LC-GRPO: BRIDGING TRAIN-INFERENCE GAP FOR FLOW-BASED GRPO WITH LANGEVIN CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-based generative models are typically sampled by solving a deterministic ordinary differential equation (ODE), whereas reinforcement learning (RL) requires stochastic rollouts for exploration and optimization. Existing online RL methods for flow models therefore replace the inference-time ODE with a stochastic differential equation (SDE) during training. Although the ODE and SDE share the same marginal distribution in continuous time, they can behave differently after finite-step discretization. In particular, SDE rollouts often become blurry as exploration noise increases, creating a mismatch between training rollouts and inference-time ODE samples. We introduce **LC-GRPO**, a flow-based GRPO framework with Langevin correction. Each rollout transition takes an inference-aligned ODE Euler step followed by a stochastic Langevin correction targeting the marginal at the resulting timestep. The required score is recovered directly from the flow velocity, requiring no additional score model, while the transition remains an isotropic Gaussian with tractable likelihood. We theoretically show that, under suitable conditions, a single Langevin correction step reduces the error of an imperfect ODE Euler step and, under matched randomness, yields a more accurate transition than standard Euler–Maruyama discretization of the reverse SDE. Additionally, experiments on SD3.5-Medium, FLUX.1-Dev, and HunyuanVideo show that LC-GRPO consistently improves reward optimization across text-to-image and text-to-video tasks, preserves generation quality, and substantially narrows the gap between stochastic training rollouts and deterministic ODE inference.

1 INTRODUCTION

Flow matching and rectified-flow models have emerged as a powerful foundation for visual generation, supporting high-quality text-to-image and text-to-video synthesis through deterministic ordinary differential equation (ODE) sampling (Liu et al., 2022; Lipman et al., 2022; Esser et al., 2024; Kong et al., 2024). Compared with stochastic diffusion sampling, the near-straight transport paths learned by these models enable accurate generation with relatively few sampling steps, making them especially attractive for large-scale visual models. As generative models have become increasingly capable, post-training with external feedback has emerged as a particularly effective paradigm for aligning model behavior with complex objectives that are difficult to capture through supervised learning alone. It has shown strong success across diverse, often non-differentiable objectives, including human preferences, visual quality, compositional correctness, and accurate text rendering (Xu et al., 2026; Gao et al., 2025a;b; Wu et al., 2025a). As a common approach in this alignment setting, online reinforcement learning (RL) methods such as DDPO (Black et al., 2024) and, more recently, Group Relative Policy Optimization (GRPO) (Liu et al., 2026a; Xue et al., 2025), directly optimize reward signals on model-generated samples.

However, one important issue emerges: *the mismatch between model-generated samples during training and inference*, which can make the optimization signal used during training misaligned with inference-time behavior. This issue is not unique to visual generation: a similar training-inference gap has also been observed in LLM RL post-training, where separate inference and training stacks can introduce numerical and algorithmic inconsistencies, causing the rollout policy used for data collection to differ from the policy being optimized (Qi et al., 2025; Wasti et al.; Qiu et al., 2026; Yao et al., 2025; Liang et al., 2026).

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.

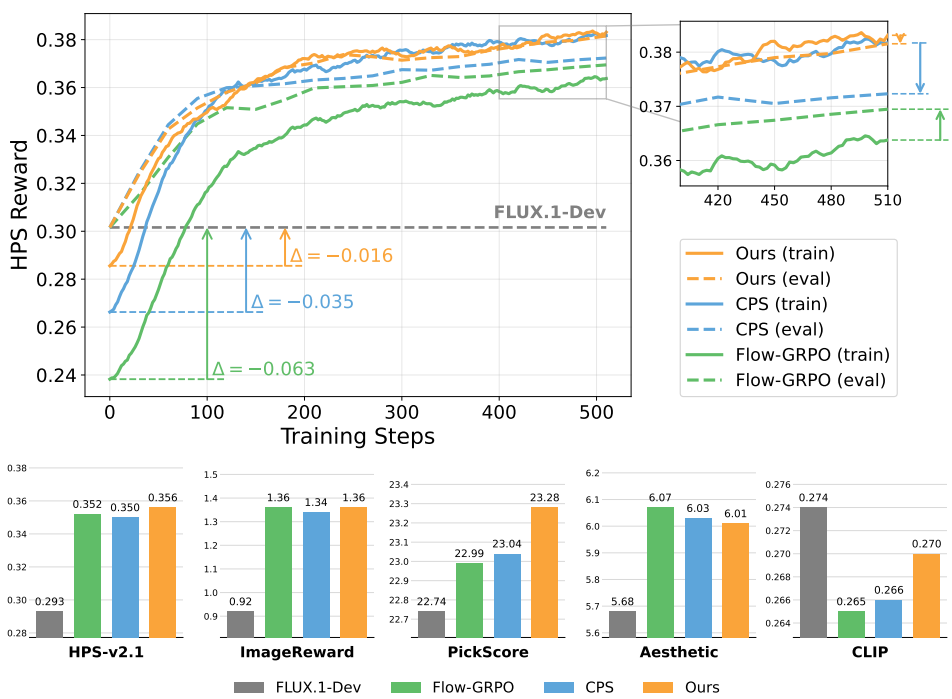


Figure 1: Performance of LC-GRPO. **Top:** Langevin correction narrows the gap between stochastic RL rollouts and ODE inference at test time, as shown at step 0, enabling LC-GRPO to achieve the smallest training–evaluation gap. **Bottom:** Evaluation results on DrawBench (Saharia et al., 2022).

The training–inference mismatch is even more fundamental in online RL for flow models. At test time, flow models are typically sampled by solving a deterministic reverse-time ODE. Reinforcement learning, however, requires *stochasticity for exploration*, allowing the policy to sample multiple actions from the same state. Existing flow-based GRPO methods therefore replace the ODE with a theoretically equivalent stochastic differential equation (SDE) for rollout generation (Liu et al., 2026a). In continuous time, the ODE and SDE share the same marginal distributions. However, in practice, sampling is typically performed with only a small number of discretization steps. The stochastic Brownian increments in SDE rollouts can introduce additional finite-step error, causing their behavior to deviate from deterministic ODE sampling (Kloeden et al., 2012). Empirically, we observe that applying SDE sampling to flow models produces noticeably blurrier and noisier samples than the ODE sampling (Liu et al., 2022; Wang & Yu, 2025).

Several recent works have attempted to reduce this training–inference mismatch in flow-based online RL. MixGRPO (Li et al., 2025) combines SDE and ODE sampling during rollout generation, but leaves the underlying SDE dynamics unchanged; its benefit is therefore largely orthogonal to the fundamental SDE–ODE mismatch. Coefficients-preserving sampling (CPS) (Wang & Yu, 2025) instead adopts a DDIM-style stochastic sampler, but its empirical gains remain limited.

We address this limitation with **LC-GRPO**, a flow-based GRPO framework based on Langevin correction (Alg. 1). Each rollout transition first takes the same ODE Euler step used by the inference-time sampler, followed by a single Langevin correction step at the resulting timestamp. This predictor–corrector construction separates transport from exploration: the ODE step preserves the inference-time dynamics, whereas the Langevin step injects stochasticity while preserving the target marginal distribution. Importantly, the required score is recovered directly from the flow velocity through Tweedie's formula, requiring no additional score model, while the resulting transition remains an isotropic Gaussian with a tractable density for policy optimization.

The key motivation is that Langevin correction provides a more accurate rollout update than directly simulating the reverse SDE, thereby reducing the gap to the inference-time ODE. Intuitively, the Langevin step acts as a correction that mitigates local sampling error while retaining the stochasticity needed for exploration. We support this intuition both theoretically and empirically. Our analysis

shows that, under matched randomness and the same exploration strength, Langevin-corrected sampling incurs lower error than the standard reverse-SDE Euler–Maruyama update (Theorem 2). Consistently, Fig. 1 shows that LC-GRPO already has the smallest gap to evaluation-time performance at training step 0. By the end of training, it maintains the smallest train–evaluation gap and achieves the best evaluation performance among the GRPO baselines.

We validate LC-GRPO on image and video generation using SD3.5-Medium (Esser et al., 2024), FLUX.1-Dev (Labs, 2024), and HunyuanVideo (Kong et al., 2024), covering visual text rendering, human-preference alignment, multi-reward optimization, and video-quality optimization (Sec. 5). Across these settings, LC-GRPO consistently outperforms Flow-GRPO, DanceGRPO, and coefficients-preserving sampling on the optimized rewards, while preserving auxiliary image and video quality metrics under rollout-computation budgets that match or favor the baselines.

To summarize, our main contributions are as follows:

1. We identify finite-step SDE discretization as a key source of the training–inference gap in flow-based GRPO and characterize its impact on rollout quality and reward alignment (Figs. 1, 2).
2. We introduce LC-GRPO, which combines inference-aligned ODE transport with a distribution-preserving Langevin correction to enable accurate, stochastic, likelihood-tractable RL rollouts.
3. We establish theoretical guarantees for the proposed correction (Theorems 1, 2) and demonstrate consistent empirical improvements across text-to-image and text-to-video tasks (Sec. 5).

2 Related Work

We first review alignment methods for diffusion and flow models, including post-training with external feedback and inference-time guidance, and then focus on closing the training–inference gap in flow-based GRPO and LLM RL post-training.

RL for Diffusion and Flow Models. Aligning diffusion and flow models during post-training generally follows one of three strategies. (1) Direct reward backpropagation, which fine-tunes the model by backpropagating through a differentiable reward function (Clark et al., 2024; Prabhudesai et al., 2023). (2) Reward-augmented regression losses, which combine external feedback with a diffusion regression objective. These include Reward-Weighted Regression (RWR) (Lee et al., 2023; Dong et al., 2023); Direct Preference Optimization (DPO), which leverages paired preference data with binary human-feedback rewards (Rafailov et al., 2023; Wallace et al., 2024; Yang et al., 2024); and DiffusionNFT (Zheng et al., 2025), which generalizes DPO by incorporating negative samples with dense, continuous rewards. (3) PPO-style policy-gradient methods, ranging from early approaches such as DPOK (Fan et al., 2023) and DDPO (Black et al., 2024) to recent GRPO extensions for diffusion and flow models (Liu et al., 2026a; Xue et al., 2025; Deng et al., 2026; Savani et al., 2026; Ding & Ye, 2025). Beyond model post-training, complementary inference-time alignment methods guide diffusion and flow sampling using time-dependent classifiers or gradients derived from downstream objectives (Song et al., 2020b; Dhariwal & Nichol, 2021; Chung et al., 2022; Yuan et al., 2023; Guo et al., 2024; 2026).

Training–Inference Gap in Flow-Based GRPO. Recent work explores reducing the mismatch between stochastic training rollouts and deterministic inference samplers in flow-based GRPO. MixGRPO (Li et al., 2025) integrates SDE and ODE sampling within the rollout to improve training efficiency. CPS (Wang & Yu, 2025) instead adopts DDIM sampling to improve sample quality. SAGE-GRPO (Zheng et al., 2026) proposes a concise formulation of the noise standard deviation that better aligns the sampling process with the data manifold for video generation.

Training–Inference Mismatch in LLM RL Post-Training. Practical LLM RL systems often use separate inference engines for rollout generation and separate frameworks for training, which can introduce a mismatch. One line of work addresses numerical inconsistencies (Qi et al., 2025; Qiu et al., 2026; Wasti et al.); for example, Qi et al. (2025) switch from BF16 to FP16 during RL fine-tuning. A complementary line of work takes an algorithmic approach by incorporating sampler-side information into the training update (Yao et al., 2025; Liu et al., 2025; Liang et al., 2026). Notably, Yao et al. (2025) use the training-to-sampler probability ratio as a clipped correction weight.

3 PRELIMINARIES

Flow Matching. Let $\boldsymbol{x}_0 \sim X_0$ denote a sample from the data distribution and $\boldsymbol{x}_1 \sim X_1$ denote a noise sample from the prior distribution, such as a standard Gaussian. Rectified Flow (Liu et al., 2022; Lipman et al., 2022) constructs a time-dependent interpolation between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \quad t \in [0,1].$$

Under this convention, $t = 0$ corresponds to data and $t = 1$ to noise. Generation is therefore performed by solving the reverse-time ODE from the prior distribution back to the data distribution:

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{v}_t \mathrm{d}t, \tag{1}$$

with $\boldsymbol{x}_1 \sim X_1$ and $t : 1 \to 0$. The velocity field $\boldsymbol{v}_t$ transports samples from the prior distribution to the data distribution. In practice, it is approximated by a transformer $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)$, trained with the flow matching regression objective: $\min_\theta \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1}\left[\|\boldsymbol{v}_\theta(\boldsymbol{x}_t,t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|_2^2\right]$. The learned velocity field is then used to solve the reverse-time ODE (Eq. 1) for generation.

SDE Sampling as RL Rollout. The inference process of flow matching can be formulated as a Markov decision process (MDP) (Black et al., 2024). At denoising step t, the state is defined as $\boldsymbol{s}_t := (\boldsymbol{c}, t, \boldsymbol{x}_t)$, the action is the next denoised sample $\boldsymbol{a}_t := \boldsymbol{x}_{t-1}$ predicted by the model, and the policy is given by $\pi(\boldsymbol{a}_t \mid \boldsymbol{s}_t) := p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c})$.

However, ODE-based sampling is deterministic and induces no action-level stochasticity, which blocks the stochastic exploration required by reinforcement learning. To resolve this mismatch, Flow-GRPO (Liu et al., 2026a) and DanceGRPO (Xue et al., 2025) replace the reverse-time ODE with an equivalent SDE:

$$\mathrm{d}\boldsymbol{x}_t = \left[\boldsymbol{v}_t(\boldsymbol{x}_t,t) + \frac{\sigma_t^2}{2t}\left(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_t(\boldsymbol{x}_t,t)\right)\right]\mathrm{d}t + \sigma_t \mathrm{d}\boldsymbol{w}_t, \tag{2}$$

which adds noise at each denoising step, enabling trajectories to serve as stochastic RL rollouts.

GRPO on Flow Models. Given a prompt $\boldsymbol{c}$, the flow model p_θ generates G independent denoising trajectories $\{(\boldsymbol{x}_t^i)_{t=0}^T\}_{i=1}^G$, each ending at sample $\boldsymbol{x}_0^i$. The advantage of the i-th sample is computed by normalizing its final reward $R(\boldsymbol{x}_0^i, \boldsymbol{c})$ against the rewards of other samples in the same group:

$$\hat{A}_t^i = \frac{R(\boldsymbol{x}_0^i,\boldsymbol{c}) - \mathrm{mean}(\{R(\boldsymbol{x}_0^j,\boldsymbol{c})\}_{j=1}^G)}{\mathrm{std}(\{R(\boldsymbol{x}_0^j,\boldsymbol{c})\}_{j=1}^G)}. \tag{3}$$

The policy is trained by maximizing the GRPO objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{\boldsymbol{c}\sim\mathcal{C},\{\boldsymbol{x}^i\}_{i=1}^G \sim \pi_{\theta_{\mathrm{old}}}(\cdot|\boldsymbol{c})} f(r, \hat{A}, \theta, \varepsilon, \beta), \tag{4}$$

where

$$f(r,\hat{A},\theta,\varepsilon,\beta) = \frac{1}{G}\sum_{i=1}^G \frac{1}{T}\sum_{t=1}^T \left(\min\left(r_t^i(\theta)\hat{A}_t^i,\ \mathrm{clip}\Big(r_t^i(\theta), 1-\varepsilon, 1+\varepsilon\Big)\hat{A}_t^i\right) - \beta D_{\mathrm{KL}}(\pi_\theta||\pi_{\mathrm{ref}})\right),$$

$$r_t^i(\theta) = \frac{p_\theta(\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i, \boldsymbol{c})}{p_{\theta_{\mathrm{old}}}(\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i, \boldsymbol{c})}.$$

In flow models, an action is one reverse-time sampling step from $\boldsymbol{x}_t^i$ to $\boldsymbol{x}_{t-1}^i$, obtained by discretizing Eq. 2 (see Eq. 6). The resulting transition probability $p_\theta(\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i, \boldsymbol{c})$ is an isotropic Gaussian, whose density can be computed in closed form.

4 METHOD

In this section, we first revisit the training–inference gap in standard GRPO for flow models, where SDE sampling is used to generate RL rollouts. We then propose LC-GRPO, which introduces an additional Langevin correction step to reduce the gap between stochastic RL rollouts and deterministic ODE inference. We validate this approach both theoretically and empirically.



Figure 2: **Sampling Method Comparison**. As the noise level[1] increases, SDE-based sampling degrades sample quality, whereas ODE sampling with Langevin correction preserves sample clarity.

Training–Inference Gap from SDE Rollouts. During inference, although the ODE (Eq. 1) and SDE (Eq. 2) reach the same marginal distribution at each timestep in continuous time (Kloeden et al., 2012), their finite-step numerical samplers can differ. ODE Euler step is usually taken for current advanced image and video generation (Esser et al., 2024; Wu et al., 2025b; Wan et al., 2025):

$$\boldsymbol{x}_{t-\Delta t} = \boldsymbol{x}_t - \boldsymbol{v}_t(\boldsymbol{x}_t, t)\Delta t. \tag{5}$$

Euler–Maruyama discretization is commonly used for SDE rollouts in GRPO for flow models:

$$\boldsymbol{x}_{t-\Delta t} = \boldsymbol{x}_t - \left[\boldsymbol{v}_t(\boldsymbol{x}_t, t) + \frac{\sigma_t^2}{2t}\left(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_t(\boldsymbol{x}_t, t)\right)\right]\Delta t + \sigma_t\sqrt{\Delta t}\,\xi, \tag{6}$$

where $\xi \sim \mathcal{N}(0, I_d)$. For the diffusion coefficient, Flow-GRPO sets $\sigma_t = \eta\sqrt{t/(1-t)}$. Here, η controls the noise level, and hence the amount of stochasticity and exploration during RL training. Both methods show that larger η values often accelerate GRPO training and lead to higher rewards.

As shown in Fig. 2, SDE-based sampling tends to produce lower-quality samples than ODE-based sampling, yielding blurrier images; this degradation becomes more severe as noise level increases. Similar behavior has been observed in prior work (Lipman et al., 2022; Song et al., 2020b). This discrepancy creates a training–inference gap in RL for flow models: SDE rollouts are worse than inference-time ODE samples, making their rewards a less accurate and less effective learning signal.

Langevin Step: Distribution-Preserving Stochastic Correction. Motivated by the training–inference gap, we aim to design a sampler that narrows this mismatch. SDE-based sampling accumulates error from both drift discretization and Brownian increments. To avoid introducing additional error through direct SDE simulation, we first take an ODE Euler step from $\boldsymbol{x}_t$ to obtain an intermediate proposal $\boldsymbol{x}'$ at time $t - \Delta t$. We then seek a stochastic update that satisfies two requirements: (i) it injects stochasticity while preserving the target distribution at time $t - \Delta t$; (ii) it improves sample quality by mitigating local numerical and discretization errors; (iii) it admits an analytically tractable transition likelihood for policy optimization in RL.

We adopt a simple yet effective strategy: after each ODE step, we apply one Langevin dynamics step initialized at $\boldsymbol{x}'$, using the score estimate at time $t - \Delta t$:

$$\boldsymbol{x}_{t-\Delta t} = \boldsymbol{x}' + \epsilon_t\,\boldsymbol{s}(\boldsymbol{x}', t - \Delta t) + \sqrt{2\epsilon_t}\,\xi, \tag{7}$$

where ϵ_t is the step size. The score function is obtained via Tweedie's formula (Efron, 2011) as $\boldsymbol{s}(\boldsymbol{x}, t) = -\left(\boldsymbol{x} + (1-t)\boldsymbol{v}(\boldsymbol{x}, t)\right)/t$. Following Song et al. (2020b), we set $\epsilon_t \propto \left(\|\boldsymbol{z}\| \,/\, \|\boldsymbol{s}\|\right)^2$. For stability, we use its expectation, which gives $\sqrt{2\epsilon_t} = \eta(t - \Delta t)$, where η controls the noise level.

This correction has the desired properties. First, Langevin dynamics driven by the score at time $t - \Delta t$ leaves the marginal $p_{t-\Delta t}$ invariant, thereby introducing randomness while preserving that marginal. Second, it serves as a corrector step: prior work on score-based generative modeling has

shown that such stochastic correction can improve sampling quality (Song et al., 2020b). As Fig 2 shows, sampling with Langevin correction preserves sample quality even under large noise levels.

Algorithm 1 Langevin Correction Sampling

Require: flow model $\boldsymbol{v}_\theta$, Langevin step size ϵ_t, time schedule $\{t_i\}_{i=1}^N$ with $t_1 = 1,\ t_N = 0$

$\boldsymbol{x}_1 \sim \mathcal{N}(\mathbf{0}, I_d)$
for $i = 1, \dots, N-1$ **do**
  $\boldsymbol{x}' \leftarrow \boldsymbol{x}_i - (t_i - t_{i+1})\boldsymbol{v}_\theta(\boldsymbol{x}_i, t_i)$ ▷ *Euler step*
  if $i < N-1$ **then**
    $\boldsymbol{s}_{t_{i+1}}(\boldsymbol{x}') \leftarrow -\dfrac{\boldsymbol{x}' + (1-t_{i+1})\boldsymbol{v}_\theta(\boldsymbol{x}', t_{i+1})}{t_{i+1}}$ ▷ *Score estimation*
    $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, I_d)$
    $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}' + \epsilon_{t_i}\boldsymbol{s}_{t_{i+1}}(\boldsymbol{x}') + \sqrt{2\epsilon_{t_i}}\,\boldsymbol{z}$ ▷ *Langevin correction*
  else
    $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}'$ ▷ *Final ODE step*
  end if
end for
return $\boldsymbol{x}_N$

As for the MDP we formulate for GRPO, the transition kernel $p_\theta(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_t, \boldsymbol{c})$ remains an isotropic Gaussian $\mathcal{N}(\boldsymbol{m}_\theta(\boldsymbol{x}_t), 2\epsilon_t I_d)$, whose mean $\boldsymbol{m}_\theta(\boldsymbol{x}_t) = \boldsymbol{x}' + \epsilon_t \boldsymbol{s}_\theta(\boldsymbol{x}', t-\Delta t)$ is the deterministic part of the Langevin update in Alg. 1, where $\boldsymbol{x}' = \boldsymbol{x}_t - \Delta t\, \boldsymbol{v}_\theta(\boldsymbol{x}_t, t)$ is the Euler step. Thus, the Langevin step size ϵ_t also controls the extent of exploration of the action policy.

The following theorem justifies the correction step in Alg. 1: whenever the Euler step is imperfect, one Langevin step provably moves the sample closer to the target marginal. The full statement and proof are deferred to Appx. A and Appx. A.1, respectively.

Theorem 1 (Informal). *Suppose the target $p_{t-\Delta t}$ is α-strongly log-concave with L-Lipschitz score. Let $\boldsymbol{x}_{\text{ode}}$ be the result of one backward Euler step to time $t - \Delta t$ (Eq. 5), and let $\varepsilon := W_2\big(\text{law}(\boldsymbol{x}_{\text{ode}}), p_{t-\Delta t}\big) > 0$ be its error. If the Langevin step size satisfies $\epsilon_t \le \alpha/L^2$ and $\epsilon_t \lesssim \alpha^2\varepsilon^2/(L^2 d)$, then one Langevin correction step strictly reduces the error:*

$$W_2\big(\text{law}(\boldsymbol{x}_{\text{lc}}), p_{t-\Delta t}\big) \;<\; W_2\big(\text{law}(\boldsymbol{x}_{\text{ode}}), p_{t-\Delta t}\big),$$

where $\boldsymbol{x}_{\text{lc}}$ is the Langevin-corrected sample (Eq. 7) and W_2 is the Wasserstein distance.

In flow-based GRPO, increasing the noise level, and thus the extent of policy exploration, typically improves reward optimization (Liu et al., 2026a) but degrades the sample quality of rollouts. Preserving quality at a fixed randomness level is therefore important. Below we show that when our Langevin-corrected sampler is matched to the Euler–Maruyama SDE at the same randomness level, that is, comparing Eq. 6 and Eq. 7 and setting the Langevin step size so that $\sqrt{2\epsilon_t} = \sigma_t\sqrt{\Delta t}$, i.e., $\epsilon_t = \sigma_t^2 \Delta t/2$, Langevin correction sampling is provably more accurate than the SDE. The formal statement and its proof are given in Appx. A and Appx. A.2, respectively.

Theorem 2 (Informal). *Under mild regularity (Assumption 2) and with the step size matched to the noise level, $\epsilon_t = \sigma_t^2\Delta t/2$, both the corrector $\boldsymbol{x}_{\text{lc}}$ and the reverse-SDE step $\boldsymbol{x}_{\text{sde}}$ (Eq. 6) have leading-order Wasserstein error of order Δt^2 (Eq. 12). If the comparison condition Eq. 13 holds, the Langevin correction sample step is strictly more accurate:*

$$W_2\big(\text{law}(\boldsymbol{x}_{\text{lc}}), p_{t-\Delta t}\big) \;<\; W_2\big(\text{law}(\boldsymbol{x}_{\text{sde}}), p_{t-\Delta t}\big),$$

for all sufficiently small Δt.

Beyond the above theoretical result, Fig. 2 provides empirical verification. Note that at the same noise level η, the sum of the variance of the Gaussian noise over all timesteps is actually larger for Langevin than for the SDE, implying at least as much exploration; nevertheless, our Langevin-corrected sampling produces substantially clearer samples than standard SDE-based sampling.

[1]The noise level refers to η in the SDE (Eq. 6), where $\sigma_t = \eta\sqrt{t/(1-t)}$ following Flow-GRPO, and in the Langevin step (Eq. 7), where we set $\sqrt{2\epsilon_t} = \eta(t - \Delta t)$. For the same η, the sum of the variance of the Gaussian noise over all timesteps is larger for Langevin than for the SDE.

5 EXPERIMENTS

In this section, we empirically evaluate LC-GRPO on text-to-image and text-to-video generation.

5.1 EXPERIMENTAL SETUP

Image Generation. We experiment with two flow-based text-to-image models, SD3.5-Medium (Esser et al., 2024) and FLUX.1-Dev (Labs, 2024), both at 512×512 resolution. These models cover two guidance regimes: SD3.5 uses classifier-free guidance (CFG), while FLUX.1-Dev is guidance-distilled and uses no CFG at inference. We consider three reward-optimization settings:

1. *Visual text rendering (OCR).* This is a verifiable reward that measures how accurately the model renders a target string. We follow Flow-GRPO (Liu et al., 2026a) for both the reward-assignment strategy and the training and evaluation datasets.
2. *Human preference alignment.* We use HPS-v2.1 (Wu et al., 2023) as the reward model, a preference model trained on human comparisons to predict which generated image humans prefer for a given prompt. We use the training and evaluation prompts from Pick-a-Pic (Kirstain et al., 2023).
3. *Multi-reward optimization.* To preserve text-image alignment while improving human preference, we jointly optimize CLIP score (Hessel et al., 2021) and HPS-v2.1. We use the same training and evaluation prompts as in the single-reward HPS-v2.1 setting.

Video Generation. For text-to-video generation, we experiment with HunyuanVideo (Kong et al., 2024). Following Xue et al. (2025), we use VideoAlign (Liu et al., 2026b) as the reward model and optimize its visual aesthetic quality component. We also report its motion quality and text–video alignment components for evaluation. As in DanceGRPO, training prompts are curated from VidProM (Wang & Yang, 2024); the evaluation dataset is from Zheng et al. (2026).

Quality Evaluation Metrics. When optimizing a target reward, we monitor quality metrics to detect reward-induced degradation or reward hacking. For images, we report two quality metrics, Aesthetic Score and CLIP score, and three human-preference metrics, ImageReward (Xu et al., 2023), HPS-v2.1, and PickScore (Kirstain et al., 2023). All image metrics are computed on DrawBench (Saharia et al., 2022). For videos, we report VBench (Huang et al., 2024) as quality metrics.

Baselines. We compare LC-GRPO against online reinforcement learning baselines based on GRPO. These methods share the same GRPO objective and differ primarily in how rollouts are sampled. The first baseline uses the standard SDE sampler in Eq. 6: Flow-GRPO (Liu et al., 2026a) for text-to-image generation and DanceGRPO (Xue et al., 2025) for text-to-video generation. The second baseline, CPS (Wang & Yu, 2025), instead samples rollouts with DDIM (Song et al., 2020a).

Fair Comparison on Computation. Since the compared methods mainly differ in rollout sampling, we match or favor the baselines in rollout computation. Each Langevin-corrected sampler step requires two NFEs; therefore, for each baseline, we run both the same number of rollout steps and twice that number, and report the better result. For SD3.5-M, FLUX.1-Dev, and HunyuanVideo, LC-GRPO uses 10, 6, and 8 rollout steps during training, while the baselines use 10/20, 6/12, and 16 steps, respectively. For evaluation, all methods use the same sampler: 40-step ODE Euler for SD3.5-M, 28 steps for FLUX.1-Dev, and 50 steps for HunyuanVideo. Thus, the proposed method does not use more computation than the baselines. Other hyperparameters are kept nearly identical, with details in Appendix B.

5.2 MAIN RESULTS

As shown in Tables 1 and 2, LC-GRPO consistently outperforms the baselines on both image and video generation tasks while preserving sample quality across auxiliary metrics. Qualitative results are shown in Fig. 3, with additional examples provided in Appendix C.

Gap between Training and Evaluation. We examine the gap between rewards on RL training rollouts and those obtained during evaluation inference. Figure 1 compares training rewards, computed on rollouts, with evaluation rewards, computed with ODE Euler sampling on the evaluation

Table 1: **Evaluation Results on Image Generation.** Task metric columns denote the reward(s) optimized in each setting. **Bold** is the best result. LC-GRPO consistently outperforms all baselines.

Model	Task Metric			Quality Metric				
	OCR	HPS-v2.1	CLIP	Aesthetic	CLIP	ImgRwd	HPS-v2.1	PickScore
SD3.5-M								
SD3.5-M	0.569	0.295	0.288	5.38	0.283	0.83	0.279	22.35
(a) Visual Text Rendering: OCR								
Flow-GRPO	0.914	—	—	5.31	0.288	0.94	**0.281**	22.43
CPS	0.935	—	—	5.20	0.287	0.75	0.265	22.16
LC-GRPO	**0.960**	—	—	**5.33**	**0.291**	**1.00**	0.280	**22.46**
(b) Human Preference Alignment: HPS-v2.1								
Flow-GRPO	—	0.381	—	**6.30**	0.267	1.41	0.357	22.89
CPS	—	0.371	—	6.12	0.270	1.38	0.344	22.62
LC-GRPO	—	**0.393**	—	6.11	**0.280**	**1.44**	**0.367**	**23.03**
(c) Multi-Reward: HPS-v2.1 + CLIP Score								
Flow-GRPO	—	0.351	0.296	**5.97**	0.289	1.28	0.327	22.76
CPS	—	0.345	0.291	5.73	0.285	1.31	0.323	22.68
LC-GRPO	—	**0.356**	**0.302**	5.73	**0.297**	**1.33**	**0.332**	**22.93**
FLUX.1-Dev								
FLUX.1-Dev	—	0.303	0.276	5.68	0.274	0.92	0.293	22.74
(a) Human Preference Alignment: HPS-v2.1								
Flow-GRPO	—	0.378	—	**6.07**	0.265	**1.36**	0.352	22.99
CPS	—	0.373	—	6.03	0.266	1.34	0.350	23.04
LC-GRPO	—	**0.384**	—	6.01	**0.270**	**1.36**	**0.356**	**23.28**
(b) Multi-Reward: HPS-v2.1 + CLIP Score								
Flow-GRPO	—	0.371	0.279	**5.94**	0.273	1.37	0.347	23.18
CPS	—	0.370	0.284	5.90	0.281	**1.44**	0.350	23.28
LC-GRPO	—	**0.376**	**0.285**	5.88	**0.284**	1.43	**0.352**	**23.40**

dataset throughout training. Flow-GRPO tends to produce blurry, lower-quality rollouts, so its training rewards are often lower than those of cleaner evaluation samples. In contrast, CPS improves rollout quality but leads to a large train–evaluation gap, with high training rewards dropping during inference. LC-GRPO also improves rollout quality while maintaining a smaller train–test gap.

Table 2: **Evaluation Results on Video Generation.** HunyuanVideo is trained with the **Visual Quality** reward from VideoAlign. **Bold** indicates the best result.

Model	VideoAlign			VBench				
	Visual Quality	Motion Quality	Text Alignment	Total Score	Quality Score	Semantic Score	Dynamic Degree	Aesthetic Quality
HunyuanVideo	-0.368	- 0.351	1.247	78.74	80.94	**69.96**	**52.8**	61.28
DanceGRPO	-0.205	-0.279	**1.312**	78.85	81.18	69.51	12.5	64.03
CPS	-0.276	**-0.178**	1.048	76.86	79.70	65.49	19.4	59.02
LC-GRPO	**0.063**	-0.223	1.151	**79.10**	**81.92**	67.83	45.8	**64.70**

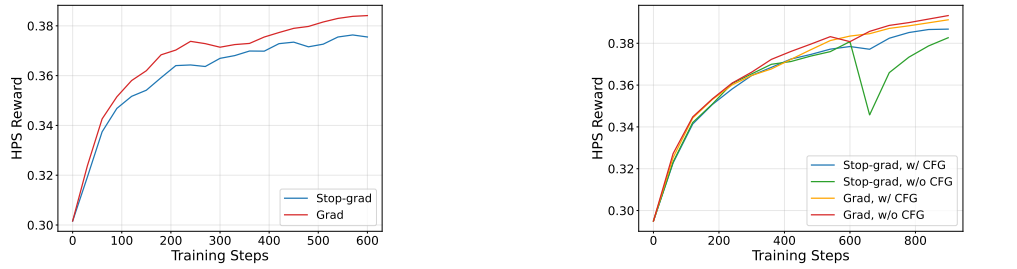
5.3 Ablation Studies

In this section, we analyze the design choices of the Langevin correction step: **(a) Gradients through Langevin Correction.** Since the Langevin correction step is not used at test time, we study whether gradients should be propagated through it during training. As shown in Fig. 4, back-propagating through the Langevin step leads to better performance and more stable training. We therefore keep this gradient path in all experiments. **(b) Classifier-Free Guidance.** Classifier-free guidance (CFG) is commonly used to improve sample quality, raising the question of whether it should also be applied in the Langevin correction step. However, the main role of Langevin correction is to introduce stochasticity and define a sampling distribution, rather than directly steering samples toward higher visual quality. We therefore ablate this design choice. As shown in Fig. 4(b),



Figure 3: **Qualitative Comparison**. Left: SD3.5-M-based models trained with OCR reward; Right: FLUX.1-Dev-based models trained with HPS-v2.1 and CLIP score. Prompts are taken from OCR and DrawBench, respectively.

applying Langevin correction without CFG achieves better performance on SD3.5-M while also reducing computation.



(a) FLUX.1-Dev: gradients in Langevin step

(b) SD3.5-M: gradients and CFG in Langevin step

Figure 4: Effect of design choices in the Langevin correction step.

6 Conclusion

We introduced LC-GRPO, a Langevin-corrected framework for online RL with flow-based generative models that reduces the mismatch between stochastic training rollouts and deterministic ODE inference. By combining inference-aligned ODE transport with a distribution-preserving correction, LC-GRPO enables exploration while maintaining accurate rollouts and tractable policy likelihoods. Our analysis shows that, at matched exploration strength, the proposed transition can achieve lower sampling error than standard SDE discretization. Experiments across text-to-image and text-to-video alignment tasks demonstrate consistent improvements in rollout quality and reward optimization, highlighting the benefit of aligning training-time sampling with inference-time generation.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement core methods, support qualitative and thematic data analysis, or interpret results. Assistance with generating synthetic datasets, translation, and dataset cleaning or reformatting is not applicable to this work.

Additionally, we used generative AI tools to create or edit utility software code and to edit the research paper to improve readability. We have reviewed all AI-assisted work.

LLMs were used only to assist with some connecting steps in the proofs in App. A.2; the main proof line and all other proof components were created manually by the authors. All AI-assisted proof content was verified by two authors. LLMs were also used only to generate utility code, which was verified and tested for correctness by two authors. Code for the core methods was written manually.

We take responsibility for the final content of this work, including any text, claims, code, or other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or the collection of sensitive personal data. Our work focuses on methodological improvements for reinforcement learning of flow-based generative models. We are not aware of any significant ethical concerns specific to the proposed methodology beyond the general considerations associated with generative image and video models.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, training procedure, experimental settings, and hyperparameters in the main paper and appendix. The appendix includes implementation details and complete proofs of the theoretical results. These materials are intended to support reproduction and verification of the main results. We plan to release the source code upon acceptance.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

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.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

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.

Haoyou Deng, Keyu Yan, Chaojie Mao, Xiang Wang, Yu Liu, Changxin Gao, and Nong Sang. Densegrpo: From sparse to dense reward for flow matching model alignment. *arXiv preprint arXiv:2601.20218*, 2026.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Zheng Ding and Weirui Ye. Treegrpo: Tree-advantage grpo for online rl post-training of diffusion models. *arXiv preprint arXiv:2512.08153*, 2025.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

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*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025a.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025b.

Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *arXiv preprint arXiv:2404.14743*, 2024.

Yingqing Guo, Yukang Yang, Hui Yuan, and Mengdi Wang. Training-free guidance beyond differentiability: Scalable path steering with tree search in diffusion and flow models. *Advances in Neural Information Processing Systems*, 38:73343–73384, 2026.

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*, pp. 7514–7528, 2021.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

Peter Eris Kloeden, Eckhard Platen, and Henri Schurz. *Numerical solution of SDE through computer experiments*. Springer Science & Business Media, 2012.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 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.

Jing Liang, Hongyao Tang, Yi Ma, Yancheng He, Weixun Wang, Xiaoyang Li, Ju Huang, Wenbo Su, Jinyi Liu, Yan Zheng, et al. The mirage of optimizing training policies: Monotonic inference policies as the real objective for llm reinforcement learning. *arXiv preprint arXiv:2606.29526*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jiacai Liu, Yingru Li, Yuqian Fu, Jiawei Wang, Qian Liu, and Yu Shen. When speed kills stability: Demystifying rl collapse from the training-inference mismatch. *Notion Blog*, 2025.

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, 2026a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. 2023.

Penghui Qi, Zichen Liu, Xiangxin Zhou, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Defeating the training-inference mismatch via fp16. *arXiv preprint arXiv:2510.26788*, 2025.

Zhaopeng Qiu, Shuang Yu, Jingqi Zhang, Shuai Zhang, Xue Huang, Jingyi Yang, and Junjie Lai. Fp8-rl: A practical and stable low-precision stack for llm reinforcement learning. *arXiv preprint arXiv:2601.18150*, 2026.

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.

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*, 35:36479–36494, 2022.

Yash Savani, Branislav Kveton, Yuchen Liu, Yilin Wang, Jing Shi, Subhojyoti Mukherjee, Nikos Vlassis, and Krishna Kumar Singh. Stepwise credit assignment for grpo on flow-matching models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42007–42017, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

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.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. *arXiv preprint arXiv:2509.05952*, 2025.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

Bram Wasti, Wentao Ye, Teja Rao, Michael Goin, et al. No more train-inference mismatch: Bitwise consistent on-policy reinforcement learning with vllm and torchtitan. 2025. *URl: https://blog. vllm. ai/2025/11/10/bitwise-consistent-train-inference. html*.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025a.

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*, 2025b.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Yixian Xu, Kaiyuan Gao, Yuxiang Chen, Yilei Chen, Zecheng Tang, Zihao Liu, Zikai Zhou, Deqing Li, Hao Meng, Kuan Cao, et al. Qwen-image-2.0-rl technical report. *arXiv preprint arXiv:2606.27608*, 2026.

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*, 2025.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024.

Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Your efficient rl framework secretly brings you off-policy rl training, august 2025. *URL https://fengyao. notion. site/off-policy-rl*, 2025.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *Advances in Neural Information Processing Systems*, 36:60599–60635, 2023.

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.

Mingzhe Zheng, Weijie Kong, Yue Wu, Dengyang Jiang, Yue Ma, Xuanhua He, Bin Lin, Kaixiong Gong, Zhao Zhong, Liefeng Bo, et al. Manifold-aware exploration for reinforcement learning in video generation. *arXiv preprint arXiv:2603.21872*, 2026.

A Theoretical Analysis

We focus on a single denoising step of the flow model, running from time t to $r := t - h$ with step size $h > 0$. The probability flow ODE is

$$\mathrm{d}\boldsymbol{x}_u = \boldsymbol{v}_u\left(\boldsymbol{x}_u\right)\,\mathrm{d}u, \tag{8}$$

with marginals p_u for $u \in [r, t]$.

Under this assumption, we show that a single Langevin step provably improves upon the Euler step. This result is stated formally as Theorem 3, with the proof given in Sec. A.1.

Assumption 1. *Let $f := \log p_r$. Then $-LI_d \preceq \nabla^2 f \preceq -\alpha I_d$ for some $0 < \alpha \leq L$; that is, f is α-strongly concave and ∇f is L-Lipschitz.*

Theorem 3 (Langevin Step Improves the Euler Step). *Let Assumption 1 hold. Starting from $\boldsymbol{x} \sim q_t$ at time t, take one backward Euler step of the probability flow ODE Eq. 8 with step size $h > 0$,*

$$\boldsymbol{x}_{\mathrm{ode}} = \boldsymbol{x} - h\, \boldsymbol{v}_t(\boldsymbol{x}),$$

and denote the resulting error by $\varepsilon := W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{ode}}),\, p_r\big)$, assumed positive. Then follow with one Langevin correction (LC) step targeting p_r,

$$\boldsymbol{x}_{\mathrm{lc}} = \boldsymbol{x}_{\mathrm{ode}} + \epsilon\, \nabla \log p_r(\boldsymbol{x}_{\mathrm{ode}}) + \sqrt{2\epsilon}\, \xi, \qquad \xi \sim \mathcal{N}(0, I_d), \quad \xi \perp \boldsymbol{x}_{\mathrm{ode}}.$$

If the step size satisfies

$$\epsilon \;\le\; \frac{\alpha}{L^2} \qquad \textit{and} \qquad \epsilon \;<\; \frac{9}{100} \cdot \frac{\alpha^2 \varepsilon^2}{L^2 d}, \tag{9}$$

then the Langevin step strictly improves upon the Euler step:

$$W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{lc}}),\, p_r\big) \;<\; W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{ode}}),\, p_r\big).$$

Theorem 3 shows that the corrector improves on the deterministic Euler step. We now compare it with the standard stochastic alternative: one Euler–Maruyama step of the reverse SDE. Denote $\boldsymbol{s}_u := \nabla \log p_u$ for the score. Given $\boldsymbol{x} \sim p_t$ and $\xi \sim \mathcal{N}(0, I_d)$, the SDE step at noise level $\sigma > 0$ is

$$\boldsymbol{x}_{\mathrm{sde}} = \boldsymbol{x} - h\, \boldsymbol{v}_t(\boldsymbol{x}) + \frac{\sigma^2 h}{2}\, \boldsymbol{s}_t(\boldsymbol{x}) + \sigma \sqrt{h}\, \xi, \tag{10}$$

while the Langevin correction step is exactly that of Theorem 3. We further define vector fields at time t:

$$\dot{\boldsymbol{v}} := \partial_t \boldsymbol{v}_t + \nabla \boldsymbol{v}_t\, \boldsymbol{v}_t, \qquad \dot{\boldsymbol{s}} := \partial_t \boldsymbol{s}_t + \nabla \boldsymbol{s}_t\, \boldsymbol{v}_t, \qquad \boldsymbol{G} := \nabla\Big(\tfrac{1}{2}\, \|\boldsymbol{s}_t\|^2 + \Delta \log p_t\Big), \tag{11}$$

Here, $\dot{\boldsymbol{v}}$ and $\dot{\boldsymbol{s}}$ are the rates of change of the velocity and score along the flow, while $\boldsymbol{G}$ is the direction of the bias introduced by a single Langevin step.

Under the following regularity assumption, we can compute the exact leading-order W_2 error of both the Langevin correction and the SDE step, and show that the Langevin correction step is more accurate under an explicit condition (Theorem 4); see Sec. A.2 for the proof.

Assumption 2. *Let $\boldsymbol{v}, \log p \in C^{2,4}\big([r, t] \times \mathbb{R}^d\big)$, with all derivatives of polynomial growth, and p_t has finite moments of all orders.*

Theorem 4 (Comparison with the SDE Step). *Let Assumption 2 hold and $\epsilon \le \Lambda h$ for a fixed constant $\Lambda > 0$. With $\dot{\boldsymbol{v}}, \dot{\boldsymbol{s}}, \boldsymbol{G}$ as defined in Eq. 11, we have*

$$\begin{aligned} W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{lc}}),\, p_r\big) &= \frac{h^2}{2}\, \Big\| \dot{\boldsymbol{v}} - \big(\tfrac{\epsilon}{h}\big)^2\, \boldsymbol{G} \Big\|_{p_t} + o(h^2), \\ W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{sde}}),\, p_r\big) &= \frac{h^2}{2}\, \Big\| \dot{\boldsymbol{v}} - \sigma^2 \dot{\boldsymbol{s}} - \tfrac{\sigma^4}{4} \boldsymbol{G} \Big\|_{p_t} + o(h^2). \end{aligned} \tag{12}$$

In particular, at $\epsilon = \sigma^2 h/2$ matching the noise level, let $\boldsymbol{E} := \dot{\boldsymbol{v}} - \frac{\sigma^4}{4} \boldsymbol{G}$, if the following holds:

$$\big\langle \boldsymbol{E}, \dot{\boldsymbol{s}} \big\rangle_{p_t} < \frac{\sigma^2}{2}\, \|\dot{\boldsymbol{s}}\|_{p_t}^2\,, \tag{13}$$

Langevin correction step is strictly more accurate:

$$W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{lc}}),\, p_r\big) \;<\; W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{sde}}),\, p_r\big),$$

for all sufficiently small h.

A.1 Proof of Theorem 3

Proof. Applying Lemma 1 at time r with initial law $q = q_r$ gives the upper bound:

$$W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{lc}}),\, p_r\big) \;\le\; \Big(1 - \frac{\alpha\epsilon}{2}\Big)\, \varepsilon + \frac{5}{3}\, L \sqrt{d}\, \epsilon^{3/2}.$$

The second condition in Eq. 9 is equivalent to $\sqrt{\epsilon} < \frac{3\alpha\varepsilon}{10\,L\sqrt{d}}$, hence

$$\frac{5}{3}\,L\sqrt{d}\,\epsilon^{3/2} \;<\; \frac{5}{3}\,L\sqrt{d}\cdot\frac{3\alpha\varepsilon}{10\,L\sqrt{d}}\cdot\epsilon \;=\; \frac{\alpha\epsilon}{2}\,\varepsilon,$$

so the discretization error is strictly dominated by the contraction gain, and

$$W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{lc}}),\,p_r\big) \;<\; \Big(1-\frac{\alpha\epsilon}{2}\Big)\,\varepsilon + \frac{\alpha\epsilon}{2}\,\varepsilon \;=\; \varepsilon.$$

Therefore, we complete the proof. □

Lemma 1 (Langevin Corrector). *Let Assumption 1 hold and given $\boldsymbol{z} \sim q$, langevin correction step:*

$$\boldsymbol{z}^+ = \boldsymbol{z} + \epsilon\nabla \log p_r(\boldsymbol{z}) + \sqrt{2\epsilon}\,\xi, \quad \xi \sim \mathcal{N}(0, I_d) \tag{14}$$

with $\epsilon \le \alpha/L^2$. Then,

$$W_2\big(\mathrm{law}(\boldsymbol{z}^+),\,p_r\big) \;\le\; \Big(1-\frac{\alpha\epsilon}{2}\Big)\,W_2(q,\,p_r) \;+\; \frac{5}{3}\,L\sqrt{d}\,\epsilon^{3/2}. \tag{15}$$

Proof. Denote $f := \log p_r$, so that the corrector step reads $\boldsymbol{z}^+ = \boldsymbol{z} + \epsilon\nabla f(\boldsymbol{z}) + \sqrt{2\epsilon}\,\xi$. Let $(\boldsymbol{z}, \boldsymbol{w}_0)$ be an optimal W_2 coupling of (q, p_r), and let $(B_\tau)_{\tau\ge 0}$ be a standard Brownian motion independent of $(\boldsymbol{z}, \boldsymbol{w}_0)$. Since ∇f is L-Lipschitz by Assumption 1, the Langevin dynamics

$$\mathrm{d}\boldsymbol{w}_\tau = \nabla f(\boldsymbol{w}_\tau)\,\mathrm{d}\tau + \sqrt{2}\,\mathrm{d}B_\tau, \qquad \tau \in [0,\epsilon], \quad \boldsymbol{w}_0 \sim p_r,$$

admits a unique strong solution with $p_r \propto e^f$ as its stationary distribution, so $\boldsymbol{w}_\epsilon \sim p_r$. Let $\xi := B_\epsilon/\sqrt{\epsilon} \sim \mathcal{N}(0, I_d)$, which is independent of $\boldsymbol{z}$. Therefore, we have $(\boldsymbol{z}^+, \boldsymbol{w}_\epsilon)$ a coupling of $\mathrm{law}(\boldsymbol{z}^+)$ and p_r, and thus

$$W_2\big(\mathrm{law}(\boldsymbol{z}^+),\,p_r\big) \le \big\|\boldsymbol{z}^+ - \boldsymbol{w}_\epsilon\big\|_{L^2}.$$

We first decompose the difference $\boldsymbol{z}^+ - \boldsymbol{w}_\epsilon$. Writing $\boldsymbol{w}_\epsilon = \boldsymbol{w}_0 + \int_0^\epsilon \nabla f(\boldsymbol{w}_\tau)\,\mathrm{d}\tau + \sqrt{2}\,B_\epsilon$ and subtracting it from the corrector step, the Brownian terms cancel and

$$\boldsymbol{z}^+ - \boldsymbol{w}_\epsilon = \underbrace{(\boldsymbol{z}-\boldsymbol{w}_0) + \epsilon\big(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{w}_0)\big)}_{=:\,A} + \underbrace{\int_0^\epsilon \big(\nabla f(\boldsymbol{w}_0) - \nabla f(\boldsymbol{w}_\tau)\big)\,\mathrm{d}\tau}_{=:\,B},$$

where A compares the two drifts at the initial points, and B is the local discretization error.

Next, we show that A contracts. Expanding the square and using $\langle \nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{w}_0),\, \boldsymbol{z} - \boldsymbol{w}_0\rangle \le -\alpha\,\|\boldsymbol{z}-\boldsymbol{w}_0\|^2$ together with $\|\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{w}_0)\| \le L\,\|\boldsymbol{z}-\boldsymbol{w}_0\|$, where the strong log-concavity and the Lipschitz property of ∇f are from Assumption 1, we have

$$\|A\|^2 \le \big(1 - 2\alpha\epsilon + L^2\epsilon^2\big)\,\|\boldsymbol{z}-\boldsymbol{w}_0\|^2 \le (1-\alpha\epsilon)\,\|\boldsymbol{z}-\boldsymbol{w}_0\|^2,$$

where the last inequality uses $L^2\epsilon \le \alpha$; note also $\alpha\epsilon \le \alpha^2/L^2 \le 1$, so the right-hand side is nonnegative. Since $\sqrt{1-x} \le 1 - x/2$ for $x \in [0,1]$ and the coupling $(\boldsymbol{z}, \boldsymbol{w}_0)$ is optimal, it holds

$$\|A\|_{L^2} \le \Big(1-\frac{\alpha\epsilon}{2}\Big)\,W_2(q,\,p_r).$$

Then, we bound the local error B. Since Langevin dynamics is stationary, $\boldsymbol{w}_r \sim p_r$ for every r. Using $(a+b)^2 \le 2a^2 + 2b^2$, the Cauchy–Schwarz inequality on the drift integral, Lemma 3, and $\mathbb{E}\,\big\|\sqrt{2}\,B_\tau\big\|^2 = 2\tau d$, we obtain

$$\mathbb{E}\,\|\boldsymbol{w}_\tau - \boldsymbol{w}_0\|^2 \le 2\tau\int_0^\tau \mathbb{E}\,\|\nabla f(\boldsymbol{w}_r)\|^2\,\mathrm{d}r + 4\tau d \le 2Ld\,\tau^2 + 4d\,\tau.$$

By the Cauchy–Schwarz inequality on the time integral and the Lipschitz property of ∇f,

$$\mathbb{E}\,\|B\|^2 \le \epsilon\int_0^\epsilon \mathbb{E}\,\|\nabla f(\boldsymbol{w}_0) - \nabla f(\boldsymbol{w}_\tau)\|^2\,\mathrm{d}\tau \le \epsilon L^2\int_0^\epsilon \mathbb{E}\,\|\boldsymbol{w}_\tau - \boldsymbol{w}_0\|^2\,\mathrm{d}\tau \le \epsilon L^2\Big(\frac{2Ld\,\epsilon^3}{3} + 2d\,\epsilon^2\Big).$$

Since $\epsilon \le \alpha/L^2 \le 1/L$, we have $L\epsilon \le 1$ and hence

$$\mathbb{E}\,\|B\|^2 \le \frac{8}{3}\,L^2 d\,\epsilon^3, \qquad \|B\|_{L^2} \le \sqrt{\tfrac{8}{3}}\,L\sqrt{d}\,\epsilon^{3/2} \le \frac{5}{3}\,L\sqrt{d}\,\epsilon^{3/2}.$$

Finally, combining the above bounds yields

$$W_2\big(\mathrm{law}(\boldsymbol{z}^+),\,p_r\big) \le \|A\|_{L^2} + \|B\|_{L^2} \le \Big(1-\frac{\alpha\epsilon}{2}\Big)\,W_2(q,\,p_r) + \frac{5}{3}\,L\sqrt{d}\,\epsilon^{3/2}.$$ □

A.2 Proof of Theorem 4

Proof. For a law q and a smooth test function f, define $\delta(f;q) := \int f \,\mathrm{d}q - \int f \,\mathrm{d}p_r$. For brevity, we write $\delta_{\mathrm{lc}}(f) := \delta(f;\mathrm{law}(\boldsymbol{x}_{\mathrm{lc}}))$ and $\delta_{\mathrm{sde}}(f) := \delta(f;\mathrm{law}(\boldsymbol{x}_{\mathrm{sde}}))$.

We first expand the Langevin correction step weakly. Let $f_\epsilon(\boldsymbol{y}) := \mathbb{E}_\xi f\big(\boldsymbol{y} + \epsilon\, \boldsymbol{s}_r(\boldsymbol{y}) + \sqrt{2\epsilon}\,\xi\big)$ denote the expected value of f after one corrector step from $\boldsymbol{y}$, so that $\mathbb{E}\, f(\boldsymbol{x}_{\mathrm{lc}}) = \mathbb{E}_{p_t} f_\epsilon(\boldsymbol{x}_{\mathrm{ode}})$ and

$$\delta_{\mathrm{lc}}(f) = \underbrace{\mathbb{E}_{p_t} f_\epsilon(\boldsymbol{x}_{\mathrm{ode}}) - \mathbb{E}_{p_r} f_\epsilon}_{\text{predictor error}} + \underbrace{\mathbb{E}_{p_r} f_\epsilon - \mathbb{E}_{p_r} f}_{\text{corrector bias}}. \tag{16}$$

For the predictor error, let $\Phi := \Phi_{t\to r}$ denote the exact flow map, which satisfies $\Phi \# p_t = p_r$. Since the trajectory $u \mapsto \boldsymbol{x}(u)$ has acceleration exactly $\dot{\boldsymbol{v}}$, a Taylor expansion gives $\Phi(\boldsymbol{x}) = \boldsymbol{x} - h\,\boldsymbol{v}_t(\boldsymbol{x}) + \frac{h^2}{2}\,\dot{\boldsymbol{v}}(\boldsymbol{x}) + O(h^3)$. Hence $\boldsymbol{x}_{\mathrm{ode}} - \Phi(\boldsymbol{x}) = -\frac{h^2}{2}\,\dot{\boldsymbol{v}}(\boldsymbol{x}) + O(h^3)$, and for any test function ψ,

$$\mathbb{E}_{p_t}\,\psi(\boldsymbol{x}_{\mathrm{ode}}) - \mathbb{E}_{p_r}\,\psi = \mathbb{E}_{p_t}\big[\psi(\boldsymbol{x}_{\mathrm{ode}}) - \psi(\Phi(\boldsymbol{x}))\big] = -\frac{h^2}{2}\,\mathbb{E}_{p_t}\langle \dot{\boldsymbol{v}},\, \nabla\psi\rangle + O(h^3).$$

We apply this with $\psi = f_\epsilon$; Taylor-expanding in ϵ gives $\nabla f_\epsilon = \nabla f + O(\epsilon)$, and since $\epsilon \le \Lambda h$, the substitution costs only $O(h^3)$.

The corrector bias is that of one Langevin step started from its own target p_r. Write $\boldsymbol{s} := \boldsymbol{s}_r$ and let $Lg := \langle \boldsymbol{s}, \nabla g\rangle + \Delta g$ be the Langevin generator; Due to integration by parts, and recalling $\boldsymbol{s} = \nabla \log p_r$, we get invariance of p_r, which means $\mathbb{E}_{p_r} Lg = 0$ for all g. Taylor expanding f in the increment $\epsilon\,\boldsymbol{s} + \sqrt{2\epsilon}\,\xi$ and taking Gaussian moments,

$$\mathbb{E}_{p_r} f_\epsilon - \mathbb{E}_{p_r} f = \epsilon\,\mathbb{E}_{p_r}\big[Lf\big] + \epsilon^2\,\mathbb{E}_{p_r}\Big[\tfrac{1}{2}\boldsymbol{s}^\top \nabla^2 f\,\boldsymbol{s} + \langle \boldsymbol{s}, \nabla \Delta f\rangle + \tfrac{1}{2}\Delta^2 f\Big] + O(\epsilon^3).$$

The order-ϵ term vanishes by invariance. For the order-ϵ^2 term, call the bracket B. By invariance with $g = Lf$, yielding $\mathbb{E}_{p_r}[\frac{1}{2}L^2 f] = 0$, we can subtract it for free. Expanding $\frac{1}{2}L^2 f$ and cancelling the common terms leaves only

$$B - \tfrac{1}{2}L^2 f = -\big\langle \nabla \boldsymbol{s},\, \nabla^2 f\big\rangle - \tfrac{1}{2}\big\langle (\boldsymbol{s}\cdot\nabla)\boldsymbol{s} + \Delta \boldsymbol{s},\, \nabla f\big\rangle,$$

i.e. B differs from $\frac{1}{2}L^2 f$ by the Hessian term $\langle \nabla \boldsymbol{s}, \nabla^2 f\rangle$ plus a first-order piece. One integration by parts, using $\nabla p_r = p_r\,\boldsymbol{s}$, removes the Hessian:

$$\mathbb{E}_{p_r}\big\langle \nabla \boldsymbol{s}, \nabla^2 f\big\rangle = -\,\mathbb{E}_{p_r}\big\langle (\boldsymbol{s}\cdot\nabla)\boldsymbol{s} + \Delta \boldsymbol{s},\, \nabla f\big\rangle,$$

so the two first-order pieces combine into $\frac{1}{2}\,\mathbb{E}_{p_r}\langle (\boldsymbol{s}\cdot\nabla)\boldsymbol{s} + \Delta\boldsymbol{s}, \nabla f\rangle$. Because $\boldsymbol{s} = \nabla \log p_r$ is a gradient, $(\boldsymbol{s}\cdot\nabla)\boldsymbol{s} = \nabla(\frac{1}{2}\|\boldsymbol{s}\|^2)$ and $\Delta\boldsymbol{s} = \nabla\Delta \log p_r$, so this field is itself a gradient. Hence

$$\mathbb{E}_{p_r} f_\epsilon - \mathbb{E}_{p_r} f = \frac{\epsilon^2}{2}\,\mathbb{E}_{p_r}\Big\langle \nabla\Big(\tfrac{1}{2}\,\|\boldsymbol{s}\|^2 + \Delta \log p_r\Big),\, \nabla f\Big\rangle + O(\epsilon^3) = \frac{\epsilon^2}{2}\,\mathbb{E}_{p_t}\langle \boldsymbol{G},\, \nabla f\rangle + O(h^3),$$

where moving from time r to time t costs $O(\epsilon^2 h) = O(h^3)$. Combining the two parts,

$$\delta_{\mathrm{lc}}(f) = -\frac{h^2}{2}\,\mathbb{E}_{p_t}\Big\langle \dot{\boldsymbol{v}} - \big(\tfrac{\epsilon}{h}\big)^2\,\boldsymbol{G},\, \nabla f\Big\rangle + O(h^3). \tag{17}$$

Next, we show the SDE step adds exactly one term. Let $\tilde{\boldsymbol{x}}$ be the Langevin step run with $\epsilon = \sigma^2 h/2$ and the same noise ξ. Then $\sqrt{2\epsilon} = \sigma\sqrt{h}$, so the two schemes share their noise and differ only in where the score is evaluated:

$$\boldsymbol{x}_{\mathrm{sde}} - \tilde{\boldsymbol{x}} = \frac{\sigma^2 h}{2}\,\big[\boldsymbol{s}_t(\boldsymbol{x}) - \boldsymbol{s}_r(\boldsymbol{x}_{\mathrm{ode}})\big] = \frac{\sigma^2 h^2}{2}\,\dot{\boldsymbol{s}}(\boldsymbol{x}) + O(h^3), \tag{18}$$

by first-order Taylor expansion of $(u, \boldsymbol{y}) \mapsto \boldsymbol{s}_u(\boldsymbol{y})$ at $(t, \boldsymbol{x})$ in the direction $(-h,\, -h\,\boldsymbol{v}_t(\boldsymbol{x}))$, which produces exactly $\dot{\boldsymbol{s}}$. This difference Eq. 18 is a function of $\boldsymbol{x}$ alone, so expanding f between the two points, the cross terms with the noise vanish in expectation (ξ is centered and independent of $\boldsymbol{x}$), and

$$\delta_{\mathrm{sde}}(f) = \delta_{\mathrm{lc}}(f)\big|_{\epsilon=\sigma^2 h/2} + \frac{\sigma^2 h^2}{2}\,\mathbb{E}_{p_t}\langle \dot{\boldsymbol{s}}, \nabla f\rangle = -\frac{h^2}{2}\,\mathbb{E}_{p_t}\Big\langle \dot{\boldsymbol{v}} - \sigma^2\dot{\boldsymbol{s}} - \tfrac{\sigma^4}{4}\boldsymbol{G},\, \nabla f\Big\rangle + O(h^3). \tag{19}$$

Finally, applying Lemma 4 to Eq. 17 with $\boldsymbol{A} = \dot{\boldsymbol{v}} - (\frac{\epsilon}{h})^2 \boldsymbol{G}$, and to Eq. 19 with $\boldsymbol{A} = \dot{\boldsymbol{v}} - \sigma^2 \dot{\boldsymbol{s}} - \frac{\sigma^4}{4}\boldsymbol{G}$. In both cases, $\boldsymbol{A}$ is a gradient field: $\boldsymbol{v}_t$ and $\boldsymbol{s}_t = \nabla \log p_t$ are gradients (hence so are $\dot{\boldsymbol{v}}$ and $\dot{\boldsymbol{s}}$), and $\boldsymbol{G}$ is a gradient by construction. This yields Eq. 12.

For the matched step $\epsilon = \sigma^2 h/2$ we have $(\epsilon/h)^2 = \sigma^4/4$, so by Eq. 12 the two leading coefficients are $\frac{h^2}{2}\|\boldsymbol{E}\|_p$ and $\frac{h^2}{2}\left\|\boldsymbol{E} - \sigma^2 \dot{\boldsymbol{s}}\right\|_p$ with $\boldsymbol{E} := \dot{\boldsymbol{v}} - \frac{\sigma^4}{4}\boldsymbol{G}$. Both norms are nonnegative, so comparing them is equivalent to comparing their squares; expanding,

$$\|\boldsymbol{E}\|_p^2 \;<\; \left\|\boldsymbol{E} - \sigma^2 \dot{\boldsymbol{s}}\right\|_p^2 = \|\boldsymbol{E}\|_p^2 - 2\sigma^2 \left\langle \boldsymbol{E},\, \dot{\boldsymbol{s}} \right\rangle_p + \sigma^4 \|\dot{\boldsymbol{s}}\|_p^2 \quad \Longleftrightarrow \quad \left\langle \boldsymbol{E},\, \dot{\boldsymbol{s}} \right\rangle_p \;<\; \frac{\sigma^2}{2}\, \|\dot{\boldsymbol{s}}\|_p^2 .$$

Hence the assumed condition makes the leading coefficient strictly of Langevin step smaller, and since both errors enter at order h^2, $W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{lc}}),\, p_r\big) < W_2\big(\mathrm{law}(\boldsymbol{x}_{\mathrm{sde}}),\, p_r\big)$ for all small h. □

A.3 Auxiliary Lemmas

Assumption 3. *For all $u \in [r, t]$, the velocity $\boldsymbol{v}_u$ is L-Lipschitz.*

Assumption 4. *For all $u \in [r, t]$ and $\boldsymbol{x} \in \mathbb{R}^d$, $\|\partial_u \boldsymbol{v}_u(\boldsymbol{x}) + \nabla \boldsymbol{v}_u(\boldsymbol{x})\, \boldsymbol{v}_u(\boldsymbol{x})\| \le K$; that is, trajectories of Eq. 8 have acceleration bounded by K.*

Lemma 2 (One-Step Euler Error for ODEs)**.** *Under Assumptions 3–4, let $\boldsymbol{x} \sim q_t$ and define the backward Euler step $\hat{\boldsymbol{x}} = \boldsymbol{x} - h\boldsymbol{v}_t(\boldsymbol{x})$, it holds*

$$W_2\big(\mathrm{law}(\hat{\boldsymbol{x}}),\, p_{t-h}\big) \;\le\; e^{Lh}\, W_2(q_t, p_t) + \frac{K}{2} h^2. \tag{20}$$

Proof. Let $\Phi := \Phi_{t \to t-h}$ denote the exact flow map of $\mathrm{d}\boldsymbol{x}_u = \boldsymbol{v}_u(\boldsymbol{x}_u)\,\mathrm{d}u$ from time t to time $t-h$.

We first show that Φ is e^{Lh}-Lipschitz. Let $\boldsymbol{x}_u$ and $\tilde{\boldsymbol{x}}_u$ be two trajectories of the ODE, and define $g(u) := \|\boldsymbol{x}_u - \tilde{\boldsymbol{x}}_u\|^2$. Then,

$$\begin{aligned} g'(u) &= 2\langle \boldsymbol{x}_u - \tilde{\boldsymbol{x}}_u,\, \boldsymbol{v}_u(\boldsymbol{x}_u) - \boldsymbol{v}_u(\tilde{\boldsymbol{x}}_u)\rangle \ge -2\,\|\boldsymbol{x}_u - \tilde{\boldsymbol{x}}_u\| \cdot \|\boldsymbol{v}_u(\boldsymbol{x}_u) - \boldsymbol{v}_u(\tilde{\boldsymbol{x}}_u)\| \\ &\ge -2L\,\|\boldsymbol{x}_u - \tilde{\boldsymbol{x}}_u\|^2 = -2Lg(u), \end{aligned}$$

where Cauchy–Schwarz inequality and the L-Lipschitz of velocity $\boldsymbol{v}_u$ from Assumption 3. Hence $\frac{\mathrm{d}}{\mathrm{d}u}\big(e^{2Lu} g(u)\big) \ge 0$. Therefore, we have $e^{2L(t-h)} g(t-h) \le e^{2Lt} g(t)$. Equivalently,

$$\|\Phi(\boldsymbol{x}_t) - \Phi(\tilde{\boldsymbol{x}}_t)\| \le e^{Lh}\,\|\boldsymbol{x}_t - \tilde{\boldsymbol{x}}_t\|.$$

Next, we compare the exact flow with the Euler step map. By Assumption 4, the map $u \mapsto \boldsymbol{v}_u(\boldsymbol{x}_u)$ is C^1 with derivative $\partial_u \boldsymbol{v}_u(\boldsymbol{x}_u) + \nabla \boldsymbol{v}_u(\boldsymbol{x}_u)\, \boldsymbol{v}_u(\boldsymbol{x}_u)$, bounded by K. Hence $\|\boldsymbol{v}_u(\boldsymbol{x}_u) - \boldsymbol{v}_t(\boldsymbol{x}_t)\| \le K(t-u)$ for $u \in [t-h, t]$, and since $\Phi(\boldsymbol{x}) = \boldsymbol{x}_t - \int_{t-h}^{t} \boldsymbol{v}_u(\boldsymbol{x}_u)\,\mathrm{d}u$,

$$\|\Phi(\boldsymbol{x}) - \hat{\boldsymbol{x}}\| = \left\| \int_{t-h}^{t} \big(\boldsymbol{v}_u(\boldsymbol{x}_u) - \boldsymbol{v}_t(\boldsymbol{x}_t)\big)\,\mathrm{d}u \right\| \le \int_{t-h}^{t} K(t-u)\,\mathrm{d}u = \tfrac{K}{2} h^2.$$

Let $\boldsymbol{y} \sim p_t$, thus we have $\Phi(\boldsymbol{y}) \sim p_{t-h}$. By the triangle inequality for Wasserstein distance, it holds

$$W_2(\mathrm{law}(\hat{\boldsymbol{x}}), p_{t-h}) \le \|\hat{\boldsymbol{x}} - \Phi(\boldsymbol{y})\|_{L^2} \le \|\hat{\boldsymbol{x}} - \Phi(\boldsymbol{x})\|_{L^2} + \|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{y})\|_{L^2} \le \tfrac{K}{2} h^2 + e^{Lh}\, W_2(q_t, p_t).$$

□

Lemma 3 (Stationary Score Moment)**.** *Under Assumption 1, it holds*

$$\mathbb{E}_{p_r}\big[\,\|\nabla f\|^2\,\big] \;\le\; Ld.$$

Proof. Since $\nabla p_r = p_r\, \nabla f$, integration by parts gives

$$\int \|\nabla f\|^2\, p_r = \int \langle \nabla f,\, \nabla p_r \rangle = -\int (\Delta f)\, p_r = \mathbb{E}_{p_r}\big[\,\mathrm{tr}(-\nabla^2 f)\big] \le Ld,$$

where the last inequality uses $-\nabla^2 f \preceq L I_d$ from Assumption 1, and the boundary terms vanish because p_r has sub-Gaussian tails under strong log-concavity while ∇f grows at most linearly. □

Lemma 4. *Let $p, (\mu_h)_{h>0} \in \mathcal{P}_2(\mathbb{R}^d)$, and suppose $\mu_h \to p$ in W_2 as $h \to 0$. Suppose $\boldsymbol{A} = \nabla\phi \in L^2(p)$ is a gradient field satisfying*

$$\int f\, d\mu_h - \int f\, dp = -\frac{h^2}{2} \int \langle \boldsymbol{A}, \nabla f \rangle\, p\, dx + o(h^2),$$

for all smooth test functions f. Then $W_2(\mu_h, p) = \frac{h^2}{2}\, \|\boldsymbol{A}\|_{L^2(p)} + o(h^2)$.

Proof. Write $\|\cdot\| := \|\cdot\|_{L^2(p)}$ and $T := \mathrm{Id} - \frac{h^2}{2}\boldsymbol{A}$.

Upper bound. Coupling p with $T_{\#}p$ through $x \mapsto T(x)$, $W_2(\mu_h, p) \le W_2(\mu_h, T_{\#}p) + W_2(T_{\#}p, p) \le o(h^2) + \frac{h^2}{2}\|\boldsymbol{A}\|$, where the first term is $o(h^2)$ because μ_h and $T_{\#}p$ share the same weak expansion, and the second uses $\int \|x - T(x)\|^2 p = \frac{h^4}{4}\|\boldsymbol{A}\|^2$.

Lower bound. Testing the hypothesis with $f = -\phi$ gives

$$\int (-\phi)\, d\mu_h - \int (-\phi)\, dp = \frac{h^2}{2} \int \|\boldsymbol{A}\|^2\, p\, dx + o(h^2) = \frac{h^2}{2}\|\boldsymbol{A}\|^2 + o(h^2).$$

On the other hand, integrating $\nabla\phi$ along the W_2-geodesic from p to μ_h and applying the first-order (Otto) calculus on $(\mathcal{P}_2, W_2)$ (Otto, 2001; Ambrosio et al., 2005),

$$\int (-\phi)\, d\mu_h - \int (-\phi)\, dp \le W_2(\mu_h, p)\, \big(\|\boldsymbol{A}\| + o(1)\big),$$

where the $o(1)$ holds because $\mu_h \to p$ in W_2, so the geodesic contracts to p. Combining the two displays,

$$\frac{h^2}{2}\|\boldsymbol{A}\|^2 + o(h^2) \le W_2(\mu_h, p)\, \big(\|\boldsymbol{A}\| + o(1)\big),$$

whence $W_2(\mu_h, p) \ge \frac{h^2}{2}\|\boldsymbol{A}\| + o(h^2)$. Together with the upper bound, this proves the claim.

□

B Experimental Setup Details

Except for the noise level η and KL coefficient β, all other hyperparameters are kept the same across methods. The image-generation configuration largely follows Flow-GRPO.

For SD3.5-M-based image generation, we use 48 prompts per epoch, a group size of 24 generations per prompt, LoRA with rank $r = 32$ and $\alpha = 64$, a learning rate of 3×10^{-4}, and a CFG scale of 4.5. For multi-reward optimization, each reward is normalized separately before being combined. We use weights of 0.4 for HPS-v2.1 and 0.6 for CLIP score. The noise level is set to $\eta = 0.7$ for Flow-GRPO, $\eta = 0.9$ for CPS, and $\eta = 0.8$ for LC-GRPO in the OCR setting and $\eta = 0.7$ in all other human-preference settings. The KL coefficient β is set to 0 for all human-preference reward experiments. For verifiable rewards such as OCR, we use $\beta > 0$ to preserve image quality: $\beta = 0.04$ for Flow-GRPO, $\beta = 10^{-4}$ for CPS, and $\beta = 4 \times 10^{-3}$ for LC-GRPO.

For FLUX.1-Dev-based image generation, we use the same prompt and group sizes as above, set $\beta = 0$, and use LoRA with rank $r = 64$ and $\alpha = 128$ and a learning rate of 3×10^{-4}. For multi-reward optimization, rewards are normalized separately and combined with weights 0.7 for HPS-v2.1 and 0.3 for CLIP score. We set noise level $\eta = 0.9$ for Flow-GRPO and CPS; for LC-GRPO, $\eta = 0.8$ for HPS and $\eta = 0.9$ for the combined HPS–CLIP setting.

For video generation, we use 16 prompts per epoch, a group size of 8 generations per prompt, a resolution of 480×480, 53 frames at 8 fps, and a learning rate of 10^{-5}. The noise level is set to $\eta = 0.25$ for DanceGRPO, $\eta = 0.8$ for CPS, and $\eta = 0.6$ for LC-GRPO.

C Additional Results

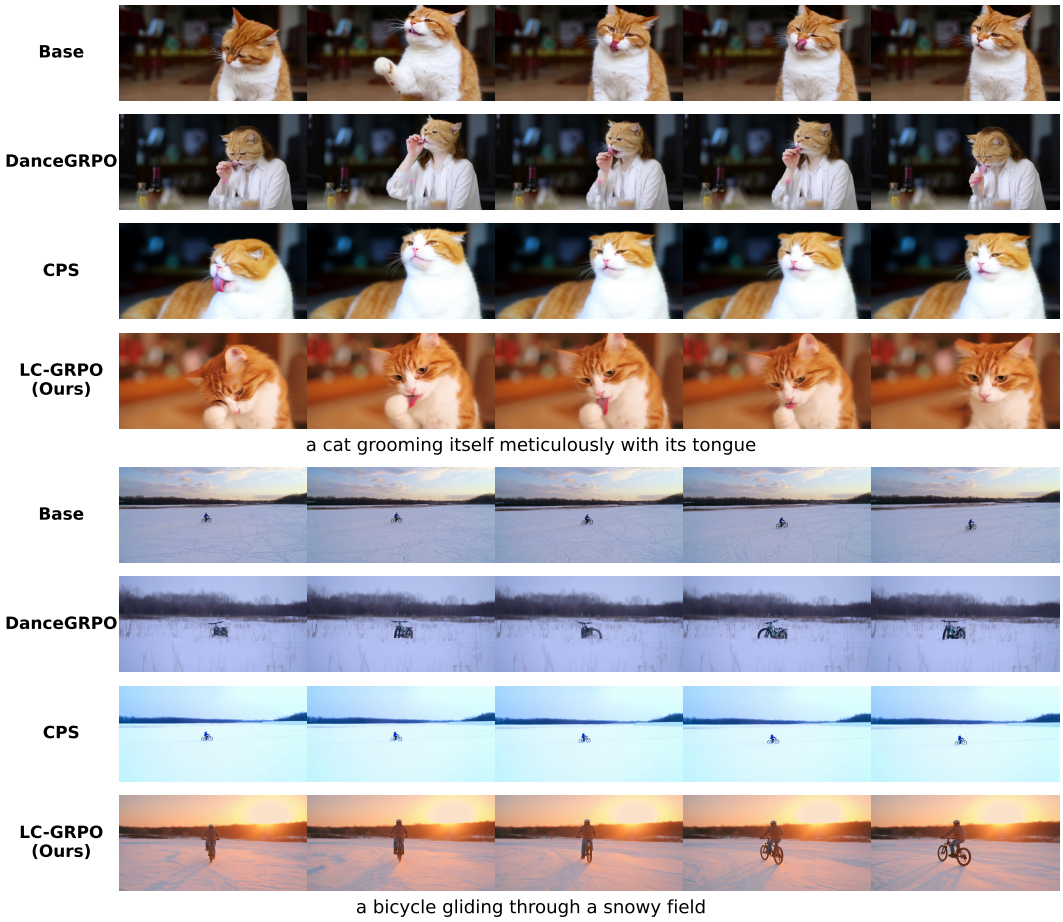


Figure 5: Qualitative comparison based on HunyuanVideo trained with the visual quality metric in VideoAlign.

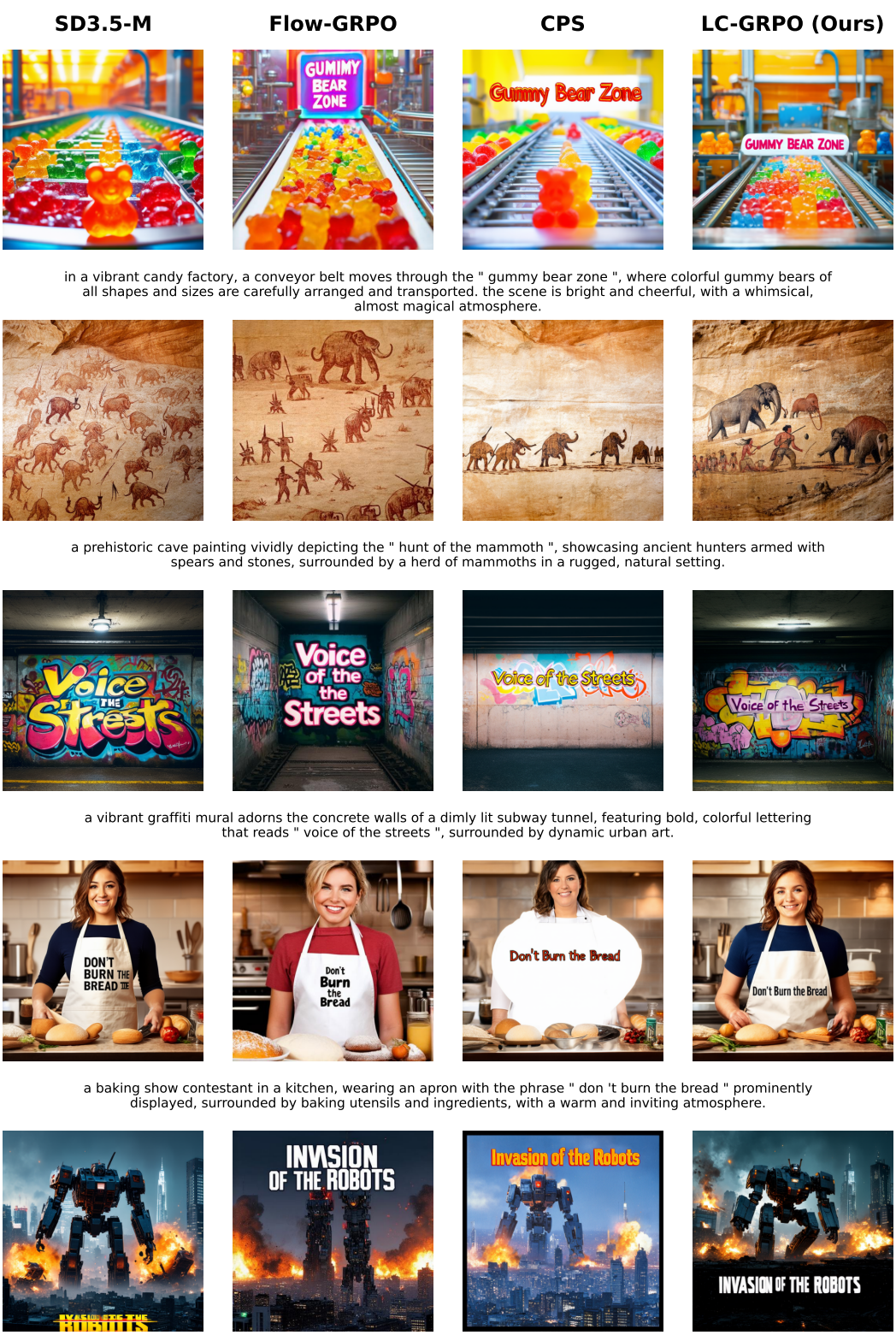


Figure 6: Additional qualitative comparison based on SD3.5-M trained with OCR reward.



Figure 7: Additional qualitative comparison based on FLUX.1-Dev trained with HPS-v2.1 reward, testing on DrawBench.