# EXPLORENET: LEARNING WHERE TO EXPLORE IN DIFFUSION GRPO

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

## ABSTRACT

Group-relative RL methods such as Flow-GRPO post-train image generators by exploring with isotropic Gaussian noise added at every denoising step. This noise decides which rollouts the model learns from, yet it perturbs every channel and spatial position of the latent equally. Our key insight is that latent elements differ in how much they change the generated image, so exploration should adapt to these differences. We introduce EXPLORENET to learn an adaptive exploration distribution. EXPLORENET is a policy that predicts a noise scale for every latent element from the current latent, the denoising step, and the prompt, before any reward is observed; it is trained on the reward spread of each rollout group and discarded after training, leaving inference unchanged. On Stable Diffusion 3.5 Medium, EXPLORENET improves held-out GenEval2 by $14\%$ over Flow-GRPO, transfers to two independent compositional benchmarks and five preference and image-quality models, and reaches a $67.2\%$ human preference win-rate. We show that in group-relative diffusion RL, exploration is learnable, the shape of the exploration distribution outweighs its magnitude, and rollout quality is more effective than rollout quantity.

## 1 INTRODUCTION

Reinforcement learning unleashes the potential of image generators for alignment (Black et al., 2023; Fan et al., 2023), from instruction following (Li et al., 2026b) and human preference (Xue et al., 2025), to personalization (Dang et al., 2025) and reasoning (Jiang et al., 2025). GRPO is a common online RL algorithm that samples a group of rollouts for each prompt and learns from the differences in their rewards (Shao et al., 2024). Flow-GRPO enables GRPO in flow-matching models by adding isotropic Gaussian noise at every otherwise deterministic denoising step, and is the only source of exploration during denoising (Liu et al., 2025a; Xue et al., 2025). This noise alone decides the rollouts and therefore the learning signal.

One line of work targets the noise, showing that the isotropic noise injection of Flow-GRPO results in excessive or insufficient stochasticity, and proposing to adjust noise levels by denoising steps (Deng et al., 2026; Li et al., 2026a). Within each step, however, the same noise scale still applies to every channel and spatial position of the latent. Our key insight is that *the latent is unevenly sensitive*. Perturbing some latent channels restructures the scene, while perturbing others changes only details (Section 4.5). **Exploration should therefore adapt to how sensitive each latent element is**, yet isotropic noise allocates it equally over fine details and scene structure.

Another line of work directly manipulates rollout groups for a better learning signal by oversampling rollouts, observing their rewards, and training on an informative subset (Ge et al., 2026; Yu et al., 2025). However, post-hoc rollout selection relies on knowing the reward *before* choosing which rollouts to train on, and each update pays for rollouts it discards. Instead, we frame **learning the exploration distribution as a reinforcement learning problem**. The noise drawn at each denoising step is an action, and its reward arrives when the image is scored *after* the last denoising step.

To address both, we propose EXPLORENET, a diffusion post-training method that learns an adaptive noise scale for each latent element, sets it before each denoising step without observing any reward, and keeps every rollout it samples for training. EXPLORENET is a separate transformer network that predicts noise scales from the current latent, the denoising step, and the prompt (Figure 1), and



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.

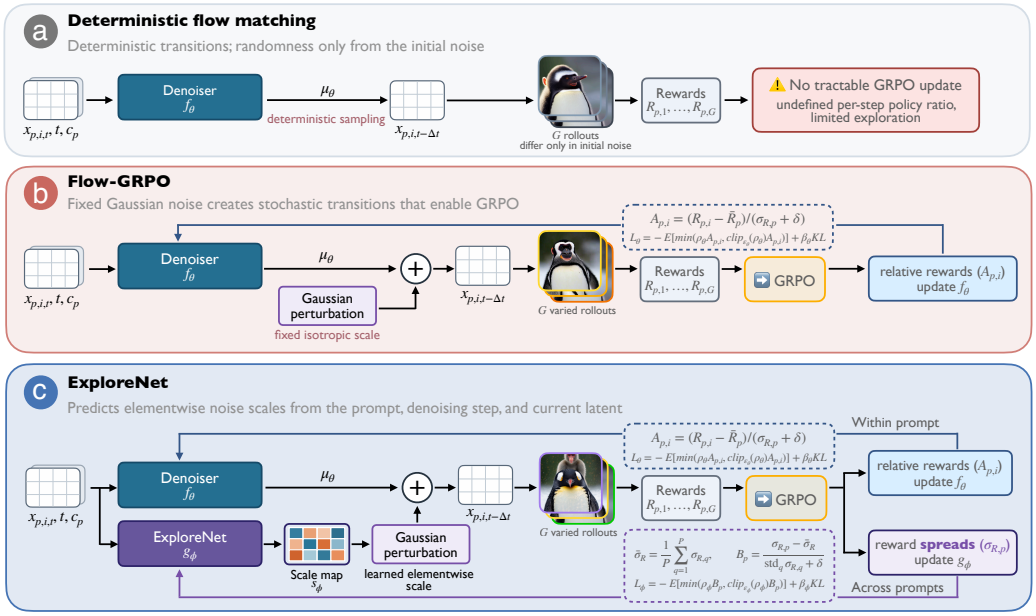


Figure 1: a Deterministic flow-matching sampling adds no randomness beyond the initial noise, which limits exploration and leaves the policy ratio without a tractable transition density. b Flow-GRPO samples rollouts with isotropic Gaussian perturbations and updates the denoiser $f_\theta$ with group-relative advantages $A_{p,i}$. c EXPLORENET $g_\phi$ predicts the scale map $s_\phi$ before each denoising step. The denoiser keeps $\mathcal{L}_\theta$, and EXPLORENET optimizes $\mathcal{L}_\phi$ with the normalized reward spread $B_p$. Inference uses only the trained denoiser.

is trained jointly with the denoiser on the reward spread of each prompt group. It is removed after training, so the trained denoiser keeps its sampler and inference cost.

On Stable Diffusion 3.5 Medium (Esser et al., 2024, SD3.5-M) trained with GenEval2 (Kamath et al., 2025), EXPLORENET outperforms Flow-GRPO across a broad image-generation evaluation suite. Learning the exploration distribution improves held-out GenEval2 by 14%, and the gain transfers to two independent compositional benchmarks and five preference and image-quality models. Human raters prefer EXPLORENET to Flow-GRPO 66.4% of the time. Further ablations show that the gain comes from the *shape* of the exploration distribution. Noise raised uniformly to EXPLORENET's average variance performs worse than Flow-GRPO, and EXPLORENET's scales are largest in the latent channels whose perturbation changes the image most. Rollout quality is also more effective than rollout quantity. With 8 rollouts per prompt, EXPLORENET exceeds Flow-GRPO with 24, and with half as many rollouts, it matches privileged post-hoc selection, which observes every reward. Our contributions are as follows:

- EXPLORENET, which learns a noise scale for every latent element at each denoising step, set from the current latent, the denoising step, and the prompt before any reward is observed, at unchanged inference cost.
- Experiments on SD3.5-M showing that exploration is learnable, with gains over Flow-GRPO across a broad image-generation evaluation suite and in human preference.
- Controlled comparisons showing that the shape of the exploration distribution outweighs its magnitude and that rollout quality is more effective than rollout quantity.

## 2 PRELIMINARIES: GROUP-RELATIVE RL FOR FLOW MODELS

Following DDPO and Flow-GRPO, we treat denoising as a Markov decision process (Black et al., 2023; Liu et al., 2025a). At denoising step $t$, the state consists of the latent $x_t$, the step $t$, and the prompt $c$; the action is the next latent $x_{t-\Delta t}$; and the reward is zero at every step except the last, where the finished image receives $R$. Because a flow-matching transition is deterministic, Flow-GRPO samples the action from a Gaussian centered on the deterministic update, which supplies exploration and a tractable transition density.

As shown in Figure 1b, Each training iteration samples $P$ prompt groups. Prompt group $p$ has text conditioning $c_p$ and $G$ rollouts indexed by $i$. We write $x_{p,i,t}$ for the latent at denoising step $t$ and $R_{p,i}$ for its terminal reward. Let $f_\theta$ denote the denoiser, a diffusion transformer, and $\mu_\theta(x_{p,i,t}, t, c_p)$ the transition mean it induces. With the noise coefficient $a_t = \sigma_t\sqrt{|\Delta t|}$, where $\sigma_t = \lambda\sqrt{t/(1-t)}$ and $\lambda$ is a fixed noise-level hyperparameter, Flow-GRPO samples

$$x_{p,i,t-\Delta t} = \mu_\theta(x_{p,i,t}, t, c_p) + a_t z_{p,i,t}, \qquad z_{p,i,t} \sim \mathcal{N}(0, I). \tag{1}$$

The noise scale at a given denoising step is therefore the same for every prompt, latent, and latent element. Latent elements differ, however, in how much a perturbation changes the generated image (Section 4.5), and a shared scale cannot adapt exploration to these differences.

Flow-GRPO normalizes terminal rewards within each prompt group:

$$\bar{R}_p = \frac{1}{G}\sum_{i=1}^{G} R_{p,i}, \qquad \sigma_{R,p} = \left(\frac{1}{G}\sum_{i=1}^{G}(R_{p,i} - \bar{R}_p)^2\right)^{1/2}, \qquad A_{p,i} = \frac{R_{p,i} - \bar{R}_p}{\sigma_{R,p} + \delta}, \tag{2}$$

where $\sigma_{R,p}$ is the standard deviation of the rewards in group $p$, which we call its *reward spread*, and $\delta$ provides numerical stability. The terminal advantage $A_{p,i}$ is assigned to every transition in rollout $i$. Next, let $p_\theta(x_{t-\Delta t} \mid x_t, t, c) = \mathcal{N}(\mu_\theta,\, a_t^2 I)$ denote the Gaussian transition density of Eq. 1. For data collected by $\theta_{\text{old}}$, the Flow-GRPO denoising ratio and clipped loss are

$$\rho_{\theta,p,i,t} = \frac{p_\theta(x_{p,i,t-\Delta t} \mid x_{p,i,t}, t, c_p)}{p_{\theta_{\text{old}}}(x_{p,i,t-\Delta t} \mid x_{p,i,t}, t, c_p)}, \tag{3}$$

$$\mathcal{L}_\theta = -\mathbb{E}_{p,i,t}\left[\min\left(\rho_{\theta,p,i,t}A_{p,i}, \text{clip}_{\epsilon_\theta}(\rho_{\theta,p,i,t})A_{p,i}\right)\right] + \beta_\theta\, \mathbb{E}_{p,i,t}\left[D_{\text{KL}}(p_\theta \,\|\, p_{\text{ref}})\right], \tag{4}$$

where $\text{clip}_\epsilon(\rho)$ truncates the ratio to $[1-\epsilon, 1+\epsilon]$ and $p_{\text{ref}}$ is the transition density of the pretrained denoiser. EXPLORENET retains this denoising objective and learns the rollout noise.

## 3 EXPLORENET

In the decision process of Section 2, each action $x_{t-\Delta t} = \mu_\theta + a_t z$ has two parts: the mean, set by the denoiser, and the noise. EXPLORENET is a second policy whose action is the noise: before the reward is known, it predicts a scale for each latent element from the current latent, the denoising step, and the prompt, and it is trained afterward from the reward spread of each rollout group. Figure 1c shows how the two policies interact during training.

### 3.1 PREDICTING ELEMENTWISE NOISE SCALES

A separate scale for each latent element lets exploration selectively concentrate on different elements, which the shared scale of Eq. 1 cannot do. Because the noise must be set before each step, each scale can depend only on the inputs the denoiser receives at that step, namely the current latent, the denoising step, and the prompt. So we parameterize EXPLORENET as a second SD3.5-M transformer $g_\phi$, initialized from the pretrained weights and trained through LoRA adapters (Appendix Section A.1). At each denoising step, it maps the current state to one score per latent element:

$$h_{\phi,p,i,t} = g_\phi(x_{p,i,t}, t, c_p), \tag{5}$$

where $h_{\phi,p,i,t,j}$ is the score of latent element $j$, indexed over the channel and spatial dimensions of the latent. We standardize these scores within each latent and convert them to bounded log standard deviations:

$$\widetilde{h}_{\phi,p,i,t,j} = \frac{h_{\phi,p,i,t,j} - \text{mean}_{j'}\, h_{\phi,p,i,t,j'}}{\max(\text{std}_{j'}\, h_{\phi,p,i,t,j'}, \delta_h)}, \tag{6}$$

$$\ell_{\phi,p,i,t,j} = \text{clip}\left(\widetilde{h}_{\phi,p,i,t,j}, \ell_{\min}, \ell_{\max}\right), \qquad s_{\phi,p,i,t,j} = \exp(\ell_{\phi,p,i,t,j}). \tag{7}$$

The two operations constrain how the policy sets the exploration distribution: it controls the *shape*, how noise is allocated across latent elements, and cannot shift the *magnitude*, the overall level of the noise. Standardization removes any shift shared by all scores of a latent, so the policy cannot raise the noise uniformly: a larger scale on some elements comes with a smaller scale on others. Clipping then bounds the scale of any single element.

Let $s_{\phi,p,i,t}$ denote the *scale map*, the vector of scales $s_{\phi,p,i,t,j}$ over all latent elements. EXPLORENET samples

$$\begin{aligned} z_{p,i,t} &\sim \mathcal{N}(0, I), & \eta_{p,i,t} &= s_{\phi,p,i,t} \odot z_{p,i,t}, \\ q_\phi(\eta_{p,i,t} \mid x_{p,i,t}, t, c_p) &= \mathcal{N}\big(0, \mathrm{diag}(s^2_{\phi,p,i,t})\big), & x_{p,i,t-\Delta t} &= \mu_\theta(x_{p,i,t}, t, c_p) + a_t \eta_{p,i,t}. \end{aligned} \tag{8}$$

The denoiser sets the transition mean and EXPLORENET reshapes the noise around it, so rollouts come from an exploration distribution centered on the denoiser's trajectories. The denoising ratio $\rho_\theta$ (Eq. 3) evaluates each sampled transition under the Gaussian of Eq. 1 for both $\theta$ and $\theta_{\mathrm{old}}$, so the scales $s_\phi$ do not enter the denoiser's update. EXPLORENET therefore changes which transitions the denoiser learns from, and the denoiser learns from them exactly the same way as in Flow-GRPO.

### 3.2 TRAINING FROM REWARD SPREAD

A noise action has no reward of its own: the reward arrives only for the finished image, and the denoiser learns from the reward differences among rollouts of the same prompt. A noise choice can therefore be credited only through the group it helps produce, and what a group offers the denoiser is its reward spread. A group whose rewards are equal gives every rollout zero advantage, and Flow-GRPO divides each group's advantages by its reward spread $\sigma_{R,p}$ (Eq. 2), so a group whose rewards are nearly equal yields advantages that amplify small, uninformative differences. To measure the effect of spread directly, we generate 16 rollouts per prompt and train on 8 selected to have the smallest reward spread, selected at random, or selected to have the greatest reward spread, with the rollout pool, update group size, and optimizer fixed. Greatest-spread selection learns fastest and smallest-spread selection slowest (Figure 2), and the ordering holds on eight of nine evaluation metrics (Appendix Table 5). We therefore take the reward spread $\sigma_{R,p}$ of each group (Eq. 2) as the return of EXPLORENET and standardize it across the $P$ prompt groups of an iteration,

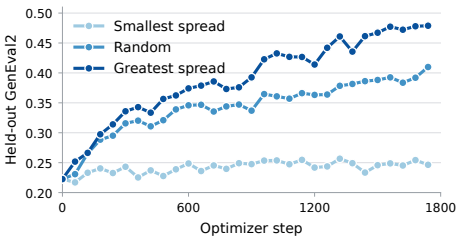


Figure 2: With all else fixed, selecting high-spread groups yields the fastest learning.

$$B_p = \frac{\sigma_{R,p} - \mathrm{mean}_q\, \sigma_{R,q}}{\mathrm{std}_q\, \sigma_{R,q} + \delta}, \tag{9}$$

where $q$ ranges over the $P$ groups. Every sampled noise vector $\eta_{p,i,t}$ in prompt group $p$ receives the same advantage $B_p$, so training raises the probability of the noise vectors sampled in groups whose spread exceeds the average of the iteration. For a rollout sampled with $\phi_{\mathrm{old}}$, the ratio for EXPLORENET is

$$\rho_{\phi,p,i,t} = \frac{q_\phi(\eta_{p,i,t} \mid x_{p,i,t}, t, c_p)}{q_{\phi_{\mathrm{old}}}(\eta_{p,i,t} \mid x_{p,i,t}, t, c_p)}. \tag{10}$$

We optimize the clipped exploration objective

$$\begin{aligned} \mathcal{L}_\phi = & -\mathbb{E}_{p,i,t}\Big[\min\Big(\rho_{\phi,p,i,t} B_p, \mathrm{clip}_{\epsilon_\phi}(\rho_{\phi,p,i,t}) B_p\Big)\Big] \\ & + \beta_\phi\, \mathbb{E}_{p,i,t}\big[D_{\mathrm{KL}}(q_\phi \,\|\, \mathcal{N}(0, I))\big]. \end{aligned} \tag{11}$$

The KL term regularizes the predicted scales toward $\mathcal{N}(0, I)$, the unscaled transition noise of Flow-GRPO in Eq. 1.

### 3.3 JOINT TRAINING AND INFERENCE

Each training iteration samples $P$ prompts and $G$ rollouts per prompt using the current denoiser and EXPLORENET, and the same rewards train both policies at two granularities. The rollout advantage $A_{p,i}$ measures each rollout's reward relative to its prompt group and weights the policy gradient of the denoiser, and the group advantage $B_p$ measures each group's reward spread relative to the other groups in the iteration and weights the policy gradient of EXPLORENET. The two networks minimize $\mathcal{L}_\theta$ and $\mathcal{L}_\phi$ with separate optimizers. At inference, we discard $g_\phi$ and sample the trained denoiser with the standard flow-matching sampler.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

We design experiments to test our central hypothesis: allocating rollout noise across latent elements improves group-relative diffusion RL. Our experiments address three questions: (1) Does EXPLORENET improve the trained denoiser? (2) Why does it work? (3) What does it learn?

**Training.** All training experiments fine-tune SD3.5-M (Esser et al., 2024) on the GenEval2 training split (Kamath et al., 2025) with the Flow-GRPO denoising objective (Liu et al., 2025a). Within each comparison, methods use the same training prompts, reward implementation, denoising schedule, rollout group size, optimizer settings for the denoiser, training duration, and checkpoint cadence. Appendix Section A.1 gives the complete training hyperparameters and computational cost.

**Baselines and controls.** We compare EXPLORENET with Flow-GRPO and two controls:

- **Flow-GRPO** uses the isotropic noise of Eq. 1 with the default noise level $\lambda = 0.7$ and 24 rollouts per prompt group, the same group size as EXPLORENET.
- **Matched isotropic control** replaces EXPLORENET's learned scale map with a single constant, $1.41$, for every latent element, which matches the average variance of EXPLORENET's noise (Appendix A.4). It removes only the allocation to test whether the gain comes from the shape or the magnitude of EXPLORENET's exploration.
- **Privileged post-hoc selection** generates 48 rollouts per prompt, observes every reward, and trains on the 24 with the greatest reward spread. It uses twice the rollouts and reward evaluations of EXPLORENET, so it tests whether learned allocation, set before any reward is observed, can match or exceed selection after sampling.

**Automatic evaluation.** On the held-out GenEval2 prompts, we score the trained models with the GenEval2 metric and with a VLM judge based on GPT-4o (OpenAI et al., 2024). To evaluate transfer performance, we use GenEval (Ghosh et al., 2023), VVRBench-Fast (Li et al., 2026b), PickScore (Kirstain et al., 2023), HPSv2.1 (Wu et al., 2023), HPSv3 (Ma et al., 2025), Aesthetic (Schuhmann, 2022), and ImageReward (Xu et al., 2023). Appendix Section B.1 specifies the prompt sets, metric implementations, generation settings, and other evaluation details.

**Human evaluation.** Three annotators each make 216 pairwise preference judgments (A, B, or Tie) over 72 unique prompts from pairs drawn from SD3.5-M, Flow-GRPO, and EXPLORENET. Win-rate is calculated as each model's total score divided by the number of comparisons it appears in (win receives a score of 1, tie receives a score of $0.5$). The prompts are sampled equally from DrawBench (Saharia et al., 2022), PartiPrompts (Yu et al., 2022), DPG-Bench (Hu et al., 2024), T2I-CompBench (Huang et al., 2023), the OCR prompt set from Flow-GRPO (Liu et al., 2025a), and VVRBench (Li et al., 2026b). Appendix Section B.2 gives the annotation protocol details.

## 4.2 EXPLORENET IMPROVES THE TRAINED DENOISER

We begin by comparing EXPLORENET with Flow-GRPO, with the unmodified SD3.5-M as a pretrained reference. Across three matched training seeds, **EXPLORENET outperforms Flow-GRPO on every metric in the evaluation suite** (Table 1).

**The gains transfer beyond the training reward.** EXPLORENET is trained only on GenEval2, a multi-constraint composition task scored by visual question answering. It also improves other benchmarks unseen during training, including GenEval, a template-based composition task scored by an object detector, and VVRBench-Fast, a spatial and topological layout task scored by a model-free verifier. On VVRBench-Fast, the gain is largest for grounding and spatial relations categories, which check the content and placement of objects (Appendix Table 12). The gains extend to preference and image-quality models. Relative to the pretrained model, Flow-GRPO leaves Aesthetic and HPSv2.1 essentially unchanged ($-0.021$ and $-0.001$), whereas EXPLORENET raises both ($+0.147$ and $+0.006$). EXPLORENET also improves HPSv3, Aesthetic, and ImageReward on each of four prompt suites (Appendix Section C.3). Consistent gains on the full evaluation suite suggest that EXPLORENET improves the generated images themselves rather than exploiting the training reward.

**Human raters prefer EXPLORENET.** Across its comparisons with both baselines, EXPLORENET reaches a $67.2\%$ tie-adjusted preference rate, compared with $45.0\%$ for Flow-GRPO and $37.7\%$

Table 1: Final-checkpoint results, mean and sample standard deviation over three training seeds. Appendix Table 7 reports each seed.

| Metric | Pretrained | Flow-GRPO | ExploreNet | Δ |
|---|---|---|---|---|
| GenEval2 | 0.2375 | $0.4957_{\pm 0.0025}$ | $\mathbf{0.5661}_{\pm \mathbf{0.0318}}$ | +0.0704 |
| VLM Judge | 4.0250 | $4.5083_{\pm 0.0946}$ | $\mathbf{4.7292}_{\pm \mathbf{0.0402}}$ | +0.2208 |
| GenEval | 0.6157 | $0.6955_{\pm 0.0095}$ | $\mathbf{0.7350}_{\pm \mathbf{0.0096}}$ | +0.0395 |
| VVR-Fast | 0.756 | $0.7646_{\pm 0.0062}$ | $\mathbf{0.7902}_{\pm \mathbf{0.0188}}$ | +0.0256 |
| PickScore | 0.8405 | $0.8448_{\pm 0.0019}$ | $\mathbf{0.8512}_{\pm \mathbf{0.0006}}$ | +0.0064 |
| HPSv2.1 | 0.3001 | $0.2993_{\pm 0.0023}$ | $\mathbf{0.3056}_{\pm \mathbf{0.0040}}$ | +0.0063 |
| HPSv3 | 7.6892 | $8.2112_{\pm 0.1784}$ | $\mathbf{8.8043}_{\pm \mathbf{0.1535}}$ | +0.5931 |
| Aesthetic | 5.5165 | $5.4960_{\pm 0.0261}$ | $\mathbf{5.6638}_{\pm \mathbf{0.0596}}$ | +0.1678 |
| ImageReward | 0.9291 | $1.1137_{\pm 0.0263}$ | $\mathbf{1.2774}_{\pm \mathbf{0.0154}}$ | +0.1637 |

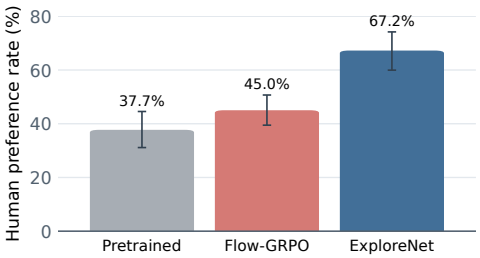


Figure 3: Human preference win-rate across three annotators on 216 comparisons; avg. pairwise agreement is 64.4%.




Figure 4: ExploreNet satisfies the count, color, text, and material constraints that SD3.5-M and Flow-GRPO miss. All three annotators preferred ExploreNet to Flow-GRPO on each prompt.

for pretrained SD3.5-M (Figure 3; Appendix Table 6). In a separate judgment task, annotators find that ExploreNet changes scene content and composition more than Flow-GRPO relative to matched pretrained outputs, at an 86.1% tie-adjusted rate (Appendix B.2). Content and composition match the two most-improved VVRBench-Fast families of grounding and spatial relations. The favorable human preference and the VVRBench category correlation together indicate that the human-perceived large divergence of ExploreNet images is toward a positive direction. Figure 4 shows example generations from the three models over five diverse prompt domains.

**The gain accumulates during training.** Figure 5 shows how the gain develops. On held-out GenEval2, ExploreNet trails Flow-GRPO through step 600, leads at every evaluation after, and its lead widens through the end of training (Appendix Section C.4). The gain therefore accumulates over co-adaptation of $f_\theta$ and $g_\phi$, consistent with ExploreNet supplying more informative rollout groups at every update, and the widening lead indicates that longer training would not close the gap. Each ExploreNet update takes 1.25× as long (Appendix Section A.1), and at equal wall-clock time ExploreNet still leads: its GenEval2 at update 2400 (0.532) exceeds Flow-GRPO's at update 3000 (0.496).

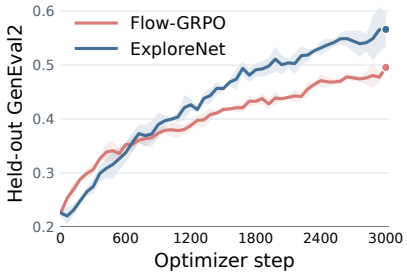


Figure 5: Held-out GenEval2 during training, avg. over three seeds.

**Finding 1.** Exploration in diffusion RL is learnable, and learning it improves the model itself.

Table 2: Controls and reduced-budget settings. The main setting uses 24 rollouts per prompt group and 25 denoising steps; the other settings change one of the two.

| Setting | Method | GenEval2 | VLM Judge | GenEval | VVR-Fast | PickScore | HPSv2.1 | HPSv3 | Aesthetic | ImageReward |
|---|---|---|---|---|---|---|---|---|---|---|
| Main | Flow-GRPO | 0.4982 | 4.5500 | 0.6923 | 0.7624 | 0.8447 | 0.3009 | 8.3142 | 5.5024 | 1.1232 |
| Main | Privileged selection | 0.5551 | **4.7500** | 0.7206 | 0.7724 | 0.8463 | 0.2961 | 8.1998 | 5.4916 | 1.1572 |
| Main | Matched isotropic | 0.4372 | 4.5000 | 0.7351 | 0.7791 | 0.8485 | **0.3034** | 8.4229 | 5.5104 | 1.1884 |
| Main | ExploreNet | **0.5674** | 4.7000 | **0.7452** | **0.8075** | **0.8509** | 0.3017 | **8.6602** | **5.5997** | **1.2597** |
| Group size 8 | Flow-GRPO | 0.4502 | 4.1000 | 0.6771 | 0.7499 | 0.8478 | 0.3006 | 8.2676 | 5.5168 | 1.0916 |
| Group size 8 | ExploreNet | **0.5389** | **4.3500** | **0.7422** | **0.7952** | **0.8499** | **0.3100** | **8.8990** | **5.6228** | **1.2806** |
| 10 steps | Flow-GRPO | 0.4153 | **3.8875** | 0.6597 | 0.7380 | **0.8371** | **0.2804** | 6.9931 | 5.5365 | 0.9521 |
| 10 steps | ExploreNet | **0.4269** | 3.8500 | **0.7109** | **0.7562** | 0.8313 | 0.2771 | **7.1955** | **5.5765** | **1.1133** |

### 4.3 Decomposing Magnitude and Shape of Exploration

ExploreNet explores more than Flow-GRPO, with about twice its average noise variance (Appendix Section A.4), a consequence of exponentiating log-scales that are spread around zero, and shapes its exploration by allocating noise unevenly across latent elements. We ask whether the gain comes from the magnitude or the shape of this exploration. To separate the two, we compare ExploreNet with two alternatives from Section 4.1 that increase exploration without shaping it: the matched isotropic control raises its magnitude, and privileged post-hoc selection raises the number of rollouts.

**Magnitude without shape lowers performance.** The matched isotropic control adds noise with the same average variance as ExploreNet but spreads it uniformly across latent elements. Raising the noise uniformly lowers GenEval2 below Flow-GRPO (0.437 versus 0.498), while allocating the same variance raises it to 0.567 (Table 2). ExploreNet also exceeds the control on eight of nine metrics and has higher GenEval2 at each of the final five checkpoints (Appendix Table 14). Two alternative parameterizations let ExploreNet shift the overall level of its noise, one clipping $g_\phi$'s output without standardization and one applying a learned scale and bias after standardization, and both end training below ExploreNet (Appendix Section A.2).

**ExploreNet matches privileged post-hoc selection with half the rollouts and reward evaluations.** Privileged selection samples 48 rollouts per prompt, observes every reward, and trains on the 24 with the greatest reward spread. ExploreNet instead sets its noise scales before any reward is observed. It matches this baseline on GenEval2 (0.567 versus 0.555) and exceeds it on VVRBench-Fast (0.808 versus 0.772), while generating half as many rollouts and requiring half as many reward evaluations; its added cost is one $g_\phi$ evaluation per denoising step (Appendix Table 3).

**Finding 2.** The shape of the exploration distribution outweighs its magnitude.

### 4.4 ExploreNet Remains Effective with Fewer Rollouts and Steps

Rollouts and denoising steps set the cost of each update, so we repeat the comparison with 8 rollouts per prompt group instead of 24 and with 10 denoising steps instead of 25 (Table 2). With a third of the rollouts, ExploreNet leads on every metric, and its GenEval2 (0.539) exceeds that of Flow-GRPO at both group size 8 (0.450) and the full group size of 24 (0.498). With 10 denoising steps, ExploreNet leads on six of nine metrics, including all three compositional benchmarks.

**Finding 3.** Rollout quality is more important than rollout quantity in diffusion RL.

### 4.5 ExploreNet Concentrates Noise on Visually Influential Channels

Section 4.3 shows that the gain comes from the shape of the exploration distribution, so we analyze the shape of ExploreNet's scale maps at the level of the 16 latent channels, to study how the predicted noise scales relate to each channel's visual effect on the generated image (Appendix D).

**Perturbation setup.** Given a denoiser and a prompt, we generate a reference image with the deterministic ODE sampler. For each latent channel, we resample twice more from the same initial seed while adding (or subtracting) a fixed Gaussian noise to only that channel at every denoising step with

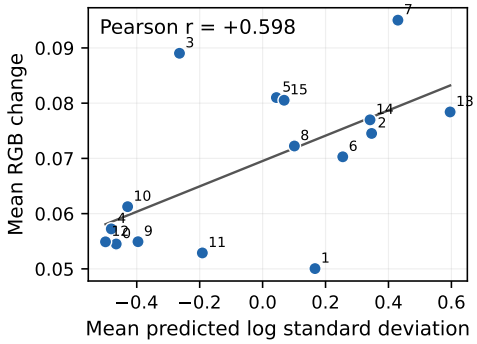


Figure 6: Predicted noise scale tracks channel sensitivity in pixel space. Each point averages one channel over 100 prompts.

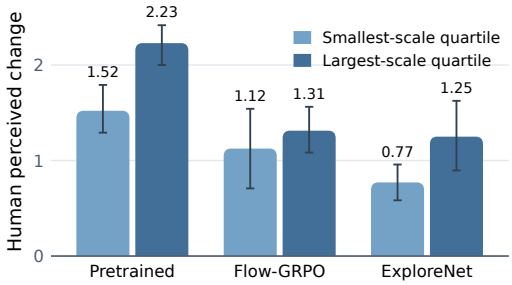


Figure 7: Perturbing EXPLORENET's largest-scale channels produces greater human-rated changes than perturbing its smallest-scale channels, in each perturbed denoiser.

all other channels following the ODE update without added noise. Change is quantified 1) automatically as the mean absolute RGB difference from the reference, and 2) with humans rating how much each perturbed image differs from the reference, from 0 (no visible difference) to 4 (a very large difference). We compare the change with the noise scale EXPLORENET assigns to each channel, obtained by evaluating $g_\phi$ on the latents of the unperturbed reference trajectory and averaging the scale over spatial positions and denoising steps, giving one value per prompt and channel.

**EXPLORENET assigns larger scales to channels that change the image more.** We perturb every channel of the trained EXPLORENET denoiser on 100 GenEval2 training prompts, giving 3,200 perturbed images. Averaged over prompts, per-channel noise scale and RGB change correlate at $r = +0.60$ across the 16 channels (Figure 6; Appendix Table 15). EXPLORENET thus concentrates exploration on the channels that most affect the generated image.

**Human raters see the same ranking in the pretrained model.** For each of the pretrained SD3.5-M, Flow-GRPO, and EXPLORENET denoisers, we perturb every channel on six prompts, giving 576 perturbed images. On SD3.5-M images, the mean rating rises from 1.52 for the channels in the bottom quartile of noise scales to 2.23 for those in the top quartile, a paired increase of $+0.708$ (Figure 7). Pixel-space change on the pretrained denoiser shows the same pattern at the scale of the main analysis, with a rank correlation between each channel's noise scale and RGB change that is positive in 99 of 100 prompts (mean $\rho = +0.42$; Appendix Table 17). Channel sensitivity is therefore a property of the SD3.5-M latent space, and EXPLORENET's noise scales identify the channels that carry it. The same pattern holds for the Flow-GRPO and EXPLORENET denoisers (Figure 7; Appendix Table 16). **The difference is visible in individual images.** Figure 8 shows interventions on the two largest-scale and two smallest-scale channels of each prompt. Perturbing the largest-scale channels produces visibly larger changes.

**Finding 4.** Generative latent spaces are unevenly sensitive, and exploration should follow that structure.

## 5 RELATED WORK

**Reward-based post-training of diffusion and flow models.** Diffusion and flow models have been post-trained with reward-weighted likelihood (Lee et al., 2023), policy gradients (Black et al., 2023; Fan et al., 2023; Zhang et al., 2024), differentiable rewards (Clark et al., 2024; Prabhudesai et al., 2023), preference optimization (Wallace et al., 2024; Yang et al., 2024a;b), and KL- or entropy-regularized stochastic control (Uehara et al., 2024; Domingo-Enrich et al., 2025). GRPO (Shao et al., 2024) was adapted to flow models by Flow-GRPO and DanceGRPO, which convert the deterministic sampler into a stochastic one (Liu et al., 2025a; Xue et al., 2025), and later methods learn the reward adversarially or optimize group preferences directly (Mao et al., 2026; Luo et al., 2026). Subsequent work changes at which denoising steps rollouts branch, are trained, or are rewarded (Li et al., 2026a; Zhang et al., 2026; He et al., 2025; Li et al., 2025; Tong et al., 2026; Savani et al., 2026; Zheng et al., 2026; Yan et al., 2026), and DenseGRPO calibrates a single noise level for each denoising step in a separate pass over sampled trajectories (Deng et al., 2026). These methods change which denoising

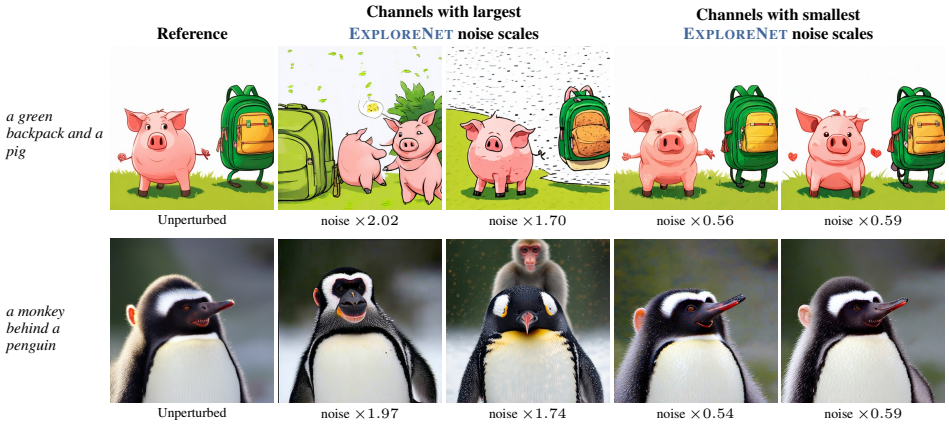


Figure 8: Channels to which ExploreNet assigns more noise change the image more. Perturbing the two channels with the largest ExploreNet noise scales (middle) changes objects, layout, and background. Perturbing the two with the smallest (right) changes only fine details. Labels give the average noise multiplier ExploreNet assigns to each channel.

steps explore, how much each step explores, or how rewards are credited to steps. ExploreNet instead learns the shape of the exploration distribution across latent elements from the current latent, the denoising step, and the prompt, and because its scale map multiplies the per-step noise coefficient $a_t$, it can be combined with any per-step noise schedule.

**Constructing informative rollouts.** Post-hoc selection methods observe the rewards of additional rollouts, with ProGRPO keeping subsets of high reward spread and Smart-GRPO searching over candidate noise (Ge et al., 2026; Yu et al., 2025). DiffExp perturbs classifier-free guidance and prompt-phrase weights, and E$^2$PO optimizes prompt embeddings for semantic spread (Chae et al., 2025; Hu et al., 2026). Initial-noise methods optimize the starting latent of a sample or learn prompt-conditioned initial-noise distributions (Guo et al., 2024; Eyring et al., 2024; Chen et al., 2024; Miao et al., 2025; Eyring et al., 2025). These approaches spend additional rollouts and reward evaluations, perturb the conditioning, or act before denoising begins. ExploreNet sets the noise at every denoising step, before any reward is observed, and is trained jointly with the denoiser.

**Learned noise and exploration.** Diffusion models have learned reverse variances and covariances (Nichol & Dhariwal, 2021; Ou et al., 2025), data-dependent forward processes (Sahoo et al., 2024), and spatially adaptive or reweighted noise (Lavda et al., 2026; Wu et al., 2026), and inference-time samplers add flow-orthogonal perturbations or learn noise schedules (Wu et al., 2025; Lim & Gandelsman, 2026); these methods shape noise to model the data or to improve samples. In flow-model RL, Coefficients-Preserving Sampling, related analyses, and Precise study how much stochasticity the sampler should inject, with the noise fixed by design (Wang & Yu, 2025; Sheng et al., 2025; Zou et al., 2026), and diversity-preserving objectives regularize what the model is trained toward (Liu et al., 2025b; Sorokin et al., 2025; Liu et al., 2026a;b; Chen et al., 2026). In deep RL, parameter noise and state-dependent exploration replace fixed action noise with structured perturbations (Plappert et al., 2018; Fortunato et al., 2018; Raffin et al., 2022). ExploreNet brings state-dependent exploration to diffusion RL, learning the exploration distribution from the reward spread of each rollout group and discarding it at inference.

## 6 Conclusion

Group-relative diffusion RL learns only from the alternatives its sampler produces, yet the exploration distribution that produces them follows a fixed schedule that treats every latent element alike. Latent elements differ in how much they change the generated image, and ExploreNet makes the exploration distribution trainable so that it can adapt to these differences. It predicts a noise scale for every latent element from the current state of generation, learns from the reward spread of each rollout group jointly with the denoiser, and is discarded after training, so the final model is sampled exactly as before. Across our experiments, exploration proves learnable, the shape of the exploration distribution outweighs its magnitude, and rollout quality is more effective than rollout quantity. Exploration is therefore a design axis of diffusion RL in its own right, alongside reward design and per-step noise schedules.

## LIMITATIONS

$q_\phi$ is a diagonal Gaussian; correlated noise across latent elements, such as low-rank or spatially structured covariances, could explore coherent changes such as moving an object as a whole. EXPLORENET learns only the transition noise, and the same framework could learn other sources of rollout variation, such as the guidance scale or the initial noise, jointly with it.

EXPLORENET multiplies the per-step noise coefficient of Flow-GRPO, so it can be combined with per-step noise schedules and step-level credit assignment, which act across denoising steps while EXPLORENET acts across latent elements. It can also be combined with post-hoc selection, so that rollouts shaped by EXPLORENET are further filtered by their observed rewards.

EXPLORENET learns from the reward spread of each rollout group. Other group statistics, such as the spread of individual reward components or of verifier constraint families, could train separate exploration distributions for different objectives.

Video latents add a temporal axis, and EXPLORENET can allocate exploration across frames as well as channels and spatial positions. Group-relative RL for autoregressive and discrete-diffusion generators explores by sampling tokens at a fixed temperature, and a learned per-position temperature could play the role of EXPLORENET's scale map.

EXPLORENET is a second SD3.5-M transformer; a lightweight head on the denoiser's features could predict the same scale maps from the denoiser's own features. Channel sensitivity is a property of the SD3.5-M latent space, so a trained EXPLORENET could be reused across rewards and fine-tuning runs of the same base model. The scale maps identify which latent channels control the content of a generated image, which could guide diverse sampling, localized editing, and the perturbations of best-of-$N$ and other inference-time search methods.

A theoretical account of when the shape of the exploration distribution outweighs its magnitude, relating latent sensitivity to the reward spread of rollout groups, remains open.

## ETHICS STATEMENT

EXPLORENET improves reward-based fine-tuning of text-to-image models and inherits the risks of the models it fine-tunes, including the generation of misleading or harmful images. Because EXPLORENET is discarded after training, the deployed model has the same inference procedure as a Flow-GRPO-trained denoiser. The human evaluation asked annotators to compare generated images and to rate visual changes; the annotators were graduate students who volunteered, and Appendix Section B.2 describes the protocol.

## REPRODUCIBILITY STATEMENT

Section 3 and Appendix Section A.1 specify the objectives, architecture, and all training hyperparameters. Appendix Section B.1 specifies the prompt sets, metric implementations, generation settings, and aggregation, and Appendix Section B.2 specifies the human evaluation protocol. We will release the code upon acceptance.

## AI USE STATEMENT

We used generative AI tools to edit the writing of this manuscript and to assist with analysis code. The authors reviewed all AI-assisted content and take responsibility for the final text, claims, and results.

## REFERENCES


Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. URL https://arxiv.org/abs/2305.13301.

Daewon Chae, June Suk Choi, Jinkyu Kim, and Kimin Lee. DiffExp: Efficient exploration in reward fine-tuning for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://doi.org/10.1609/aaai.v39i15.33723.

Changgu Chen, Libing Yang, Xiaoyan Yang, Lianggangxu Chen, Gaoqi He, Changbo Wang, and Yang Li. FIND: Fine-tuning initial noise distribution with policy optimization for diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. doi: 10.1145/3664647.3681047. URL https://arxiv.org/abs/2407.19453.

Chubin Chen, Sujie Hu, Jiashu Zhu, Meiqi Wu, Jintao Chen, Yanxun Li, Nisha Huang, Chengyu Fang, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Taming preference mode collapse via directional decoupling alignment in diffusion reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_Taming_Preference_Mode_Collapse_via_Directional_Decoupling_Alignment_in_Diffusion_CVPR_2026_paper.html.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17400.

Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8020–8030, 2025. URL https://arxiv.org/abs/2501.06655.

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.20218.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xQBRrtQM8u.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, 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 *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2403.03206.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing one-step text-to-image models through reward-based noise optimization. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.04312.

Luca Eyring, Shyamgopal Karthik, Alexey Dosovitskiy, Nataniel Ruiz, and Zeynep Akata. Noise hypernetworks: Amortizing test-time compute in diffusion models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2508.09968.

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. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.16381.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alex Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *The Sixth International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rywHCPkAW.

Shiran Ge, Chenyi Huang, Yuang Ai, Qihang Fan, Huaibo Huang, and Ran He. Expand and prune: Maximizing trajectory diversity for effective GRPO in generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ge_Expand_and_Prune_Maximizing_Trajectory_Diversity_for_Effective_GRPO_in_CVPR_2026_paper.html.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2310.11513.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. InitNO: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guo_InitNO_Boosting_Text-to-Image_Diffusion_Models_via_Initial_Noise_Optimization_CVPR_2024_paper.html.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. TempFlow-GRPO: When timing matters for GRPO in flow models. *arXiv preprint arXiv:2508.04324*, 2025. URL https://arxiv.org/abs/2508.04324.

Sujie Hu, Chubin Chen, Jiashu Zhu, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Embedding-perturbed exploration preference optimization for flow models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.15803.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. URL https://arxiv.org/abs/2403.05135.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2307.06350v2.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2I-R1: Reinforcing image generation with collaborative semantic-level and token-level CoT. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.00703.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025. URL https://arxiv.org/abs/2512.16853.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.01569.

Frantzeska Lavda, Maciej Falkiewicz, Van Khoa Nguyen, and Alexandros Kalousis. Spatially adaptive noise injection. *arXiv preprint arXiv:2609.18466*, 2026. URL https://arxiv.org/abs/2609.18466.

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. URL https://arxiv.org/abs/2302.12192.

Junzhe Li, Yutao Cui, Tao Huang, Weijie Kong, Yiming Cheng, Chuxuan Zeng, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, and Liefeng Bo. MixGRPO: Unlocking flow-based GRPO efficiency with mixed ODE-SDE. In *European Conference on Computer Vision*, 2026a. URL https://arxiv.org/abs/2507.21802.

Shuyue Stella Li, Xiaochuang Han, Yulia Tsvetkov, and Luke Zettlemoyer. VVRBench: Verifiable visual rewards transfer from synthetic scenes to natural prompts. Hugging Face dataset, 2026b. URL https://huggingface.co/datasets/stellalisy/VVRBench.

Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. BranchGRPO: Stable and efficient GRPO with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*, 2025. URL https://arxiv.org/abs/2509.06040.

Arisrei Lim and Yossi Gandelsman. Learning sampling parameters for diffusion models. *arXiv preprint arXiv:2607.23488*, 2026. URL https://arxiv.org/abs/2607.23488.

Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. DiverseGRPO: Mitigating mode collapse in image generation via diversity-aware GRPO. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_DiverseGRPO_Mitigating_Mode_Collapse_in_Image_Generation_via_Diversity-Aware_GRPO_CVPR_2026_paper.html.

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. In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2505.05470.

Jinmei Liu, Haoru Li, Zhenhong Sun, Chaofeng Chen, Yatao Bian, Bo Wang, Daoyi Dong, Chunlin Chen, and Zhi Wang. Beyond the dirac delta: Mitigating diversity collapse in reinforcement fine-tuning for versatile image generation. *arXiv preprint arXiv:2601.12401*, 2026b. URL https://arxiv.org/abs/2601.12401.

Zhen Liu, Tim Z. Xiao, Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. Efficient diversity-preserving diffusion alignment via gradient-informed GFlowNets. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2412.07775.

Yihong Luo, Tianyang Hu, and Jing Tang. Reinforcing diffusion models by direct group preference optimization. In *International Conference on Learning Representations*, volume 2026, pp. 21438–21454, 2026. URL https://arxiv.org/abs/2510.08425.

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*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_HPSv3_Towards_Wide-Spectrum_Human_Preference_Score_ICCV_2025_paper.html.

Weijia Mao, Hao Chen, Zhenheng Yang, and Mike Zheng Shou. The image as its own reward: Reinforcement learning with adversarial reward for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5999–6009, 2026. URL https://arxiv.org/abs/2511.20256.

Yanting Miao, William Loh, Pascal Poupart, and Suraj Kothawade. A minimalist method for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2506.12036*, 2025. URL https://arxiv.org/abs/2506.12036.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia,

Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Zijing Ou, Mingtian Zhang, Andi Zhang, Tim Z. Xiao, Yingzhen Li, and David Barber. Improving probabilistic diffusion models with optimal diagonal covariance matching. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PxTIG12RRHS.

Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y. Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. In *The Sixth International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByBAl2eAZ.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. URL https://arxiv.org/abs/2310.03739.

Antonin Raffin, Jens Kober, and Freek Stulp. Smooth exploration for robotic reinforcement learning. In *Proceedings of the 5th Conference on Robot Learning*, 2022. URL https://proceedings.mlr.press/v164/raffin22a.html.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2205.11487.

Subham Sekhar Sahoo, Aaron Gokaslan, Christopher De Sa, and Volodymyr Kuleshov. Diffusion models with learned adaptive noise. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2312.13236.

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*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Savani_Stepwise_Credit_Assignment_for_GRPO_on_Flow-Matching_Models_CVPR_2026_paper.html.

Christoph Schuhmann. LAION-aesthetics predictor (improved-aesthetic-predictor). GitHub repository, 2022. URL https://github.com/christophschuhmann/improved-aesthetic-predictor.

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

Jiayuan Sheng, Hanyang Zhao, Haoxian Chen, David D. Yao, and Wenpin Tang. Understanding sampler stochasticity in training diffusion models for RLHF. *arXiv preprint arXiv:2510.10767*, 2025. URL https://arxiv.org/abs/2510.10767.

Dmitrii Sorokin, Maksim Nakhodnov, Andrey Kuznetsov, and Aibek Alanov. ImageReFL: Balancing quality and diversity in human-aligned diffusion models. *arXiv preprint arXiv:2505.22569*, 2025. URL https://arxiv.org/abs/2505.22569.

Yunze Tong, Mushui Liu, Canyu Zhao, Didi Zhu, Wanggui He, Shiyi Zhang, Hongwei Zhang, Peng Zhang, Jinlong Liu, and Hao Jiang. Alleviating sparse rewards by modeling step-wise and long-term sampling effects in flow-based GRPO. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.06422.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M. Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024. URL https://arxiv.org/abs/2402.15194.

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*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wallace_Diffusion_Model_Alignment_Using_Direct_Preference_Optimization_CVPR_2024_paper.html.

Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. *arXiv preprint arXiv:2509.05952*, 2025. URL https://arxiv.org/abs/2509.05952.

Fang Wu, Haokai Zhao, Da Xing, Hanqun Cao, Tinson Xu, Yanchao Li, Xiangru Tang, Zehong Wang, Aaron Tu, Kuan Pang, Hanchen Wang, Hongbin Lin, Zeqi Zhou, Yinxi Li, Peng Xia, Li Erran Li, Molei Tao, Jure Leskovec, Aditya Joshi, and Yejin Choi. NoiseRater: Meta-learned noise valuation for diffusion model training. *arXiv preprint arXiv:2605.08144*, 2026. URL https://arxiv.org/abs/2605.08144.

Jingxuan Wu, Zhenglin Wan, Xingrui Yu, Yuzhe Yang, Bo An, Ivor Tsang, and Yang You. Letting trajectories spread: Quality-preserving control for diverse flow matching. *arXiv preprint arXiv:2510.09060*, 2025. URL https://arxiv.org/abs/2510.09060.

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. URL https://arxiv.org/abs/2306.09341.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.05977.

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. URL https://arxiv.org/abs/2505.07818.

Renye Yan, Jikang Cheng, You Wu, Wei Peng, Zongwei Wang, Ling Liang, and Yimao Cai. Explore or converge? stage-guided per-step optimization for diffusion models. *arXiv preprint arXiv:2608.06768*, 2026. URL https://arxiv.org/abs/2608.06768.

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*, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Using_Human_Feedback_to_Fine-tune_Diffusion_Models_without_Any_Reward_CVPR_2024_paper.html.

Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b. URL https://proceedings.mlr.press/v235/yang24e.html.

Benjamin Yu, Jackie Liu, and Justin Cui. Smart-GRPO: Smartly sampling noise for efficient RL of flow-matching models. *arXiv preprint arXiv:2510.02654*, 2025. URL https://arxiv.org/abs/2510.02654.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2206.10789.

Shengjun Zhang, Zhang Zhang, Chensheng Dai, and Yueqi Duan. E-GRPO: High entropy steps drive effective reinforcement learning for flow models. *arXiv preprint arXiv:2601.00423*, 2026. URL https://arxiv.org/abs/2601.00423.

Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. In *European Conference on Computer Vision*, 2024. URL https://doi.org/10.1007/978-3-031-73036-8_1.

Mingzhe Zheng, Weijie Kong, Yue Wu, Dengyang Jiang, Yue Ma, Xuanhua He, Bin Lin, Kaixiong Gong, Zhao Zhong, Liefeng Bo, Qifeng Chen, and Harry Yang. Manifold-aware exploration for reinforcement learning in video generation. *arXiv preprint arXiv:2603.21872*, 2026. URL https://arxiv.org/abs/2603.21872.

Jade Zou, Tao Huang, Weijie Kong, Junzhe Li, Yue Wu, Qi Tian, Jiangfeng Xiong, Jianwei Zhang, Liefeng Bo, and Zhao Zhong. Precise: SDE-consistent stochastic sampling for RL post-training of flow-matching models. *arXiv preprint arXiv:2605.23522*, 2026. URL https://arxiv.org/abs/2605.23522.

# A Training Details

## A.1 Hyperparameters and Compute

The image experiments use SD3.5-M at $512 \times 512$ resolution with 25 sampling and evaluation denoising steps, classifier free guidance 4.5, Flow-GRPO noise level $\lambda = 0.7$, and 24 rollouts per prompt in the main setting. Training uses 64 H100 GPUs over eight nodes. Each rank samples microbatches of 3, with 8 sampling batches per epoch, one inner epoch, and gradient accumulation 4. In the main setting, each training iteration therefore samples $P = 64$ prompts with $G = 24$ rollouts each (1,536 images), and both policies take two optimizer updates on these rollouts. The update counts reported throughout the paper are optimizer updates. Both the denoiser and ExploreNet use rank 32 LoRA adapters with LoRA scale 64 on the SD3.5-M attention projections. The denoising optimizer uses Adam with learning rate $3 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $10^{-4}$, and gradient norm clipping at 1.0.

ExploreNet initializes a separate SD3.5-M transformer from the pretrained model, freezes its base weights, and trains its rank 32 LoRA parameters in fp32. Its Adam optimizer uses learning rate $10^{-4}$ and zero weight decay. The final parameterization normalizes the output per sample and clips log standard deviation directly to $[-1, 1]$, with a standard-deviation floor of $10^{-6}$ during normalization. Reward normalization uses an additive constant of $10^{-4}$. Both clipped objectives use clip range $\epsilon = 10^{-4}$ and advantages clipped to $[-5, 5]$. The exploration objective applies a KL penalty of weight $\beta_\phi = 10^{-3}$ toward the unit Gaussian, and the denoising objective applies a KL penalty of weight $\beta_\theta = 0.04$ toward the pretrained denoiser, for both Flow-GRPO and ExploreNet. The denoising and exploration policy optimizers are stepped after the same backward pass.

The 48-to-24 privileged post-hoc selection baseline doubles rollout generation and GenEval2 reward calls relative to the 24 rollout runs, then keeps the 12 lowest- and 12 highest-reward samples, the subset of 24 with the greatest reward spread.

Measured over steps 2160 through 2580, Flow-GRPO requires 151.7 seconds per optimizer update and ExploreNet requires 189.4 seconds. ExploreNet therefore takes $1.25\times$ as long per update, corresponding to 161.8 versus 202.0 GPU hours per 60 updates, including periodic evaluation and checkpointing. During training, ExploreNet holds a second 2.5-billion-parameter SD3.5-M transformer in memory, and both methods fit a per-rank microbatch of 3 on 80 GB H100 GPUs. At inference, both methods use 25 denoiser evaluations and ExploreNet is not used.

| Setting | Rollouts | Reward calls | Denoising steps | ExploreNet evaluations |
|---|---|---|---|---|
| Main Flow-GRPO training | 24 | 24 | 600 | 0 |
| Main ExploreNet training | 24 | 24 | 600 | 600 |
| Privileged selection 48 to 24 training | 48 | 48 | 1200 | 0 |
| Matched isotropic training | 24 | 24 | 600 | 0 |
| Group size 8 training | 8 | 8 | 200 | 200 |
| 10 step training | 24 | 24 | 240 | 240 |
| Flow-GRPO inference | 1 | 0 | 25 | 0 |
| ExploreNet inference | 1 | 0 | 25 | 0 |

Table 3: ExploreNet adds auxiliary transformer evaluations during training while both trained denoisers retain the same inference cost. Counts describe the sampling passes for one training prompt group or one inference image.

## A.2 Noise-Scale Parameterization

This ablation tests how the output of $g_\phi$, $h_\phi$, is converted into the elementwise log standard deviation $\ell_\phi$. The *normalized affine* variant standardizes $h_\phi$ as in Eq. 6, applies a learned scalar scale and bias, and clips the result to $[-1, 1]$. The *raw clipped* variant uses the transformer output directly and clips it to the same range. The final ExploreNet parameterization standardizes the output and clips it directly, without the learned scalar scale and bias, as in Eq. 7. In each case, exponentiating $\ell_\phi$ gives the standard deviation of the zero-mean Gaussian transition noise.

All four rows use the same prompts, 25-step sampler, rollout group size, and microbatch-3 optimization schedule. Table 4 reports one 80-prompt GenEval2 evaluation at an intermediate checkpoint and at step 3060; each row is a single training run. We use normalized-and-clipped noise in the main experiments because it performs best at both checkpoints in this controlled comparison.

| Noise parameterization | Step 1140 | Step 3060 |
|---|---|---|
| Flow-GRPO (isotropic) | 0.3825 | 0.5146 |
| EXPLORENET, normalized affine | 0.4009 | 0.5057 |
| EXPLORENET, raw clipped | 0.4080 | 0.5358 |
| EXPLORENET, normalized and clipped | **0.4179** | **0.5679** |

Table 4: GenEval2 scores for Flow-GRPO and three mappings from the exploration transformer output to log standard deviation.

### A.3 REWARD-SPREAD SELECTION

The selection experiment in Section 3 draws 16 rollouts per prompt and trains on 8, chosen by smallest reward spread, at random, or by greatest reward spread. Greatest-spread selection keeps the four lowest- and four highest-reward rollouts, which maximizes the reward spread of the kept rollouts; smallest-spread selection keeps the eight rollouts with consecutive rewards of lowest variance. The three conditions share the rollout pool, group size, optimizer schedule, and training length, with one training run each; this schedule is separate from the main experiments, so its values compare the selection rules with one another. Table 5 evaluates the three final models on the full benchmark suite with fixed prompts and generation seed.

| Metric | Smallest | Random | Greatest |
|---|---|---|---|
| GenEval2 | 0.2579 | 0.3859 | **0.4648** |
| VLM Judge | 3.8750 | 4.4625 | **4.5125** |
| GenEval | 0.6171 | 0.6537 | **0.6948** |
| VVR-Fast | 0.7210 | 0.7380 | **0.7614** |
| PickScore | 0.8414 | 0.8433 | **0.8456** |
| HPSv2.1 | 0.3016 | 0.3054 | **0.3073** |
| HPSv3 | 7.8131 | 8.1597 | **8.3543** |
| Aesthetic | 5.5201 | **5.5228** | 5.5122 |
| ImageReward | 0.9409 | 1.0517 | **1.1075** |

Table 5: Benchmark results of the three selection rules. Greatest-spread selection is best on eight of nine metrics.

### A.4 MATCHED ISOTROPIC CONTROL

The control samples each transition with Eq. 8, setting $s_j = 1.410490688$ for every latent element, so its noise covariance is $1.989483981\, a_t^2 I$. This matches the mean variance predicted across latent elements and the three main EXPLORENET seeds over training steps 2760–2999. As in EXPLORENET, the transition mean is Flow-GRPO's $\mu_\theta$. The control contains no noise network and uses the seed 1 training schedule through step 3000. Appendix Table 14 compares it with Flow-GRPO and EXPLORENET across the final five checkpoints.

## B EVALUATION PROTOCOL

### B.1 AUTOMATIC EVALUATION

**Benchmarks and prompt sets.** GenEval2 (Kamath et al., 2025) contains 720 training prompts and a fixed 80-prompt held-out split covering object presence, count, attributes, and spatial relations. Each prompt is decomposed into atomic visual questions. Qwen3-VL-2B assigns probability to the accepted short answers, and the image score is the geometric mean across questions. GenEval2 is

the sole reward used to train the main models. On the same 80-prompt test set, VLM Judge scores prompt alignment from 1 to 5.

Official GenEval (Ghosh et al., 2023) uses 553 prompts with four generations per prompt, giving 2,212 evaluated images. VVRBench-Fast (Li et al., 2026b) contains 820 prompts sampled from VVRBench Core across its attainable complexity levels. Its frozen, model-free verifier evaluates grounding, cardinality, spatial, size, and topology constraints from executable scene specifications. We report dense VVR, the mean dense verifier score over all 820 prompts.

We evaluate PickScore (Kirstain et al., 2023) on 500 unique prompts from the Pick-a-Pic v1 validation split. We evaluate HPSv2.1 (Wu et al., 2023) on its complete 3,200-prompt benchmark, comprising 800 prompts each for anime, concept art, paintings, and photo. We report the mean PickScore, the four HPS domain means, and their unweighted macro average.

We evaluate HPSv3 (Ma et al., 2025), Aesthetic score, and ImageReward on the full 200-prompt DrawBench (Saharia et al., 2022) set and fixed 1,000-prompt lists from PartiPrompts (Yu et al., 2022), DPG-Bench (Hu et al., 2024), and T2I-CompBench (Huang et al., 2023). The aesthetic score uses the LAION improved aesthetic predictor (Schuhmann, 2022), with a CLIP ViT-L/14 image encoder and its released linear MSE head; ImageReward uses ImageReward-v1.0 (Xu et al., 2023). We average each metric within a prompt list and report the unweighted mean of the four list means.

**Generation and aggregation.** Checkpoints are saved and evaluated every 60 optimizer updates. For the main comparison, we evaluate the checkpoint at step 3000 without selecting by evaluation score. Evaluation uses the trained denoiser with the deterministic flow-matching sampler without ExploreNet. The primary comparison comprises three independent training seeds per method and three GenEval2 evaluation passes per seed. All other automatic benchmarks use one pass per training seed. Each model generates one image per prompt except for official GenEval, which uses four. Matched methods receive the same prompts and initial latent seeds. We average evaluation passes within each training seed, then report the mean and sample standard deviation across seeds. Controlled and robustness experiments use one training seed unless their corresponding result states otherwise. Flow-GRPO and ExploreNet runs are paired by training seed (42, 123, and 456, labeled 1 to 3); Table 7 reports every seed.

### B.2 Human Evaluation

The preference and displacement studies use the same 72 matched triplets of pretrained, Flow-GRPO, and ExploreNet outputs described in Section 4. Each of the six prompt suites contributes 12 examples, and each training seed contributes four examples within a suite. For DrawBench, PartiPrompts, DPG-Bench, T2I-CompBench, and the Flow-GRPO OCR prompt set (Liu et al., 2025a), we draw 12 distinct prompts uniformly without replacement using a fixed study seed. For VVR-Bench, we draw one prompt from each of 12 prespecified strata spanning composition tier and difficulty. We then shuffle each suite with the same fixed seed and assign prompts cyclically across the three training seeds. The panel is fixed before image generation, and the three models use the same generation seed for each prompt.

The preference interface presents the prompt and two unlabeled candidates. It instructs annotators to prioritize satisfaction of the full prompt and then visual quality, reserving a tie for cases in which neither image is meaningfully preferable. Trial order and candidate position are independently randomized. Three annotators evaluate all 216 pairwise comparisons, giving 648 judgments. The primary endpoint is tie-adjusted preference for ExploreNet over Flow-GRPO, with each tie contributing one half to each model. Confidence intervals resample prompts while retaining all ratings associated with each sampled prompt.

Pairwise exact agreement among the three annotators ranges from $56.9\%$ to $72.7\%$, and $47.7\%$ of comparisons receive unanimous labels.

| Preferred model | Wins | Ties | Losses | Rate | 95% CI |
|---|---|---|---|---|---|
| Flow-GRPO over pretrained | 118 | 8 | 90 | 56.5% | [48.6, 64.4] |
| EXPLORENET over pretrained | 143 | 8 | 65 | 68.1% | [60.0, 75.9] |
| EXPLORENET over Flow-GRPO | 140 | 7 | 69 | 66.4% | [58.6, 74.1] |

Table 6: EXPLORENET wins each pairwise preference comparison. Each row contains 216 ratings over 72 prompts. Rates split ties equally, and confidence intervals resample prompts with all associated ratings.

**Qualitative examples.** Figure 4 shows four study prompts on which all three annotators preferred EXPLORENET to Flow-GRPO, with each model's image at the shared generation seed, and one prompt from the held-out GenEval2 set. Its OCR header abbreviates the prompt "A wizard holds a staff with glowing runes, casting a bright light as he utters the spell 'Lumos Maximus', illuminating a dark, mystical forest."

**Change from pretrained outputs.** The displacement study uses the same 72 matched triplets and three complete annotators. The interface presents the pretrained output as a labeled reference between blinded Flow-GRPO and EXPLORENET candidates. Annotators identify which candidate differs more from the reference in scene content or composition. The instructions define this criterion through entities, attributes, counts, spatial arrangement, viewpoint, and framing, while excluding differences limited to color, lighting, texture, rendering style, or image quality. Trial order and candidate position are independently randomized for each annotator. Across 216 ratings, EXPLORENET is selected 181 times, Flow-GRPO 25 times, and 10 judgments are ties. The tie-adjusted EXPLORENET rate is 86.1%, with a 95% CI of [80.1, 91.4]. EXPLORENET receives a majority on 62 of 72 prompts, Flow-GRPO on 6, and 4 prompts have no majority. Pairwise exact agreement ranges from 77.8% to 80.6%, with unanimous labels on 70.8% of prompts.

# C ADDITIONAL RESULTS

## C.1 PER-SEED RESULTS

| Method | Seed | GenEval2 | VLM Judge | GenEval | VVR-Fast | PickScore | HPSv2.1 | HPSv3 | Aesthetic | ImageReward |
|---|---|---|---|---|---|---|---|---|---|---|
| Flow-GRPO | 1 | 0.4982 | 4.5500 | 0.6923 | 0.7624 | 0.8447 | 0.3009 | 8.3142 | 5.5024 | 1.1232 |
| Flow-GRPO | 2 | 0.4933 | 4.5750 | 0.7062 | 0.7715 | 0.8467 | 0.3004 | 8.3142 | 5.5184 | 1.1339 |
| Flow-GRPO | 3 | 0.4956 | 4.4000 | 0.6879 | 0.7598 | 0.8430 | 0.2966 | 8.0051 | 5.4673 | 1.0840 |
| EXPLORENET | 1 | 0.5674 | 4.7000 | 0.7452 | 0.8075 | 0.8509 | 0.3017 | 8.6602 | 5.5997 | 1.2597 |
| EXPLORENET | 2 | 0.5972 | 4.7750 | 0.7335 | 0.7928 | 0.8509 | 0.3056 | 8.7869 | 5.6744 | 1.2841 |
| EXPLORENET | 3 | 0.5337 | 4.7125 | 0.7261 | 0.7703 | 0.8519 | 0.3097 | 8.9658 | 5.7175 | 1.2883 |

Table 7: Per-seed results at step 3000 underlying Table 1. GenEval2 averages three evaluation passes; other metrics use one pass.

## C.2 LATE-TRAINING STABILITY AND PROMPT-LEVEL UNCERTAINTY

We measure stability over the final five checkpoints by averaging steps 2760, 2820, 2880, 2940, and 3000 within each evaluation pass and training seed. The result then averages three evaluation passes within each seed and reports the mean and sample standard deviation across the three training seeds.

| Metric | Flow-GRPO | EXPLORENET | Gain |
|---|---|---|---|
| GenEval2 | $0.4809_{\pm 0.0093}$ | $\mathbf{0.5524_{\pm 0.0300}}$ | +0.0715 |

Table 8: The GenEval2 gain persists across the final five checkpoints. Each seed value averages checkpoints 2760 through 3000 over three evaluation passes. Subscripts report sample standard deviation across training seeds.

We also estimate sensitivity to the composition of each evaluation set with paired prompt bootstraps. For GenEval2, two evaluation passes retained complete prompt-level records for all three training seeds, both methods, and checkpoints 2760 through 3000; each replicate preserves these pairings and averages over training seeds, checkpoints, and evaluation passes. For VVRBench-Fast, each replicate resamples the 820 matched checkpoint-3000 prompts and averages the paired difference across the three training-seed pairs. Both analyses use 200,000 replicates and condition on the six trained models.

| Evaluation | Paired gain | 95% prompt bootstrap interval |
|---|---|---|
| GenEval2 | $+0.0748$ | $[+0.0405, +0.1085]$ |
| VVR-Fast | $+0.0256$ | $[+0.0202, +0.0311]$ |

Table 9: Prompt resampling gives positive gain intervals for GenEval2 and VVRBench-Fast.

### C.3 Per-Benchmark Results

Table 10 reports PickScore and the four HPSv2.1 styles on their benchmark prompts, and Table 11 reports HPSv3, Aesthetic, and ImageReward on each of the four prompt suites. Each entry is the mean and sample standard deviation across the three training seeds, and each seed generates one image per prompt with evaluation seed 42.

| Prompt set and metric | Flow-GRPO | ExploreNet | Gain |
|---|---|---|---|
| Pick-a-Pic v1 (500), PickScore | $0.8448_{\pm 0.0019}$ | $\mathbf{0.8512_{\pm 0.0006}}$ | $+0.0064$ |
| HPSv2 anime prompts (800) | $0.3130_{\pm 0.0018}$ | $\mathbf{0.3204_{\pm 0.0038}}$ | $+0.0073$ |
| HPSv2 concept art prompts (800) | $0.3034_{\pm 0.0026}$ | $0.3092_{\pm 0.0054}$ | $+0.0058$ |
| HPSv2 paintings prompts (800) | $0.3038_{\pm 0.0027}$ | $0.3085_{\pm 0.0039}$ | $+0.0047$ |
| HPSv2 photo prompts (800) | $0.2770_{\pm 0.0027}$ | $\mathbf{0.2845_{\pm 0.0032}}$ | $+0.0075$ |
| HPSv2.1 macro (3,200) | $0.2993_{\pm 0.0023}$ | $0.3056_{\pm 0.0040}$ | $+0.0063$ |

Table 10: ExploreNet scores higher on PickScore and on all four HPSv2.1 prompt styles. Bold marks gains whose displayed intervals do not overlap.

| Prompt suite | Method | HPSv3 | Aesthetic | ImageReward |
|---|---|---|---|---|
| DrawBench | Flow-GRPO | $9.5718_{\pm 0.2295}$ | $5.3664_{\pm 0.0279}$ | $1.0602_{\pm 0.0503}$ |
| | ExploreNet | $\mathbf{10.3331_{\pm 0.2491}}$ | $\mathbf{5.5109_{\pm 0.0625}}$ | $\mathbf{1.2683_{\pm 0.0210}}$ |
| PartiPrompts | Flow-GRPO | $7.4324_{\pm 0.1504}$ | $5.5783_{\pm 0.0275}$ | $1.2560_{\pm 0.0110}$ |
| | ExploreNet | $\mathbf{7.8012_{\pm 0.0808}}$ | $\mathbf{5.7438_{\pm 0.0507}}$ | $\mathbf{1.3872_{\pm 0.0174}}$ |
| DPG-Bench | Flow-GRPO | $9.2953_{\pm 0.2166}$ | $5.6607_{\pm 0.0287}$ | $0.8724_{\pm 0.0316}$ |
| | ExploreNet | $\mathbf{9.8481_{\pm 0.0996}}$ | $\mathbf{5.8578_{\pm 0.0401}}$ | $\mathbf{1.0256_{\pm 0.0153}}$ |
| T2I-CompBench | Flow-GRPO | $6.5454_{\pm 0.1530}$ | $5.3788_{\pm 0.0253}$ | $1.2663_{\pm 0.0285}$ |
| | ExploreNet | $\mathbf{7.2348_{\pm 0.2011}}$ | $\mathbf{5.5428_{\pm 0.0885}}$ | $\mathbf{1.4283_{\pm 0.0110}}$ |

Table 11: HPSv3, Aesthetic, and ImageReward results on the shared four-suite prompt panel. ExploreNet improves all three metrics on every prompt suite.

| Constraint family | Prompts | Flow-GRPO | EXPLORENET | Gain |
|---|---|---|---|---|
| Grounding | 302 | $0.6293_{\pm 0.0311}$ | $\mathbf{0.7683_{\pm 0.0175}}$ | $+0.1390$ |
| Spatial | 665 | $0.2924_{\pm 0.0094}$ | $\mathbf{0.4187_{\pm 0.0164}}$ | $+0.1263$ |
| Size | 285 | $0.1816_{\pm 0.0361}$ | $\mathbf{0.2722_{\pm 0.0270}}$ | $+0.0906$ |
| Topology | 242 | $0.1525_{\pm 0.0313}$ | $\mathbf{0.2385_{\pm 0.0543}}$ | $+0.0860$ |
| Cardinality | 367 | $0.6790_{\pm 0.0080}$ | $0.6644_{\pm 0.0568}$ | $-0.0146$ |

Table 12: EXPLORENET improves grounding, spatial, size, and topology constraints on VVRBench-Fast. Each entry is the mean verifier score on the prompts that contain the constraint family, averaged within each training seed and reported as mean and sample standard deviation across the three seeds. Bold marks gains whose displayed intervals do not overlap.

### C.4 HELD-OUT GENEVAL2 DURING TRAINING

Table 13 reports held-out GenEval2 averaged over the three training seeds, using the evaluations plotted in Figure 5. EXPLORENET trails Flow-GRPO at every evaluation through update 600, and each training seed trails its Flow-GRPO counterpart at every evaluation through update 420. From update 660 to 3000, EXPLORENET leads at every evaluation, and its lead grows from 0.0028 at update 660 to 0.0704 at update 3000.

| Update | Flow-GRPO | EXPLORENET | Difference |
|---|---|---|---|
| 60 | 0.2535 | 0.2202 | $-0.0333$ |
| 120 | 0.2703 | 0.2316 | $-0.0386$ |
| 180 | 0.2883 | 0.2484 | $-0.0399$ |
| 240 | 0.2989 | 0.2648 | $-0.0341$ |
| 300 | 0.3064 | 0.2749 | $-0.0315$ |
| 360 | 0.3262 | 0.2989 | $-0.0272$ |
| 420 | 0.3387 | 0.3085 | $-0.0302$ |
| 480 | 0.3411 | 0.3157 | $-0.0255$ |
| 540 | 0.3361 | 0.3264 | $-0.0097$ |
| 600 | 0.3526 | 0.3378 | $-0.0148$ |
| 660 | 0.3530 | 0.3559 | $+0.0028$ |
| 720 | 0.3606 | 0.3730 | $+0.0123$ |
| 900 | 0.3738 | 0.3898 | $+0.0160$ |
| 1200 | 0.3880 | 0.4265 | $+0.0385$ |
| 1800 | 0.4329 | 0.4913 | $+0.0584$ |
| 2400 | 0.4713 | 0.5318 | $+0.0605$ |
| 3000 | 0.4957 | 0.5661 | $+0.0704$ |

Table 13: Held-out GenEval2 during training, averaged over three training seeds. EXPLORENET trails Flow-GRPO through update 600 and leads from update 660 onward.

### C.5 MATCHED ISOTROPIC CONTROL ACROSS CHECKPOINTS

| Evaluation | Flow-GRPO | Matched isotropic | EXPLORENET |
|---|---|---|---|
| GenEval2, step 2760 | 0.4682 | 0.4022 | **0.5490** |
| GenEval2, step 2820 | 0.4658 | 0.3788 | **0.5335** |
| GenEval2, step 2880 | 0.4568 | 0.4211 | **0.5484** |
| GenEval2, step 2940 | 0.4722 | 0.4189 | **0.5448** |
| GenEval2, step 3000 | 0.4982 | 0.4372 | **0.5674** |
| VVR-Fast, step 3000 | 0.7624 | 0.7791 | **0.8075** |

Table 14: EXPLORENET exceeds isotropic noise of the same average variance on GenEval2 at every checkpoint and on VVRBench-Fast. All methods use seed 1, the same prompts, and the same initial latent seed within each evaluation. Flow-GRPO and EXPLORENET GenEval2 values average three evaluation passes.

## D LATENT CHANNEL ANALYSIS

The analysis uses the seed 1 EXPLORENET denoiser and $g_\phi$ at the end of training, together with the first 100 prompts in the fixed GenEval2 training manifest. For each prompt, we sample one initial latent and generate a reference image with the deterministic flow-matching ODE. At every one of the 25 denoising steps, we draw a spatial Gaussian field, normalize it to zero mean and unit standard deviation, and add it to one selected latent channel with either positive or negative sign. The perturbation uses the Flow-GRPO noise coefficient $a_t$ with $\lambda = 0.7$. All other channels remain on the deterministic trajectory. Repeating this procedure for 16 channels gives 1,600 antithetic pairs, or 3,200 perturbed generations.

We evaluate each perturbed image by its mean absolute RGB difference from the reference. For the same prompt, EXPLORENET is evaluated along the unperturbed reference trajectory. Its predicted log scale is averaged over spatial positions and then over denoising steps for each channel. The two signs are averaged to obtain one visual change value per prompt and channel.

Table 15 reports the correlations at three aggregation levels. The four channels assigned the largest scales are 14, 2, 7, and 13 and average $0.0812$ RGB change. The four channels assigned the smallest scales are 12, 4, 0, and 10 and average $0.0570$.

| Aggregation | $N$ | Pearson $r$ | Spearman $\rho$ | 95% Pearson CI |
|---|---|---|---|---|
| Prompt and channel combinations | 1600 | $+0.306$ | $+0.332$ | $[+0.273, +0.342]$ |
| Centered within prompt | 1600 | $+0.461$ | $+0.490$ | $[+0.438, +0.485]$ |
| Channel means | 16 | $+0.598$ | $+0.491$ | $[+0.145, +0.844]$ |
| Within prompt, then averaged | 100 | $+0.485$ | $+0.490$ | $[+0.462, +0.506]$ |

Table 15: EXPLORENET's predicted scale tracks intervention sensitivity across prompt and channel aggregation levels. The first two intervals and the last resample prompts with replacement over 20,000 replicates; the channel-means interval is a Fisher interval over the 16 channels. The within-prompt correlation is positive for all 100 prompts, with minimum Spearman $\rho = +0.156$.

**Human perceptual evaluation.** The human study uses prompts at fixed manifest indices 0, 20, 26, 47, 67, and 98 from the GenEval2 training set and the pretrained, Flow-GRPO, and EXPLORENET denoisers. For every model and prompt, we generate a deterministic reference and intervene on each of the 16 latent channels with both signs. This gives 288 trials, each containing a reference and two perturbed images, and 576 perturbed images in total. Model, channel, and perturbation sign are hidden from annotators. Nine annotators participated, and each trial contributes exactly three independent ratings. When four ratings were available, we dropped one at random for each trial using a fixed seed before aggregation. Each image receives an ordinal change magnitude rating from 0 for no obvious visible difference to 4 for a very large difference.

We use the median magnitude across annotators for each perturbed image. Within each prompt, channels are ranked by EXPLORENET's mean predicted log standard deviation along the reference trajectory. Table 16 compares the top and bottom four channels. The pooled estimate averages both perturbation signs and all three denoisers; confidence intervals resample the six prompts over 20,000 replicates.

| Denoiser | Bottom quartile | Top quartile | Difference | 95% CI |
|---|---|---|---|---|
| Pretrained | 1.521 | 2.229 | $+0.708$ | $[+0.271, +0.958]$ |
| Flow-GRPO | 1.125 | 1.313 | $+0.188$ | $[-0.188, +0.521]$ |
| EXPLORENET | 0.771 | 1.250 | $+0.479$ | $[+0.250, +0.708]$ |
| Pooled | 1.139 | 1.597 | $+0.458$ | $[+0.319, +0.583]$ |

Table 16: Cross denoiser comparison of human change ratings for EXPLORENET's largest and smallest scale quartiles. The separation is present before training and remains after EXPLORENET training. Quartiles are defined within each prompt, and ratings use a scale from 0 to 4.

Magnitude ratings form the perceptual analysis. Across the twelve available comparisons between pairs of annotators, ratings are within one scale point in $84.7\%$ of cases, with mean Spearman correlation $0.565$ and mean quadratic weighted $\kappa = 0.440$.

**Pixel-space change across denoisers.** The perturbations of the human study also record the RGB change of every perturbed image, which repeats the pixel-space analysis on the pretrained and Flow-GRPO denoisers. We use the perturbations at the strength of the main analysis, average the two signs and the six prompts for each channel, and correlate the result across the 16 channels with EXPLORENET's channel mean log scale from the main analysis. Table 17 also reports the rank correlation computed within each prompt, using that prompt's scales, and the number of prompts for which it is positive. We also repeat the full pixel-space analysis on the pretrained denoiser with the 100 prompts, seeds, and perturbations of the main analysis, giving 3,200 perturbed generations. On both prompt sets, the correlation on the pretrained denoiser is close to that on the EXPLORENET denoiser, so the relation between EXPLORENET's scales and pixel-space sensitivity is present before RL training.

| Perturbed denoiser | Pearson $r$ | Spearman $\rho$ | 95% Pearson CI | Within-prompt $\rho$ | Positive prompts |
|---|---|---|---|---|---|
| *100 prompts of the main analysis* | | | | | |
| Pretrained | $+0.516$ | $+0.488$ | $[+0.027, +0.806]$ | $+0.421$ | 99/100 |
| EXPLORENET | $+0.598$ | $+0.491$ | $[+0.145, +0.844]$ | $+0.490$ | 100/100 |
| *Six prompts of the human study* | | | | | |
| Pretrained | $+0.562$ | $+0.535$ | $[+0.092, +0.827]$ | $+0.451$ | 6/6 |
| Flow-GRPO | $+0.509$ | $+0.494$ | $[+0.018, +0.802]$ | $+0.425$ | 6/6 |
| EXPLORENET | $+0.576$ | $+0.618$ | $[+0.113, +0.834]$ | $+0.523$ | 6/6 |

Table 17: Correlation between EXPLORENET's predicted scale and pixel-space change when each denoiser is perturbed. The first three columns use channel means; intervals are Fisher intervals over the 16 channels. Within-prompt $\rho$ averages the per-prompt rank correlation.