香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Creative, Controllable & Cognitive Computing Group

KlingAI

NormGuard: Reward-Preserving Norm Constraints in Flow-Matching Reinforcement Learning

Tianlin Pan[*1,2,3] **Lianyu Pang**[*1,2] **Cheng Da**[2] **Huan Yang**[2]
Changqian Yu[2] **Kun Gai**[2] **Wenhan Luo**[†1]

[1]The Hong Kong University of Science and Technology [2]Kuaishou Technology
[3]University of Chinese Academy of Sciences

Abstract

Reinforcement learning (RL) post-training improves the reward alignment of flow-based generators, but often degrades perceptual quality in ways that are not captured by the reward proxy. We identify a simple structural signature of this drift: across three post-training methods (NFT, AWM, DPO), RL fine-tuning inflates the per-step velocity norm $\|v_\theta\|$ by 5% to 15% relative to the reference. A form of norm inflation has been studied in classifier-free guidance (CFG), where rescaling the velocity back to a reference norm at inference time can mitigate the resulting artifacts. However, this inference-time correction does not transfer cleanly to RL: rescaling v_θ to match $\|v_{\text{ref}}\|$ at inference time neither improves reward nor fixes the quality degradation, because the inflation is co-adapted into the model weights. Furthermore, an adjoint sensitivity analysis shows that velocity magnitude rescaling carries no coherent first-order reward signal at the batch level, indicating that suppressing norm inflation is unlikely to remove a consistently reward-carrying component. Since inference-time renormalization fails while norm suppression carries no reward cost, training-time intervention is the appropriate strategy. Together, these findings motivate NormGuard, a hinge penalty that activates only when $\|v_\theta\|$ exceeds $\|v_{\text{ref}}\|$ and composes additively with any velocity-local base loss. Across two base models, three post-training methods, and two reward proxies, NormGuard consistently improves MLLM-judged image quality and forensic realism while preserving reward, with gains that amplify under few-step inference and are not explained by early stopping.

Figure 1 We propose **NormGuard**, a simple norm-budget regularizer that suppresses reward-irrelevant norm inflation during RL training of flow-matching models. NormGuard reduces the over-sharpening, color oversaturation, and unnatural lighting seen in the baseline, producing more photo-realistic results. Prompts see Appendix D.1.

*Equal contribution. † Corresponding author.

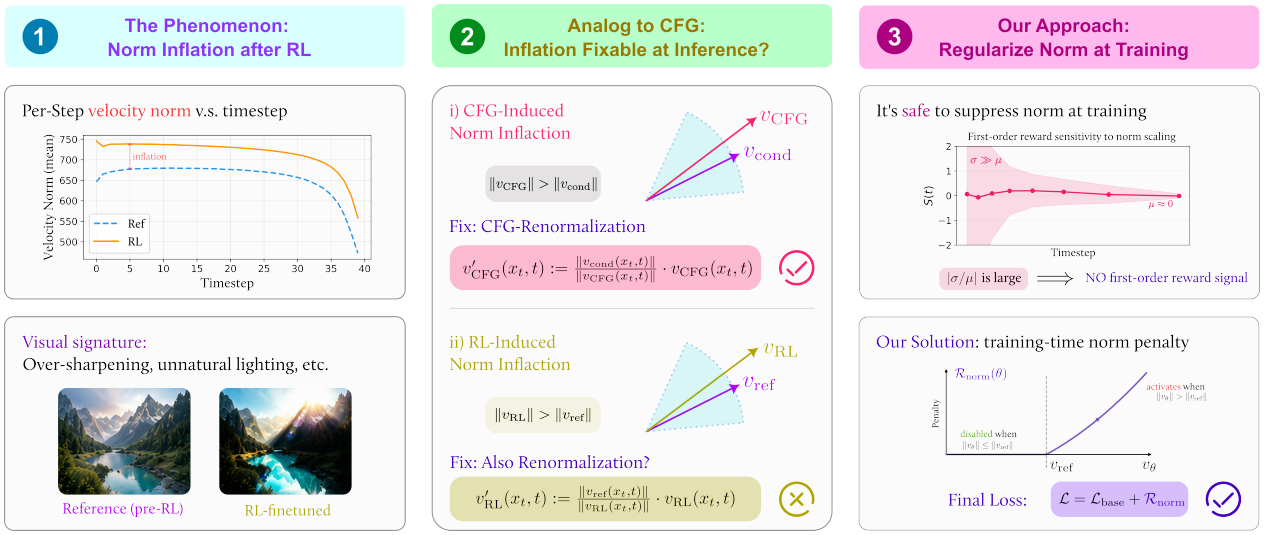

Figure 2 Motivation: RL post-training inflates the per-step velocity norm and produces visual artifacts (**left**). Unlike CFG, inference-time renormalization fails for RL because the inflation is co-adapted into the model weights (**middle**). An adjoint sensitivity analysis confirms that suppressing norm inflation carries no coherent first-order reward signal, motivating our training-time hinge penalty on excess velocity norm (**right**).

1 Introduction

Reinforcement learning (RL) post-training has become a standard tool for aligning flow-based generative models [8, 14, 22, 26] with human preferences [10, 31]. Reward gains, however, are consistently accompanied by *reward over-optimization* [4, 6, 34]: perceptual quality degrades in ways that are not captured by the reward proxy, including over-sharpening, color shift, unnatural lighting, and loss of fine texture. The standard mitigations, such as early stopping and KL regularization, treat the post-training drift as a single aggregate quantity. Let v_θ denote the fine-tuned velocity and v_{ref} the pre-trained reference. KL regularization typically takes the form of an MSE penalty $\|v_\theta - v_{\text{ref}}\|^2$. These methods constrain how much the velocity deviates in total but do not distinguish which component of that deviation is associated with the artifact, leaving no basis for a targeted fix.

Figure 2 outlines the full motivation. Without imposing any structural assumption on the drift, a simple starting point is to inspect the per-step velocity norm $\|v_\theta(x_t,t)\|$ and its ratio to the reference norm $\|v_{\text{ref}}(x_t,t)\|$. Take SD3.5-Medium with PickScore as an example: across NFT, AWM, and DPO, RL post-training consistently inflates this norm by 5% to 15% relative to the reference, uniformly along the denoising trajectory (Figure 3).

A similar form of norm inflation has been well documented in inference-time classifier-free guidance (CFG) [7]. For guidance scale $\omega > 1$, the CFG-modified velocity has a substantially larger magnitude than the conditional velocity ($\|v_{\text{CFG}}\| > \|v_{\text{cond}}\|$), and this norm growth has been shown to drive sampling trajectories to overshoot beyond the learned data distribution [13].The resulting artifacts, such as over-sharpening and unnatural white balance, is also observed in RL fine-tuned models. As shown in Figure 2 (middle), [13] address the CFG case by rescaling the deployed velocity back to a reference norm at inference time while preserving its direction, and this simple fix successfully resolves the CFG artifacts.

However, this CFG-renormalization technique does not apply to the norm inflation induced by RL post-training. We scale the RL-fine-tuned velocity back to match $\|v_{\text{ref}}\|$ at every step. Reward is preserved after renormalization, yet the resulting images exhibit over-sharpening and unnatural lighting (Table 1 and Figure 5). The contrast with CFG is diagnostic: CFG inflation is an explicit inference-time combination of two velocity heads, so rescaling removes it cleanly; RL inflation, by contrast, is trained into the model weights, and rescaling it back at inference distorts the co-adapted dynamics.

Then, we ask whether suppressing the inflation during training would disturb reward gains. The first-order reward change under a multiplicative velocity scaling $v_\theta \mapsto (1+\varepsilon)v_\theta$ is governed by a per-timestep norm-scaling sensitivity $S(t) = v_\theta(x_t,t)^\top a(t)$, where $a(t)$ is the reward adjoint state. On 6,400 samples, $S(t)$ exhibits substantial per-sample magnitude and sign heterogeneity across prompts, while its batch mean remains close to zero; the noise-to-signal ratio

$\sigma/|\mu|$ ranges from 3× to 100×. Thus, at first order and at the batch level, velocity magnitude rescaling does not reveal a coherent reward signal, suggesting that suppressing norm inflation is unlikely to remove reward gains at training.

The two findings point in the same direction. **1)** Inference-time renormalization fails, so the intervention must happen at training time. **2)** Velocity magnitude rescaling does not reveal a coherent first-order reward signal at the batch level, suggesting that training-time norm suppression is unlikely to systematically reduce reward gains . Based on these observations, we propose **NormGuard**, a training-time penalty on the velocity norm that activates only when $\|v_\theta\|$ exceeds $\|v_{\text{ref}}\|$, as shown in Figure 2 (right). The objective sits alongside the existing velocity-local post-training loss and only adds a single scalar hyperparameter.

We validate this regularizer across two base flow models, three post-training methods (NFT, AWM, DPO), and two reward models. Using both MLLM-based image quality assessment and forensic realism detection, we find that the regularizer consistently improves visual quality and, in most settings, improves realism while preserving the majority of the reward gain. Notably, the gains are particularly pronounced at few-step inference, and the improvement cannot be explained by early stopping and is complementary to KL regularization. These results suggest that the aggregate drift penalized by KL can be fruitfully decomposed: norm inflation drives perceptual artifacts while carrying little reward signal, whereas directional realignment accounts for most of the reward gain. Distinguishing these two components offers a finer diagnostic for reward over-optimization.

Our contributions are threefold. **1)** We identify a consistent velocity-norm inflation effect in RL post-training for flow-based generative models, which may cause visual artifacts. **2)** We show that a CFG-style inference-time renormalization does not transfer cleanly to RL-trained models, while batch-level first-order sensitivity does not reveal a stable reward effect aligned with uniform norm scaling. **3)** We introduce **NormGuard**, a simple training-time regularizer that suppresses excess velocity-norm growth and improves perceptual quality across a range of models, post-training methods, and reward functions.

2 Related Work

2.1 RL Post-Training of Generative Models

Text-to-image generation has seen significant improvements in recent years [3, 8, 11, 22]. Inspired by the success of RLHF in language models [21, 23, 24], a growing body of work has adapted these techniques to flow-based generative models, including Diffusion-DPO [27], DDPO [1], Flow-GRPO [15], Dance-GRPO [33], Diffusion-NFT [35], and AWM [32]. All of these methods rely on a reward model such as PickScore [10], GenEval [5], or HPS [16, 31] to provide the training signal.

2.2 Reward Over-optimization

A well-documented pathology of RL fine-tuning is reward over-optimization [4, 17, 25], where models exploit imperfections in the reward proxy at the expense of true quality. In text-to-image generation, optimizing rewards such as PickScore or HPS can produce artifacts including over-sharpening, color bias, and unnatural lighting [2, 34], so proxy reward may improve even as perceptual quality deteriorates.

Several recent works address this issue from different angles. GRPO-Guard [28] stabilizes optimization with ratio normalization and gradient reweighting. RSA-FT [9] attributes reward hacking to sharp reward landscapes and mitigates it by flattening the landscape in image and parameter space. RealGen [34] replaces human-preference rewards with detector-based rewards better aligned with photorealism. RewardDance [30] reduces reward hacking by scaling up the reward model, Pref-GRPO [29] replaces pointwise score maximization with pairwise preference fitting to avoid illusory advantages, and GARDO [6] introduces gated adaptive regularization and diversity-aware optimization to balance anti-hacking and exploration.

3 Norm Inflation: Phenomenon and Diagnostics

This section sharpens the empirical motivation for our method. We first document the norm inflation induced by RL post-training and relate it to a similar phenomenon in classifier-free guidance (Section 3.1). We then examine whether the corresponding inference-time renormalization transfers to RL-trained models, and whether suppressing the inflated norm exhibits a stable first-order reward effect at the batch level (Section 3.2).

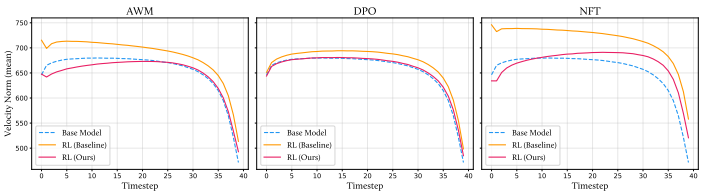

Figure 3 RL-induced norm inflation. Mean velocity norm across denoising timesteps for AWM, DPO, and NFT on SD3.5-Medium. RL post-training (orange) uniformly inflates the norm above the pretrained reference (blue dashed). NormGuard regularizer (magenta) suppresses the excess without sacrificing reward.

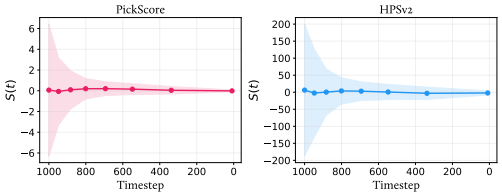

Figure 4 Norm-Scaling sensitivity $S(t)$ across denoising steps (SD3.5-M + 6.4K samples). The noise-to-signal ratio $\sigma/|\mu|$ ranges from 3× to 100×, indicating noisy signal along the norm inflation direction.

3.1 The Phenomenon: Norm Inflation and Its Visual Signature

Empirical observation. We measure the per-step velocity norm $\|v_\theta(x_t, t)\|$ before and after RL post-training on SD3.5-Medium + PickScore, across three RL methods (NFT [35], AWM [32] and DPO [27]). RL fine-tuning consistently shifts the velocity norm distribution upward by 5% to 15%, uniformly along the denoising trajectory and across all configurations (Figure 3).

Connection to a known failure mode: CFG-induced velocity inflation. A similar form of norm inflation is well documented in inference-time classifier-free guidance (CFG) [7]. In CFG, the modified velocity is

$$\hat{v}_{\text{CFG}}(x_t, t) = v_{\text{uncond}}(x_t, t) + \omega\big(v_{\text{cond}}(x_t, t) - v_{\text{uncond}}(x_t, t)\big), \quad \omega > 1, \tag{1}$$

where ω is the guidance scale. STIV [13] report that for $\omega > 1$, $\hat{v}_{\text{CFG}}$ shows substantially larger magnitude than the conditional velocity, $\|\hat{v}_{\text{CFG}}\| > \|v_{\text{cond}}\|$, especially in early integration stages. This norm growth drives sampling trajectories to overshoot beyond the learned data distribution, resulting in artifacts. Their fix, *CFG-Renormalization*, rescales $\hat{v}_{\text{CFG}}$ to $\|v_{\text{cond}}\|$ at inference while preserving its direction:

$$\hat{v}'_{\text{CFG}}(x_t, t) := \frac{\|v_{\text{cond}}(x_t, t)\|}{\|\hat{v}_{\text{CFG}}(x_t, t)\|} \cdot \hat{v}_{\text{CFG}}(x_t, t). \tag{2}$$

As a result, this inference-time correction effectively mitigates the artifacts in the CFG setting. The next subsection asks whether the same intervention can be transferred to inference time for an RL-trained model.

3.2 Can Inference-Time Renormalization Transfer to RL?

Table 1 Inference-time renormalization does not imporve reward. ΔR denotes the change in reward w/ and w/o velocity renormalization.

Metric	Config.	Bas.	RL	+Renorm	ΔR
Pick-Score	SD3.5 (NFT)	0.768	0.869	0.869	+0.000
	SD3.5 (AWM)	0.768	0.869	0.865	−0.004
	SD3.5 (DPO)	0.768	0.849	0.850	+0.001
	FLUX (NFT)	0.802	0.905	0.904	−0.001
HPSv2	SD3.5 (NFT)	0.289	0.308	0.300	−0.008
	FLUX (NFT)	0.271	0.315	0.313	−0.002

Figure 5 Inference-time renormalization introduces additional artifacts.

We present two measurements. Observation 1 shows that the inference-time correction used in CFG does not transfer cleanly to the RL setting. Observation 2 examines whether suppressing velocity magnitude exhibits a stable first-order reward effect at the batch level.

Observation 1: inference-time renormalization does not transfer cleanly to RL-trained models. We test whether the CFG-renorm-style inference fix transfers. Given a reward-fine-tuned v_θ, we apply the per-step inference-time norm scaling

$$v'_\theta(x_t, t) := \frac{\|v_{\text{ref}}(x_t, t)\|}{\|v_\theta(x_t, t)\|} \cdot v_\theta(x_t, t), \tag{3}$$

which preserves the direction of v_θ and rescales its magnitude to the reference norm. Table 1 and Figure 5 report the result. Reward does not improve after renormalization, and the sampled images exhibit over-sharpening and unnatural lighting. This differs from the CFG case, where norm inflation is introduced explicitly at inference time and can be removed by renormalizing the deployed velocity. In the RL setting, the same intervention appears insufficient once the inflated norm has been absorbed by the fine-tuned model, suggesting that a more effective intervention should be applied during training rather than only at inference time.

Observation 2: Velocity magnitude rescaling does not exhibit a stable first-order reward effect at the batch level. Independently of Observation 1, we ask whether suppressing the inflated norm would directly remove useful reward signal. The first-order reward change under a velocity-space perturbation $v_\theta \mapsto v_\theta + \varepsilon\,\eta$ is

$$\frac{\mathrm{d}R}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} = \int_0^1 a(t)^\top \eta(x_t, t)\,\mathrm{d}t, \tag{4}$$

where the adjoint state $a(t)$ satisfies $-\dot{a}(t) = [\nabla_{x_t} v_\theta]^\top a(t)$ with terminal condition $a(1) = \nabla_{x_1} R(x_1)$. Setting $\eta = v_\theta$ (i.e., a multiplicative magnitude scaling $v_\theta \mapsto (1+\varepsilon)v_\theta$) yields the *norm-scaling sensitivity*

$$S(t) := v_\theta(x_t, t)^\top a(t), \tag{5}$$

which measures the marginal first-order reward effect of rescaling velocity magnitude at time t.

We compute $S(t)$ on 6,400 samples from SD3.5-Medium with PickScore and HPSv2. Across all timesteps, S exhibits substantial per-sample magnitude and sign-heterogeneous behavior across prompts, while its batch mean remains close to zero; the noise-to-signal ratio $\sigma/|\mu|$ ranges from 3× to 100× (Figure 4). These measurements do not reveal a stable batch-level first-order reward signal aligned with uniform norm scaling. This suggests that suppressing the inflated norm is unlikely to systematically remove a reward-carrying direction.

Taken together, these observations point to a training-time solution. Direct inference-time renormalization is insufficient in the RL setting, while the sensitivity analysis does not reveal a stable batch-level first-order reward effect tied to uniform norm scaling. This motivates the regularizer introduced next, which suppresses excess velocity-norm growth during post-training.

4 NormGuard

We now present NormGuard. We first identify a local output-space structure shared by the post-training losses considered in this paper (Section 4.1), and then introduce a norm-budget regularizer that operates in the same space (Section 4.2).

4.1 Velocity-Local Post-Training Losses

We focus on post-training objectives whose gradients act through local velocity residuals in the same output space as flow matching. For a noisy state x_t at timestep t under condition c, let $v_\theta(x_t, t, c)$ denote the model velocity and let $\tilde{v}(x_t, t, c)$ denote a local target velocity. The following definition formalizes this shared structure.

Definition 1 (Velocity-local post-training loss)**.** *A post-training objective $\mathcal{L}_{post}$ is* velocity-local *if its parameter gradient admits the form*

$$\nabla_\theta \mathcal{L}_{post} = \mathbb{E}\left[w(x_t, t, c)\,\nabla_\theta \|v_\theta(x_t, t, c) - \tilde{v}(x_t, t, c)\|_2^2\right], \tag{6}$$

where $w(x_t, t, c)$ is a scalar weight that may depend on the sampled state x_t, timestep t, condition c, or reward label. This definition isolates whether the update is driven by per-timestep residuals in the velocity output space.

The three post-training methods used in our experiments fit this template. **NFT** uses reward-weighted flow-matching residuals to a target velocity. **AWM** reduces, under the FM-ELBO surrogate, to an advantage-weighted flow-matching residual with a KL term instantiated as $\|v_\theta - v_{\text{ref}}\|_2^2$. **DPO** injects preference signal through sigmoid-weighted local

deviations from the reference velocity on winner–loser pairs. By contrast, **Flow-GRPO** is a non-example: its gradient passes through a trajectory-level likelihood ratio over reverse transitions, rather than a reweighting of single-step velocity residuals. Full derivations are deferred to Appendix A.

4.2 NormGuard: Norm Regularizer

Our goal is to suppress excess velocity-norm growth during post-training while interfering as little as possible with directional updates. We therefore add a one-sided penalty that activates only when the current velocity norm exceeds the reference-model norm.

Regularizer. For each sampled (x_t, t, c), we define

$$\mathcal{R}_{\text{norm}}(\theta; x_t, t, c) = \lambda \cdot \max\left\{0,\ \frac{\|v_\theta(x_t,t,c)\|_2^2 - \|v_{\text{ref}}(x_t,t,c)\|_2^2}{\|v_{\text{ref}}(x_t,t,c)\|_2^2}\right\}, \tag{7}$$

and optimize the batch-averaged objective

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \mathbb{E}\big[\mathcal{R}_{\text{norm}}(\theta; x_t, t, c)\big], \tag{8}$$

where $\lambda > 0$ controls the regularization strength. The hinge leaves updates unconstrained as long as $\|v_\theta\|_2^2 \le \|v_{\text{ref}}\|_2^2$, and penalizes only the excess norm beyond this reference budget.

Compatibility with velocity-local losses. The penalty acts in the same velocity output space as the base loss. Whenever the hinge is active, its gradient is proportional to $J_\theta^\top v_\theta$, so it modifies the update through the local velocity at the sampled state rather than through a separate parameter-space constraint. As a result, $\mathcal{R}_{\text{norm}}$ composes naturally with any velocity-local base loss, including the NFT, AWM, and DPO objectives.

5 Experiments

Table 2 Training configuration. Hyperparameters for all (model, reward, method) combinations. λ denotes the velocity-norm regularization strength, and r denotes the LoRA rank. Full details are provided in Appendix B.

Model	Reward	Method	LR	LoRA r	λ
SD3.5-M	PickScore	NFT	1×10^{-4}	32	1.0
		DPO	1×10^{-5}	32	1.0
		AWM	3×10^{-4}	32	0.01
	HPSv2	NFT	1×10^{-4}	32	1.0
FLUX.2-4B	PickScore	NFT	1×10^{-4}	64	1.0
		AWM	3×10^{-4}	64	0.01
	HPSv2	NFT	1×10^{-4}	64	1.0

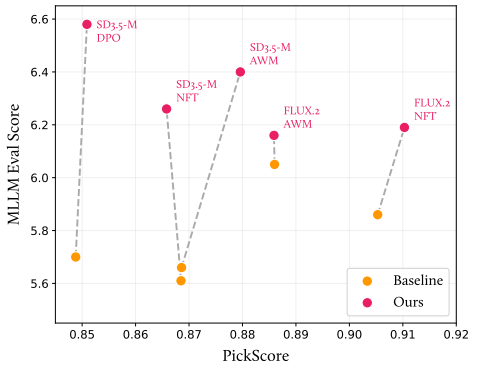

Figure 6 Reward vs. MLLM quality score. Our regularizer improves quality with minimal reward change.

5.1 Evaluation Benchmarks

Experimental Setup. We evaluate our NormGuard regularizer across two base flow models (SD3.5-Medium [3] and FLUX.2-klein-base-4B [11]), three RL post-training methods (NFT [35], AWM [32], DPO [27]), and two reward models (PickScore [10], HPSv2 [31]). We conduct all experiments via Flow-Factory [19] on 8× GPUs. Training prompts are drawn from the PickScore dataset, and test prompts are drawn from HPDv3 [16]. Training details are provided in Appendix B and summarized in Table 2. For fair comparison, all hyperparameters other than λ follow the defaults of Flow-Factory.

MLLM Image Quality Assessment. Following recent work on high-quality RL [2, 34], we use multimodal LLMs as judges to assess image *quality*. We employ both Qwen3.5-35B-A3B [20] and GPT-4.1 [18] to score generated images on

six axes: physical plausibility, texture and material fidelity, edge and boundary coherence, color and tone consistency, semantic coherence, and artifact detection (detailed prompts see Appendix C). For each prompt, both methods generate an image and the judge provides a pairwise preference (Win / Tie / Loss) as well as an absolute score ($1 \sim 10$ scale). For each configuration, the rates are evaluated over 3 seeds with 300 samples, where the average rates are reported.

Forensic Realism Detection. Following RealGen [34], we additionally evaluate with Forensic-Chat [12], an AIGC detector that scores how closely a generated image resembles a real photograph versus a synthetic output. This provides a complementary measure of *realism* independent of aesthetic preference, capturing artifacts such as unnatural lighting and over-sharpening that reward models may overlook.

5.2 Main Results

Qualitative Results. Figure 7 shows representative samples from our method and the unregularized baseline across configurations. The baseline images exhibit a range of quality issues, including texture degradation, edge blurring, and artifact introduction. In contrast, our regularized models produce natural edges, more coherent textures, and fewer artifacts, suggesting that suppressing radial inflation is associated with tangible quality improvements in the generated images.

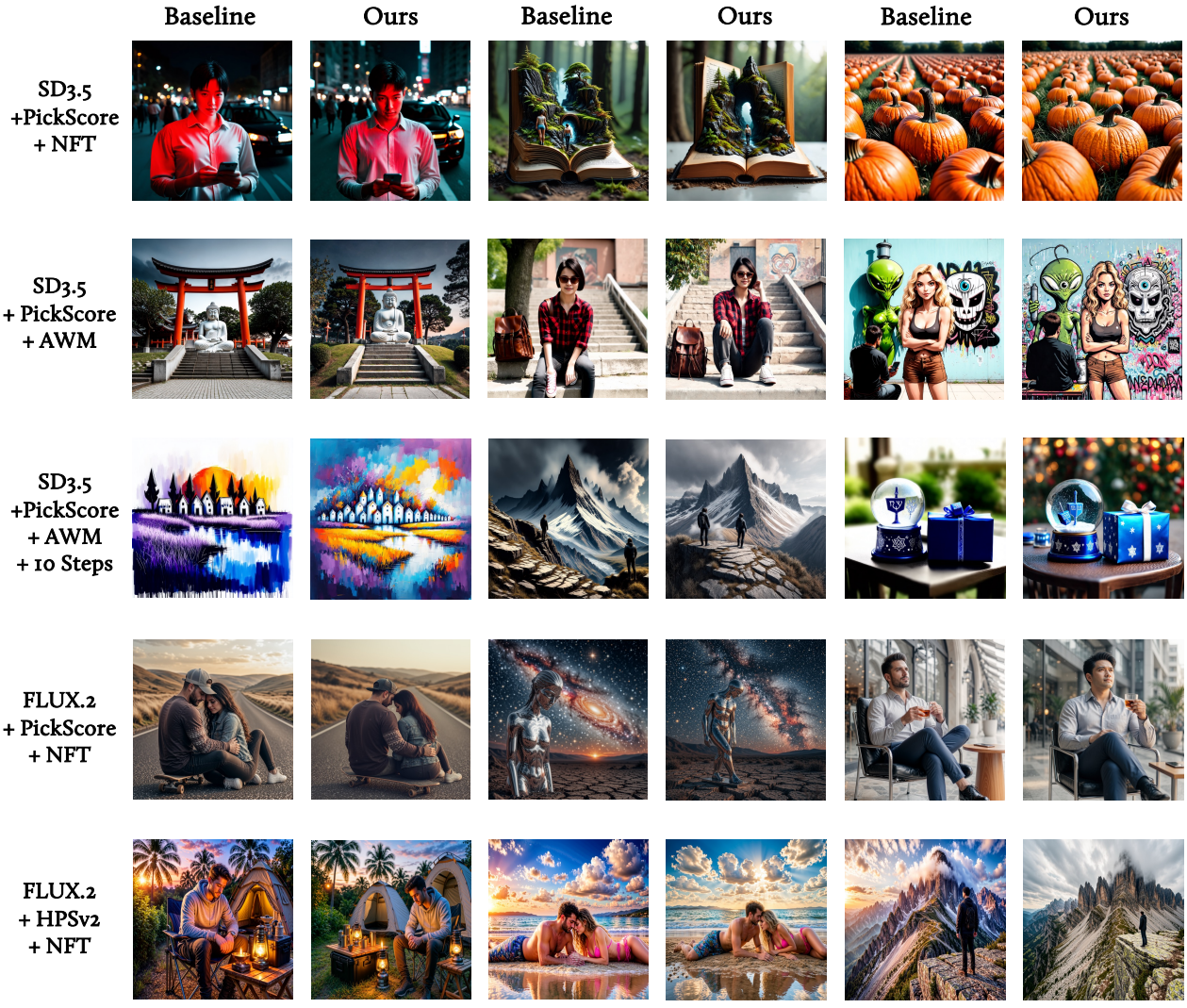

Figure 7 Qualitative results. We compare samples with and without NormGuard regularizer. Prompts see Appendix D.2.

Table 3 Image Quality & Realism. NormGuard improves MLLM-judged quality (Qwen3.5 + GPT-4.1) and forensic realism (RealScore) while largely retaining reward gains.

			Reward↑		Qwen3.5 Win Rate			GPT-4.1 Win Rate			RealScore↑	
Reward	Model	Method	Bas.	Ours	Bas.	**Ours**	Tie	Bas.	**Ours**	Tie	Bas.	**Ours**
PickScore	SD3.5-M	NFT	0.869	0.866	35%	**64%**	1%	30%	**67%**	2%	0.248	**0.310**
	SD3.5-M	DPO	0.849	0.851	28%	**68%**	4%	20%	**47%**	33%	0.468	**0.472**
	SD3.5-M	AWM	0.869	0.880	28%	**72%**	0%	23%	**73%**	4%	0.283	**0.269**
	FLUX.2-4B	NFT	0.905	0.910	38%	**47%**	15%	46%	**51%**	3%	0.239	**0.274**
	FLUX.2-4B	AWM	0.886	0.886	41%	**59%**	0%	38%	**55%**	7%	0.277	**0.320**
HPSv2	SD3.5-M	NFT	0.308	0.309	42%	**58%**	0%	47%	**52%**	1%	0.218	**0.299**
	FLUX.2-4B	NFT	0.315	0.311	36%	**63%**	1%	36%	**60%**	4%	0.240	**0.274**

Table 4 Few-step ablation (FLUX.2-4B, NFT, PickScore). Our regularizer's advantage grows as steps decrease.

	RealScore		MLLM Win%		
Steps	Bas.	**Ours**	Bas.	**Ours**	Tie
28 (Default)	0.239	**0.274**	38%	**47%**	15%
10	0.233	**0.264**	39%	**61%**	0%
4	0.189	**0.221**	40%	**60%**	0%

Table 5 Early stopping ablation (FLUX.2-4B, NFT, PickScore). Our method at iter 200 outperforms all earlier checkpoints.

Baseline Ckpt.	PickScore	RealScore	MLLM
Iter 160	0.8978	0.259	6.08
Iter 180	0.9010	0.252	5.99
Iter 200	0.9053	0.240	6.05
Ours (200)	**0.9103**	**0.274**	**6.16**

Image Quality and Realism. Table 3 reports both MLLM-judged pairwise win rates and forensic realism scores across all configurations. **1)** On image quality, our regularizer is preferred over the unregularized baseline by both Qwen3.5-35B and GPT-4.1 in all seven settings, spanning two base models, three fine-tuning algorithms, and two reward signals. The two judges agree on preference direction in every case, suggesting that the improvements reflect genuine quality gains rather than evaluator bias. **2)** On forensic realism, our method improves RealScore in six out of seven configurations. Under AWM, NormGuard appears to trade a small amount of detector-based realism for a larger gain in MLLM perceptual quality, suggesting that these axes are not perfectly aligned. **3)** On reward, our method largely preserves the gains of the unregularized baseline, with PickScore differences of $-0.003 \sim +0.011$ and HPSv2 differences of $-0.004 \sim +0.001$. This suggests that controlling radial inflation does not appear to materially hinder the directional changes that are more reward-aligned in our measurements, consistent with our diagnosis.

Figure 6 visualizes the relationship between reward and quality across all configurations. The arrows from baseline to ours are nearly vertical: MLLM quality improves substantially while PickScore remains largely unchanged.

5.3 Ablations

Few-step robustness. Table 4 reports quality metrics as inference steps are reduced from 28 to 4. The advantage of our regularizer grows monotonically with fewer steps: the MLLM win-rate gap widens from 9% at 28 steps to 20% at 4 steps, while the baseline's RealScore degrades sharply ($0.239 \rightarrow 0.189$) and ours remains comparatively stable. This is consistent with our diagnosis: fewer steps mean larger step sizes in the ODE integrator, amplifying the effect of velocity-norm inflation on discretization error.

Not explained by early stopping. A potential confound is that our regularizer simply slows training, and the same quality could be obtained by stopping the baseline earlier. Table 5 rules this out: we compare baseline checkpoints at iterations 160, 180, and 200 against our method at iteration 200. Our regularized model achieves higher reward, higher RealScore, and higher MLLM score than *any* earlier baseline checkpoint. The improvement is not accounted for by the early-stopping baselines we test.

Complementarity with KL regularization. Table 6 shows that our regularizer is complementary to KL regularization rather than redundant with it. At every level of β_{KL}, adding our norm regularization ($\lambda_{\text{VN}} = 1.0$) consistently improves RealScore. Our regularizer targets excess radial growth specifically, whereas KL does not selectively suppress this norm inflation and may also constrain directional changes that appear more reward-aligned in our measurements. The two

Table 6 KL complementarity (FLUX.2-4B, NFT, PickScore). Our regularizer (λ) improves RealScore at every KL strength (β_{KL}).

β_{KL}	λ	PickScore	RealScore
0.001	0	0.9034	0.256
0.001	1.0	0.8993	**0.329**
0.0001	0	0.9004	0.270
0.0001	1.0	**0.9110**	**0.278**
0	0	0.8978	0.259
0	1.0	**0.9103**	**0.274**

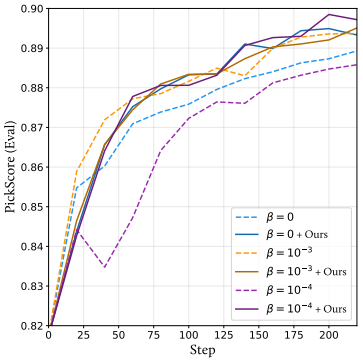

Figure 8 Training curve of PickScore evaluation corresponding to Table 6.

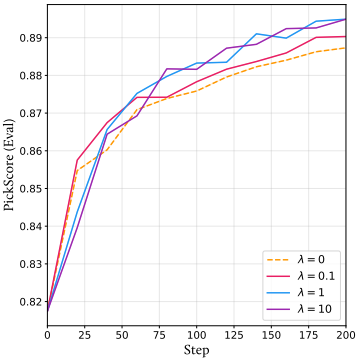

Figure 9 PickScore evaluation with different regularizer strengths (λ).

penalties appear to address distinct failure modes and combine additively in our measurements.

Sensitivity to λ. We vary the regularization strength over $\lambda \in \{0, 0.1, 1, 10\}$ and track PickScore during training. Nonzero regularization consistently outperforms $\lambda = 0$, suggesting that controlling radial growth is important. Among the tested values, intermediate strengths work best: $\lambda = 1$ achieves the highest score at 200 iterations, while $\lambda = 10$ appears slightly over-regularized early in training. Overall, performance is stable across a broad range of nonzero λ, indicating that the method is not sensitive to precise tuning.

6 Conclusion

RL post-training inflates the per-step velocity norm by 5 to 15%, producing the same artifact family as CFG over-amplification. Unlike CFG, however, RL inflation is trained into the weights: inference-time renormalization fails because the network has co-adapted to the inflated norm, while an adjoint sensitivity analysis shows that suppressing velocity magnitude carries no coherent first-order reward signal at the batch level. Because training-time intervention is both necessary and safe, we propose NormGuard, a one-sided hinge penalty on excess velocity norm that composes additively with any velocity-local base loss. Across two base models, three post-training methods, and two reward proxies, NormGuard consistently improves perceptual quality and realism while preserving reward, with gains that amplify under few-step inference and are not explained by early stopping or redundant with KL regularization.

Limitations and future work. The dynamic origin of the radial inflation remains an open question. Our reward-neutrality claim concerns only first-order signal at the batch level; per-sample radial perturbations do affect reward, but their effects average to zero across prompts. Our analysis applies to velocity-local objectives (NFT, AWM, DPO); extending the norm-budget perspective to trajectory-level objectives such as Flow-GRPO, where gradients flow through likelihood ratios over reverse transitions, is left for future investigation.

Acknowledgments

This work was supported by Kuaishou Technology and partially supported by the Beijing Natural Science Foundation under Grant No. QY25188.

References

[1] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301, 2023.

[2] Guo Cheng, Danni Yang, Ziqi Huang, Jianlou Si, Chenyang Si, and Ziwei Liu. Realdpo: Real or not real, that is the preference. arXiv preprint arXiv:2510.14955, 2025.

[3] Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In International Conference on Machine Learning, 2024.

[4] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In International Conference on Machine Learning, pp. 10835–10866. PMLR, 2023.

[5] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. arXiv preprint arXiv:2310.11513, 2023.

[6] Haoran He, Yuxiao Ye, Jie Liu, Jiajun Liang, Zhiyong Wang, Ziyang Yuan, Xintao Wang, Hangyu Mao, Pengfei Wan, and Ling Pan. Gardo: Reinforcing diffusion models without reward hacking. arXiv preprint arXiv:2512.24138, 2025.

[7] Jonathan Ho. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

[9] Kwanyoung Kim and Byeongsu Sim. Reward sharpness-aware fine-tuning for diffusion models. arXiv preprint arXiv:2603.21175, 2026.

[10] 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. arXiv preprint arXiv:2305.01569, 2023.

[11] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

[12] Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Junyan Ye, Ke-Yue Zhang, Yue Zhou, Peng Jin, Bin Li, Taiping Yao, and Shouhong Ding. Seeing before reasoning: A unified framework for generalizable and explainable fake image detection. arXiv preprint arXiv:2509.25502, 2025.

[13] Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, Cha Chen, Yi Fei, Yifan Jiang, Le-Qun Li, Yizhou Sun, Kai-Wei Chang, and Yinfei Yang. Stiv: Scalable text and image conditioned video generation. 2025 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 16249–16259, 2024.

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

[15] 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. arXiv preprint arXiv:2505.05470, 2025.

[16] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025.

[17] Ted Moskovitz, Aaditya Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained rlhf. In International Conference on Learning Representations, volume 2024, pp. 21998–22025, 2024.

[18] OpenAI. Gpt-4.1, 2025. URL https://openai.com/index/gpt-4-1/.

[19] Bowen Ping, Chengyou Jia, Minnan Luo, Hangwei Qian, and Ivor Tsang. Flow-factory: A unified framework for reinforcement learning in flow-matching models. arXiv preprint arXiv:2602.12529, 2026.

[20] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

[21] 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. NeurIPS, 36:53728–53741, 2023.

[22] 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, pp. 10684–10695, 2022.

[23] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

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

[25] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. Advances in Neural Information Processing Systems, 35:9460–9471, 2022.

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In ICLR, 2021.

[27] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq R. Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8228–8238, 2023.

[28] Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, Meng Wang, Pengfei Wan, and Xiaodan Liang. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. arXiv preprint arXiv:2509.25502, 2025.

[29] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. arXiv preprint arXiv:2508.20751, 2025.

[30] Jie Wu, Yu Gao, Zi-Nuo Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, Yangyang Zeng, and Weilin Huang. Rewarddance: Reward scaling in visual generation. arXiv preprint arXiv:2509.08826, 2025.

[31] 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.

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

[33] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818, 2025.

[34] Junyan Ye, Leiqi Zhu, Yuncheng Guo, Dongzhi Jiang, Zilong Huang, Yifan Zhang, Zhiyuan Yan, Haohuan Fu, Conghui He, and Weijia Li. Realgen: Photorealistic text-to-image generation via detector-guided rewards. arXiv preprint arXiv:2512.00473, 2025.

[35] 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.

Appendix

A Theoretical Analysis of Velocity-Local Post-Training Losses

This appendix provides the formal derivations underlying the velocity-local property in Section 4.1. We unify four representative RL post-training methods (Diffusion-DPO, Diffusion-NFT, AWM, Flow-GRPO) under a single notation and identify which of them admit a velocity-residual gradient form.

A.1 Notation

Shared symbols.

- $x_t = (1-t)x_0 + t\,\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$: noisy latent at time t.
- $v_\theta(x_t, t, c)$: trainable velocity predictor.
- $v^{\text{target}} := \epsilon - x_0$: flow-matching target.
- v_{ref}: frozen reference velocity field (the pretrained model).
- $\omega(t)$: a timestep-dependent weighting.
- $J_\theta(x_t, t, c) := \nabla_\theta v_\theta(x_t, t, c)$: per-sample velocity Jacobian.

The flow-matching pretraining objective is

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{x_0,\epsilon,t,c}\left[\omega(t)\,\|v_\theta(x_t,t,c) - v^{\text{target}}\|_2^2\right]. \tag{9}$$

A.2 Method Decompositions in Velocity Coordinates

We now express each method's training signal in the coordinate system above and verify the velocity-local property of Definition 1.

A.2.1 Diffusion-DPO

Starting from the Bradley–Terry preference model and reparameterizing the reward through the diffusion likelihood, the practical Diffusion-DPO loss can be rewritten in a per-timestep comparison form between a winning sample x_0^w and a losing sample x_0^l:

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{t,\epsilon^w,\epsilon^l}\log\sigma\Big(-\beta T\omega(\lambda_t)\big[&\underbrace{\|v_\theta(x_t^w,t) - v_{\text{ref}}(x_t^w,t)\|_2^2}_{\text{winner deviation}}\\ &-\underbrace{\|v_\theta(x_t^l,t) - v_{\text{ref}}(x_t^l,t)\|_2^2}_{\text{loser deviation}}\big]\Big).\end{aligned} \tag{10}$$

The gradient is a sigmoid-weighted sum of two velocity-residual gradients with target $\tilde{v} = v_{\text{ref}}$, hence velocity-local.

A.2.2 Diffusion-NFT

NFT constructs implicit positive and negative policies from the same trainable parameters:

$$v_\theta^+(x_t,c,t) = (1-\beta)\,v^{\text{old}}(x_t,c,t) + \beta\,v_\theta(x_t,c,t), \tag{11}$$

$$v_\theta^-(x_t,c,t) = (1+\beta)\,v^{\text{old}}(x_t,c,t) - \beta\,v_\theta(x_t,c,t). \tag{12}$$

The training objective is a supervised regression on the forward process:

$$\mathcal{L}_{\text{NFT}}(\theta) = \mathbb{E}_{c,x_0,t}\Big[r\,\|v_\theta^+(x_t,c,t) - v^{\text{target}}\|_2^2 + (1-r)\,\|v_\theta^-(x_t,c,t) - v^{\text{target}}\|_2^2\Big], \tag{13}$$

where $r \in \{0,1\}$ is a binary optimality label derived from the reward. The gradient is a reward-weighted FM residual to v^{target}, hence velocity-local.

A.2.3 Advantage-Weighted Matching (AWM)

AWM replaces the per-step reverse-transition likelihood used in DDPO with a sequence-level policy whose likelihood is approximated through the flow-matching ELBO. The GRPO-style objective is

$$\mathcal{J}_{\text{AWM}}(\theta) = \mathbb{E}_{c,\{x_0^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}} \frac{1}{G} \sum_{i=1}^{G} \left(\frac{\hat{\pi}_\theta(x_0^i \mid c)}{\hat{\pi}_{\theta_{\text{old}}}(x_0^i \mid c)} \cdot A_i - \beta \, D_{\text{KL}}(\hat{\pi}_\theta \,\|\, \hat{\pi}_{\text{ref}}) \right). \tag{14}$$

Under the FM-ELBO surrogate, the likelihood ratio is estimated through the difference of two flow-matching losses, and the per-sample gradient simplifies to

$$\nabla_\theta \mathcal{J}_{\text{AWM}} \propto -\nabla_\theta \, \mathbb{E}_t \left[\omega(t) \, \| v_\theta(x_t, t, c) - v^{\text{target}} \|_2^2 \right] \cdot A_i. \tag{15}$$

The KL term is also instantiated as a velocity-space MSE:

$$D_{\text{KL}} \propto \omega(t) \, \| v_\theta(x_t, t, c) - v_{\text{ref}}(x_t, t, c) \|_2^2. \tag{16}$$

AWM makes the structural symmetry between pretraining and RL explicit: at the optimization level, the policy update is the same per-timestep FM residual, modulated only by an advantage weight A_i. Both the policy and KL terms are therefore velocity-local.

A.2.4 Flow-GRPO

Flow-GRPO formulates the denoising process as an MDP with state $s_t = (c, t, x_t)$, action $a_t = x_{t-1}$, and policy equal to the reverse transition $\pi(a_t \mid s_t) = p_\theta(x_{t-1} \mid x_t, c)$. The objective is a clipped policy gradient with importance ratio

$$r_t^i(\theta) = \frac{p_\theta(x_{t-1}^i \mid x_t^i, c)}{p_{\theta_{\text{old}}}(x_{t-1}^i \mid x_t^i, c)}. \tag{17}$$

To enable stochastic exploration, Flow-GRPO converts the deterministic ODE into an SDE and discretizes via Euler–Maruyama, yielding a Gaussian reverse transition

$$p_\theta(x_{t+\Delta t} \mid x_t, c) = \mathcal{N}\left(\bar{x}_{t+\Delta t, \theta}, \; \sigma_t^2 \Delta t \cdot I \right), \tag{18}$$

where the mean $\bar{x}_{t+\Delta t,\theta}$ depends on $v_\theta(x_t, t)$ through the drift term. Because the variance is shared between π_θ and $\pi_{\theta_{\text{old}}}$, the importance ratio reduces to a function of the two Gaussian means:

$$r_t^i(\theta) = \exp\left(\frac{\| x_{t+\Delta t}^i - \bar{x}_{t+\Delta t,\text{old}} \|^2 - \| x_{t+\Delta t}^i - \bar{x}_{t+\Delta t,\theta} \|^2}{2\sigma_t^2 \Delta t} \right). \tag{19}$$

The gradient flows through this trajectory-level likelihood ratio rather than through a per-timestep velocity residual; v_θ enters only indirectly through the Gaussian transition mean. Flow-GRPO therefore violates Definition 1, and we treat Flow-GRPO as out of scope for this work.

A.3 Summary

Table 7 Classification of post-training methods by velocity-locality (Definition 1).

Method	Velocity-local?	Reason
NFT	✓	Gradient is a reward-weighted sum of FM residuals to v^{target}.
AWM	✓	FM-ELBO surrogate reduces the policy update to an advantage-weighted FM residual; KL is also a velocity MSE to v_{ref}.
DPO	✓	Sigmoid-weighted local quadratic deviations $\|v_\theta - v_{\text{ref}}\|_2^2$.
Flow-GRPO	✗	Gradient flows through a trajectory-level policy ratio; v_θ enters only via the Gaussian transition mean.

B Training Details

Table 8 provides the full training hyperparameters for each (model, method, reward) combination. Table 9 details method-specific hyperparameters, and Table 10 lists shared optimization settings. For detailed implementation, see Flow-Factory [19] codes.

Table 8 Main experiments. Full hyperparameters for each (model, method, reward) combination. Here *Train Steps* means the number of sampling steps during rollout, and *GS* means group size.

Model	Method	NormGuard	LR	LoRA r	λ	Train Res.	Train Steps	GS	Batch/dev
SD3.5-M	$\text{NFT}_{\text{PickScore}}$	✗	1×10^{-4}	32	–	512	14	16	8
	$\text{NFT}_{\text{PickScore}}$	✓	1×10^{-4}	32	1.0	512	14	16	8
	$\text{NFT}_{\text{HPSv2}}$	✗	1×10^{-4}	32	–	512	14	16	8
	$\text{NFT}_{\text{HPSv2}}$	✓	1×10^{-4}	32	1.0	512	14	16	8
	$\text{AWM}_{\text{PickScore}}$	✗	3×10^{-4}	32	–	512	14	16	8
	$\text{AWM}_{\text{PickScore}}$	✓	3×10^{-4}	32	0.01	512	14	16	8
	$\text{DPO}_{\text{PickScore}}$	✗	1×10^{-5}	32	–	512	40	4	24
	$\text{DPO}_{\text{PickScore}}$	✓	1×10^{-5}	32	1.0	512	40	4	24
FLUX.2-4B	$\text{NFT}_{\text{PickScore}}$	✗	1×10^{-4}	64	–	384	8	16	2
	$\text{NFT}_{\text{PickScore}}$	✓	1×10^{-4}	64	1.0	384	8	16	2
	$\text{NFT}_{\text{HPSv2}}$	✗	1×10^{-4}	64	–	384	8	16	2
	$\text{NFT}_{\text{HPSv2}}$	✓	1×10^{-4}	64	1.0	384	8	16	2
	$\text{AWM}_{\text{PickScore}}$	✗	3×10^{-4}	64	–	384	8	16	1
	$\text{AWM}_{\text{PickScore}}$	✓	3×10^{-4}	64	0.01	384	8	16	1

Table 9 Method-specific hyperparameters.

Method	Adv. Agg.	Timestep Sampler	$\#t_{\text{train}}$	Time Shift	Adv. Clip	Other
NFT	sum / gdpo	discrete	4 (SD3.5) / 2 (FLUX)	3.0	5.0	EMA decay 0.5
AWM	sum	discrete_wo_init	6 (SD3.5) / 4 (FLUX)	3.0	1.0	ghuber power 0.25; clip 1.0
Online-DPO	–	logit_normal	1	3.0	–	$\beta = 100$; pair best-vs-worst

Table 10 Shared optimization settings.

Setting	Optimizer	Weight Decay	Betas	ϵ	Max Grad Norm
All runs	AdamW	1×10^{-4}	$(0.9, 0.999)$	1×10^{-8}	1.0

C MLLM Evaluation Prompts

Image Realism Judge Prompt

```
You are an expert "Image Realism Judge".

Input:
- image_A: the first image
- image_B: the second image
{prompt_section}

Objective:
Compare the two images and determine which one appears more realistic (i.e., more like a genuine, unmanipulated
     photograph of a real-world scene).

Evaluation Criteria (apply in order of importance):
1. Physical plausibility: lighting, shadows, reflections, perspective, and scale must be consistent with real-world
     physics.
2. Texture & material fidelity: surfaces (skin, fabric, metal, wood, etc.) should exhibit natural micro-detail, noise
     , and variation. Penalize images that look overly smooth, plastic, painted, or unnaturally "oily/greasy"
3. Edge & boundary coherence: object boundaries should be natural; look for blurring halos, jagged masks, unnatural
     cutouts, or excessive / artificial sharpening (e.g., crisp white/black halos around edges, unnaturally hard
     transitions)
4. Color & tone consistency: global and local color grading should be coherent; watch for inconsistent saturation,
     clipped highlights, or artificial color casts.
5. Semantic coherence: all depicted objects, scenes, and interactions must be logically plausible (e.g., no floating
     objects, impossible reflections, anatomical errors).
6. Artifact detection: check for common AI-generated artifacts: repeated patterns, watermarks, noise inconsistencies,
      duplicate elements, ghost limbs, distorted text.

Judgment Protocol:
A) For each image, list specific realism strengths and weaknesses with concrete visual evidence.
B) Score each image on a 1-10 realism scale (10 = indistinguishable from a real photograph).
C) Decide which image is more realistic: "A" or "B". If both are equally realistic, output "tie".
D) Provide a brief justification focusing on the most decisive differences.

Output format:
Return STRICT JSON only (no markdown, no extra text), using this schema:

{
  "image_A_analysis": {
    "realism_score": <int 1-10>,
    "strengths": ["..."],
    "weaknesses": ["..."],
    "key_evidence": ["..."]
  },
  "image_B_analysis": {
    "realism_score": <int 1-10>,
    "strengths": ["..."],
    "weaknesses": ["..."],
    "key_evidence": ["..."]
  },
  "winner": "A" | "B" | "tie",
  "justification": "..."
}

Now compare these two images and judge which is more realistic.
```

D Prompts for Main-Paper Figures

This appendix lists the text prompts used to generate qualitative examples that appear in the main paper. We group prompts by figure for ease of reference. All the prompts are sampled from HPDv3 [31].

D.1 Teaser

Tropical coastline scene.

```
Captured from an elevated perspective, the image showcases a tropical paradise, where lush greenery meets the vibrant
      turquoise ocean. The coastline is marked by a cliff face covered in dense foliage, interspersed with large,
     rocky formations jutting out into the sea.  A pristine white sand beach curves along the coast, gently kissed
     by the foamy waves. Palm trees sway rhythmically along the shoreline, casting shadows over what appears to be a
      small building with a thatched roof. The ocean's color gradient is striking, transitioning from a deep azure
     in the distance to a clear turquoise near the shore, revealing sandy patches beneath the surface.  The scene
     evokes a sense of tranquility and seclusion, ideal for a tropical getaway. The composition highlights the
     natural beauty of the landscape, emphasizing the contrast between the verdant vegetation, the golden sand, and
     the crystal-clear waters.
```

Bird over mountain vista.

```
The image captures a breathtaking mountain landscape under a bright, clear sky. A bird, silhouetted against the light
     , is the focal point of the composition. It appears to be perched on a weathered wooden railing, wings
     partially extended as if ready to take flight. The bird's dark feathers contrast sharply with the radiant
     sunlight behind it, highlighting its form.  In the background, a range of majestic mountains stretches towards
     the horizon. The peaks are rendered in soft shades of blue and gray, fading into the distance as they recede.
     The terrain below the mountains is rocky and sparsely vegetated, suggesting a rugged, high-altitude environment
     .  The wooden railing in the foreground adds a sense of depth and perspective to the image. The warm tones of
     the wood provide a textural contrast to the cool blues and grays of the mountains. The overall effect is a
     serene and awe-inspiring scene that captures the beauty and grandeur of nature.
```

Tree silhouette at sunset.

```
The prompt in English translates to:  This image is a detailed computer-generated artwork depicting a tranquil sunset
      scene. The foreground shows a nearly black horizon that gradually transitions into blue-gray tones as the
     distance increases. At the center of the picture stands a slender tree, with its sparse leaves forming a
     striking silhouette against the vibrant sky. The tree's trunk is thin and long, with sparsely distributed
     leaves on its branches, creating a delicate contrast with the colorful background.
```

D.2 Main Results

SD3.5 + PickScore + NFT.

Person with smartphone, red-lit urban night.

```
The image captures a person holding a smartphone in an urban setting, likely at night. The individual is the primary
     focus, positioned slightly off-center, wearing a light-colored button-down shirt. The red lighting casts a
     distinct glow on their face and parts of their clothing, creating a dramatic contrast.  Their short, dark hair
     is casually styled, framing their face. The person's expression is somewhat neutral, perhaps thoughtful or
     observant. The smartphone is held in their hand, suggesting they may be texting, browsing, or waiting for
     something.  The background is blurred, indicating a shallow depth of field. It consists of city streets with
     blurry pedestrians, vehicles, and building lights. A black car, possibly a taxi, is faintly visible behind the
     person, adding to the urban feel. The bokeh from the lights creates a soft, dreamy effect. The overall tone is
     modern and slightly mysterious due to the lighting and blurred background.
```

3D sculpted book landscape.

```
The image presents an open book, standing upright with its pages forming a vibrant, three-dimensional landscape. The
     pages have been sculpted to resemble a steep, rocky terrain, teeming with lush greenery. Verdant foliage,
     including moss, ferns, and small trees, clings to the rocky surfaces, creating a miniature, intricate ecosystem
     .   A lone figure, appearing small in the grand scale of the scene, is walking through the landscape. This
     individual is set against a backdrop of soft light, which suggest the break of dawn, or a hidden clearing ahead
     . The lighting adds to the dreamy, surreal quality of the image, enhancing the feeling of entering a
     fantastical world.   The book's cover and visible pages have a rustic, weathered look, adding to the sense of
     history and discovery. The scene sits on what seems to be a sandy or earthy surface, grounding the book and its
      internal world in a tangible reality. The overall composition invokes a sense of wonder, inviting the viewer
```

to explore the boundless possibilities contained within the pages of a book.

Pumpkin patch in autumn.

The image shows a field of pumpkins, closely packed together, suggesting a pumpkin patch or harvest scene. The pumpkins vary in size and color, with most appearing in shades of orange. Their textures seem rough, and the shapes are somewhat irregular. The stems are thick and curved, appearing dark green or brown. Some green grass blades are visible between the pumpkins in the foreground. The focus is on the pumpkins in the front, making them appear larger, while the pumpkins in the background are slightly out of focus, creating depth and perspective in the image. The lighting seems soft and natural, enhancing the warm, earthy tones of the pumpkins . The overall impression is one of abundance and autumn.

SD3.5 + PickScore + AWM.

Buddha statue with torii gate, twilight.

The image depicts a tranquil outdoor scene featuring a prominent white Buddha statue seated with legs crossed and hands resting on its knees. The statue, made of marble or similar material, is situated on a small hill surrounded by several trees. In the foreground, there are three concrete steps leading up to a red torii gate, which is a traditional Japanese wooden entrance arch commonly found at Shinto shrines. The torii gate consists of two vertical posts and two horizontal beams, with the top beam being longer than the bottom one. The sky transitions from deep blue near the top to a lighter shade of blue towards the horizon, suggesting that the time of day might be either dusk or dawn. Overall, the composition evokes a serene and culturally rich atmosphere.

Woman on steps, urban mural background.

The image features a woman sitting on a concrete surface, likely a step or platform, with stone steps rising behind her. She is wearing a red and black plaid flannel shirt and dark jeans, paired with white Converse sneakers. Her dark hair is styled short, and she is wearing brown sunglasses. To her left is a brown leather backpack. The background includes a building with a faded mural or painting. A tree partially obscures the building, adding a touch of nature to the urban setting. The overall impression is one of a casual moment, perhaps a pause during exploration or travel. The lighting is bright, suggesting a sunny day, and the composition is balanced, with the woman as the central focus and the surrounding elements adding context and interest.

Graffiti wall, woman and alien character.

This painting depicts a vibrant and detailed graffiti wall that blends elements of cartoon and realism. At the center of the image is a woman with long wavy blonde hair, wearing a black tank top and brown shorts, her arms crossed as she stares directly at the viewer. To her left is a comical green alien figure with a cylindrical head, a large single eye, and a spray paint can in hand, seemingly in the process of creating more graffiti. On the right side of the wall is a complex black-and-white design featuring geometric shapes and mask-like faces, adding to the intricacy of the work. The background is a light blue, dotted with additional graffiti elements. In the foreground, a man dressed in black sits on the sidewalk, engrossed in his phone, seemingly oblivious to the bustling scene behind him. Overall, the composition is rich in color and detail, capturing the essence of urban street art.

SD3.5 + PickScore + AWM + 10 Steps.

Abstract painted landscape, stylized village.

Certainly! Here is a description of the image: The image presents an abstract landscape painting in a vibrant color palette. In the foreground, a stylized field or body of water dominates, rendered in shades of purple and blue with visible brushstrokes. Centered in the midground is a cluster of simplified, white-walled buildings or structures, reminiscent of a small village or town. These buildings are clustered together, creating a compact silhouette against the sky. Above the structures are stylized trees with pointy shapes that create a decorative feel. The background is a blend of warm, muted tones, suggesting a sunset or sunrise. The overall composition is balanced, with the dark field in the foreground grounding the lighter, more colorful elements above. The style is simplified and somewhat naive, emphasizing color and shape over realistic representation.

Solitary figure on rocky mountain precipice.

The image depicts a solitary figure standing on the edge of a rocky precipice, with a towering mountain range as the backdrop. The person, dressed in dark clothing, is positioned on a flat, rocky outcrop, looking out towards the majestic landscape. The mountains are rugged and imposing, with sharp, jagged peaks that reach towards the sky. The upper portion of the highest peak is shrouded in mist or clouds, adding an element of mystery and grandeur to the scene. The slopes of the mountain are steep and rocky, with visible scree and paths, indicating the challenging terrain. The foreground features a rocky landscape with patches of vegetation, providing a textural contrast to the smooth, pale slopes of the mountain. The overall tone of the image is one of awe and solitude, emphasizing the smallness of the individual in the face of nature's immense power and beauty. The muted color palette and the presence of mist suggest a cool, perhaps even slightly cold, climate.

Hanukkah gifts, snow globe with dreidel.

This photo captures a heartwarming scene of Hanukkah gifts on a table. On the left, a charming snow globe featuring a dreidel is the focal point. The globe is clear, encapsulating a blue and white dreidel with Hebrew letters on each side, set against a blurred backdrop. The base of the globe is dark blue, adorned with white Jewish stars and swirls, trimmed with a delicate white border. To the right, a gift is wrapped in shiny blue paper, decorated with a festive pattern of white Hanukkah-themed designs, including stars and menorahs. The wrapping adds a touch of elegance and anticipation to the scene. The table itself is textured, adding depth to the composition. The background is softly blurred, suggesting an outdoor setting, possibly a patio or garden. Overall, the image conveys the warmth and joy of the Hanukkah season, emphasizing the beauty of tradition and the excitement of gift-giving. The lighting is soft and natural, creating a cozy and inviting atmosphere.

FLUX.2 + PickScore + NFT.

Couple on skateboard, sepia road at dusk.

The image captures a warm, intimate moment between a young couple sitting on a skateboard in the middle of a deserted road. The color palette leans towards sepia tones, giving the scene a nostalgic, almost vintage feel. The man wears a dark long-sleeved shirt with a patterned sleeve and a backwards baseball cap. He's seated with his back mostly exposed, one arm wrapped around the woman, as they lean their heads towards each other. The woman is dressed in a denim jacket and dark pants. Her hair is dark and flows down her back as she leans slightly into the man. They sit on a skateboard, casually placed on the asphalt road. The road stretches into the distance, flanked by golden, grassy hills. The sky is a soft, muted yellow, giving the impression of either dawn or dusk. The overall feeling is one of peaceful companionship and shared solitude.

Metallic sculpture, Milky Way backdrop.

The image showcases a mesmerizing juxtaposition of earthly art and celestial wonder. In the foreground, a metallic, human-like sculpture stands on a barren, cracked landscape. The figure appears to be introspective, its head tilted downwards, seemingly lost in contemplation. The reflective surface of the sculpture gleams subtly, picking up the faint light from the night sky. Dominating the background is a breathtaking view of the Milky Way, stretching like a luminous river across the deep blue canvas of the night. The core of our galaxy blazes with countless stars, its swirling patterns creating a sense of depth and immensity. A faint reddish hue hints at distant nebulae. The stars shimmer and scatter across the sky, creating a magical, almost surreal atmosphere. The stark contrast between the grounded sculpture and the boundless expanse of the cosmos invites reflection on humanity's place in the universe.

Man in modern chair, whiskey glass, airy interior.

The image shows a well-dressed man sitting in a stylish, modern chair, exuding an air of relaxed sophistication. He wears a light grey, button-down shirt, neatly tucked into navy-blue trousers, complemented by black leather shoes. His dark hair is neatly styled. The man is holding a glass of amber-colored liquid, possibly whiskey, in his hands, but not drinking, his eyes are gazed upwards. The chair he is seated in has a sleek, black leather cushion and a minimalist metal frame. His legs are crossed, adding to the casual and comfortable posture. Beside him, a wooden table holds a smartphone, suggesting a moment of disconnecting from the digital world. The image has a bright, airy feel. The background is somewhat blurred, with hints of modern architectural elements, further emphasizing the contemporary and sophisticated atmosphere. The soft lighting enhances the scene, creating a sense of calm and introspection.

FLUX.2 + HPSv2 + NFT.

Camper at dusk, lantern and portable stove.

The image captures a serene moment of a camper enjoying the peaceful ambiance of an outdoor setting as dusk settles in. A young man is seated comfortably in a portable camping chair, sporting a casual grey hoodie and dark hiking pants. His gaze is directed downwards, giving off a contemplative vibe. Next to him, a small, portable stove is lit, casting a warm glow. Nearby, a makeshift table setup consisting of a black storage box topped with a wooden surface displays various camping essentials. A lantern provides ample light, illuminating the immediate area. In the background, a tent stands as a simple shelter. The surrounding landscape features lush greenery and palm trees under a twilight sky. The composition is balanced and creates a sense of tranquility and the simple pleasures of camping.

Couple on wet beach, ocean backdrop, golden hour.

The image captures a serene scene of a couple lying on a sandy beach with a backdrop of the ocean and a partly cloudy sky. The couple is positioned in the foreground, their bodies aligned horizontally. The man is wearing patterned swim trunks, while the woman is in a pink bikini. They appear relaxed and content, with their heads close together, creating an intimate atmosphere. The beach appears to be wet, reflecting the soft, diffused light, which adds to the tranquility of the scene. In the background, the ocean stretches out to the horizon, where distant landmasses can be seen. The sky transitions from a deep blue at the top of the frame to a lighter hue near the horizon, with a few fluffy clouds adding depth and texture to the composition. The overall tone of the image is peaceful and romantic.

Solitary figure on rocky precipice, misty peaks.

The image depicts a solitary figure standing on the edge of a rocky precipice, with a towering mountain range as the backdrop. The person, dressed in dark clothing, is positioned on a flat, rocky outcrop, looking out towards the majestic landscape. The mountains are rugged and imposing, with sharp, jagged peaks that reach towards the sky. The upper portion of the highest peak is shrouded in mist or clouds, adding an element of mystery and grandeur to the scene. The slopes of the mountain are steep and rocky, with visible scree and paths, indicating the challenging terrain. The foreground features a rocky landscape with patches of vegetation, providing a textural contrast to the smooth, pale slopes of the mountain. The overall tone of the image is one of awe and solitude, emphasizing the smallness of the individual in the face of nature's immense power and beauty. The muted color palette and the presence of mist suggest a cool, perhaps even slightly cold, climate.