.3468	3.4786	3.6870	15.5127
DRTune	50	0.3882	23.5185	1.3384	3.4793	3.6679	15.5606
DRaFT-LV	50	0.3859	23.6437	1.3562	3.4868	3.6887	15.3699
LeapAlign	50	<u>0.4096</u>	<u>23.7239</u>	1.5105	3.4837	3.7284	15.7548
RTDMD	4	0.3724	22.8279	1.3020	3.4015	3.7452	15.1737
RTDMD	8	0.3923	23.3059	1.4134	3.4779	<u>3.7482</u>	15.0931
AdvFidelity	4	0.4055	23.6211	1.5164	<u>3.4981</u>	3.7338	15.4299
AdvFidelity	8	**0.4125**	**23.7685**	<u>1.5255</u>	**3.5024**	**3.7509**	15.3197

2023), ImageReward (Xu et al., 2023a), and the alignment and image-quality sub-scores of UnifiedReward (Wang et al., 2025); HPSv3 (Ma et al., 2025) is shaded and reported for reference only, as its scores can be unreliable; see the appendix A for details. We compare against RL methods including DanceGRPO (Xue et al., 2025b), MixGRPO (Li et al., 2025), ReFL (Xu et al., 2023b), DRTune (Wu et al., 2024), DRaFT-LV (Clark et al., 2024), and LeapAlign (Liang et al., 2026), while RTDMD (Huang et al., 2026) serves as the joint RL–distillation baseline. We conduct experiments under both 4-step and 8-step sampling settings. The general fidelity reward D_ϕ is a frozen DINOv2-Large (Oquab et al., 2023) encoder and a trainable MLP head. We use crop sizes of 256 and 384 pixels and set the patch-level loss weight w_{patch} to 0.3. We optimize the generator with AdamW using a learning rate of 1×10^{-4} and the discriminator with AdamW using a learning rate of 5×10^{-5}. We set the gradient discount factor α to 0.3, the satisfaction threshold m to 0.55, and the fidelity loss weight w_{crop} to 0.05.

Paired Data. Positives are target-anchored (Sec. 3.4): as the HPSv2 reward drives the generator toward stylized outputs, we take 50-step samples from the official LeapAlign (Liang et al., 2026) checkpoint as positives. Negatives are constructed by adding forward noise with $\sigma \in \{0.3, 0.5\}$ to a positive and completing the remaining denoising with the current generator in one or two steps.

Results. Table 1 shows that the 8-step AdvFidelity model surpasses full-step LeapAlign, and the 4-step model nearly matches it. At the same sampling budgets, AdvFidelity consistently outperforms RTDMD.Compared with RTDMD, AdvFidelity better preserves the preferred style, follows the input prompts more closely, and retains finer visual details, as further illustrated in Figure 3(a).

4.2 Compositional Generation on GenEval

Implementation Details and Paired Data. To verify that our method enables effective few-step distillation without relying on data from an RL-trained teacher model, we conduct compositional generation experiments on GenEval (Ghosh et al., 2023), using HPSv2 (Wu et al., 2023) as the preference reward. All positive samples are 50-step outputs from the base FLUX model. We report the overall GenEval score together with its sub-categories. The source of negative samples and all other settings are the same as those in Section 4.1.

Results. Table 2 shows that AdvFidelity surpasses full-step LeapAlign at both 8 and 4 steps, with gains concentrated on challenging categories such as color and attribute binding. Figure 3(b) further shows that, compared with full-step RL methods and RTDMD, our method preserves object relations, image–text consistency, and fine details as the sampling budget decreases.

4.3 Virtual Dressing on Real Data

Implementation Details and paired data. We evaluate AdvFidelity on the virtual dressing task using Qwen-Image (Wu et al., 2025), testing its generalization to a practical application. We follow the discriminator architecture and objective described in Section 4.1. The ground-truth try-on images serve directly as native positives. During training, we use the DINO feature similarity between each generated image and its ground-truth as the reward. For evaluation, we report DINO similarity and

Table 2: Compositional generation results on FLUX.1-dev, with HPSv2 as the training reward. **Bold** and underlined values denote the best and second best results. Higher is better.

Methods	NFE	Overall	Single Obj	Two Obj	Count	Color	Pos	AttrB
Flux.1-dev	50	0.6481	99.38	85.10	66.56	76.06	19.00	42.75
DanceGRPO	50	0.6775	99.38	90.15	69.38	76.33	22.25	49.00
MixGRPO	50	0.7232	99.69	93.69	80.00	80.05	24.25	56.25
ReFL	50	0.7011	99.38	92.68	69.06	75.80	26.75	57.00
DRTune	50	0.7101	99.38	93.69	73.12	76.86	27.50	55.50
DRaFT-LV	50	0.7024	99.69	92.42	74.06	75.53	24.00	55.75
LeapAlign	50	0.7420	99.38	96.46	72.50	80.59	30.25	66.00
RTDMD	4	0.6794	99.38	92.68	73.75	68.83	33.25	44.75
RTDMD	8	0.7236	99.69	95.45	72.81	74.73	**36.25**	55.25
AdvFidelity	4	0.7535	99.38	96.46	74.06	82.45	30.75	69.00
AdvFidelity	8	**0.7759**	**100.0**	**98.99**	**82.50**	**84.57**	26.75	**72.25**

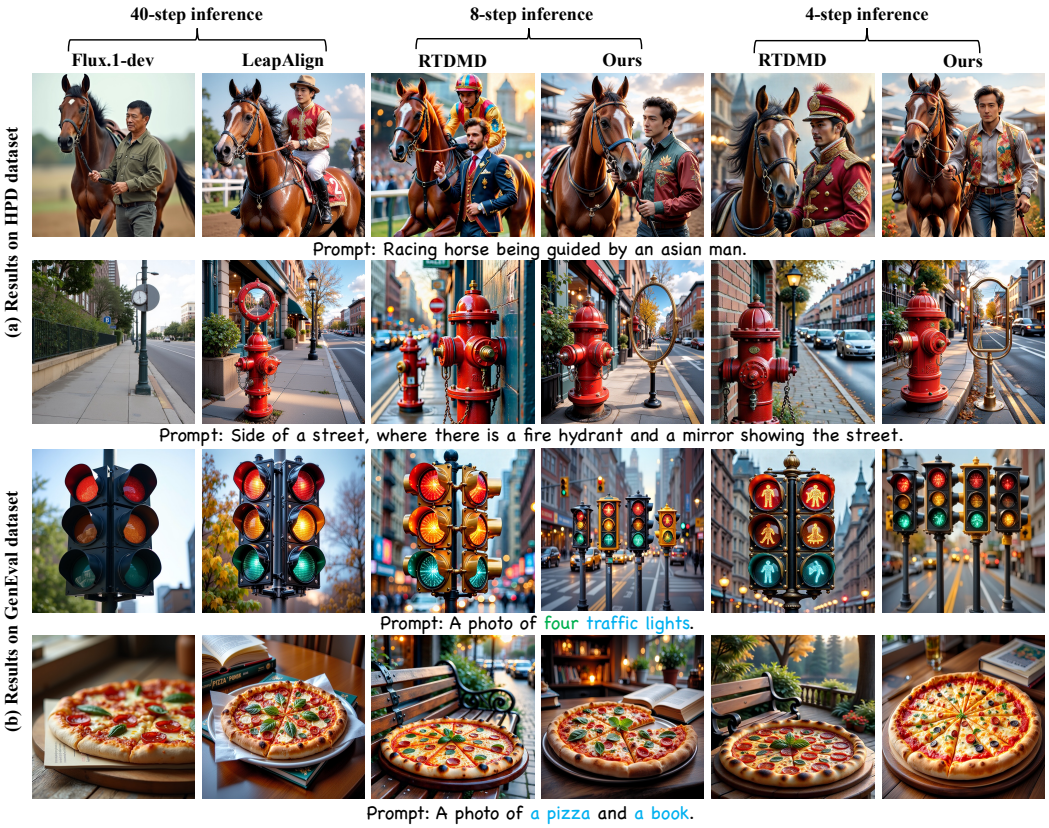


Figure 3: Qualitative comparison across preference-alignment and compositional prompts.

Gemini VLM scores for Realism, Consistency, Alignment, and Coherence, each on a scale from 0 to 10.

Results. Table 3 shows that the 8-step AdvFidelity model outperforms all three full-step baselines across DINO similarity and the four Gemini evaluation aspects. The qualitative comparison are provided in the appendix.

4.4 Ablation Studies

We evaluate five configurations of AdvFidelity under 8-step inference on HPDv2, together with two references. FLUX with 8-step inference serves as the baseline, and LeapAlign+DMD, which distills the official full-step LeapAlign checkpoint to 8 steps, represents the sequential RL-then-distillation pipeline. Our variants are built cumulatively from a vanilla configuration, which trains a frozen discriminator on native positives and unpaired rollout negatives. Adding the online update

Table 3: Quantitative comparison on virtual dressing with Qwen-Image.

Method	CFG	NFE	DINO	Realism	Consistency	Alignment	Coherence
Qwen-Image	✓	40	0.7426	6.96$\pm$1.10	8.64$\pm$0.50	7.36$\pm$0.62	5.52$\pm$1.04
LeapAlign	×	40	0.7793	7.58$\pm$0.60	8.69$\pm$0.42	7.36$\pm$0.70	5.77$\pm$1.07
LeapAlign	✓	40	0.7874	7.45$\pm$0.72	8.69$\pm$0.44	7.46$\pm$0.62	5.78$\pm$1.10
AdvFidelity	×	8	**0.7929**	**7.66**$\pm$0.49	**8.74**$\pm$0.42	**7.51**$\pm$0.63	**5.84**$\pm$1.10

Table 4: Ablation study at 8-step inference under the HPDv2 setting.

Method	Paired	Anchor	r_{fid}	HPSv2	PickScore	ImgRwd	UR-Align	UR-Img
Flux.1-dev	–	–	–	0.2929	22.3912	0.9532	3.4227	3.7024
LeapAlign + DMD	–	–	–	0.3978	22.9654	1.3976	3.4259	3.7164
Vanilla	×	Base	Frozen	0.3416	22.6735	1.1650	3.4404	3.7123
+Adaptive R_{fid}	×	Base	✓	0.3558	22.8637	1.2171	3.4494	3.7215
+Teacher-anchored pairs	×	Target	✓	0.3920	23.2588	1.3567	3.4456	3.7141
Ours (Full)	✓	target	✓	**0.4125**	**23.7685**	**1.5255**	**3.5024**	**3.7509**

(+adaptive R_{fid}) examines whether the fidelity reward must track the evolving generator. Adding target-anchored positives (+target-anchored positives) examines whether positives in the target style provide better supervision. Adding content-matched pairs then yields the full AdvFidelity. One more variant freezes the discriminator at the full configuration (Ours w/ frozen R_{fid}) to isolate the contribution of the online update.

Table 4 shows that sequentially applying DMD to the full-step LeapAlign model weakens the alignment achieved by RL, while AdvFidelity preserves it in a single stage. Each component of AdvFidelity contributes to the final results. Updating the discriminator online already improves over the vanilla configuration, indicating that a frozen fidelity reward is gradually hacked by the evolving generator. Switching to target-anchored positives brings a further gain, because positives in the target style remove the style difference that the discriminator could otherwise exploit as a shortcut. Adding content-matched pairs gives another improvement, confirming that pairing the two samples in content and structure directs the discriminator toward artifacts. Figure 4 presents the same trend visually: the vanilla configuration produces visible grid-like artifacts, which are gradually suppressed as the components accumulate, and the full AdvFidelity generates the cleanest images. Freezing the discriminator at the full configuration falls behind the full method, showing that the online update remains necessary even with paired target-anchored data. The complete AdvFidelity achieves the strongest results, supporting jointly performing RL and distillation in a single stage.



Figure 4: Qualitative comparison of ablation study on Flux with HPD dataset.

5 Conclusion

We presented AdvFidelity, a single-stage framework that unifies preference alignment and few-step distillation for rectified flow models. It builds on the observation that reward optimization on few-step trajectories already distills implicitly, while the artifacts accumulated along such trajectories escape the supervision of preference rewards. AdvFidelity therefore pairs the preference reward with a fidelity reward, an online discriminator trained on content-matched pairs that isolates artifacts from semantic and stylistic content, and propagates both rewards to any generation step through the leap mechanism. Experiments across diverse rewards, settings, and base model families show that AdvFidelity produces preference-aligned, high-fidelity images in a few steps and without classifier-free guidance, on par with or beyond the full-step CFG-guided references.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, literature and reference verification, and code implementation for comparative experiments. One of the authors reviewed and tested the AI-assisted code for correctness. We reviewed and revised all AI-assisted manuscript content and take full responsibility for the final text, claims, code, and data presented in this study.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, volume 2024, pp. 4793–4822, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Yushi Huang, Xiangxin Zhou, Ruoyu Wang, Chi Zhang, Jun Zhang, and Tianyu Pang. Reinforcing few-step generators via reward-tilted distribution matching. *arXiv preprint arXiv:2605.26108*, 2026.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Zhen Li, Bo Zhang, et al. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025.

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

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13199–13208, 2025.

Zhanhao Liang, Tao Yang, Jie Wu, Chengjian Feng, and Liang Zheng. Leapalign: Post-training flow matching models at any generation step by building two-step trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23238–23248, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Weijia Mao, Hao Chen, Zhenheng Yang, and Mike Zheng Shou. The image as its own reward: Reinforcement learning with adversarial reward for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5999–6009, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 32211–32252, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Fu-Yun Wang, Zhaoyang Huang, Alexander W Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 108–124. Springer, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023a.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023b.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning rl with pretraining in diffusion models. *arXiv preprint arXiv:2509.25050*, 2025a.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025b.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.