# SOAR: SELF-CORRECTION FOR OPTIMAL ALIGNMENT AND REFINEMENT IN DIFFUSION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The post-training pipeline for diffusion models currently has two stages: supervised fine-tuning (SFT) on curated data and reinforcement learning (RL) with reward models. A fundamental gap separates them. SFT optimizes the denoiser only on ground-truth states sampled from the forward noising process; once inference deviates from these ideal states, subsequent denoising relies on out-of-distribution generalization rather than learned correction, exhibiting the same *exposure bias* that afflicts autoregressive models, but accumulated along the denoising trajectory instead of the token sequence. RL can in principle address this mismatch, yet its terminal reward signal is sparse, suffers from credit-assignment difficulty, and risks reward hacking. We propose **SOAR** (Self-Correction for Optimal Alignment and Refinement), a bias-correction post-training method that fills this gap. Starting from a real sample, SOAR performs a single stop-gradient rollout with the current model, re-noises the resulting off-trajectory state, and supervises the model to steer back toward the original clean target. The method is on-policy, reward-free, and provides dense per-timestep supervision with no credit-assignment problem. On SD3.5-Medium, SOAR improves GenEval from 0.70 to 0.78 and OCR from 0.64 to 0.67 over SFT, while simultaneously raising all model-based preference scores. In controlled reward-specific experiments, SOAR surpasses Flow-GRPO in final metric value on both aesthetic and text-image alignment tasks, despite having no access to a reward model. Since SOAR's base loss subsumes the standard SFT objective, it can directly replace SFT as a stronger first post-training stage after pretraining, while remaining fully compatible with subsequent RL alignment.

## 1 INTRODUCTION

Diffusion and flow-matching models have become the dominant paradigm for high-quality visual generation, powering text-to-image (Rombach et al., 2022; Esser et al., 2024; Black Forest Labs, 2024), text-to-video, and 3D synthesis systems. Despite rapid progress in scaling and architecture, the *post-training* recipe for these generators remains far less mature than its counterpart in large language models. In current practice a model is first pretrained at scale, then fine-tuned on curated data (SFT), and optionally aligned with reward-based optimization (RL). Yet even strong models continue to suffer from recurring failures: wrong object counts, malformed hands, broken text rendering, and unstable identity preservation.

We argue that many of these failures share a common contributor: *exposure bias* in the denoising trajectory. In autoregressive language models, exposure bias refers to the mismatch between ground-truth-conditioned training and autoregressive inference (Bengio et al., 2015). Diffusion models suffer from the same phenomenon along a different axis. Standard SFT constructs noisy training states via the forward process of real data and optimizes the denoiser on these ideal, ground-truth states. During inference, however, the model conditions on its own earlier predictions. Once an early denoising step makes a small error, subsequent states enter regions that were not directly optimized; recovery then depends on out-of-distribution generalization rather than a learned correction mechanism. The error compounds over dozens of sampling steps and can become irreversible well before the final image is rendered. Prior work has identified this exposure bias in diffusion models and reduced it with input perturbation, inference-time epsilon scaling, and consistency training (Ning et al., 2023; 2024; Daras et al., 2023); we build on this diagnosis.



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.

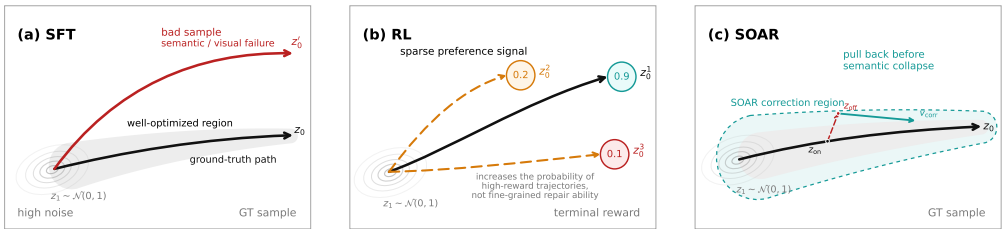


Figure 1: Conceptual comparison of post-training regimes from a trajectory perspective. **Left: SFT** trains only on ground-truth denoising states. Once an inference rollout exits the narrow corridor of well-optimized states, subsequent steps sharpen the wrong structure. **Middle: RL** separates good and bad final outcomes with a terminal reward, but the sparse signal arrives after the erroneous rollout has already formed. **Right: SOAR** performs a one-step rollout to generate off-trajectory states and supervises the model to correct back toward the clean target, acting before semantic errors fully materialize.

A natural response is to apply preference-based or reward-based optimization. Offline methods such as Diffusion-DPO (Wallace et al., 2024) adapt direct preference optimization to diffusion models by learning from human-annotated image pairs, while online methods such as Flow-GRPO (Liu et al., 2025) convert the deterministic ODE sampler into a stochastic SDE and apply policy-gradient exploration, achieving impressive improvements on compositional benchmarks. Similarly, DiffusionNFT (Zheng et al., 2026) contrasts reward-split positive and negative generations on the forward process to define an implicit policy improvement direction. These methods can optimize preferences that paired-data SFT cannot express. However, their preference or reward signal is sparse and terminal, only available after a complete rollout, making credit assignment across dozens of denoising steps difficult. Aggressive single-reward optimization also risks reward hacking, where the model improves the targeted metric at the cost of diversity or other quality dimensions (Liu et al., 2025).

This paper proposes **SOAR** (Self-Correction for Optimal Alignment and Refinement), a reward-free post-training method that can serve as a trajectory-corrected alternative to standard SFT directly following pretraining, while remaining compatible with later reward-aligned fine-tuning. The core idea is simple: instead of waiting for a terminal reward after a full rollout, we directly teach the model how to correct its own trajectory errors at the timestep where they occur. Starting from a real training pair, SOAR performs a single stop-gradient ODE step with the current model to generate an off-trajectory state, re-noises this state toward the same sampled noise at multiple auxiliary noise levels, and supervises the denoiser to steer each auxiliary state back toward the original clean target. The resulting training signal is *locally on-policy*, since the one-step off-trajectory states evolve with the model's own parameters; *dense*, as the correction target is derived from the ground-truth sample itself rather than collapsed into a scalar reward or a relative preference; and *reward-free*, requiring no external reward model or preference annotation. The correction target is derived analytically from the training pair, avoiding terminal reward credit assignment. Because shared-noise re-noising contracts the one-step deviation (Section 3.2), the original clean sample remains a local anchor for this target.

Placing SOAR in the landscape of existing post-training paradigms (Figure 1) clarifies its role. SFT trains exclusively on ground-truth denoising states; it is off-policy by construction because the model never sees its own inference-time errors during training. RL-based methods introduce on-policy, reward-driven signals, subject to the credit-assignment and reward-hacking risks above. By design, SOAR avoids dependence on terminal rewards and external reward models, while still requiring paired data and a reliable clean target. Moreover, when the training corpus is curated according to a target preference, such as filtering for high aesthetic or high text-image alignment, SOAR can use that curated data without directly optimizing a scalar reward; in our experiments, this improves the target metric without the measured non-target drop observed for Flow-GRPO, also when both methods train the same LoRA adapters. In the current work we position SOAR as a first-stage alternative to SFT that can be combined with later reward-aligned fine-tuning, although the best combination strategy remains future work. Because it changes only which states and targets are supervised, SOAR keeps the architecture and sampler of the SFT model, adds no inference cost, and improves over SFT on both broad and curated data, under full-parameter as well as LoRA training (Section 4).

Our main contributions are:

- Building on prior analyses of exposure bias in diffusion models (Ning et al., 2023; 2024), we treat the distributional mismatch between ground-truth training states and model-induced inference states as a target for supervised trajectory correction.
- We propose SOAR, a practical post-training algorithm that constructs off-trajectory states via a single ODE rollout, re-noises them toward the same noise endpoint, and optimizes the model with an analytically derived correction target anchored to the original clean sample; a geometric analysis shows how guidance and re-noising control the resulting deviation.
- We provide a controlled experimental comparison and component ablations showing that SOAR improves both rule-based (GenEval, OCR) and model-based (PickScore, HPSv2.1, Aesthetic, ImageReward) metrics over SFT, with CLIPScore tied, and that at an equal number of supervised states it outperforms SFT trained for seven times as many updates on five of six metrics.
- We show that on curated high-reward data subsets, SOAR outperforms both vanilla SFT and explicit RL (Flow-GRPO) in final metric value, also with matched LoRA adapters, supporting trajectory correction as a competitive reward-free alternative in the studied settings.

## 2 Related Work

**Diffusion models and flow matching.** Diffusion models learn to reverse a noise-corruption process (Ho et al., 2020; Song et al., 2021); latent diffusion (Rombach et al., 2022) and flow matching (Lipman et al., 2023; Liu et al., 2023) make high-resolution sampling practical through compressed latents and ODE velocity fields. Current systems such as SD3.5 (Esser et al., 2024) and FLUX (Black Forest Labs, 2024) combine rectified-flow training with DiT-style backbones (Peebles & Xie, 2023). Our contribution is orthogonal to architecture: it changes the post-training state distribution seen by the denoiser.

**Post-training for diffusion models.** SFT improves targeted capabilities by continuing the diffusion or flow-matching objective on curated data, but it still trains only on forward-process states. Preference and reward methods optimize external signals: Diffusion-DPO (Wallace et al., 2024) learns from preferred/dispreferred image pairs; DiffusionNFT (Zheng et al., 2026) splits online generations by reward; and Flow-GRPO (Liu et al., 2025) turns the deterministic sampler into a stochastic policy for group-relative policy gradients (Shao et al., 2024). These methods are powerful but inherit sparse terminal supervision and reward-design risks, whereas SOAR supplies dense correction targets before reward optimization is introduced.

**Exposure bias in iterative generation.** The distinction between training on reference inputs and inference on model-generated inputs is familiar from autoregressive sequence generation, where scheduled sampling mixes reference and generated tokens during training (Bengio et al., 2015). In diffusion models, Ning et al. (2023) add input perturbations to reduce sensitivity to the resulting distribution shift, and Ning et al. (2024) analyze prediction-error accumulation and propose inference-time epsilon scaling. SOAR builds on this diagnosis, but its perturbations come from the current denoiser's guided local rollout followed by shared-noise re-noising, rather than from an independently chosen input-noise distribution; Section 4.4 compares it with both methods.

**Consistency on model-generated trajectories.** Consistent Diffusion Models (Daras et al., 2023) complement denoising score matching with a consistency property relating the model's predictions on its own generated states across time, and connect it to accurate scores on drifted states. SOAR also trains on model-induced states, but regresses each auxiliary state toward an endpoint target derived from the original paired clean sample.

**Prompt optimization and external feedback.** Promptist (Hao et al., 2023), Prompt Auto-Editing (Mo et al., 2024), Idea2Img (Yang et al., 2024), and Self-correcting LLM-controlled Diffusion (Wu et al., 2024) refine prompts or apply generation-time corrections, whereas SOAR updates the denoiser itself. Appendix C.3 discusses the exposure-bias methods further.

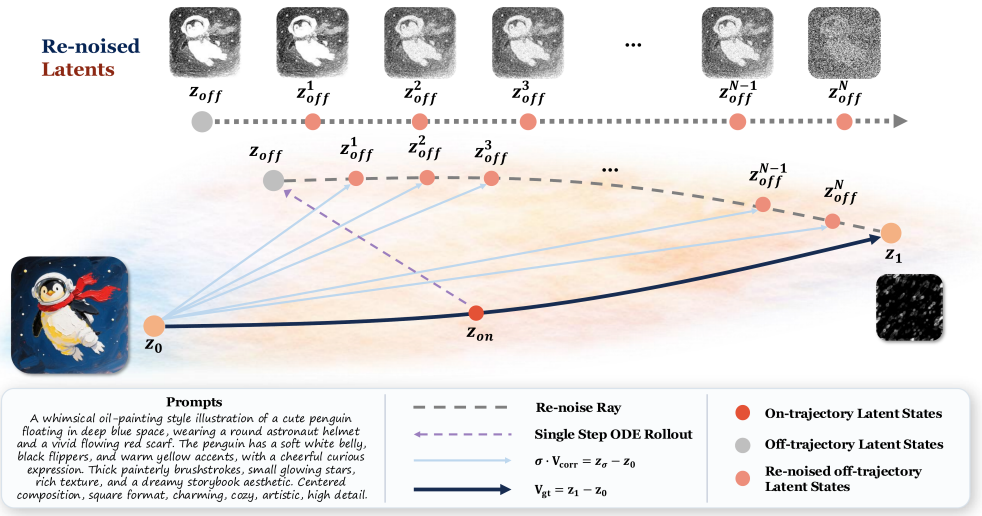


Figure 2: Schematic overview. SOAR constructs a local off-trajectory state with one detached ODE step, re-noises it toward the same sampled noise endpoint, and supervises its paired clean-endpoint estimate alongside the base flow-matching loss.

# 3 Methodology

## 3.1 Flow Matching and Local State Construction

For a paired clean latent and condition $(z_0, c)$, draw $z_1 \sim \mathcal{N}(0, I)$. The rectified-flow training path and its conditional velocity are

$$z_\sigma = (1-\sigma)z_0 + \sigma z_1, \qquad u = z_1 - z_0, \qquad \sigma \in [0,1]. \tag{1}$$

Flow matching fits $v_\theta(z_\sigma, c, \sigma)$ to $u$ by weighted squared error (Liu et al., 2023; Esser et al., 2024). Sampling follows $dz/d\sigma = v_\theta$ from noise to data: an Euler step is $z_{\sigma_1} = z_{\sigma_0} + (\sigma_1 - \sigma_0)v_\theta(z_{\sigma_0}, c, \sigma_0)$, with $\sigma_1 < \sigma_0$. Standard SFT trains on the paired interpolation states, whereas inference conditions on model-generated states. SOAR augments supervised training with local states constructed using the current model. These states approximate aspects of inference drift.

## 3.2 One-Step Rollout and Shared-Noise Re-Noising

Let $t_0$ index the noise schedule, $t_1 = \max(t_0 - 1/K, 0)$, and $\sigma_i = \sigma(t_i)$. At the ideal starting state $z_{\sigma_0}$, evaluate the conditional and unconditional velocities $v_c, v_u$. With classifier-free guidance (CFG) (Ho & Salimans, 2022) scale $w$, form $v_{\text{cfg}} = \text{sg}[v_u + w(v_c - v_u)]$, where sg stops gradients. The local rollout is

$$\widetilde{z}_{\sigma_1} = z_{\sigma_0} + \Delta\sigma\, v_{\text{cfg}}, \qquad \Delta\sigma = \sigma_1 - \sigma_0. \tag{2}$$

For each of $N$ auxiliary levels $\sigma' \sim \text{Uniform}[\sigma_1, 1]$, we re-noise this state using the *same* $z_1$:

$$\alpha = \frac{\sigma' - \sigma_1}{1 - \sigma_1}, \qquad z'_{\sigma'} = (1-\alpha)\widetilde{z}_{\sigma_1} + \alpha z_1. \tag{3}$$

These auxiliary states are detached. By default we use one rollout step with schedule resolution $K = 40$ (a step size, not a step count), $N = 6$, and constructor CFG $w = 4.5$.

The geometry makes the role of CFG explicit. Let $\epsilon_c = v_c - u$ and $g = v_c - v_u$. Relative to the paired ideal state at $\sigma_1$, the rollout deviation and auxiliary deviation satisfy

$$\delta_w = \Delta\sigma[\epsilon_c + (w-1)g], \qquad z'_{\sigma'} - z_{\sigma'} = (1-\alpha)\delta_w. \tag{4}$$

Thus shared-noise re-noising contracts the existing one-step deviation, while stronger CFG can enlarge it. This supports a *local paired-anchor* interpretation: $z_0$ remains the designated endpoint of the sampled training pair. Appendix A of the supplementary material gives the bound and its assumptions; Section 4 evaluates the construction empirically.

### 3.3 PAIRED-ENDPOINT CORRECTION OBJECTIVE

The model's one-step clean-endpoint estimate is $\widehat{z}_0(z,\sigma) = z - \sigma v_\theta(z,c,\sigma)$; it is not generally the endpoint of the full nonlinear ODE. We impose $z'_{\sigma'} - \sigma' v_{\mathrm{corr}} = z_0$ as a supervised constraint, giving

$$v_{\mathrm{corr}} = \frac{z'_{\sigma'} - z_0}{\sigma'}, \qquad v_{\mathrm{corr}} - u = \frac{(1-\alpha)\delta_w}{\sigma'}. \tag{5}$$

The target reduces to the standard conditional velocity when the deviation vanishes.

Write $\ell(z,\sigma,y) = \omega(\sigma)\|v_\theta(z,c,\sigma) - y\|^2$, where $\omega$ is the base flow-matching weight. For $B$ base states and $P$ valid auxiliary states, the objective is

$$\mathcal{L}_{\mathrm{SOAR}} = \frac{\sum_{b=1}^{B} \ell(z^{(b)}_{\sigma_0}, \sigma_0^{(b)}, u^{(b)}) + \lambda \sum_{p=1}^{P} \ell(z'^{(p)}_{\sigma'_p}, \sigma'_p, v^{(p)}_{\mathrm{corr}})}{B + \lambda P}. \tag{6}$$

With one ODE path and six auxiliary points, each sample contributes seven supervised states; Section 4.5 examines $\lambda \in \{0, 0.5, 1, 2\}$.

SOAR is thus supervised, data-anchored fine-tuning; Algorithms S1 and S2 give the ODE implementation and stochastic branches, and Appendix A.4 an endpoint-regression interpretation.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETUP

**Data and evaluation.** We use SD3.5-Medium (Esser et al., 2024) and a corpus of 286,119 image–caption pairs. The high-Aesthetic subset contains 3,725 pairs with aesthetic score $\geq 6.8$; the high-CLIP subset contains 6,857 pairs with CLIP score $\geq 0.40$. These subsets reflect preferences specified through offline filtering, while SOAR trains without querying a reward model. Most of the corpus comes from publicly available sources (Appendix F.1 describes the data release). Within every comparison, SFT and SOAR train on identical pairs, so the gap between them reflects the objective rather than data curation. We report PickScore (Kirstain et al., 2023), CLIPScore (Hessel et al., 2021), HPSv2.1 (Wu et al., 2023), the LAION aesthetic predictor (Schuhmann, 2022), and ImageReward (Xu et al., 2023) on out-of-domain DrawBench prompts (Saharia et al., 2022), and GenEval (Ghosh et al., 2023) and OCR accuracy on the corresponding Flow-GRPO test sets (Liu et al., 2025). Evaluation prompts are not used for training. Images are generated at $512 \times 512$ with evaluation CFG 4.5; exceptions for external reference models are marked in Table 1.

**Experimental protocols.** The broad-data experiment trains full-parameter SFT and SOAR for 10k updates with batch size 512 and learning rate $2 \times 10^{-5}$; SOAR uses one ODE step, $K = 40$, six auxiliary points, and $\lambda = 1$. The reward-specific experiment trains full-parameter SFT/SOAR and LoRA Flow-GRPO on the curated subsets for 5k gradient updates with batch size 32 and learning rate $2 \times 10^{-5}$. Because its parameterization differs across methods, we add controls for supervised-state count, matched LoRA adapters, and individual components. Each control has its own protocol. Appendix B of the supplementary material lists all settings.

### 4.2 BROAD-DATA RESULTS

Table 1 summarizes the broad-data comparison results. SOAR improves over SFT on six of seven reported metrics, with CLIPScore tied at the displayed precision: on the Flow-GRPO test sets, GenEval rises from 0.70 to 0.78 (+11% relative) and OCR from 0.64 to 0.67; on OOD DrawBench, the other four model-based scores improve simultaneously. Relative to the pretrained model, SOAR raises GenEval by 0.15 and OCR by 0.08, whereas SFT raises them by 0.07 and 0.05 with the same data and number of updates. The rule-based gains are particularly relevant because GenEval and OCR directly measure compositional accuracy and text-rendering fidelity, failure modes tied to early semantic decisions in the denoising trajectory; the improvements are consistent with trajectory-level correction reducing error propagation from high-noise stages. Notably, the OOD DrawBench model-based improvements (PickScore +0.15, HPSv2.1 +0.005, Aesthetic +0.11, ImageReward +0.05 over SFT) are achieved *without* any reward model during training. A plausible interpretation is that cleaner trajectory states improve global structure, which preference models may reward

Table 1: Broad-data results. Bold and underline mark the best and second-best values *within the three SD3.5-M rows*. †Official checkpoint; ‡generation at $1024^2$.

| Model | Steps | Rule-based | | Model-based (DrawBench) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | GenEval | OCR | PickScore | CLIPScore | HPSv2.1 | Aesthetic | ImageReward |
| *Reference models* | | | | | | | | |
| SD-XL‡ | – | 0.55 | 0.14 | 22.42 | 0.287 | 0.280 | 5.60 | 0.76 |
| SD3.5-L‡ | – | 0.71 | 0.68 | 22.91 | 0.289 | 0.288 | 5.50 | 0.96 |
| FLUX.1-Dev† | – | 0.66 | 0.59 | 22.84 | 0.295 | 0.274 | 5.71 | 0.96 |
| *SD3.5-Medium post-training* | | | | | | | | |
| SD3.5-M | – | 0.63 | 0.59 | 22.34 | 0.285 | 0.279 | <u>5.36</u> | 0.85 |
| + SFT | 10k | <u>0.70</u> | <u>0.64</u> | <u>22.71</u> | <u>0.295</u> | <u>0.284</u> | 5.35 | <u>1.04</u> |
| + SOAR | 10k | **0.78** | **0.67** | **22.86** | **0.295** | **0.289** | **5.46** | **1.09** |

when scoring the final image. Among the reference models (SD-XL (Podell et al., 2024), SD3.5-L, and FLUX.1-Dev (Black Forest Labs, 2024)), SOAR on SD3.5-M surpasses the larger SD3.5-L on Flow-GRPO test-set GenEval (0.78 vs. 0.71) and OOD DrawBench HPSv2.1 (0.289 vs. 0.288), and exceeds FLUX.1-Dev on OOD DrawBench PickScore (22.86 vs. 22.84), suggesting that trajectory correction can narrow selected gaps between a medium-sized model and larger reference systems.

### 4.3 Supervision Count

SOAR supervises seven states per sample per update, one base state and six auxiliary states, whereas SFT supervises one. Table 2 therefore compares SOAR 5k with SFT 35k, which supervise the same number of states, to separate the benefit of constructed supervision from simply training SFT on more on-trajectory states.

At an equal number of updates, SOAR 5k exceeds SFT 5k on all six metrics. At an equal number of supervised states, it outperforms SFT 35k on five of six: on high-CLIP it exceeds SFT 35k on both GenEval (0.7517 versus 0.7166) and CLIPScore (0.2999 versus 0.2971), and on high-Aesthetic it attains higher PickScore, HPSv2.1, and ImageReward, while SFT 35k reaches a higher Aesthetic score.

Table 2: State-count control (full-parameter). SOAR 5k matches SFT 35k in supervised states. Bold/underline: best/second best per metric.

| Metric | SFT 5k | SOAR 5k | SFT 35k |
|---|---|---|---|
| *High-CLIP subset* | | | |
| GenEval | 0.6983 | **0.7517** | <u>0.7166</u> |
| CLIPScore | <u>0.2980</u> | **0.2999** | 0.2971 |
| *High-Aesthetic subset* | | | |
| PickScore | <u>22.9464</u> | **23.0093** | 22.3779 |
| Aesthetic | 5.7504 | <u>5.9426</u> | **6.0664** |
| HPSv2.1 | <u>0.2954</u> | **0.2978** | 0.2771 |
| ImageReward | <u>1.1431</u> | **1.1771** | 0.9277 |

**Lighter configurations.** A single auxiliary state (two supervised states per sample instead of seven) retains most of the Aesthetic gain, with 5.8866 versus 5.9045 for six points and 5.7903 for LoRA-SFT, and still exceeds LoRA-SFT on CLIPScore (0.2941 versus 0.2915; Table 3 and Table S2). Inference cost is unchanged in every configuration.

### 4.4 Comparisons under Matched LoRA Parameterization

**Is the comparison with reward optimization fair?** The reward-specific curves in Figure 3 compare full-parameter SFT/SOAR with LoRA Flow-GRPO. Table 3 removes this capacity difference: all three methods train the same LoRA adapters (Hu et al., 2022) (rank 32, $\alpha = 64$, on all attention projections) and share one evaluation protocol. Without access to the reward model, SOAR reaches higher target values than Flow-GRPO on high-Aesthetic (5.9045 versus 5.8803) and high-CLIP (0.2971 versus 0.2965). On high-Aesthetic, Flow-GRPO falls well below LoRA-SFT on GenEval (0.5718 versus 0.6744) and OCR (0.3974 versus 0.6226), whereas SOAR stays close to LoRA-SFT on both (0.6783 and 0.6138). Flow-GRPO also falls below LoRA-SFT on PickScore, HPSv2.1, and ImageReward on this subset, consistent with over-optimizing its target reward, while SOAR

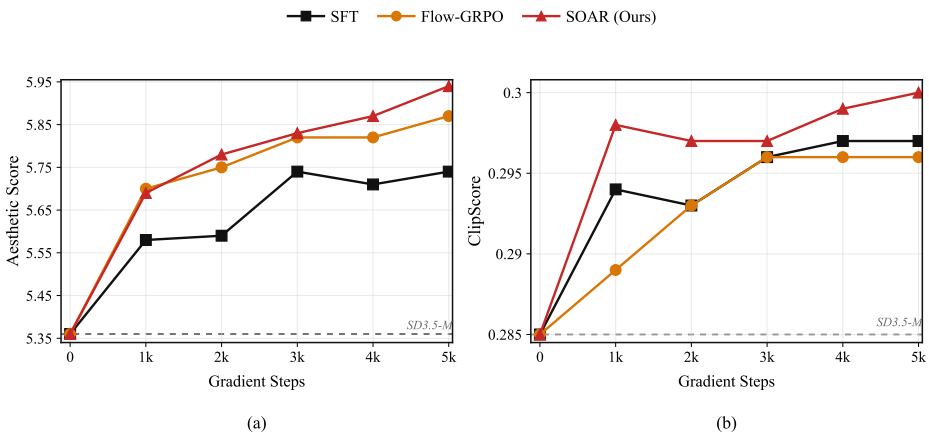


Figure 3: Reward-specific training dynamics on OOD DrawBench: Aesthetic after high-Aesthetic training (left) and CLIPScore after high-CLIP training (right). SFT/SOAR update all parameters and Flow-GRPO uses LoRA; Table 3 provides a comparison under a shared LoRA configuration.

Table 3: Matched-LoRA comparison with a shared adapter parameterization and evaluation protocol. All methods use 5k checkpoints. Bold and underline mark the best and second-best values within each subset.

| Method | Rule-based | | Model-based | | | |
|---|---|---|---|---|---|---|
| | GenEval | OCR | PickScore | Target | HPSv2.1 | ImageReward |
| *High-Aesthetic subset (target: Aesthetic)* | | | | | | |
| LoRA-SFT | <u>0.6744</u> | **0.6226** | <u>22.8062</u> | 5.7903 | <u>0.2898</u> | **1.1040** |
| LoRA-SOAR | **0.6783** | <u>0.6138</u> | **22.8949** | **5.9045** | **0.2927** | <u>1.0677</u> |
| LoRA-GRPO | 0.5718 | 0.3974 | 22.5909 | <u>5.8803</u> | 0.2809 | 0.8916 |
| *High-CLIP subset (target: CLIPScore)* | | | | | | |
| LoRA-SFT | <u>0.6980</u> | **0.6952** | <u>22.6980</u> | 0.2915 | **0.2878** | 1.0194 |
| LoRA-SOAR | **0.7303** | 0.6812 | **22.7463** | **0.2971** | <u>0.2877</u> | **1.0917** |
| LoRA-GRPO | 0.6687 | <u>0.6902</u> | 22.5161 | <u>0.2965</u> | 0.2804 | <u>1.0654</u> |

exceeds LoRA-SFT on PickScore and HPSv2.1. On high-CLIP, SOAR attains the best GenEval (0.7303), PickScore (22.7463), and ImageReward (1.0917) of the three methods, and across the two subsets it improves most non-target scores relative to Flow-GRPO. These results motivate a staged view of post-training: first use SOAR to correct trajectory-level bias on broad data, improving measured general quality without directly optimizing a reward model, and then study targeted reward optimization (e.g., GRPO) when a specific metric deficit remains.

**Training dynamics and visual comparisons.** In Figure 3, SOAR reaches the best final value at 5,000 steps in both reward-specific experiments: 5.94 aesthetic score (vs. 5.74 for SFT, 5.87 for Flow-GRPO) and 0.300 CLIPScore (vs. 0.297 for SFT, 0.296 for Flow-GRPO). Its Aesthetic score rises at every checkpoint and exceeds Flow-GRPO's from 2k updates on, and its CLIPScore is the highest at every checkpoint. Figure 4 follows three high-Aesthetic prompts over ten checkpoints for all three methods.

### 4.5 Mechanism and Constructor Ablations

**Which components drive the gain?** Table 4 identifies metric-dependent effects. Removing re-noising lowers all six metrics relative to $\lambda = 1$ and leaves five of them below $\lambda = 0$ SFT, showing that the re-noised states are essential. Replacing the paired correction target with the standard flow target still improves five of six metrics over $\lambda = 0$ SFT. At $\lambda = 1$, paired correction yields higher GenEval, CLIPScore, and Aesthetic, while the standard target yields higher PickScore, HPSv2.1, and ImageReward. Re-noised model-induced states therefore account for most of the gain, while

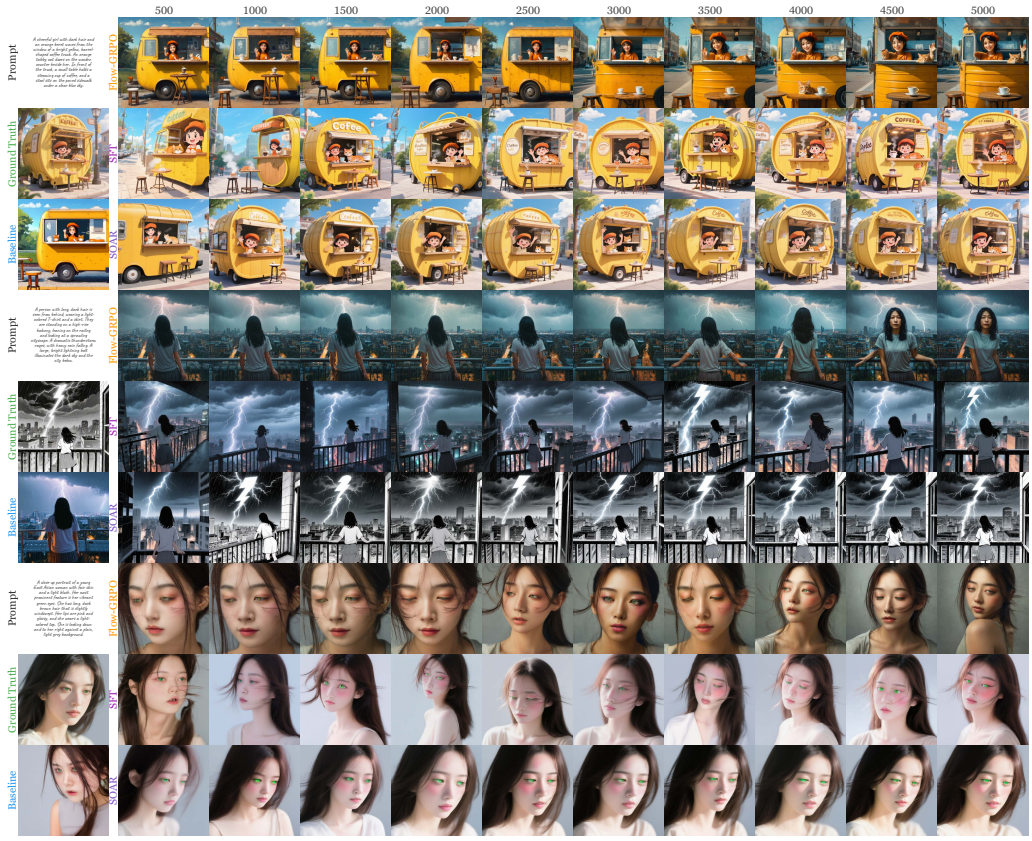


Figure 4: Training progression for three high-Aesthetic prompts. Additional examples are provided in Appendix G. The left column shows the prompt, paired ground truth, and pretrained output. Rows compare Flow-GRPO (LoRA), SFT, and SOAR (full parameters), and columns span checkpoints from 500 to 5,000 updates.

Table 4: Core ablations on the full 286K corpus with LoRA, batch 512, $512^2$ resolution, and 5k updates. This protocol differs in parameterization and budget from Table 1. Bold and underline mark the best and second-best values per column.

| Variant | Rule-based | Model-based | | | | |
|---|---|---|---|---|---|---|
| | GenEval | PickScore | CLIPScore | HPSv2.1 | Aesthetic | ImageReward |
| *Correction weight* | | | | | | |
| $\lambda = 0$ (SFT) | 0.6787 | 22.7566 | 0.2930 | 0.2890 | 5.4227 | <u>1.0695</u> |
| $\lambda = 0.5$ | 0.7059 | **22.8552** | <u>0.2947</u> | <u>0.2892</u> | <u>5.4600</u> | **1.0841** |
| $\lambda = 1$ | <u>0.7156</u> | 22.8076 | 0.2944 | 0.2880 | **5.4777** | 1.0440 |
| $\lambda = 2$ | **0.7175** | 22.7705 | **0.2957** | 0.2865 | 5.4395 | 1.0565 |
| *Component substitutions at* $\lambda = 1$ | | | | | | |
| No re-noising | 0.6755 | 22.6841 | 0.2931 | 0.2814 | 5.4048 | 1.0039 |
| Standard target | 0.7050 | <u>22.8421</u> | 0.2938 | **0.2893** | 5.4346 | 1.0628 |

the correction target shifts the metric profile toward GenEval, CLIPScore, and Aesthetic. Across $\lambda \in \{0.5, 1, 2\}$, the best value depends on the metric, and $\lambda = 0.5$ gives the most balanced profile across the six metrics.

Constructor sweeps (Appendix C.2) hold evaluation CFG at 4.5 while varying rollout CFG or the number of auxiliary points. Conditional-only construction (CFG 1) and a single auxiliary point both retain target-reward gains over LoRA-SFT on the two curated subsets (for example, 5.8915 and 5.8866 versus 5.7903 Aesthetic). The strongest guidance tested, CFG 7, gives the sweep's best

Aesthetic score (5.9169) and stays within 0.002 of the default on the high-CLIP subset, so the paired target remains usable under strong guidance. Increasing the auxiliary count from 1 to 3, 6, and 12 raises both target rewards at every step. We use six points by default, and constructor CFG 4.5 matches the pretrained model's inference setting.

### 4.6 Exposure-Bias Baselines and Analysis

**How does SOAR compare with exposure-bias remedies?** Input Perturbation (Ning et al., 2023) adds noise to the training inputs to simulate prediction errors, and Epsilon Scaling (Ning et al., 2024) rescales the predicted noise at inference. Table 5 compares both with SOAR under the shared high-Aesthetic LoRA protocol (batch 32, 5k updates), with perturbation coefficient 0.1 for Input Perturbation. SOAR's Aesthetic gain over SFT is more than three times that of Epsilon Scaling (0.1142 versus 0.0335), whereas Input Perturbation falls below SFT, so perturbations induced by the model itself help more than random input noise in this setting. Appendix C.3 gives the protocol details.

Table 5: Exposure-bias baselines on the high-Aesthetic subset with shared LoRA adapters. Bold/underline: best/second best.

| **Method** | Aesthetic | $\Delta$ vs. SFT |
|---|---|---|
| LoRA-SFT | 5.7903 | – |
| Input Perturbation | 5.7004 | −0.0899 |
| Epsilon Scaling | <u>5.8238</u> | <u>+0.0335</u> |
| LoRA-SOAR | **5.9045** | **+0.1142** |

**How much diversity does SOAR retain?** Over 400 prompts with eight seeds each, SOAR retains 84.8% of SFT's LPIPS, 86.7% of its DINOv2 distance, and 95.8% of its Vendi score among correct outputs (95.2% over all outputs). For DINO distance and both Vendi scores, most of the reduction from the pretrained model (73–81%) already occurs with SFT (Appendix E).

**Does the one-step construction reflect inference drift?** At CFG 4.5, the constructed drift is 0.40–0.48 of the true multi-step ODE drift for $\sigma \leq 0.725$, rises to 0.59 at $\sigma = 0.825$, and reaches parity at $\sigma = 0.925$ (Appendix D.1). SOAR thus trains on local perturbations concentrated at high noise: 96.5% of its auxiliary states lie above $\sigma = 0.5$. Under multi-step SDE sampling (CFG 4.5, noise 0.2), constructed-to-true clean-endpoint-estimation error ratios are 0.79–1.07 for $\sigma = 0.425$–0.975; because the paired target acts on this endpoint estimate, constructed states expose errors comparable to those met during sampling. Building auxiliary states from the full model-generated ODE trajectory is also feasible and improves GenEval, OCR, PickScore, and CLIPScore over its own pretrained baseline, including OCR from 0.5770 to 0.7189 (Appendix D.4).

**Why do semantic metrics improve most?** The correction objective is especially relevant at high-noise stages, where object count, layout, and text structure are still being determined and where, as shown above, most auxiliary states lie. Errors made there are hard to repair later because low-noise steps mostly refine an already chosen plan. This helps explain why GenEval and OCR improve strongly, while perceptual metrics such as aesthetics improve more modestly. It is also consistent with the steadily rising Aesthetic curve in Figure 3.

## 5 Conclusion

We proposed SOAR, a bias-correction post-training method for diffusion and flow-matching generators. By constructing nearby off-trajectory states from the model's own rollout, re-noising them toward the same sampled noise, and supervising clean-endpoint correction, SOAR directly targets the exposure-bias gap between SFT training states and inference-time states. On SD3.5-Medium, SOAR improves rule-based and model-based metrics over SFT, yields gains on curated reward-specific subsets, and surpasses Flow-GRPO in final target metric value without access to a reward model. Dense trajectory correction provides a reward-free post-training stage that is compatible with later reward-aligned fine-tuning; how best to combine the two remains an open direction. Since the mechanism only assumes a multi-step diffusion or flow trajectory with a clean target, SOAR can extend to video generation, 3D synthesis, distillation, and other generative tasks.

## AI USE STATEMENT

We used AI tools to draft and revise sections of the manuscript from the authors' notes and experimental results, and for language editing that improves readability. We also used AI tools to format the LaTeX source, tables, and figure layouts, and to search for and check the bibliographic details of cited works. AI-drafted descriptions of results were checked by the authors against the experimental outputs. We did not use generative AI to design the method or the experiments, to run experiments, or to generate or change any reported number. The authors reviewed and verified all AI-assisted text, figures, and references. We take full responsibility for the final content.

## REPRODUCIBILITY STATEMENT

Section 3 specifies the method and its default configuration, and Section 4 describes each experimental protocol. The supplementary material contains the derivations (Appendix A), training algorithms (Algorithms S1 and S2), experimental protocols (Appendix B), additional ablations and diagnostics (Appendices C–E), and anonymized training and evaluation code whose README maps the paper's hyperparameters to command-line flags. We will first release the high-Aesthetic and high-CLIP subsets with their exact manifests and preprocessing scripts, followed by the complete 286K-pair corpus (Appendix F.1).

## ETHICS STATEMENT

SOAR inherits the data biases and potential misuse of its pretrained text-to-image backbone and training corpus. Better measured alignment does not establish fairness, factual reliability, or safety of generated images. Because supervision is anchored to paired images, data provenance, caption quality, and usage rights matter for the planned dataset release.

## REFERENCES

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Black Forest Labs. FLUX, 2024. URL https://github.com/black-forest-labs/flux.

Giannis Daras, Yuval Dagan, Alexandros G. Dimakis, and Constantinos Daskalakis. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/831406cfe7e4a0aed5ac5c8a8389d1f5-Abstract-Conference.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning (ICML)*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d346d91999074dd8d6073d4c3b13733b-Abstract-Conference.html.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

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 (NeurIPS)*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. Dynamic prompt optimizing for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26627–26636, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Mo_Dynamic_Prompt_Optimizing_for_Text-to-Image_Generation_CVPR_2024_paper.html.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *International Conference on Machine Learning (ICML)*, 2023.

Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems (NeurIPS)*, 35:36479–36494, 2022.

Christoph Schuhmann. Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

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 (CVPR)*, pp. 8228–8238, 2024.

Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting LLM-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6327–6336, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_Self-correcting_LLM-controlled_Diffusion_Models_CVPR_2024_paper.html.

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

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

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2Img: Iterative self-refinement with GPT-4V for automatic image design and generation. In *European Conference on Computer Vision*, pp. 167–184. Springer, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/5515_ECCV_2024_paper.php.

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. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.16117.