Smart-GRPO: Smartly Sampling Noise for Efficient RL of Flow-Matching Models

Anonymous submission

Abstract

Recent advancements in flow-matching have enabled high-quality text-to-image generation. However, the deterministic nature of flow-matching models makes them poorly suited for reinforcement learning, a key tool for improving image quality and human alignment. Prior work has introduced stochasticity by perturbing latents with random noise, but such perturbations are inefficient and unstable. We propose Smart-GRPO, the first method to optimize noise perturbations for reinforcement learning in flow-matching models. Smart-GRPO employs an iterative search strategy that decodes candidate perturbations, evaluates them with a reward function, and refines the noise distribution toward higher-reward regions. Experiments demonstrate that Smart-GRPO improves both reward optimization and visual quality compared to baseline methods. Our results suggest a practical path toward reinforcement learning in flow-matching frameworks, bridging the gap between efficient training and human-aligned generation.

Introduction

Flow-matching models (Lipman et al. 2023) have emerged as a reliable alternative to diffusion-based generative models (Ho, Jain, and Abbeel 2020), offering deterministic sampling and stable training dynamics. While large-scale pretraining enables high-quality generation, such models often lack mechanisms to ensure consistent and dependable behavior aligned with human intent. Reinforcement learning with human feedback (RLHF) (Ouyang et al. 2022), originally designed to improve the reliability of large language models, has been extended to visual generation tasks, including diffusion-based architectures (Black et al. 2023; Yang et al. 2024).

Adapting RLHF to flow-matching models introduces unique reliability challenges. The inherently deterministic sampling process conflicts with the stochastic exploration required for robust policy optimization. Flow-GRPO (Liu et al. 2025) mitigates this by adding random perturbations to the inputs before denoising, enabling Group Relative Policy Optimization (GRPO) (Shao et al. 2024). However, such noise injection remains unreliable: most randomly sampled perturbations lead to unstable or low-reward outcomes, offering weak learning signals and inconsistent improvement across training. This limitation highlights the need for architectures and training paradigms that explicitly manage uncertainty to ensure dependable performance.

Recent advances have sought to improve the reliability of flow-based RL through better optimization (Li et al. 2025a,b), training stability (Wang et al. 2025; Xue et al. 2025), and generative fidelity (He et al. 2025). Our work instead focuses on the reliability of noise selection itself. We introduce **Smart-GRPO**, a reliability-aware extension of Flow-GRPO that integrates reward-guided noise selection. Smart-GRPO leverages a pretrained reward model to evaluate candidate noise seeds and preferentially sample those expected to yield more reliable, high-quality generations. By adaptively refining the noise distribution over time, Smart-GRPO improves the consistency, efficiency, and robustness of reinforcement learning for flow-based generative models, enhancing both model dependability and alignment with human preferences.

Related Works

Flow-matching models: Let $x_0 \in X_0$ be a sample from a true distribution and let $x_1 \in X_1$ be a sample from a known distribution (e.g. a Gaussian). Flow-matching models (Esser et al. 2024) define a path between the data and the noise as a linear interpolation:

$$x_t = (1-t)x_0 + tx_1, \quad t \in [0,1]. \tag{1}$$

Taking the derivative with respect to t yields the target velocity field:

$$\frac{dx_t}{dt} = x_1 - x_0. \tag{2}$$

The goal is then to learn a parameterized velocity predictor $v_\theta(x_t, t)$ that approximates this ground-truth field. This is achieved by minimizing the following loss (Lipman et al. 2023):

$$\mathbb{L}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\|(x_1 - x_0) - v_\theta(x_t,t)\|^2\right]. \tag{3}$$

Compared to diffusion (Ho, Jain, and Abbeel 2020) models, which learn a score function or directly predict noise, flow-matching instead learns the velocity of the probability flow ODE. This provides a more direct parameterization of the generative process, and in practice can lead to faster and more stable training.

Reinforcement Learning for Flow-matching models: Due to the deterministic nature of flow-matching models, they are not intrinsically designed for reinforcement learning. The probability flow ODE deterministically maps inputs to outputs, leaving little room for the stochastic exploration that reinforcement learning requires. This mismatch makes direct application of standard policy optimization methods ineffective.

Flow-GRPO (Liu et al. 2025) addresses this by converting the deterministic probability flow ODE into an equivalent stochastic differential equation (ODE-to-SDE), which injects randomness while preserving the model's marginal distributions, and by introducing a denoising reduction strategy that reduces the number of denoising steps during training while keeping the full schedule at inference. These modifications enable the incorporation of GRPO into flow-matching models. Empirically, Flow-GRPO achieves substantial gains in compositional image generation, text rendering, and human preference alignment, while maintaining image quality and minimizing reward hacking.

Methods

We introduce Smart-GRPO, an efficient algorithm for fine-tuning flow-matching models with reinforcement learning. Our method improves upon GRPO-style approaches by directly searching over the noise variables that determine the decoded output. Instead of perturbing latents with random noise (as in GRPO), Smart-GRPO searches for noise that maximizes reward in one-shot decoding. We treat the noise distribution as a parameterized search space. Instead of blindly perturbing latents, we iteratively refine a Gaussian noise distribution toward regions of higher reward using a Cross-Entropy Method (CEM)-like update.

Algorithm

Let X_t denote the latent at timestep t, and let $f : X \to \mathbb{R}$ be a scalar reward function. Smart-GRPO proceeds as follows: We first initialize a Gaussian distribution over noise variables, parameterized by mean $\mu = 0$ and standard deviation $\sigma = I$. In each iteration, we sample K candidate noises as:

$$m_i = \mu + \sigma n_i, n_i \sim N(0, I) \tag{4}$$

We then perturb the latent with the noise via the following equation:

$$Z_i = X_t + \sqrt{-dt}\sigma_t m_i \tag{5}$$

where σ_t is the noise scale and dt is the step size. To evaluate the effect of each perturbation, we form a one-step approximation of the decoded image using the predicted velocity v_θ.

$$x_0^{(i)} \approx z_i - t v_\theta(z_i, t) \tag{6}$$

This is intended to provide a rough estimate of the final image without requiring the full reverse process. Note that at earlier timesteps (high noise levels), the one-step approximation produces near-random outputs, making reward evaluation unreliable. Smart-GRPO is therefore most effective at later timesteps where the latent has a stronger correlation with the decoded image

Algorithm 1: Smart-GRPO

Require: Latent image X_t, number of sampled noises K, number of iterations N, saving fraction $P \in [0, 1]$, reward function $f(z) : X \to \mathbb{R}$

Ensure: Optimized latent mean μ or sampled latent $m = \mu + \sigma \cdot n$

1: Initialize $\mu = 0$, $\sigma = I$ of the shape of latent variable
2: **for** $n = 1$ **to** N **do**
3: Sample K random noises $\{n_i\}_{i=1}^{K}$ and compute modified noises $m_i = \mu + \sigma \cdot n_i$
4: Perturb the latent with noise:

$$Z_i = X + \sqrt{-dt} \cdot \sigma_t \cdot m_i$$

5: Decode latents Z_i from m_i and compute reward $R_i = f(Z_i)$
6: Select top $T = \lfloor P \cdot K \rfloor$ noises with highest rewards
7: Update mean and standard deviation:

$$\mu = \frac{1}{T}\sum_{i=1}^{T} m_i, \quad \sigma^2 = \frac{1}{T}\sum_{i=1}^{T}(m_i - \mu)^2$$

8: **end for**
9: **return** μ or $m = \mu + \sigma n$

We then decode the image, and calculate the reward $R_i = f(x_0^{(i)})$. We then select the top $T = \lfloor P \cdot K \rfloor$, where $P \in [0, 1]$ candidates with the highest rewards, using these noises to update the μ and σ used to sample. This process is repeated N times.

This update step shifts the distribution toward higher-reward regions while adaptively controlling its spread, ensuring a balance between exploration and exploitation.

Once this process is complete, either the mean noise μ or a final sample $m = \mu + \sigma n$ is drawn to be used to perturb the latent for training.

Experiments

This section describes the methods used to empirically evaluate whether Smart-GRPO improves performance of flow-matching models. To show this, we utilize two baselines and train our model on two reward functions and analyze results.

Baselines: To compare the performance of our algorithm on fine-tuning flow-matching models, we have two baselines: base Stable-Diffusion 3.5-M (Esser et al. 2024), base Stable-Diffusion 3.5-L (Esser et al. 2024), base FLUX.1-dev (Batifol et al. 2025), and Stable-Diffusion 3.5-M fine-tuned with Flow-GRPO (Liu et al. 2025) without our algorithm.

We selected ImageReward and Aesthetic Score as reward functions because they capture complementary aspects of text-to-image generation. ImageReward is a general-purpose model trained to evaluate prompt-image alignment, visual fidelity, and harmlessness, making it a broad measure of generation quality. In contrast, the Aesthetic Score directly targets visual appeal, reflecting how pleasing an image is to

human perception. Using both rewards allows us to evaluate Smart-GRPO across semantic alignment and visual quality, demonstrating its effectiveness under different alignment objectives.

We choose a prompt dataset of 3000 prompts sampled from datasets provided by Flow-GRPO, generated from GenEval scripts (Ghosh, Hajishirzi, and Schmidt 2023) to train our models on, which was randomly sampled. We split our dataset into a training and evaluation dataset of 2700 training prompts and 300 evaluation prompts.

Analysis

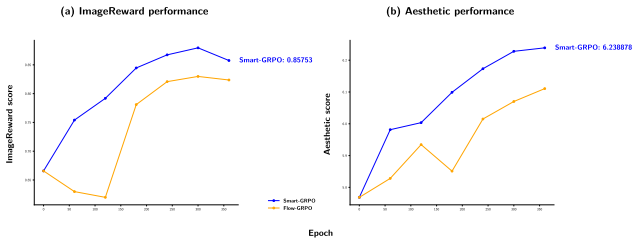

Figure 1: Training results of Smart-GRPO over 360 epochs. Figure **(a)** is trained with ImageReward, and Figure **(b)** is trained using the Aesthetic score

For both rewards we experimented on, our method has both better performance and more stable compared to base Flow-GRPO. Figure 1 shows that Smart-GRPO consistently improves ImageReward scores across training epochs, converging faster and achieving higher final reward than Flow-GRPO. Over our evaluation dataset, Smart-GRPO consistently outperforms the baseline models, as shown in Table **??**

Ablation Study

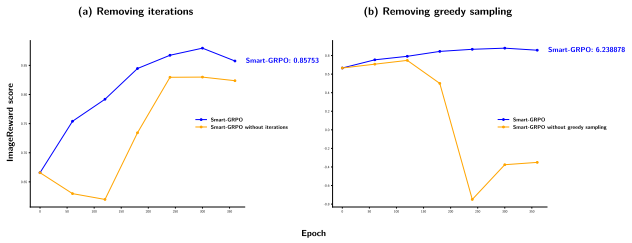

Figure 2: Figure for ablation studies

To better understand the contribution of Smart-GRPO's core mechanisms, we conduct ablation studies on its two central components: **iterative refinement** and **greedy noise selection**. All ablations were performed under the same training hyperparameters and with the ImageReward reward function.

Iterative refinement. Instead of progressively updating the noise distribution, we evaluate a one-shot alternative in which 25 noise samples are drawn, the top 12 are selected, and their mean is used to perturb the latent. We compare this baseline to Smart-GRPO with $N = 5$ iterations and $K = 5$ sampled noises per iteration. As shown in Figure 2a, both approaches achieve similar performance, but Smart-GRPO consistently outperforms the one-shot baseline. This suggests that iterative refinement enables the model to repeatedly concentrate its sampling distribution around higher-quality noise regions, yielding stronger overall performance.

Greedy noise selection. To assess the role of greedy selection, we replace high-reward noise selection with random sampling, keeping $N = 5$ and $K = 5$. As shown in Figure 2b, this variant exhibits highly unstable training: repeatedly updating with low-quality noise often leads to collapse and degraded generations. In contrast, greedy selection stabilizes training by systematically steering updates toward high-reward regions.

Limitations

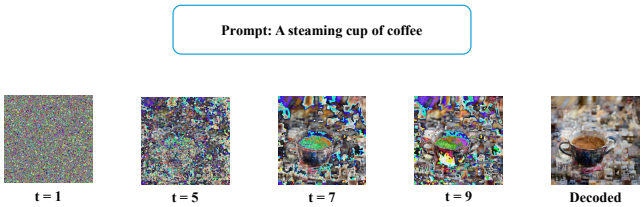

Figure 3: Intermediate approximations from Equation 6 during flow-matching generation of the prompt 'A steaming cup of coffee'. Starting from a noise level of 0.6 and decoded over 10 steps, earlier timesteps yield outputs resembling noise, while later timesteps progressively form low-quality images.

While Smart-GRPO demonstrates promising improvements, it also has several limitations. First, the effectiveness of our approach depends heavily on the choice of reward function. Many reward models are not well calibrated to evaluate poor-quality or highly noisy images, which constrains their ability to guide the noise selection process. For example, our experiments with PickScore (Kirstain et al. 2023) and CLIPScore (Hessel et al. 2021) did not yield statistically significant gains, suggesting that these metrics may be ill-suited for reinforcement learning in high-noise regimes. A further limitation arises from the greedy approximation used in equation 6: as illustrated in Figure 3, this approximation does not hold well at earlier timesteps. Two issues follow: (1) the reward model is not designed to reliably score low-quality images, and (2) the early approximations themselves often fail to capture a meaningful representation of the final image. Exploring alternative metrics to evaluate high-quality noise could potentially be beneficial in improving model generations.

Second, due to computational constraints, we were unable to fully explore larger-scale experiments, longer training schedules, or higher values of K and N, which could further clarify the method's benefits.

Conclusion

We introduced Smart-GRPO, one of the first works in optimizing the noise sampling for fine-tuning flow-matching

Model	ImageReward	Aesthetic Score
Stable Diffusion 3.5M	0.6658	5.769
Stable Diffusion 3.5L	-0.0310	5.602
FLUX.1-dev	0.5121	6.093
SD 3.5M (with Flow-GRPO)	0.8237	6.111
SD 3.5M (with Smart-GRPO)	0.8575	6.238

Table 1: Smart-GRPO model results over evaluation dataset. Dataset consists of 1000 sample prompts generated from GenEval scripts, provided in Flow-GRPO's repository. Results are means over scores of generated images from prompt dataset.

generative models with reinforcement learning. By guiding noise sampling toward higher-reward regions, Smart-GRPO reduces wasted updates and improves both stability and convergence compared to existing methods. Importantly, this is the first approach to explicitly optimize the noise process for reinforcement learning in flow-matching models, offering a novel perspective on noise-aware training. We hope this work paves the way for future research on noise optimization, reward design, and scalable reinforcement learning for generative modeling.

References

Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv e-prints*, arXiv–2506.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

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

He, X.; Fu, S.; Zhao, Y.; Li, W.; Yang, J.; Yin, D.; Rao, F.; and Zhang, B. 2025. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36: 36652–36663.

Li, J.; Cui, Y.; Huang, T.; Ma, Y.; Fan, C.; Yang, M.; and Zhong, Z. 2025a. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*.

Li, Y.; Wang, Y.; Zhu, Y.; Zhao, Z.; Lu, M.; She, Q.; and Zhang, S. 2025b. BranchGRPO: Stable and Efficient GRPO with Structured Branching in Diffusion Models. *arXiv preprint arXiv:2509.06040*.

Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. *The Eleventh International Conference on Learning Representations*.

Liu, J.; Liu, G.; Liang, J.; Li, Y.; Liu, J.; Wang, X.; Wan, P.; Zhang, D.; and Ouyang, W. 2025. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Wang, Y.; Li, Z.; Zang, Y.; Zhou, Y.; Bu, J.; Wang, C.; Lu, Q.; Jin, C.; and Wang, J. 2025. Pref-GRPO: Pairwise Preference Reward-based GRPO for Stable Text-to-Image Reinforcement Learning. *arXiv preprint arXiv:2508.20751*.

Xue, Z.; Wu, J.; Gao, Y.; Kong, F.; Zhu, L.; Chen, M.; Liu, Z.; Liu, W.; Guo, Q.; Huang, W.; et al. 2025. DanceGRPO: Unleashing GRPO on Visual Generation. *arXiv preprint arXiv:2505.07818*.

Yang, K.; Tao, J.; Lyu, J.; Ge, C.; Chen, J.; Shen, W.; Zhu, X.; and Li, X. 2024. 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*, 8941–8951.