PAIRREFL: PAIRWISE REWARD FEEDBACK LEARNING FOR DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning text-to-image models to human preferences, guided by a reward model, has become common practice. Compared with policy-gradient methods relying soly on scalar reward, reward feedback learning methods backpropagate feedback through the reward model, exploiting richer training guidance from each sample. Despite its efficiency, this learning paradigm is susceptible to *reward hacking*: maximizing a reward model's per-image score can amplify undesirable visual patterns such as oversaturated colors, without improving perceived image quality. In this paper, we propose Pairwise Reward Feedback Learning (**PairReFL**), which replaces scalar-score maximization with a differentiable comparison against another image sampled from the current model, in which the comparison reference changes as the model learns. Across four human-preference metrics, PairReFL improves the base model by $1.3\times$ as much as vanilla reward feedback learning and $1.8\times$ as much as the strongest policy-gradient baseline. Our differentiable pairwise reward model reduces the error rate in visual-quality comparisons by 62% relative to its pointwise counterpart. Further analyses trace these gains to the pairwise interface, which provides more stable feedback on closely matched samples and achieves more significant preference gains.

1 INTRODUCTION

With the rapid advancement of text-to-image generation, generative foundation models synthesize impressive and photorealistic scenes from free-form prompts (Esser et al., 2024; Black Forest Labs, 2024). To further improve prompt following, composition, and visual appeal, aligning models with human preference guided by a reward model has become standard practice (Xu et al., 2023; Black et al., 2024). Existing methods differ in how they use the reward model. On the one hand, policy-gradient methods treat it as a black box, using only the scalar scores, *i.e.*, the rewards, of sampled images to reweight samples or contrast winners against losers (Black et al., 2024; Fan et al., 2023; Liu et al., 2026; Wang et al., 2025b; Zheng et al., 2026). Specifically, such a score indicates the quality of an image but provides no explicit direction for improvement, requiring the generator to infer the optimization direction from score differences across its own samples. On the other hand, Reward Feedback Learning (ReFL) (Xu et al., 2023) methods directly propagate through the differentiable reward model into the generator (Clark et al., 2024; Prabhudesai et al., 2023), providing explicit gradient signals that directly indicate how each image should be improved. Such signals provide richer per-sample information, thereby substantially improving optimization efficiency.

Despite its efficiency, reward feedback learning remains susceptible to *reward hacking* (Kim & Sim, 2026; Wang et al., 2025b; Kang et al., 2026). In general, a reward model is an imperfect proxy for human preferences: its per-image score can increase with low-level visual patterns such as color saturation even when the image does not align well with the prompt or improve human-judged visual attributes such as composition, coherence, and aesthetics, so maximizing the score can amplify these patterns (Gao et al., 2023; Clark et al., 2024; Shen et al., 2025), as illustrated in Figure 6. Moreover, vanilla pointwise optimization uses the same absolute scoring criterion throughout training, so it keeps providing the same direction as the generator amplifies these patterns, allowing them to accumulate.

Motivated by this issue, we propose to replace isolated per-image scores with a relative comparison against another image sampled from the current generator (Shah et al., 2014; Christiano et al.,

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.

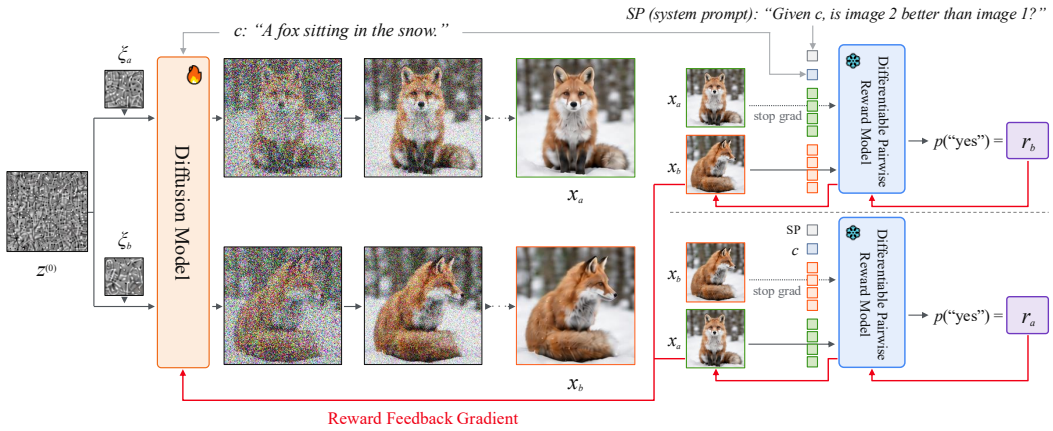


Figure 1: **Illustration of the proposed Pairwise Reward Feedback Learning (PairReFL).** Starting from shared initial noise, PairReFL evolves two generation branches with independent random perturbations to produce a matched image pair. The differentiable pairwise reward model PairHPSv3 jointly evaluates both images and backpropagates the preference signal through the image being optimized to the generator, while the other image serves as a detached anchor. Their roles are exchanged so that each image is optimized against its peer.

2017). Because the comparison depends on how the two images differ, a pattern already shared by both images cannot by itself create a preference advantage, while the sampled peer evolves with the generator. Rather than assigning each image an independent score, the joint comparison evaluates whether one image is preferred to the other, allowing the feedback to depend on the specific alternative being compared. This does not require abandoning the rich per-sample signals of reward feedback learning. Instead, by making the comparison differentiable, we can propagate preference feedback through the generated images into the generator.

Toward this end, we propose **Pairwise Reward Feedback Learning (PairReFL)**, as illustrated in Figure 1. For each prompt, we generate two images from the same initial noise, introducing independent random perturbations during generation to obtain distinct outputs. We evaluate the pair twice, each time stopping gradients through one image while backpropagating through the other. PairHPSv3 jointly evaluates the prompt and both images and outputs a differentiable preference logit, which is converted into a preference loss for the image receiving the gradient. We then swap the gradient path, so each image is optimized relative to the other. This training procedure requires a reward model that jointly evaluates the prompt and both images while remaining differentiable with respect to the image being optimized. To support the training, we develop *PairHPSv3*, a differentiable pairwise reward model that predicts the relative preference between two images, with frozen parameters during generator training and gradients propagating to the generator.

We compare PairReFL against representative post-training methods on the widely used FLUX.1-dev base model (Black Forest Labs, 2024). Across diverse human-preference metrics, PairReFL improves the base model by $1.3\times$ as much as pointwise reward feedback learning (Xu et al., 2023) and $1.8\times$ as much as the strongest policy-gradient method (Wang et al., 2025b). Trained with the same backbone (Wang et al., 2024) and preference data (Ma et al., 2025), our differentiable pairwise reward model *PairHPSv3* achieves 1.9% higher preference prediction accuracy (Ma et al., 2025; Chen et al., 2026; Wu et al., 2023; Jiang et al., 2024) than its pointwise counterpart (Ma et al., 2025). Further analyses in Section 4.5 show that joint comparison gives more consistent preference judgments under mild image perturbations, while PairReFL achieves 37% more preference improvement for a given amount of change from the base model.

Our contributions are as follows.

- We identify the optimization target as a key limitation of reward feedback learning: its gradients provide rich training signals, but maximizing per-image scores can reinforce visual

Table 1: **Preference alignment on the HPDv2 test set.** Higher is better; best and second-best are bold and underlined. Small annotations give the relative change from FLUX.1-dev, green for improvement and red for regression. Bootstrap intervals and prompt-paired differences for every method and metric are given in Tables 5 and 6.

Method	HPSv2.1	HPSv3	PickScore	ImageReward
FLUX.1-dev	0.3079	13.5577	22.7874	1.0441
ReFL	<u>0.3154</u> (↑2.4%)	<u>13.8196</u> (↑1.9%)	22.7642 (↓0.1%)	<u>1.1138</u> (↑6.7%)
Flow-GRPO	0.3113 (↑1.1%)	13.6912 (↑1.0%)	22.7898 (↑0.0%)	1.0684 (↑2.3%)
Pref-GRPO	0.3148 (↑2.2%)	13.7775 (↑1.6%)	<u>22.8170</u> (↑0.1%)	1.0851 (↑3.9%)
PairReFL	**0.3182** (↑3.4%)	**13.9677** (↑3.0%)	**22.8348** (↑0.2%)	**1.1252** (↑7.8%)

patterns that do not improve perceived quality. This motivates replacing absolute scores with differentiable pairwise comparisons.

- We propose PairReFL, a pairwise reward feedback learning method that optimizes each generated image with a differentiable pairwise reward model *PairHPSv3* comparing it against another image sampled from the current model. The preference signal is back-propagated through the image being optimized while the other serves as a detached peer, making the update depend on their relative quality.
- Experiments show that PairReFL achieves state-of-the-art performance in human-preference alignment, outperforming widely used pointwise reward feedback learning and policy-gradient methods. We further analyze the mechanisms underlying its gains.

2 Related Work

Flow-Matching RL. Reinforcement learning has become a common post-training tool for diffusion and flow-matching models. Black et al. (2024) and Fan et al. (2023) treat denoising as a stochastic trajectory and optimize reward-weighted objectives, often with KL regularization or advantage normalization. Flow-GRPO (Liu et al., 2026) extends online RL to flow matching with stochastic sampling and group-relative policy updates, while Pref-GRPO (Wang et al., 2025b) replaces pointwise score maximization with pairwise preference fitting to mitigate reward hacking. A second line of work performs direct reward backpropagation through generation. Xu et al. (2023) introduced both ImageReward and ReFL-style direct tuning, while DRaFT (Clark et al., 2024), AlignProp (Prabhudesai et al., 2023), LeapAlign (Liang et al., 2026), and FlowBP (Wang et al., 2026) further improve the stability and memory profile of reward gradients. PairReFL belongs to this branch but replaces scalar image-level feedback with a differentiable pairwise comparator over two online samples.

Reward Models. Reward models trained from human preference data have become standard substitutes for expensive human feedback in text-to-image alignment. CLIP-style metrics (Radford et al., 2021) provide useful but coarse alignment signals, while ImageReward (Xu et al., 2023), Pick-a-Pic (Kirstain et al., 2023) and its PickScore model, HPSv2 (Wu et al., 2023), and HPSv3 (Ma et al., 2025) use large-scale preference data to improve evaluation and post-training. UnifiedReward (Wang et al., 2025c) broadens the setting to a more general multimodal judge, and benchmarks such as MJ-Bench (Chen et al., 2026) and GenAI-Bench (Jiang et al., 2024) probe whether these models recover held-out human preferences. Most reward models used for generator optimization expose pointwise scores, even when they are trained from pairwise annotations. PairHPSv3 keeps the comparison interface itself: it jointly conditions on the prompt and both images and returns a differentiable preference logit.

Pairwise Preference Feedback. Pairwise comparisons support generative model alignment and evaluation. DPO (Rafailov et al., 2023) learns from preference pairs without a separate reward model, with Diffusion-DPO (Wallace et al., 2024) and D3PO (Yang et al., 2024) extending this principle to diffusion models. Wang et al. (2025a) exploit negative preferences to improve classifier-free guidance. Pref-GRPO (Wang et al., 2025b) demonstrates that pairwise rewards improve reinforcement learning stability and reduce reward hacking. Through extensive experiments, Li et al. (2026) show that pairwise scoring is more accurate, consistent, and aligned with human preferences than

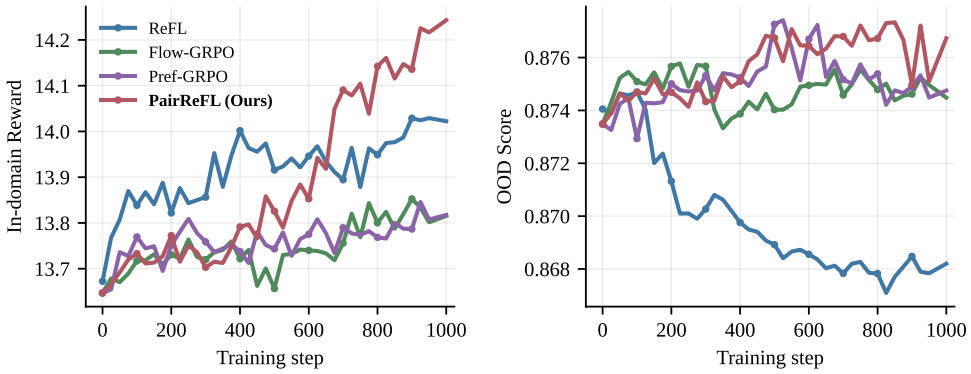


Figure 2: **Direct reward feedback shows a stronger reward-hacking trend than pairwise and policy-gradient optimization.** The in-domain reward is pointwise HPSv3 score on a fixed evaluation prompt set, while the OOD score is PickScore on the same set; OOD here denotes an independent preference scorer excluded from training. ReFL's in-domain reward rises while its OOD score falls, whereas PairReFL improve the in-domain reward without sacrificing the OOD score.

pointwise scoring. PairReFL brings pairwise comparison to reward feedback learning by directly backpropagating differentiable preference signals into the generator.

3 METHOD

We introduce **Pairwise Reward Feedback Learning (PairReFL)** for direct preference optimization of flow-matching models. PairReFL generates a matched pair from shared initial noise and updates each image against its detached peer using a differentiable pairwise reward model. It requires neither trajectory likelihoods nor group-normalized policy-gradient advantages. Figure 1 illustrates how PairReFL constructs a matched image pair and backpropagates pairwise preference feedback.

3.1 PRELIMINARIES: DIRECT REWARD FEEDBACK

Reward Feedback Learning (ReFL) (Xu et al., 2023) directly fine-tunes diffusion models by backpropagating a differentiable reward through generation. Unlike policy-gradient methods (Black et al., 2024; Liu et al., 2026), ReFL requires no log-probabilities or advantage estimates. Its reward term is

$$\mathcal{L}_{\text{reward}}(\theta) = \lambda\, \mathbb{E}_{c\sim\mathcal{C}}\left[\varphi(r_\psi(c, g_\theta(c)))\right], \tag{1}$$

where $g_\theta(c)$ is the generated image, φ maps the frozen reward r_ψ to loss, and λ scales the feedback. ReFL optimizes this term alongside the diffusion pre-training loss. PairReFL retains pathwise optimization but replaces the pointwise reward with a differentiable comparator jointly conditioned on two images. From $z^{[0]} \sim \mathcal{N}(0, I)$, we differentiate the flow-matching mean transitions (Lipman et al., 2022) after a sampled truncation point and the frozen VAE decoder D.

3.2 DIFFERENTIABLE PAIRWISE REWARD MODEL

Let $q_{\text{pair}}(c, x_a, x_b)$ denote the instruction asking whether x_b is preferred to x_a, and let $\ell_\phi(v \mid q)$ be the pre-softmax next-token logit for token v. We define

$$R_\phi(c; x_a, x_b) = \ell_\phi(\texttt{yes} \mid q_{\text{pair}}(c, x_a, x_b)), \qquad p_\phi(x_b \succ x_a \mid c) = \sigma(R_\phi(c; x_a, x_b)), \tag{2}$$

where a positive logit favors the second image. During training, we randomly swap the preferred and dispreferred images and set $y = 1$ when the preferred one is second. We minimize

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}\left[y \log p_\phi + (1-y)\log(1-p_\phi)\right]. \tag{3}$$

Random swapping prevents position from becoming a prediction shortcut. The comparator jointly encodes the prompt and both images. We call the resulting differentiable pairwise reward model PairHPSv3. During generator training, its parameters are frozen but its visual path remains differentiable with respect to pixels. Unlike $s_\phi(c, x_b) - s_\phi(c, x_a)$, this joint comparator can make the feedback for one image depend on its peer.

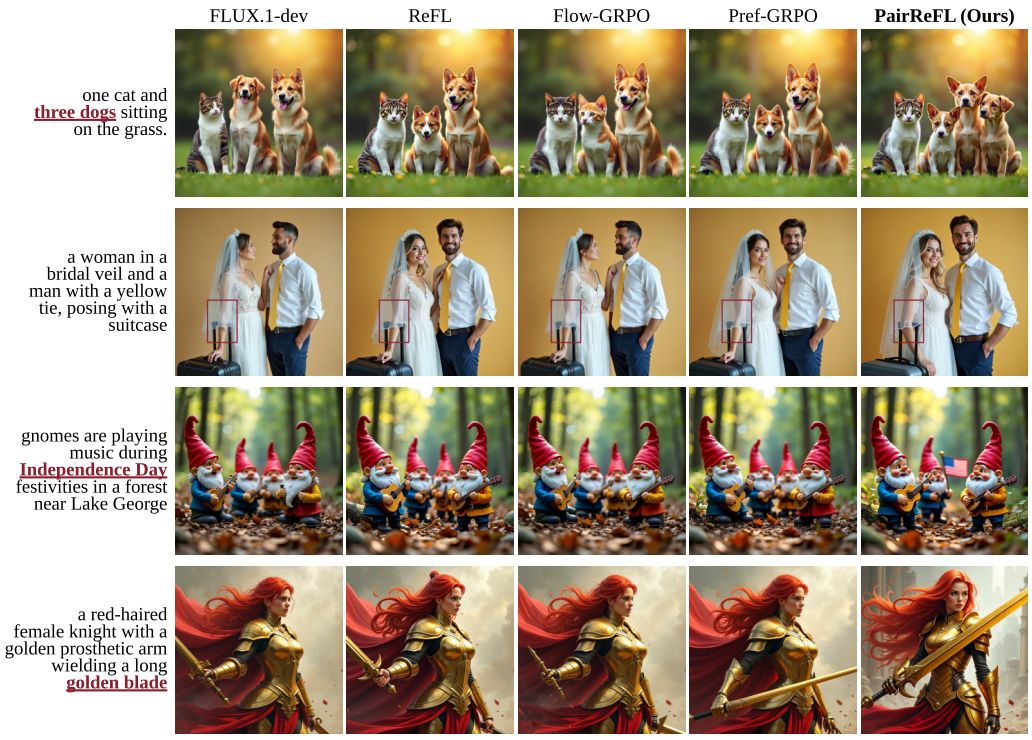


Figure 3: **Qualitative comparison of PairReFL and baseline methods.** Each row uses the same prompt and inference seed across methods. PairReFL follows the instructions more faithfully, matching the requested object counts, depicting human limbs more realistically, and showing better contextual reasoning by capturing the implied holiday theme. It also produces clearer object details, more prominent subjects, and more visually appealing compositions.

3.3 PairReFL

For each prompt and pair m, we duplicate one initial latent and evolve the two copies with independent stochastic innovations:

$$\begin{aligned} z_{m,a}^{[0]} &= z_{m,b}^{[0]} \sim \mathcal{N}(0, I), \\ z_{m,j}^{[k+1]} &= F_\theta\Big(z_{m,j}^{[k]}, t_k, c; \xi_{m,j,k}\Big), \quad j \in \{a, b\},\ k < K, \quad \xi_{m,a,k} \perp \xi_{m,b,k}. \end{aligned} \tag{4}$$

The prefix is sampled without gradients using Coefficients-Preserving Sampling (CPS) (Wang & Yu, 2025), which injects fresh noise while preserving the scheduler-prescribed sample and total-noise coefficients. From K onward, each branch follows the differentiable mean transition,

$$x_{m,j} = D\Big(\overline{F}_\theta^{[K:N]}\big(\mathrm{sg}(z_{m,j}^{[K]}), c\big)\Big), \qquad j \in \{a, b\}. \tag{5}$$

The intent of the shared latent is to hold coarse content and layout fixed while the independent CPS innovations create distinct outcomes, so that the comparator sees a pair whose variation is not dominated by unrelated initial conditions. Shared noise refers only to $z^{[0]}$, not to the stochastic noise injected along both trajectories.

Given the resulting pair, we evaluate it twice, alternating the active image and detached anchor:

$$r_{b\leftarrow a} = R_\phi(c; \mathrm{sg}(x_a), x_b), \qquad r_{a\leftarrow b} = R_\phi(c; \mathrm{sg}(x_b), x_a). \tag{6}$$

With $[u]_+ = \max(0, u)$, we define

$$\ell_\tau(r) = [\tau - \sigma(r)]_+, \tag{7}$$

Table 2: **Preference prediction accuracy (%).** Higher is better. With the same backbone, training data, and training recipe as HPSv3, PairHPSv3 replaces independent image scoring with joint pairwise comparison, improving preference prediction and visual-quality judgments.

Reward model	HPDv2	HPDv3	GenAI-Bench	MJ-Bench
CLIP ViT-H/14	65.1	48.6	58.3	68.1
ImageReward	74.0	58.6	65.1	81.8
PickScore	79.8	65.6	<u>72.4</u>	89.8
HPSv2.1	<u>83.3</u>	65.3	69.8	94.6
HPSv3	**85.4**	<u>76.9</u>	**72.7**	<u>97.1</u>
PairHPSv3 (ours)	**85.4**	**77.7**	69.4	**98.9**

and optimize M pairs with

$$\mathcal{L}_{\mathrm{PairReFL}}(\theta) = \frac{1}{2M} \sum_{m=1}^{M} \left[\ell_\tau(r_{m,b\leftarrow a}) + \ell_\tau(r_{m,a\leftarrow b}) \right]. \tag{8}$$

Here $\tau \in (0.5, 1)$ is a confidence threshold. Each image is optimized once against its detached peer, and the hinge stops its update after the target confidence is reached; no winner selection is required.

Why the peer must be detached. Bilateral stop-gradient is not a convenience. Suppose the comparator were exactly order-antisymmetric, $R(c; x_a, x_b) = -R(c; x_b, x_a) = r$. If both cap terms are active, a fully differentiated bidirectional objective satisfies

$$\ell_\tau(r) + \ell_\tau(-r) = 2\tau - \sigma(r) - \sigma(-r) = 2\tau - 1, \tag{9}$$

a constant, so its gradient with respect to both images vanishes: the better image is pushed down exactly as hard as the worse image is pushed up. Detaching the peer changes the derivative without changing the forward value, because the first term then updates only x_b and the second only x_a, and the two partial derivatives can no longer cancel through the opposite image path. Exact cancellation needs both antisymmetry and simultaneous cap activity, and Appendix B.10 measures how closely the learned comparator meets those conditions and confirms the cancellation numerically. A second motivation, a variance decomposition showing that a shared initial latent removes a between-seed term from the pairwise signal, is deferred to Appendix B.1.

4 Experiments

4.1 Experimental Setup

We fine-tune FLUX.1-dev (Black Forest Labs, 2024) on HPDv2 training prompts for 1,000 optimization steps. We compare PairReFL with the pretrained model and three post-training baselines: ReFL (Xu et al., 2023) for direct reward feedback learning, Flow-GRPO (Liu et al., 2026) for pointwise policy-gradient optimization, and Pref-GRPO (Wang et al., 2025b) for pairwise policy-gradient optimization. All fine-tuned methods use the same prompt source and optimization budget, and we evaluate their final checkpoints. PairReFL uses a frozen PairHPSv3 comparator obtained by fine-tuning Qwen2-VL-7B (Wang et al., 2024) on HPDv3 preference data (Ma et al., 2025).

We evaluate preference alignment on the HPDv2 test set (Wu et al., 2023) using HPSv2.1 (Wu et al., 2023), HPSv3 (Ma et al., 2025), PickScore (Kirstain et al., 2023), and ImageReward (Xu et al., 2023). Following the evaluation protocol of prior work (Liang et al., 2026), we use identical generation settings, prompts, and seed schedules across all methods. Full training and generation settings are provided in Appendix A.

4.2 Main Results

Table 1 shows that PairReFL achieves the strongest score on all four human preference metrics. Since every method in this comparison is trained with a reward model learned from HPDv3 preference data, HPSv3 Ma et al. (2025) is the in-domain metric, while HPSv2.1 Wu et al. (2023),

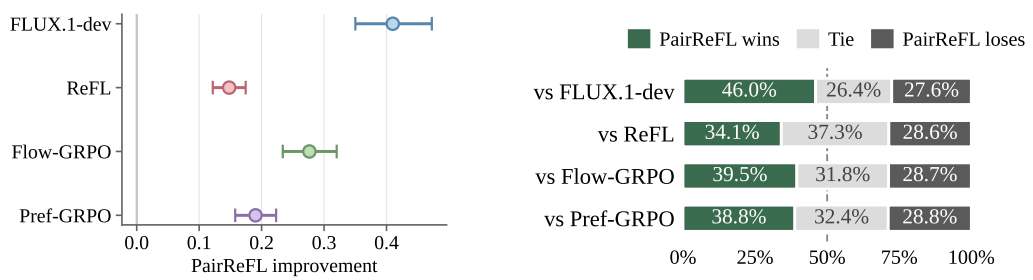


Figure 4: **Left:** PairReFL improvements in HPSv3 score over baselines. PairReFL achieves statistically significant gains over all compared models. **Right:** Human GSB evaluation of PairReFL against baselines. Human evaluators consistently favor PairReFL over all compared models.

PickScore Kirstain et al. (2023), and ImageReward Xu et al. (2023) are held-out preference models trained on different data. Averaged over the four metrics, PairReFL improves over FLUX.1-dev Black Forest Labs (2024) by 3.6% relative, against 2.7% for ReFL Xu et al. (2023), 2.0% for Pref-GRPO Wang et al. (2025b), and 1.1% for Flow-GRPO Liu et al. (2026), so the gain of PairReFL is $1.3\times$ that of ReFL and $1.8\times$ that of Pref-GRPO. The composition of that average matters as much as its size. ReFL, the strongest baseline on most metrics, falls below the unadapted model on the held-out PickScore, so its gains come at the cost of another preference metric, a pattern the training dynamics below will identify as reward hacking. PairReFL is the only method that improves the in-domain metric and all three held-out metrics simultaneously. Because every method is evaluated on the same prompts and seeds, significance can be assessed with a prompt-paired bootstrap; PairReFL's lead over the strongest baseline excludes zero on HPSv2.1, HPSv3, and ImageReward, and only the PickScore margin over Pref-GRPO is inconclusive (Appendix B.2). Figure 4 visualises the paired HPSv3 gains.

The four preference metrics in Table 1 are themselves learned models, so we complement them with a GSB human study in which annotators view a PairReFL image next to an image from the pretrained model or a baseline and judge it as good, same, or bad relative to its counterpart (protocol in Appendix B.3). PairReFL wins more comparisons than it loses against every opponent (Figure 4), with relative win rates, the good rate minus the bad rate, of 18.4 points over FLUX.1-dev, 5.5 over ReFL, 10.8 over Flow-GRPO, and 10.0 over Pref-GRPO.

The training dynamics provide a separate diagnostic of reward hacking. Figure 2 tracks dense evaluations on an additional FLUX.1-dev experiment. For a fair cross-method comparison, the left panel scores all three methods during training with the same pointwise HPSv3 rather than with PairReFL's differently scaled training logit, and the right panel tracks the held-out PickScore on the same fixed subset. ReFL's HPSv3 rises while its PickScore falls almost in mirror image, with a correlation of -0.94 across checkpoints, so its PickScore deficit in Table 1 is the end point of a steady trade rather than noise. PairReFL gains roughly $1.7\times$ as much HPSv3 as ReFL over the same budget while its PickScore holds steady, and Flow-GRPO improves on both axes but gains the least.

4.3 PairHPSv3 Accuracy

We separately evaluate whether PairHPSv3 recovers held-out human preferences. Since its score depends on image order, for an annotated pair (x^+, x^-) we use the antisymmetrized margin

$$\Delta_\phi(c, x^+, x^-) = \frac{1}{2}\left[R_\phi(c; x^-, x^+) - R_\phi(c; x^+, x^-)\right] \tag{10}$$

and count the pair as correct when $\Delta_\phi > 0$. We evaluate broad human preference on HPDv2 (Wu et al., 2023) and HPDv3 (Ma et al., 2025), the image-generation split of GenAI-Bench (Jiang et al., 2024) and MJ-Bench (Chen et al., 2026). For GenAI-Bench, we report accuracy on its 1,306 decisive left/right votes and exclude 429 tie or both-bad votes, which scalar reward models do not predict. Pointwise models score each image independently, while PairHPSv3 uses the antisymmetrized comparison above.

PairHPSv3 matches the best result on HPDv2 and improves the strongest prior result from 76.9% to 77.7% on HPDv3. On GenAI-Bench, PairHPSv3 obtains 69.4%, while HPSv3 and PickScore lead at

Table 3: **Pair-construction ablations on the HPDv2 test set.** Higher is better. The independent-noise row is greyed because its higher scores are not supported by the drift and appearance diagnostics of Section 4.5.1 and Appendix B.7.

Setting	HPSv2.1	HPSv3	PickScore	ImageReward
Reference anchor	0.3140	13.7638	22.7975	1.0934
Independent initial noise	0.3282	14.0233	22.6161	1.1781
PairReFL (Ours)	0.3182	13.9677	22.8348	1.1252

Table 4: **Reliability of a preference decision on shared-noise pairs.** Flip rate is the percentage of pairs whose preferred image changes under a preference-preserving perturbation, averaged over three perturbations; the signal ratio is the median margin divided by the median perturbation-induced change in that margin, on FLUX.1-dev pairs. HPDv3 accuracy is repeated from Table 2. Intervals are in Table 11.

	HPDv3	Flip rate (%) ↓		Signal
Scorer	acc. (%) ↑	FLUX.1-dev pairs	PairReFL pairs	ratio ↑
PairHPSv3 (joint)	**77.7**	**9.00**	8.17	4.41
HPSv3 difference	76.9	13.58	8.38	1.78
PickScore difference	65.6	7.79	**5.58**	5.79
HPSv2.1 difference	65.3	**6.08**	6.38	**6.50**
ImageReward difference	58.6	9.92	6.21	6.37
CLIP difference	48.6	8.58	8.92	5.10

72.7% and 72.4%, respectively. On MJ-Bench, PairHPSv3 reaches 98.9% against 97.1% for HPSv3 and 94.6% for HPSv2.1, reducing the error rate relative to HPSv3 by 62%, and the improvement is significant under an exact paired McNemar test ($p < 0.001$). The MJ-Bench split isolates paired differences in blur and visual artifacts, where joint comparison is particularly useful. These results support PairHPSv3 as a preference and visual-quality reward.

4.4 Ablation Studies

To isolate what each part of the pair construction contributes, we ablate the two design choices, training each variant under the same optimization budget and evaluating it with the protocol of the main results. The first ablation replaces each online stochastic peer with a fixed reference-model anchor, removing the moving target supplied by the current policy. The second samples the two images of a pair from independent initial latents, removing the shared content that keeps the comparison away from unrelated differences in layout. Table 3 reports all three variants. Stochastic pairing improves every metric over the reference anchor. The independent-noise checkpoint scores higher on HPSv2.1, HPSv3, and ImageReward and lower on PickScore, but its apparent HPSv3 lead over PairReFL is not statistically significant under the prompt-paired bootstrap (Appendix B.2). Taken alone this pattern is merely inconclusive; the drift diagnostics of Section 4.5.1 and the appearance diagnostics of Appendix B.7 show that it is in fact a reward-hacking shortcut.

4.5 Analysis

4.5.1 PairReFL Converts Drift into Preference Gain More Efficiently

PairReFL attains the best score on all four preference metrics in Table 1, but a higher score can come either from better alignment or from moving further away from a well-behaved prior, and the table cannot separate the two. We therefore measure how far PairReFL and each baseline move from FLUX.1-dev and how much sample variety they retain, reusing the images of the main results. With DINOv2 embeddings, *drift* is the mean cosine distance to the FLUX.1-dev image for the same prompt and seed, and *diversity* is the mean pairwise cosine distance among the seeds of a prompt (Appendix B.5).

Gain and drift increase together across methods, so no run dominates another on the raw axes (Figure 5, left), and the methods separate only in the rate at which they exchange one for the other.

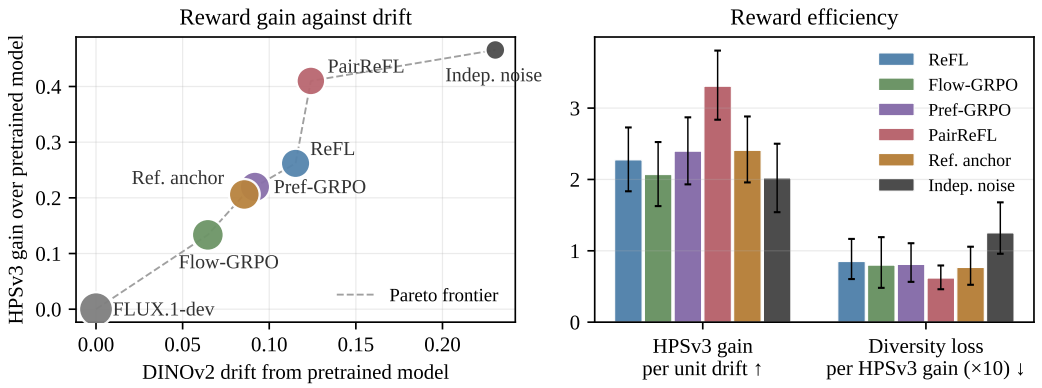


Figure 5: **Left:** HPSv3 gain against DINOv2 drift from the pretrained model; marker area grows with diversity retention. Gain and drift increase together, so no run is dominated and the dashed frontier passes through all of them. **Right:** the two efficiency ratios with 95% bootstrap intervals over prompts, which is where the methods separate. PairReFL has the highest gain per unit of drift and the lowest diversity loss per unit of gain; the independent-noise variant is worst on both.

PairReFL converts drift into preference gain at 3.31 HPSv3 points per unit of drift, 37% above the strongest baseline and significantly above every baseline under the prompt-paired bootstrap, and it pays the least diversity per unit of gain (Figure 5, right). This rate is not an artifact of drift magnitude, since Flow-GRPO drifts the least of all runs yet ranks second worst, and it holds under CLIP ViT-H/14 embeddings (Appendix B.5). The independent-noise variant drifts by far the furthest and has the lowest gain per unit of drift of any run, which shows that its higher raw scores in Table 3 come from a shift toward saturated, glossy rendering (Appendix B.7) rather than from a better model.

4.5.2 Joint Comparison Gives a More Stable Verdict than Score Differencing

Table 2 evaluates the comparator on annotated pairs drawn from different generators, whereas the pairs that train PairReFL come from one prompt and one initial latent, so they are far harder to tell apart and carry no human label. To test whether the verdict remains reliable on such pairs, we replay the training rollout and collect pairs from FLUX.1-dev and from the PairReFL checkpoint. Without labels we cannot measure accuracy, so we apply the same mild perturbation to both images of a pair, which leaves the preference relation intact, and record how often the preferred image changes. We also divide the median margin a scorer assigns by the median change that the perturbations induce in it, which gives a unitless signal-to-nuisance ratio (Appendix B.9).

PairHPSv3 and HPSv3 are both Qwen2-VL-7B fine-tuned on HPDv3, so Table 4 isolates the interface, a joint comparison against a score difference. On FLUX.1-dev pairs the joint verdict flips on 9.0% of pairs, 34% less often than the score difference at 13.6%, a gap that is significant under the prompt-paired bootstrap, and its margin stands 4.4 times above the nuisance level against 1.8 times for HPSv3. Differencing two independently calibrated scores on images this similar leaves a quantity barely larger than the noise floor, while conditioning on both images at once does not. PickScore and HPSv2.1 differences are more stable still, but they reach only about 65% accuracy on HPDv3 against 77.7% for PairHPSv3, so their stability belongs to a signal weakly related to human preference. Among scorers accurate enough to drive a generator, PairHPSv3 is the stable one.

5 Conclusion

In this paper, we propose PairReFL, which keeps the gradient pathway of direct reward feedback but replaces the pointwise target with the verdict of a differentiable pairwise comparator against a detached peer sampled from the current model. This removes the incentive to exploit calibration errors of the reward model and to chase an ever higher absolute score. On FLUX.1-dev, PairReFL outperforms pointwise reward feedback and policy-gradient baselines on most widely-adopted human preference metrics. We hope this work establishes pairwise reward feedback as a practical route for optimizing visual generative models to better align human preferences.

AI USE STATEMENT

We primarily used LLMs to polish the writing of this paper. All LLM-assisted edits were reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

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

Zhaorun Chen, Zichen Wen, Yichao Du, Yiyang Zhou, Chenhang Cui, Siwei Han, Jen Weng, Chaoqi Wang, Zhengwei Tong, Leria HUANG, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *Advances in Neural Information Processing Systems*, 38, 2026.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, volume 2024, pp. 4793–4822, 2024.

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

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

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2024.

Hyeongyu Kang, Jaewoo Lee, Woocheol Shin, Kiyoung Om, and Jinkyoo Park. Diffusion fine-tuning via reparameterized policy gradient of the soft q-function. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 78690–78725, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7f072dfceefc8e8f88e2b198529f8706-Paper-Conference.pdf.

Kwanyoung Kim and Byeongsu Sim. Reward sharpness-aware fine-tuning for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36051–36061, June 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Ruihang Li, Leigang Qu, Jingxu Zhang, Dongnan Gui, Mengde Xu, Xiaosong Zhang, Han Hu, Wenjie Wang, and Jiaqi Wang. Genarena: How can we achieve human-aligned evaluation for visual generation tasks? *arXiv preprint arXiv:2602.06013*, 2026.

Zhanhao Liang, Tao Yang, Jie Wu, Chengjian Feng, and Liang Zheng. Leapalign: Post-training flow matching models at any generation step by building two-step trajectories. *arXiv preprint arXiv:2604.15311*, 2026.

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

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Nihar B. Shah, Sivaraman Balakrishnan, Joseph K. Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. When is it better to compare than to score? *arXiv preprint arXiv:1406.6618*, 2014.

Xiangwei Shen, Zhimin Li, Zhantao Yang, Shiyi Zhang, Yingfang Zhang, Donghao Li, Chunyu Wang, Qinglin Lu, and Yansong Tang. Directly aligning the full diffusion trajectory with fine-grained human preference. *arXiv preprint arXiv:2509.06942*, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

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

Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 71440–71462, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Ruoyu Wang, Boye Niu, Xiangxin Zhou, Yushi Huang, Tongliang Liu, and Chi Zhang. Exploring the design space of reward backpropagation for flow matching. *arXiv preprint arXiv:2606.11075*, 2026.

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

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025c.

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

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

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8941–8951. IEEE, 2024.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

A Experimental Details

Generator training. We initialize all fine-tuned methods from FLUX.1-dev and train on HPDv2 training prompts for 1,000 optimization steps, reporting the final checkpoints. We use full-parameter bfloat16 training with AdamW and a learning rate of 10^{-5}. Online rollouts use 512×512 images, 10 sampling steps, and guidance scale 1.0. PairReFL uses groups of eight, 16 unique prompts per step, CPS noise level 0.7, $\tau = 0.9$, and a frozen Qwen2-VL-7B PairHPSv3 comparator. The differentiable suffix begins at a uniformly sampled normalized timestep in $[0.625, 1.0]$.

PairHPSv3 training. We fully fine-tune Qwen2-VL-7B on the HPDv3 training split for two epochs with a global batch size of 384 and a learning rate of 2×10^{-6}. We use the binary objective in Equation 3 and the final checkpoint as PairReFL's frozen comparator.

Evaluation. Following the common evaluation protocol (Xu et al., 2023; Liang et al., 2026), we generate all images at 720×720 resolution with 50 sampling steps and classifier-free guidance scale 3.5. For preference alignment, we generate four images for each of the HPDv2 test prompts and report HPSv2.1, HPSv3, PickScore, and ImageReward scores.

B Analysis Protocols and Derivations

B.1 Shared-Noise Variance Decomposition

Let $Q(Z, \xi)$ be any scalar statistic of a generated image, where Z is the initial latent and ξ collects the subsequent stochastic innovations. Write $\mu(Z) = \mathbb{E}_\xi[Q(Z, \xi) \mid Z]$ and $v(Z) = \mathrm{Var}_\xi[Q(Z, \xi) \mid Z]$. For branches that are independent and identically distributed conditional on their initial latent, the variance of their difference is

$$\begin{aligned} \mathrm{Var}(Q(Z_b, \xi_b) - Q(Z_a, \xi_a)) &= 2\,\mathrm{Var}_Z[\mu(Z)] + 2\mathbb{E}_Z[v(Z)], && Z_a \perp Z_b, \\ \mathrm{Var}(Q(Z, \xi_b) - Q(Z, \xi_a)) &= 2\mathbb{E}_Z[v(Z)], && \text{shared } Z. \end{aligned} \tag{11}$$

Sharing Z removes the nonnegative between-seed term $2\,\mathrm{Var}_Z[\mu(Z)]$ while retaining variation from independent CPS innovations. The statement applies to scalar projections of the pairwise signal, so it motivates but does not assert lower variance for every nonlinear gradient coordinate.

B.2 Prompt-Paired Bootstrap

For every method we recover the four per-image scores associated with each HPDv2 test prompt and average them before comparison, for all four preference metrics. Per-image HPSv3 scores



Figure 6: **Qualitative examples of reward hacking.** Each column shows a base-model output (top) and its corresponding output after RL (bottom) for the same prompt. Reward optimization can amplify score-increasing visual patterns such as star-like particles, ornamental decoration, floral or fragmented textures, golden contrast, and glossy plastic-like highlights without a corresponding improvement in the requested content.

are read back from saved pair predictions, ImageReward scores from the saved per-image file, and HPSv2.1 and PickScore are recomputed per image with the same checkpoints and the raw PickScore scale used in Table 1. For each baseline, we form 400 paired prompt-level differences against PairReFL. The reported interval is the 2.5th–97.5th percentile range of 100,000 bootstrap means obtained by resampling prompts with replacement. The prompt, rather than the individual image, is the resampling unit; this retains dependence among the four deterministic seeds for a prompt.

Table 5: 95% percentile bootstrap intervals for each prompt-level mean in Table 1. These intervals carry prompt-to-prompt variation and therefore overlap heavily; the paired differences in Table 6 remove that variation.

Method	HPSv2.1	HPSv3	PickScore	ImageReward
FLUX.1-dev	$[0.3044, 0.3114]$	$[13.350, 13.759]$	$[22.651, 22.923]$	$[0.976, 1.110]$
ReFL	$[0.3121, 0.3186]$	$[13.627, 14.008]$	$[22.627, 22.900]$	$[1.046, 1.179]$
Flow-GRPO	$[0.3078, 0.3147]$	$[13.491, 13.887]$	$[22.653, 22.926]$	$[1.000, 1.135]$
Pref-GRPO	$[0.3114, 0.3182]$	$[13.581, 13.970]$	$[22.679, 22.954]$	$[1.017, 1.151]$
PairReFL	$[0.3149, 0.3215]$	$[13.778, 14.153]$	$[22.696, 22.973]$	$[1.057, 1.190]$
Reference anchor	$[0.3107, 0.3174]$	$[13.567, 13.957]$	$[22.660, 22.934]$	$[1.026, 1.159]$
Independent noise	$[0.3251, 0.3313]$	$[13.824, 14.217]$	$[22.473, 22.758]$	$[1.112, 1.241]$

B.3 Human GSB Evaluation

Figure 4 follows the GSB protocol, which compares two models one image pair at a time and records whether the first image is good, same, or bad relative to the second. We sample 500 pairs from the evaluation images of the main results; each pair places a PairReFL image next to the image generated by FLUX.1-dev, ReFL, Flow-GRPO, or Pref-GRPO for the same prompt and seed, with 125 pairs per opponent, and the two images appear in randomized left and right positions. Forty annotators each label 100 pairs, so the 4,000 verdicts cover each pair eight times on average. On these overlapping annotations, two annotators viewing the same pair give the same verdict 87% of the time. The percentages in Figure 4 aggregate all verdicts for each opponent.

B.4 Order-Consistency Metrics

We run PairHPSv3 twice per image pair and save both logits. With $R_{ab} = R(c; x_a, x_b)$ and $R_{ba} = R(c; x_b, x_a)$, the antisymmetric and symmetric components are $a = (R_{ba} - R_{ab})/2$ and $s = (R_{ba} + R_{ab})/2$. Swap consistency is $\Pr[\mathrm{sign}(R_{ab}) = -\mathrm{sign}(R_{ba})]$, the reverse-logit correlation is the

Table 6: Prompt-paired difference of PairReFL against every other run, with 95% percentile bootstrap intervals over 400 prompts. Only the PickScore comparison against Pref-GRPO fails to exclude zero.

Baseline	HPSv2.1	HPSv3	PickScore	ImageReward
FLUX.1-dev	$0.0103\,[0.0095, 0.0112]$	$0.410\,[0.350, 0.473]$	$0.047\,[0.020, 0.075]$	$0.081\,[0.063, 0.099]$
ReFL	$0.0029\,[0.0025, 0.0032]$	$0.148\,[0.122, 0.175]$	$0.071\,[0.056, 0.085]$	$0.011\,[0.003, 0.020]$
Flow-GRPO	$0.0069\,[0.0063, 0.0076]$	$0.277\,[0.234, 0.320]$	$0.045\,[0.023, 0.068]$	$0.057\,[0.043, 0.071]$
Pref-GRPO	$0.0034\,[0.0029, 0.0039]$	$0.190\,[0.158, 0.223]$	$0.018\,[0.000, 0.036]$	$0.040\,[0.028, 0.052]$
Reference anchor	$0.0042\,[0.0037, 0.0047]$	$0.204\,[0.172, 0.237]$	$0.037\,[0.020, 0.055]$	$0.032\,[0.020, 0.044]$
Independent noise	$-0.0100\,[-0.0110, -0.0090]$	$-0.056\,[-0.127, 0.018]$	$0.219\,[0.178, 0.260]$	$-0.053\,[-0.072, -0.034]$

Table 7: PairHPSv3 order consistency. "Swap" is the percentage of pairs with opposite-sign logits after exchanging the images and "order gap" is the absolute difference between the two corresponding decision rates. The two shared-noise dumps have no human preference label.

Dataset	Pairs	Swap (%)	$\rho(R_{ba}, -R_{ab})$	Order gap (pp)	$\mathbb{E}\lvert s\rvert/\mathbb{E}\lvert a\rvert$
HPDv2	15,300	98.20	0.985	0.19	0.096
HPDv3	14,400	95.40	0.982	2.69	0.123
GenAI-Bench	1,735	94.76	0.983	2.53	0.109
Shared-noise pairs (FLUX.1-dev)	800	93.12	0.981	0.88	0.096
Shared-noise pairs (PairReFL)	800	94.62	0.980	2.12	0.104

Pearson correlation between R_{ba} and $-R_{ab}$, and the order gap is $|\Pr(R_{ab} < 0) - \Pr(R_{ba} > 0)|$. For annotated pairs $x_a = x^+$ and $x_b = x^-$; for generated pairs the first image is an arbitrary designation. Table 7 reports these statistics on the annotated benchmarks together with the two shared-noise dumps of Section 4.5.2. The comparator is closer to antisymmetric on the shared-noise pairs than the earlier row of independently seeded evaluation images suggested, which is the distribution that matters for the stop-gradient argument.

B.5 Drift and Diversity

All methods share the prompt list and the seed schedule, so drift is defined per image against the FLUX.1-dev image for the same prompt and seed. Diversity is computed among the four seeds of a prompt, and both quantities are averaged within a prompt before prompts are resampled. Table 8 reports the raw values and the two efficiency ratios. PairReFL moves further from FLUX.1-dev than every baseline and is compensated proportionally more for the move, while its diversity retention matches that of ReFL. Its diversity loss per unit of gain is the lowest of all runs, significantly below ReFL and Pref-GRPO under the prompt-paired bootstrap and on par with Flow-GRPO and the reference anchor.

Replacing DINOv2 with CLIP ViT-H/14 image embeddings changes the absolute scale of both distances but leaves the conclusions of Table 8 intact. Diversity retention becomes 0.900 for PairReFL against 0.925 (ReFL), 0.963 (Flow-GRPO), 0.929 (Pref-GRPO), 0.940 (reference anchor), and 0.799 (independent noise), preserving the DINOv2 order. The HPSv3 gain per unit of drift becomes 5.17 for PairReFL against 3.51, 3.63, 4.00, 3.85, and 2.71 respectively, and the diversity loss per unit of gain becomes 0.0427 for PairReFL, the lowest of all runs, against 0.0754 for the independent-noise variant, the highest. PairReFL and the independent-noise variant therefore keep the two extreme positions under either encoder, while the four remaining baselines reorder among themselves on the two efficiency ratios.

B.6 Confidence-Cap Sweep

A direction of a pair is active when $\sigma(r) < \tau$ for its directed logit, so both the number of active directions and the summed reward-space slope $\sigma(r)(1 - \sigma(r))$ can be evaluated for any τ from the two logits already computed for Section 4.5.2. Table 10 gives the sweep on both shared-noise dumps. Between $\tau = 0.8$ and $\tau = 1$ the summed slope changes by less than a tenth while the fraction of pairs contributing both directions moves from a third to all of them, which is why we describe the cap as a selector rather than a magnitude limiter.

Table 8: Drift from the unadapted model and sample diversity on the HPDv2 test set, from DINOv2 embeddings of the images used in Table 1. Retention is diversity relative to FLUX.1-dev. The last two columns divide the prompt-paired HPSv3 gain by drift and the diversity loss by that gain, so higher is better in the first and lower in the second. Bootstrap intervals are in Table 9.

Method	Diversity	Retention	Drift	HPSv3 gain / drift	Div. loss / gain
FLUX.1-dev	0.2984	1.000	—	—	—
ReFL	0.2762	0.925	0.1152	2.27	0.0849
Flow-GRPO	0.2877	**0.964**	0.0646	2.07	0.0799
Pref-GRPO	0.2806	0.940	0.0917	2.40	0.0809
Reference anchor	0.2826	0.947	0.0856	2.41	0.0766
PairReFL	0.2731	0.915	0.1240	**3.31**	**0.0618**
Independent noise	0.2402	0.805	0.2305	2.02	0.1250

Table 9: Bootstrap intervals for the DINOv2 columns of Table 8, resampling prompts.

Method	Retention	Drift	HPSv3 gain / drift
ReFL	$[0.907, 0.944]$	$[0.1070, 0.1235]$	$[1.83, 2.73]$
Flow-GRPO	$[0.951, 0.977]$	$[0.0587, 0.0708]$	$[1.63, 2.52]$
Pref-GRPO	$[0.924, 0.957]$	$[0.0841, 0.0996]$	$[1.93, 2.87]$
Reference anchor	$[0.931, 0.963]$	$[0.0787, 0.0926]$	$[1.96, 2.88]$
PairReFL	$[0.895, 0.935]$	$[0.1154, 0.1329]$	$[2.84, 3.81]$
Independent noise	$[0.778, 0.832]$	$[0.2178, 0.2432]$	$[1.54, 2.50]$

B.7 Quantifying the Reward-Hacking Signature

The failure mode we observed among the independent-noise samples with the largest HPSv3 gains is a glossy "plastic" appearance: smooth surfaces, strong specular highlights, saturated colours, and toy-like rendering. The style survives even when the prompt names a medium explicitly, with pixel art becoming a smooth illustration and a silkscreen print becoming a glossy portrait (Figure 7). Selected examples cannot establish how widespread this is, so we measure it over all $1{,}600$ images of every run. We compute mean chroma, Hasler–Süsstrunk colourfulness, the share of spectral power above a quarter of the Nyquist frequency, and two CLIP probe axes contrasting glossy against matte descriptions and photographic against illustrative ones. We also test medium fidelity directly: on the 53 test prompts that name exactly one medium, we ask whether the CLIP nearest neighbour among 20 medium descriptions is the requested one.

Every reward-tuned run moves in the same direction, and PairReFL is no exception, so the interesting quantity is the magnitude. Relative to FLUX.1-dev, PairReFL shifts chroma by $+0.0351$, colourfulness by $+0.0223$, high-frequency energy by $+0.0078$, and the glossy axis by $+0.0081$, all in line with ReFL, Pref-GRPO, and the reference anchor. The independent-noise variant shifts by $+0.0724$, $+0.0629$, $+0.0260$, and $+0.0327$: two to four times further on every statistic, with non-overlapping intervals (Figure 8). Medium fidelity falls from 0.5283 for FLUX.1-dev to 0.4811 for PairReFL (-0.0472, $[-0.0943, -0.0047]$) and to 0.4151 for the independent-noise variant (-0.1132, $[-0.1934, -0.0425]$), so the variant loses roughly 2.4 times as much prompt-specified medium. The independent-noise run therefore buys its higher HPSv3 and ImageReward with a systematic move toward saturated, high-contrast, generically attractive rendering at the cost of the constraints the prompt asked for, which is the signature of reward hacking rather than improved fidelity. The medium-fidelity subset is small, and we report it as corroboration rather than as a primary result.

B.8 Appearance Statistics

Chroma is the mean of $(\max_c - \min_c)/\max_c$ over RGB channels. Colourfulness is the Hasler–Süsstrunk statistic on the opponent channels $r - g$ and $\frac{1}{2}(r + g) - b$. High-frequency energy is the share of power spectral density beyond a quarter of the Nyquist radius of a mean-subtracted luminance image. The blown-highlight fraction counts pixels with value above 0.95 and chroma below 0.15. Images are resized to 512×512 before these statistics are computed. The glossy and photographic axes are differences of mean cosine similarity between the CLIP ViT-H/14 image

Table 10: Confidence-cap sweep on 800 shared-noise pairs per source. "Both" is the percentage of pairs with two active directions and "slope" is the summed $\sigma(1-\sigma)$ over active directions, which is proportional to the reward-space gradient the pair contributes.

	FLUX.1-dev pairs			PairReFL pairs		
τ	Active dirs	Both (%)	Slope	Active dirs	Both (%)	Slope
0.5	1.009	3.9	0.1085	1.021	3.8	0.1414
0.6	1.111	11.2	0.1339	1.154	15.5	0.1740
0.7	1.219	21.9	0.1583	1.281	28.1	0.2030
0.8	1.327	32.8	0.1787	1.425	42.5	0.2299
0.9	1.490	49.0	0.1989	1.600	60.0	0.2512
0.95	1.647	64.8	0.2095	1.715	71.5	0.2591
1.0	2.000	100.0	0.2163	2.000	100.0	0.2648

Table 11: Bootstrap intervals for the flip rates of Table 4, and the test–retest Pearson correlation between the clean and perturbed margins. The correlation is high for every scorer, including HPSv3: its margin magnitude is reproducible, but on these pairs the magnitude is small enough that the sign is not.

	FLUX.1-dev pairs		PairReFL pairs	
Scorer	Flip rate (%)	Test–retest r	Flip rate (%)	Test–retest r
PairHPSv3 (joint)	$[7.58, 10.50]$	0.966	$[6.83, 9.54]$	0.956
HPSv3 difference	$[11.92, 15.29]$	0.972	$[7.04, 9.71]$	0.974
PickScore difference	$[6.54, 9.00]$	0.973	$[4.50, 6.75]$	0.983
HPSv2.1 difference	$[5.04, 7.17]$	0.986	$[5.25, 7.54]$	0.985
ImageReward difference	$[8.58, 11.29]$	0.978	$[5.17, 7.25]$	0.980
CLIP difference	$[7.29, 9.96]$	0.962	$[7.58, 10.33]$	0.966

embedding and two sets of three or four text probes each; the axes are reported as differences so that the arbitrary offset of CLIP similarity cancels. Medium fidelity uses the 53 test prompts containing exactly one entry of a 20-item medium vocabulary and asks whether the nearest probe "an image rendered as {medium}" is the requested one. All statistics are averaged within a prompt before prompts are resampled.

B.9 Nuisance Perturbations

The pairs are obtained by replaying the training rollout without gradients under the unchanged training configuration, which yields 800 pairs each from FLUX.1-dev and from the PairReFL checkpoint. The three perturbations are a JPEG round-trip at quality 75, a centre crop retaining 96% of each side followed by bicubic resizing back to the original resolution, and a global gamma adjustment with exponent 1.08. Each is applied identically to both images of a pair, so the preference relation is preserved by construction while the pixels change. For the joint comparator the decision is the sign of the antisymmetrised margin $\frac{1}{2}[R_\phi(c; x_a, x_b) - R_\phi(c; x_b, x_a)]$; for a pointwise model it is the sign of the score difference. Flip rates are averaged over the three perturbations for each pair before prompts are resampled. The signal ratio divides the median $|\cdot|$ of the clean margin by the median absolute change that the perturbations induce in that same margin, pooled over the three perturbations, so it is invariant to the arbitrary scale of each scorer.

On pairs from the PairReFL checkpoint every scorer becomes more stable and the two interfaces converge (Table 4), so the joint verdict helps most early in training, where the pairs are hardest. A perturbation test measures self-consistency rather than correctness. It ranks two interfaces on the same backbone, while agreement with human judgement is established separately by the held-out accuracies of Table 2.

B.10 Mechanism Diagnostics

We check three design claims of Section 3.3 on the shared-noise pair distribution described in Section 4.5.2.

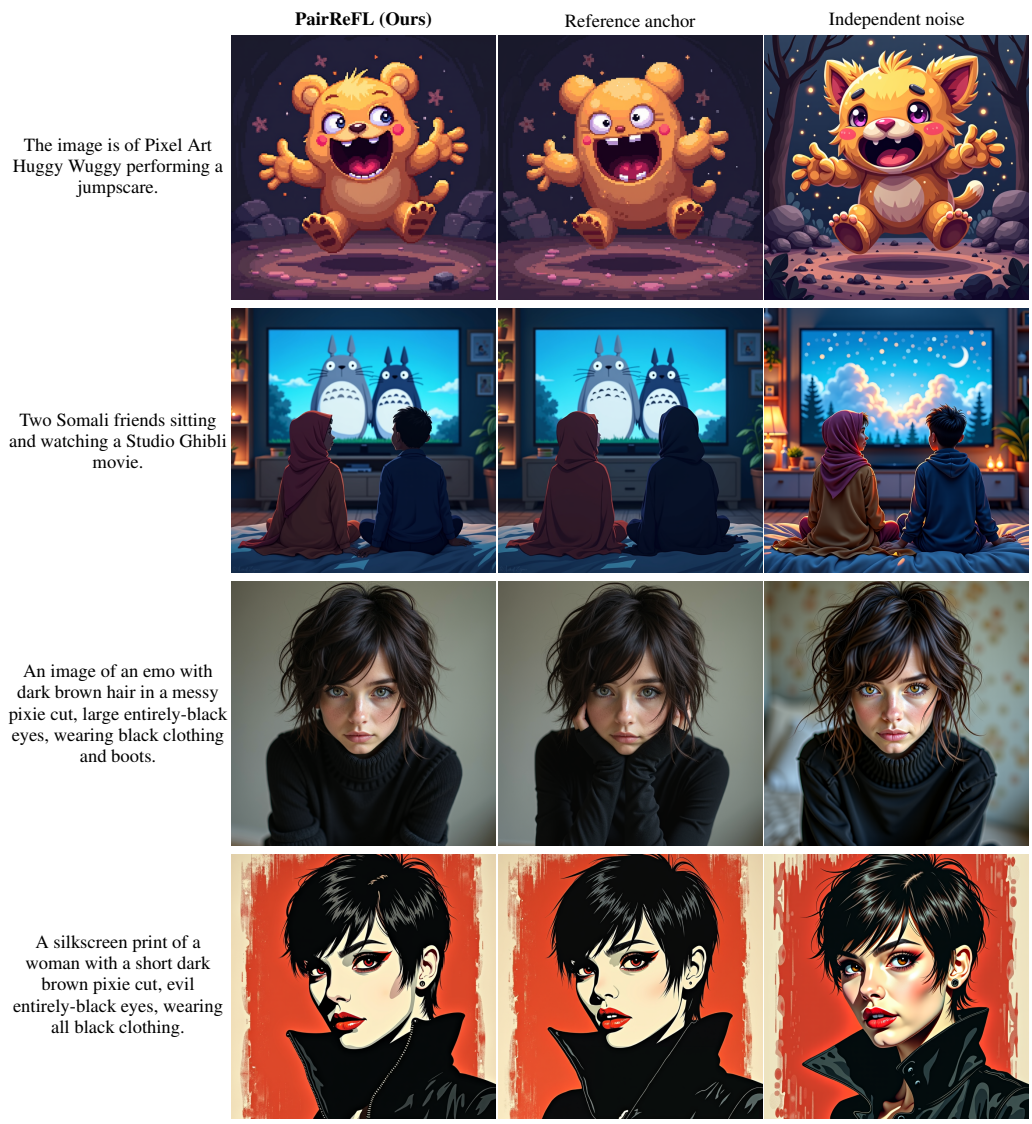


Figure 7: Qualitative comparison of the three pair-construction variants. Each row fixes the prompt and inference seed. The independent-noise outputs exhibit a recurring glossy, plastic appearance and weaker medium fidelity despite their high reward scores.

Order antisymmetry. The cancellation argument of Equation 9 assumes an order-antisymmetric comparator. On the dumped pairs PairHPSv3 produces opposite-sign logits after a swap for 93.1% of FLUX.1-dev pairs and 94.6% of PairReFL pairs, with a reverse-logit correlation of 0.980 and a symmetric component that is 9.6% and 10.4% of the antisymmetric one. The comparator is thus strongly but not exactly antisymmetric on precisely the pairs it scores during training, which is what the stop-gradient argument requires: bilateral stop-gradient removes the exact cancellation in the antisymmetric component without relying on residual position bias. Appendix B.4 reports the same statistics on the annotated benchmarks.

What the confidence cap actually selects. Because the two directed logits of a pair determine cap activity for any threshold, τ can be swept on the dumped pairs without retraining. At the $\tau = 0.9$ we use, both directions remain active for 49.0% of FLUX.1-dev pairs and 60.0% of PairReFL pairs, exactly one direction for the remainder, and no pair is fully deactivated; the mean number of active directions is 1.49 and 1.60 out of two. Removing the cap entirely ($\tau = 1$) raises this to two by

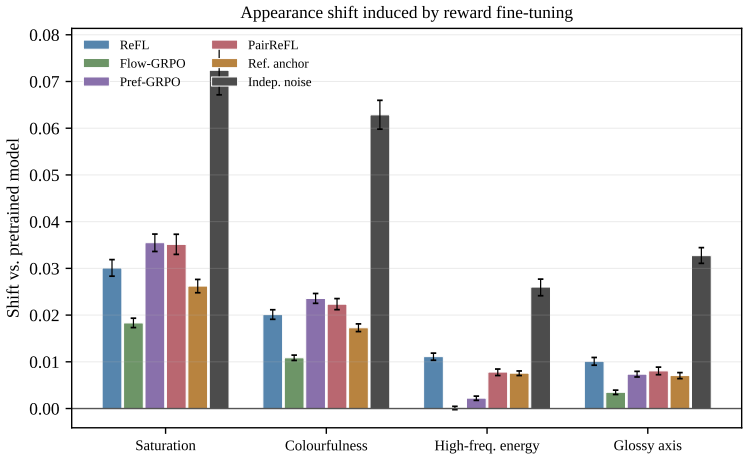


Figure 8: Shift of four appearance statistics relative to FLUX.1-dev over all 1,600 images per method, with 95% bootstrap intervals over prompts. Every reward-tuned run moves in the same direction; the independent-noise variant moves two to four times further than any other run on all four statistics.

Table 12: PairHPSv3 pixel-gradient diagnostics over 32 held-out prompts, four images each. Norm ratios use the bilateral-stop-gradient objective as denominator. Intervals are in Table 13.

Diagnostic	Value
Fully differentiated / bilateral-stop gradient norm	1.045
Same ratio for the exact antisymmetric component†	5.59×10^{-8}
Anchor-gradient cosine across three detached peers	0.828
Peer sensitivity $(1-\cos)$	0.172

†26 of 32 prompts with both cap terms active.

definition but increases the summed reward-space gradient magnitude by only 8.8% and 5.4%. For an active direction, $|\partial \ell_\tau(r)/\partial r| = \sigma(r)(1-\sigma(r)) \le \frac{1}{4}$; an inactive direction has zero derivative. The cap is therefore a selector over comparisons rather than a magnitude limiter: τ decides which of the two directed updates a pair still contributes. Pushing τ down to 0.5 collapses this to one active direction for 93% of pairs and discards roughly half the available signal, so the useful range is narrow and the value we chose sits inside it. Appendix B.6 gives the full sweep.

Stop-gradient and peer dependence in pixel space. The two remaining design claims are testable directly on the frozen comparator. Differentiating PairHPSv3 with respect to normalized RGB pixels over 32 held-out prompts (Table 12), we first replace it by its exact antisymmetric component and evaluate the fully differentiated bidirectional objective. On the 26 prompts where both cap terms are active, the resulting gradient norm is 5.59×10^{-8} of the bilateral-stop norm, which confirms Equation 9 numerically rather than only algebraically. With the unmodified comparator the ratio is 1.045, so real gradients do not vanish; but the only thing preventing the cancellation is the comparator's residual symmetric component, which is exactly the part of its output that carries position bias rather than preference. Stop-gradient removes the dependence on that accident. Second, holding the active image fixed and swapping the detached peer rotates its logit gradient by $1 - \cos = 0.172$ on average. A pointwise score difference cannot do this: its active-image gradient is by construction independent of the peer. This is the concrete sense in which the joint comparator supplies feedback that a difference of two scores cannot.

B.11 Pixel-Gradient Diagnostic

We resize saved PairReFL evaluation images to 448×448 and differentiate PairHPSv3 in bfloat16 with respect to normalized RGB pixels, using 32 held-out HPDv2 prompts, four images per prompt, and three alternative anchors per active image. For cancellation, the first two images of each prompt

Table 13: Bootstrap intervals for the pixel-gradient diagnostics of Table 12, over 10,000 prompt resamples.

Diagnostic	Mean	95% CI
Fully differentiated / bilateral-stop norm	1.045	$[0.827, 1.310]$
Antisym. component / bilateral-stop norm†	5.59×10^{-8}	$[2.64, 9.07]\times10^{-8}$
Active directions (of 2)	1.813	$[1.656, 1.938]$
Anchor-gradient cosine	0.828	$[0.774, 0.878]$
Peer sensitivity $(1 - \cos)$	0.172	$[0.122, 0.226]$

†26 prompts with both cap terms active.

form (x_a, x_b). We compare the bidirectional objective with both images attached, the same objective with each anchor detached as in PairReFL, and a no-stop objective formed from the exact antisymmetric component $(R_{ab} - R_{ba})/2$; ratios use the bilateral-stop gradient norm as denominator. For peer sensitivity, one active image is held fixed and compared against the other three, and we average the three pairwise cosines between active-image logit gradients. Because these prompts use the 720×720 evaluation images rather than the training-configuration dump, the cap activity measured here (1.813 of two directions active, both active for 84.6% of prompts) is not comparable to the sweep of Appendix B.6, which uses the shared-noise pairs; the evaluation images are drawn from independent seeds and are therefore easier to separate.