APD-FLOW: REWARD POST-TRAINING FOR FLOW-MATCHING IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-matching image restorers produce high-quality outputs with deterministic inference, but they are trained to match paired ground truth rather than human preferences or task-specific rewards. Reward post-training offers a way to close this gap. However, policy-gradient methods rely on stochastic exploration and action likelihoods, neither of which a deterministic sampler provides. We propose *APD-Flow*, a reward post-training framework that introduces exploration at the source state instead of the sampling trajectory. Given a degraded input, *APD-Flow* perturbs the source state in two opposite directions and restores both with the same deterministic sampler. The difference between the two restorations, weighted by their reward difference, defines an Antithetic Preference Direction (APD) that points toward the preferred output. The restorer is then optimized along this direction without action likelihoods or trajectory noise, and fidelity-aware reward adaptation keeps fidelity at the level of the pretrained restorer. The framework applies to restorers with arbitrary numbers of sampling steps and to non-differentiable rewards. Experiments on blind face restoration, blind image super-resolution, and scene-text restoration show that *APD-Flow* consistently improves the targeted perceptual or task-specific metrics while keeping the inference architecture and budget unchanged. Code and a video explanation are provided in the supplementary material.

1 INTRODUCTION

Image restoration seeks a perceptually plausible clean image that remains faithful to the observed content (Blau & Michaeli, 2018). Because the problem is ill-posed, a single degraded input admits many high-quality solutions that differ in texture and local detail. Recent flow-matching and rectified-flow restorers cast restoration as deterministic ordinary differential equation (ODE) transport from a source distribution to the conditional distribution of clean images. These models are trained on paired ground truth, and this supervision does not rank the feasible solutions by human preference or task utility. The resulting restorers therefore tend toward conservative, over-smoothed outputs that lack high-frequency detail (Wu et al., 2026).

Reward post-training can supply the missing preference signal (Yi et al., 2026; Wu et al., 2026). Learned quality assessors (Gu et al., 2020; Yang et al., 2022; Chen et al., 2024; Ke et al., 2021; Wang et al., 2023; Wu et al., 2024a; Zhao et al., 2026) and task-specific metrics such as text-recognition accuracy provide scalar feedback that can steer a pretrained restorer toward preferred outputs. However, likelihood-ratio methods such as PPO and GRPO assume stochastic actions with tractable likelihoods (Schulman et al., 2017; Shao et al., 2024). Once the degraded input y and the source state z_0 are fixed, a flow-matching restorer follows a deterministic ODE trajectory, and its intermediate velocities are network predictions rather than sampled actions (Liu et al., 2023; Lipman et al., 2023). The sampler therefore defines no per-step action likelihood, and a non-differentiable terminal reward also rules out backpropagation through the rollout. Flow-GRPO and FDFO explore by injecting noise along the trajectory (Liu et al., 2025; McAllister et al., 2026a), but this changes the sampler, perturbs the pretrained restoration path, and ties exploration to the solver schedule (Zhu et al., 2024). FDFO keeps the last step deterministic, so it cannot explore a one-step restorer at all.

We propose *APD-Flow*, a reward post-training framework for flow-matching image restoration built on Antithetic Preference Directions (APDs). *APD-Flow* explores at the source state and leaves the

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.

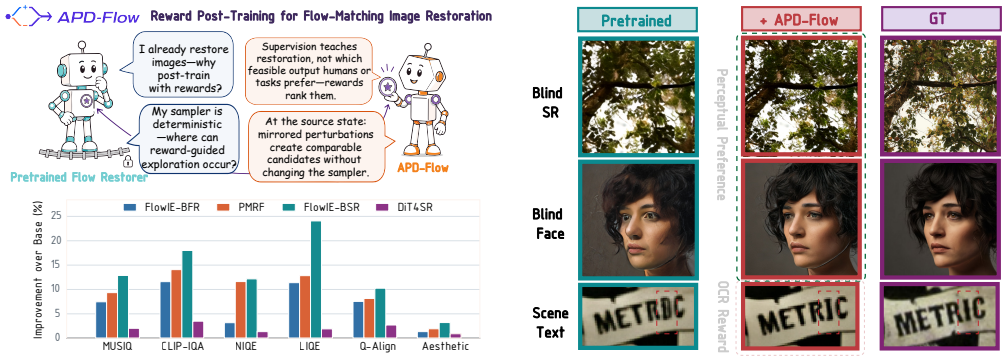

Figure 1: Overview of *APD-Flow*. Top left: source-state antithetic exploration for reward post-training. Bottom left: relative gains across four pretrained restoration backbones. Right: representative results from perceptual-preference and OCR reward optimization, with GT for reference.

ODE sampler deterministic. For each degraded input, mirrored perturbations of a shared base noise form an antithetic pair of source states, which the same sampler maps to two candidate restorations. The displacement between the two restorations, oriented and scaled by their signed reward difference, defines an APD that points toward the preferred output. Unlike group-based methods that draw several samples per input (Shao et al., 2024; Liu et al., 2025; Yi et al., 2026), *APD-Flow* needs only one antithetic pair. Because exploration happens before the rollout starts, the construction applies unchanged to restorers with one to fifteen ODE steps. Figure 1 summarizes the framework and representative results.

The APD is defined in output space, whereas model updates act on the velocity field along the ODE trajectory. A proxy ratio built from distances to APD-shifted velocity targets transfers the output-space APD to the velocity field without action likelihoods or trajectory noise. Because rewards enter only through forward evaluation, *APD-Flow* can optimize non-differentiable rewards directly.

Optimizing preference rewards alone, however, can erode restoration fidelity (Wu et al., 2026). *APD-Flow* therefore adds fidelity-aware reward adaptation, which reweights a composite fidelity reward on every batch so that fidelity stays at a target set by the pretrained restorer on validation data.

Our contributions are:

- We propose *APD-Flow*, a reinforcement learning framework for flow-matching image restoration that explores with an antithetic pair of source states instead of trajectory noise, so post-training leaves the pretrained sampler unchanged and applies to restorers with arbitrary step counts.
- We turn each pair into a velocity update whose regularizer constrains only the joint drift of the two branches, and we add fidelity-aware reward adaptation, a feedback controller that keeps fidelity at a target calibrated on the pretrained restorer.
- Across four pretrained restorers and three tasks, *APD-Flow* consistently improves the targeted perceptual or task-specific metrics with few post-training images.

2 Related Work

Flow-matching restoration. Early blind restorers balanced reconstruction fidelity and perceptual realism with adversarial and perceptual losses (Wang et al., 2021). Diffusion-based restorers later introduced stronger generative priors and one- or few-step variants (Wang et al., 2024; Lin et al., 2024; Wu et al., 2024c;b; Yue et al., 2025). Flow matching and rectified flow instead transport the source state to the restored image along short deterministic ODE trajectories (Liu et al., 2023; Lipman et al., 2023; Esser et al., 2024). FlowIE, PMRF, FlowSR, OFTSR, and DiT4SR reach strong quality with few ODE steps in blind face restoration and real-world image super-resolution (Zhu et al., 2024; Ohayon et al., 2025; Xu et al., 2025b; Zhu et al., 2026; Duan et al., 2025).

Reward post-training for generation and restoration. Reward post-training aligns generative models with human preferences through policy gradients, preference optimization, and group-relative

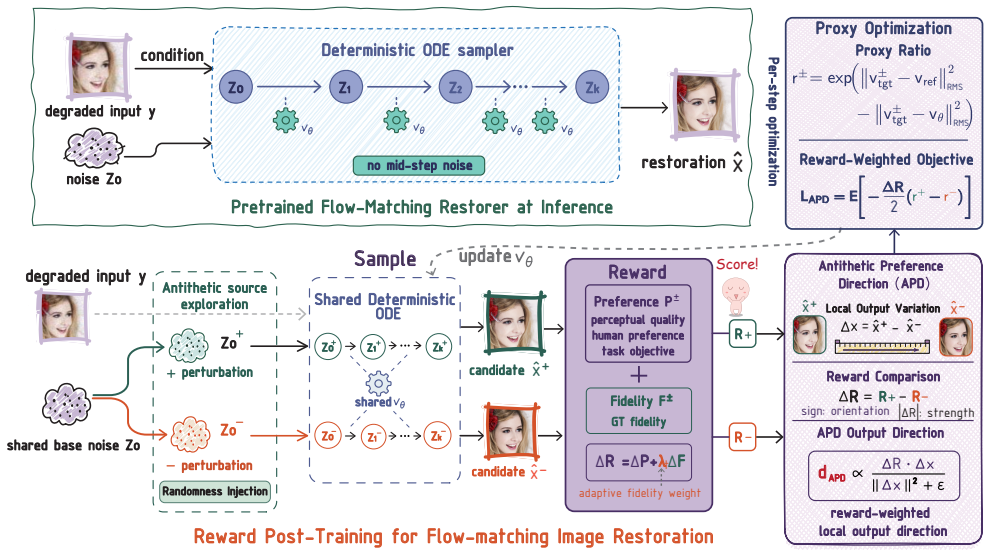

Figure 2: Overview of *APD-Flow*. It preserves deterministic ODE inference through source-state antithetic exploration rather than trajectory noise. The pair's output displacement and signed reward difference define an APD, which a training-only proxy ratio transfers to the velocity field. Because only forward reward evaluation is required, the reward may be non-differentiable.

updates (Black et al., 2024; Fan et al., 2023; Rafailov et al., 2023; Wallace et al., 2024). Flow-GRPO recasts deterministic ODE sampling as an SDE to introduce stochasticity for online reinforcement learning (Liu et al., 2025). FDFO instead pairs two trajectories that diverge through injected noise and pulls every velocity toward their normalized, reward-weighted difference (McAllister et al., 2026a). Reward Feedback Learning (ReFL) requires differentiable rewards and thus cannot directly optimize non-differentiable feedback (Wu et al., 2025). Reward post-training has also reached image restoration, but most existing methods target stochastic diffusion samplers rather than pretrained flow restorers with deterministic ODE samplers (Yi et al., 2026; Wu et al., 2026; 2025). In contrast, *APD-Flow* keeps the pretrained ODE sampler unchanged, explores with an antithetic pair at the source state, and controls fidelity against the pretrained restorer.

3 Preliminaries and Problem Setup

Deterministic flow-matching restoration. Given a degraded observation y, a conditional flow-matching restorer maps an initial state z_0, termed the source state, to a restored image. The source state may be pure Gaussian noise or an input-dependent estimate perturbed by Gaussian noise. Both $v_\theta(\cdot;\, y)$ and $\mathcal{G}_\theta(\cdot;\, y)$ are conditioned on y. The source state and the resulting deterministic rollout are given by

$$z_0 = m(y) + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad z_{k+1} = z_k + v_\theta(z_k, t_k;\, y)\Delta t_k, \quad \hat{x} = z_K = \mathcal{G}_\theta(z_0;\, y). \tag{1}$$

Here, $m(y) = 0$ gives a Gaussian source, while an input-dependent $m(y)$ gives an estimate-centered source. The states z_k and the output $\hat{x}$ lie in the restorer's own state space, which is either the pixel space or a VAE latent space. Rewards are computed on the decoded image, which we still write as $R(\hat{x}, x_{\mathrm{gt}})$. *APD-Flow* requires only that the trajectory be deterministic once (y, z_0) is fixed, and such a sampler defines no per-step action distribution for PPO- or GRPO-style likelihood ratios.

Preference among feasible restorations. Blind restoration is ill-posed, since a single degraded input admits many feasible clean images. We define the feasible restoration set and its reward-optimal solution as

$$\mathcal{X}(y) = \{x :\ D(\mathcal{A}(x), y) \le \eta,\ x \in \mathrm{supp}(p_{\mathrm{nat}})\}, \qquad x^\star(y) \in \arg\max_{x \in \mathcal{X}(y)} R(x, x_{\mathrm{gt}}). \tag{2}$$

Here, $\mathcal{A}$ is the underlying degradation process, D measures observation-space discrepancy, and p_{nat} is the natural-image distribution. In blind restoration, however, the degradation operator is unknown,

so the observation-consistency constraint cannot be used directly for post-training. We therefore measure fidelity against the ground truth with a composite score $F(\hat{x}, x_{\mathrm{gt}})$ whenever paired images are available. The terminal reward may encode any preference or task objective, from perceptual quality and identity similarity to text readability, and need not be differentiable. The fidelity objective in Sec. 4.3 keeps this adaptation at the fidelity of the pretrained restorer.

Reward post-training objective. Reward post-training maximizes the expected terminal reward:

$$\max_{\theta} \; J(\theta) = \mathbb{E}_{x_{\mathrm{gt}},y,z_0}\big[R\big(\mathcal{G}_\theta(z_0;\, y), x_{\mathrm{gt}}\big)\big]. \tag{3}$$

The ground truth x_{gt} is optional, since preference-based and task-specific rewards need not use paired references, and we omit fixed task annotations for clarity. *APD-Flow* addresses these limitations with source-state antithetic exploration and a proxy ratio that converts paired terminal reward differences into velocity-field updates.

4 METHOD

Figure 2 provides an overview of *APD-Flow*. Section 4.1 builds two comparable restorations per input, Section 4.2 turns their differences into velocity updates, and Section 4.3 keeps these updates within a fidelity target. Appendix A contains the formal derivations, while Appendix B.3 provides additional intuition and diagnostics.

4.1 SOURCE-STATE ANTITHETIC EXPLORATION

Each pretrained sampler enters only as a deterministic mapping $\mathcal{G}_\theta(\cdot;\, y)$, which covers different architectures and ODE step counts with one formulation. Figure 3(a) compares source-state antithetic exploration with Flow-GRPO-style trajectory noise (Liu et al., 2025) at several noise levels. The antithetic construction preserves input-conditioned content, whereas stronger trajectory noise increasingly distorts the restoration.

For each y, *APD-Flow* draws shared base noise ϵ and a perturbation direction δ, then constructs

$$z_0^{\pm} = m(y) + \sigma\big(\sqrt{1-w^2}\,\epsilon \;\pm\; w\,\delta\big), \qquad \epsilon, \delta \sim \mathcal{N}(0, I), \quad w \in (0, 1]. \tag{4}$$

The two states z_0^+ and z_0^- form an antithetic source pair: they share y and ϵ but are perturbed in opposite directions along δ, and the same deterministic sampler maps them to $\hat{x}^{\pm} = \mathcal{G}_\theta(z_0^{\pm};\, y)$. Each branch preserves the original source marginal $m(y) + \sigma\xi$, $\xi \sim \mathcal{N}(0, I)$, and thus remains compatible with the pretrained restorer. Because both candidates are generated from the same degraded input by the same deterministic sampler, their output difference isolates the change induced by the antithetic construction, while their reward difference indicates which direction is preferred. Independent source noises, by contrast, add unrelated output variation to the comparison, which obscures the cause of the reward difference and slows convergence (Fig. 7, right). Although the perturbations are opposite along δ, the two source states remain correlated because they share the same base noise ϵ. Because exploration acts only on the source state, the deterministic rollout stays unchanged for restorers with any ODE step count.

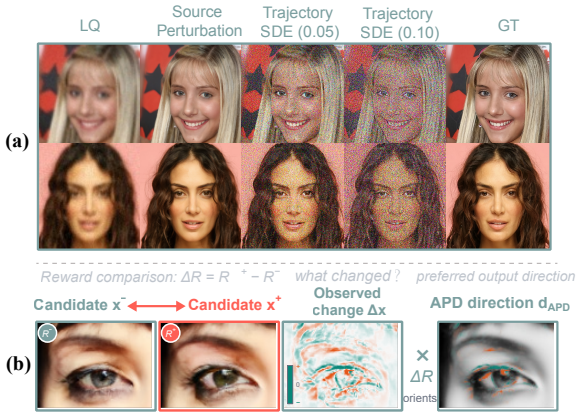

Figure 3: Source-state antithetic exploration better preserves input-conditioned content than trajectory noise at the tested noise levels (a). For an antithetic pair, Δx records the induced output change, while the signed reward difference ΔR determines its reward-oriented direction and scale (b).

4.2 Antithetic Preference Optimization

For the candidate restorations $\hat{x}^{\pm}$ generated from the antithetic pair in Eq. (4), let $R^{\pm} = R(\hat{x}^{\pm}, x_{\mathrm{gt}})$ denote their terminal rewards. For $a \in \mathbb{R}^N$, we define the dimension-normalized squared norm as $\|a\|^2_{\mathrm{RMS}} = \|a\|_2^2/N$. Let $\Delta R = R^+ - R^-$ and $\Delta x = \hat{x}^+ - \hat{x}^-$ denote the reward difference and output displacement, respectively. With a small stabilizer $\varepsilon > 0$, the Antithetic Preference Direction (APD) is defined as

$$\hat{d} = \frac{\Delta x}{\|\Delta x\|^2_{\mathrm{RMS}} + \varepsilon}, \qquad d_{\mathrm{APD}} = \Delta R\, \hat{d}. \tag{5}$$

Figure 3(b) visualizes the resulting output displacement Δx between the two candidate restorations. The displacement captures the restoration change induced by the antithetic construction, and the sign and magnitude of ΔR set its orientation and strength. The APD is thus a local pairwise preference direction rather than an estimate of the full reward gradient. Because the pair is antithetic, Δx and ΔR retain only the odd part of the sampler's response to δ, so content shared by the two branches cancels (Appendix A.1). Appendix A.2 provides a regularized minimum-norm interpretation of d_{APD}.

The APD is defined in output space, whereas optimization acts on velocity predictions along the ODE trajectory, where the deterministic sampler offers no policy likelihood ratio. *APD-Flow* bridges this gap by using the RMS-scaled displacement $\hat{d}$ to define branch-specific velocity targets and ΔR to weight the resulting updates. Because $\hat{d}$ lies in the same state space as the velocity, the rollout $\hat{x} = z_0 + \sum_k \Delta t_k\, v_\theta(z_k, t_k;\, y)$ of Eq. (1) gives $\partial \hat{x}/\partial v_k = \Delta t_k I$ when the rollout states are held fixed. An output-space direction is thus the same velocity-space direction up to the scalar step size Δt_k, which we absorb into the update scale so that all steps are weighted equally. At each ODE step k, let v_{ref} denote the frozen reference velocity from the rollout and v_θ the current prediction. We construct detached targets for the two branches as

$$v^{\pm}_{\mathrm{tgt}} = \big(v_{\mathrm{ref}} \pm \hat{d}\big)_{\mathrm{detach}}. \tag{6}$$

The target offsets depend only on $\hat{d}$, whereas ΔR only weights the objective and thus controls the update strength without moving the targets. We define a training-only proxy ratio that compares the current prediction and the rollout reference by their squared distances to a branch target:

$$r_\theta\big(v_{\mathrm{tgt}}\big) = \exp\Big(\big\|v_{\mathrm{tgt}} - v_{\mathrm{ref}}\big\|^2_{\mathrm{RMS}} - \big\|v_{\mathrm{tgt}} - v_\theta\big\|^2_{\mathrm{RMS}}\Big). \tag{7}$$

At each step k, $r_k^{\pm}$ denotes the corresponding proxy response for the two branch targets. By construction, $r_k^{\pm} = 1$ when $v_\theta = v_{\mathrm{ref}}$ and increases as v_θ approaches the corresponding branch target. Because the ratio measures progress relative to the frozen reference, several optimizer steps can reuse one rollout, as in PPO. The finite-difference direction and the shifted targets follow FDFO (McAllister et al., 2026a), and the distance-based ratio follows FPO (McAllister et al., 2026b). *APD-Flow* changes how the pair is drawn and how its two branches are regularized.

Using the two branch responses r_k^+ and r_k^-, we minimize

$$\mathcal{L}_{\mathrm{APD}}(\theta) = \mathbb{E}\left[-\frac{\Delta R}{2}\big(r_k^+ - r_k^-\big) + \beta|\Delta R| \left(\frac{r_k^+ + r_k^-}{2} - 1\right)^2 \right], \tag{8}$$

where the expectation is over antithetic pairs and ODE rollout steps and β controls the regularization strength. The first term is analogous to the PPO surrogate objective (Schulman et al., 2017), with ΔR acting as a pairwise advantage and $r_k^+ - r_k^-$ comparing the responses to the two branch targets. Unlike PPO, $r_k^{\pm}$ are distance-based training proxies rather than policy likelihood ratios. When $\Delta R > 0$, minimizing the first term increases $r_k^+ - r_k^-$, which moves the velocity prediction toward the $+$ target and away from the $-$ target. When $\Delta R < 0$, the same mechanism instead favors the $-$ target. In both cases, gradient descent moves the two branch velocities along the same direction d_{APD}, each evaluated at its own rollout state (Appendix A.4). At $v_\theta = v_{\mathrm{ref}}$, the gradient of the preference term is proportional to d_{APD} (Appendix A.4), which transfers the output-space preference direction to the velocity field. The second term is an SPO-inspired regularizer (Xie et al., 2025) that limits excessive updates away from the rollout reference. It penalizes only the mean response of the two branches, so the pair cannot drift jointly away from the reference, while their contrast, which carries the preference signal, remains unpenalized. Appendices A.3 and A.4 provide its derivation and analyze alternative forms.

Blind face restoration on CelebA-Test										
Method	**MUSIQ↑**	**CLIP-IQA↑**	**NIQE↓**	**LIQE↑**	**Q-Align↑**	**Aesthetic↑**	**FID↓**	**CLIP-Sim↑**	**Deg.↓**	**LMD↓**
FlowIE-BFR	74.10	0.684	4.52	4.842	4.597	5.961	16.65	0.932	33.69	2.259
+ *APD-Flow*	**76.24**	**0.749**	4.71	**4.941**	**4.755**	**6.023**	**16.50**	**0.935**	**32.75**	**2.178**
PMRF	64.63	0.446	5.14	3.564	3.941	5.644	13.38	0.941	30.47	2.030
+ *APD-Flow*	**71.35**	**0.539**	**4.39**	**4.249**	**4.317**	**5.719**	**10.59**	**0.943**	**30.37**	**2.006**
Blind image super-resolution on DIV2K-Val										
Method	**MUSIQ↑**	**CLIP-IQA↑**	**NIQE↓**	**LIQE↑**	**Q-Align↑**	**Aesthetic↑**	**LPIPS↓**	**DISTS↓**	**FID↓**	**CLIP-Sim↑**
FlowIE-BSR	58.47	0.476	5.125	2.915	2.768	4.243	0.457	0.232	144.1	0.880
+ *APD-Flow*	**67.26**	**0.545**	**4.292**	**3.749**	**3.200**	**4.339**	**0.451**	**0.215**	**142.6**	**0.882**
DiT4SR	67.18	0.653	4.38	3.983	3.779	4.960	0.473	0.218	119.6	0.903
+ *APD-Flow*	**69.05**	**0.684**	**4.31**	**4.095**	**3.926**	**4.989**	0.474	**0.215**	**119.3**	**0.904**

Table 1: Quantitative comparison on synthetic datasets using FR and NR metrics: blind face restoration on CelebA-Test and blind image super-resolution on DIV2K-Val.

4.3 Fidelity-Aware Reward Adaptation

When paired ground truth is available, reward post-training should raise the target reward without lowering the fidelity of the pretrained restorer. We denote the perceptual or task-specific reward by $P(\hat{x}_\theta)$ and a composite fidelity score by $F(\hat{x}_\theta, x_{\mathrm{gt}})$. For perceptual adaptation, P can be instantiated with quality models trained on human opinion scores or pairwise preference data (Gu et al., 2020; Zhao et al., 2026). The fidelity score F measures agreement with the ground truth using a task-appropriate combination of pixel, identity, perceptual, or semantic features.

For $\hat{x}_\theta = \mathcal{G}_\theta(z_0;\, y)$, we formalize this goal as

$$\max_\theta \ \mathbb{E}[P(\hat{x}_\theta)] \quad \text{s.t.} \quad \mathbb{E}[F(\hat{x}_\theta, x_{\mathrm{gt}})] \geq \tau, \tag{9}$$

where the fidelity target τ equals the validation fidelity of the pretrained restorer.

We implement this design through the paired reward difference used by APD. For antithetic pair i, let $\Delta P_i = P_i^+ - P_i^-$ and $\Delta F_i = F_i^+ - F_i^-$. The adapted reward difference is

$$\Delta R_i = \Delta P_i + \lambda_t \, \Delta F_i, \qquad \lambda_t = \left[\mu_t - \frac{\sum_i \Delta P_i \Delta F_i}{\sum_i \Delta F_i^2}\right]_+. \tag{10}$$

The fraction measures the extent to which the preference differences are already aligned with higher fidelity. It is positive when preference and fidelity favor the same branch on average and negative when they conflict. The slowly updated feedback state μ_t increases while batch fidelity stays below τ and decreases once the target is met. Consequently, λ_t compensates only for the residual fidelity deficit. It is zero when the preference differences already provide sufficient fidelity-aligned improvement and increases when that contribution is insufficient or opposed to fidelity. Unlike a fixed reward mixture, the controller adjusts the fidelity contribution to the current batch. Appendix A.5 derives this rule from a local linearization of Eq. (9) and gives its normalization, numerical safeguards, and μ_t update. Appendix A.6 gives the complete training loop, and Appendix B.4 reports controller diagnostics. Section 5 specifies the task-dependent choices of P and F.

5 Experiments

The appendix gives additional results, implementation details, and hyperparameters.

5.1 Experimental Setup

Pretrained restorers. We evaluate *APD-Flow* on four officially released flow-matching restorers: the 1-step FlowIE-BSR and 4-step FlowIE-BFR (Zhu et al., 2024), the 10-step PMRF (Ohayon et al., 2025), and the 15-step DiT4SR (Duan et al., 2025). They span blind face restoration and blind image super-resolution as well as U-Net and MM-DiT backbones. For scene-text image restoration, we use text-conditioned DiT4SR with a non-differentiable OCR reward. In all cases, post-training updates only the velocity field and leaves the deterministic sampler and inference budget unchanged.

Real-world blind face restoration												
	LFW				WebPhoto				CelebChild			
Method	MUSIQ↑	NIQE↓	Q-Align↑	Aes.↑	MUSIQ↑	NIQE↓	Q-Align↑	Aes.↑	MUSIQ↑	NIQE↓	Q-Align↑	Aes.↑
FlowIE-BFR	72.49	4.262	4.155	5.64	64.27	4.802	3.300	5.17	65.44	4.446	3.479	5.49
+ *APD-Flow*	**75.87**	**4.166**	**4.465**	**5.70**	**70.92**	**4.604**	**3.549**	**5.22**	**70.33**	**4.319**	**3.745**	**5.52**
PMRF	63.04	4.858	3.538	5.52	56.91	5.244	2.948	5.13	61.15	4.710	3.185	5.32
+ *APD-Flow*	**69.39**	**4.236**	**3.915**	**5.59**	**64.16**	**4.663**	**3.221**	**5.19**	**64.43**	**4.368**	**3.333**	**5.38**
Real-world blind image super-resolution												
	RealSRSet				RealLQ250				RealLR200			
Method	MUSIQ↑	NIQE↓	Q-Align↑	Aes.↑	MUSIQ↑	NIQE↓	Q-Align↑	Aes.↑	MUSIQ↑	NIQE↓	Q-Align↑	Aes.↑
FlowIE-BSR	61.89	6.252	3.212	4.52	61.18	5.19	3.067	4.93	61.95	5.31	3.030	4.81
+ *APD-Flow*	**69.26**	**5.839**	**3.490**	**4.64**	**68.66**	**4.69**	**3.381**	**5.02**	**68.88**	**4.86**	**3.330**	**4.89**
DiT4SR	66.02	5.84	3.528	4.69	67.10	4.77	3.731	5.32	68.51	4.79	3.781	5.19
+ *APD-Flow*	**67.30**	**5.71**	**3.594**	**4.70**	**68.36**	**4.73**	**3.830**	**5.34**	**69.55**	**4.76**	**3.872**	**5.21**

Table 2: No-reference results on real-world blind face restoration (top) and blind image super-resolution (bottom). Aes.: LAION Aesthetic.

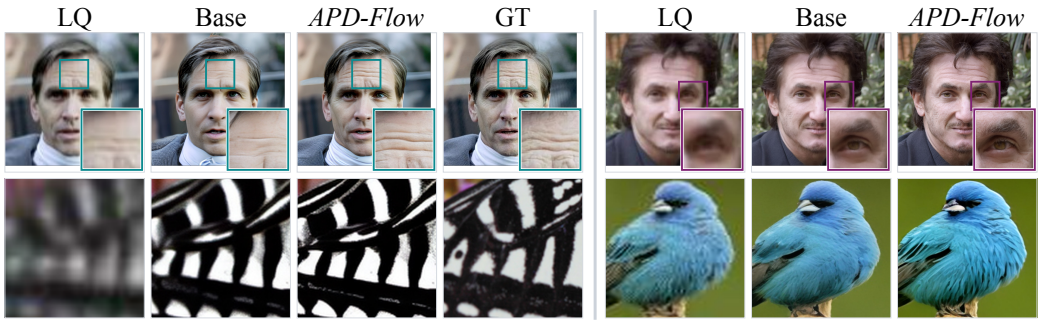

Figure 4: Qualitative comparisons (LQ / Base / *APD-Flow* / GT when available). Top: FlowIE-BFR on CelebA-Test, PMRF on LFW. Bottom: DiT4SR on DIV2K-Val, FlowIE-BSR on RealLQ250.

Post-training data. Face models are post-trained on 1,000 FFHQ images (Karras et al., 2019), degraded online with a standard blind-face pipeline (Wu et al., 2026); SR models use 800 DIV2K images (Agustsson & Timofte, 2017) and the two-stage Real-ESRGAN degradation (Wang et al., 2021). Post-training is data-efficient, with every released checkpoint adapted from a pool of at most 1,000 images. Preference rewards combine RALI (Zhao et al., 2026) with TOPIQ-Face (Chen et al., 2024) for face restoration and MANIQA (Yang et al., 2022) for SR. The text extension uses an OCR-based word-similarity reward on the TextOCR training split (Singh et al., 2021) and evaluates on its validation split.

Evaluation datasets and metrics. We evaluate on synthetic datasets (CelebA-Test for face restoration and DIV2K-Val for SR) and real-world datasets. The real-world face datasets are LFW, WebPhoto, and CelebChild; the real-world SR datasets are RealSRSet (Zhang et al., 2021), RealLQ250 (Ai et al., 2024), and RealLR200 (Wu et al., 2024c). Full-reference (FR) metrics include LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), FID (Heusel et al., 2017), and CLIP-Sim (Radford et al., 2021), together with ArcFace angular distance (Deng et al., 2019) and landmark distance (LMD) (Gu et al., 2022) for face restoration. They assess perceptual, distributional, semantic, identity, or geometric agreement with the ground truth. No-reference (NR) image-quality metrics include MUSIQ, CLIP-IQA, NIQE, LIQE, Q-Align, and LAION Aesthetic (Ke et al., 2021; Wang et al., 2023; Mittal et al., 2013; Zhang et al., 2023; Wu et al., 2024a; Schuhmann, 2022). Text restoration uses exact word accuracy, ANLS (Biten et al., 2019), and 1-NED (Levenshtein, 1966). Each model is compared with the corresponding pretrained restorer under matched deterministic inference.

5.2 Main Results

Blind face restoration. The face-restoration blocks of Tables 1 and 2 (top) show that *APD-Flow* substantially improves perceptual quality for both PMRF and FlowIE-BFR. On CelebA-Test, it also improves identity consistency, demonstrating that reward post-training enhances perceptual quality without sacrificing face fidelity.

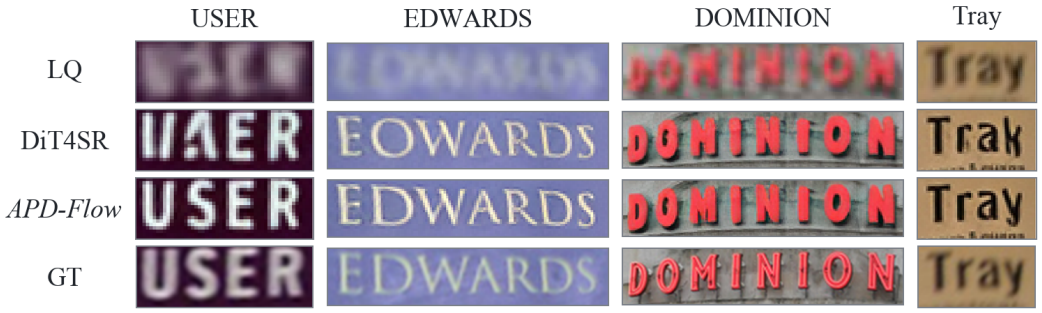

Figure 5: Matched TextOCR examples under the same text condition. Rows: LQ, text-conditioned DiT4SR, *APD-Flow*, and GT.

Comparison	Quality	Fidelity to GT
APD-Flow vs. PMRF	**64.2** / 25.4 / 10.4	**44.8** / 39.0 / 16.2
APD-Flow vs. FlowIE-BFR	**50.8** / 39.2 / 10.0	**44.0** / 37.8 / 18.2

Table 3: Human preference (%) on CelebA-Test. Entries: *APD-Flow* / pretrained / tie; GT is shown only for fidelity.

Method	Exact ↑	ANLS ↑	1-NED ↑
DiT4SR	0.0589	0.1418	0.1217
+ *APD-Flow*	**0.0644**	**0.1517**	**0.1311**

Table 4: Scene-text restoration results on TextOCR.

Blind image super-resolution. The super-resolution blocks of Tables 1 and 2 (bottom) show the same trend for blind image super-resolution. FlowIE-BSR improves every reference-dependent and NR measure on DIV2K-Val and every NR metric on all three real-world datasets. DiT4SR improves DISTS, FID, CLIP-Sim, and all NR metrics, with nearly unchanged LPIPS. These results show that *APD-Flow* applies to both one-step and multi-step flow restorers.

Human preference. Twenty participants evaluated 100 CelebA-Test images in a randomized pairwise study with balanced presentation order. Table 3 shows that *APD-Flow* was preferred over PMRF for visual quality in 64.2% of comparisons, whereas PMRF was preferred in 25.4%. Against FlowIE-BFR, the corresponding rates were 50.8% and 39.2%. Fidelity judgments also favored *APD-Flow*, with preference rates of 44.8% versus 39.0% for PMRF and 44.0% versus 37.8% for FlowIE-BFR.

Qualitative comparison. On the paired CelebA-Test and DIV2K-Val examples of Figure 4, *APD-Flow* recovers sharper and more realistic forehead wrinkles and butterfly-wing patterns. On the unpaired LFW and RealLQ250 examples, it renders more detailed eye structures and more visually pleasing feather textures than the pretrained outputs.

Text restoration. We optimize a non-differentiable reward based on character similarity. Because the pretrained and post-trained DiT4SR receive the same text condition, the gain comes from reward post-training rather than from conditioning. *APD-Flow* improves Exact, ANLS, and 1-NED by 9.3%, 7.0%, and 7.7% (Table 4). Across all words, 742 change from incorrect to correct and only 303 from correct to incorrect, and Figure 5 shows visibly corrected glyphs.

5.3 Reward Post-Training Comparisons

To compare *APD-Flow* with representative reward post-training strategies under a common setting, we adapt the image-generation methods ReFL (Xu et al., 2023; Wu et al., 2025), Flow-GRPO (Liu et al., 2025), DiffusionNFT (Zheng et al., 2026), and FDFO (McAllister et al., 2026a) to the same PMRF backbone and training protocol. All methods use the same training data and number of optimization updates. *APD-Flow* uses two candidates per input, whereas Flow-GRPO uses six. As shown in Table 5, Flow-GRPO injects noise along the ODE trajectory and forms group-relative advantages, yet improves none of the reported metrics and degrades all three reference-based measures. ReFL sharply increases MUSIQ but substantially worsens NIQE and every reference-based measure, which indicates reward hacking of the optimized proxy (Skalse et al., 2022). DiffusionNFT, which trains on the forward process from deterministic rollouts, improves MUSIQ and NIQE over the pretrained restorer, but *APD-Flow* achieves better results on all five reported metrics at the final checkpoint.

Method	MUSIQ ↑	NIQE ↓	LPIPS ↓	DISTS ↓	FID ↓
PMRF (pretrained)	64.63	5.14	0.331	0.156	13.38
+ Flow-GRPO	64.50	5.15	0.343	0.158	13.97
+ ReFL	77.21	6.50	0.439	0.203	28.79
+ DiffusionNFT	70.04	4.78	0.389	0.168	12.06
+ FDFO	65.88	4.93	0.338	0.155	12.81
+ *APD-Flow*	71.35	4.39	0.349	0.154	10.59

Table 5: Comparison of reward post-training methods on PMRF evaluated on CelebA-Test.

Variant	FID ↓	Deg. ↓	DISTS ↓	MUSIQ ↑	Q-Align ↑
Trajectory SDE	12.96	30.41	0.155	64.74	3.937
Direct regression	11.29	30.63	0.159	68.65	4.178
APD-Flow (ours)	10.59	30.37	0.154	71.35	4.317

Table 6: Post-training variant comparison on PMRF/CelebA-Test.



Figure 6: ReFL versus *APD-Flow* on CelebA-Test (columns per case: LQ input, ReFL, *APD-Flow*, GT). ReFL produces plastic-looking faces and hallucinated artifacts characteristic of reward hacking.

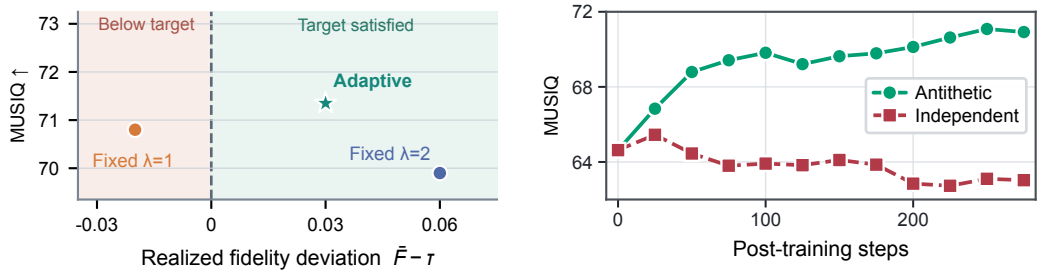

Figure 7: PMRF/CelebA-Test ablations. Left: fidelity operating points under fixed and adaptive weighting. Right: MUSIQ convergence with antithetic and independent source pairs.

Under the same update budget, *APD-Flow* achieves higher MUSIQ and lower NIQE, DISTS, and FID than FDFO. The ReFL outputs in Figure 6 show plastic-looking faces and hallucinated artifacts.

5.4 Ablation Studies

To isolate the contributions of source-state antithetic exploration and proxy optimization, we replace them with two alternatives in Table 6. Trajectory SDE moves exploration from the source state into the sampling trajectory through marginal-preserving noise injection (Liu et al., 2025), and direct regression performs one reward-weighted update per rollout. Both variants underperform *APD-Flow* on identity and perceptual metrics.

Figure 7 (left) compares adaptive and fixed fidelity weights. The adaptive controller satisfies the fidelity target and achieves higher MUSIQ than either fixed setting. Fixed λ=1 falls below the fidelity target, whereas λ=2 satisfies it at a lower MUSIQ score. Figure 7 (right) shows that training with antithetic pairs improves steadily, whereas independent source pairs make reward attribution ambiguous and do not converge reliably.

6 Conclusion

We presented *APD-Flow*, a reward post-training method for flow-matching image restorers with deterministic ODE trajectories. Instead of perturbing the sampling trajectory, *APD-Flow* constructs antithetic source-state pairs, converts their terminal reward difference into an output-space preference direction, and optimizes the velocity field through a training-only distance-based proxy while adaptively controlling fidelity. Across blind face restoration, blind image super-resolution, and scene-text restoration, *APD-Flow* improves perceptual or task-specific metrics using at most 1,000 post-training images, without changing the pretrained sampler or inference budget.

Ethics statement

All datasets used in this work are publicly available research benchmarks. We do not anticipate ethical concerns specific to APD-Flow beyond those associated with generative image restoration, including potentially inaccurate reconstructed details and misuse of restored images.

AI use statement

In this work, we used large-language-model assistants for error checking, for sentence-level editing and polishing of the manuscript, and for help with typesetting. The research idea, the method, the derivations, the implementation, and the experiments are the authors' own work, and we have reviewed all AI-assisted text. We take responsibility for the final content of this work, including its text and claims.

Reproducibility statement

The code is included in the supplementary materials submitted with this paper.

All pretrained restorers, post-training datasets, reward models, and evaluation datasets used in this paper are publicly available. The complete training loop is given as pseudocode in Appendix A.6, and the derivations of the antithetic construction, proxy ratio, common-mode regularizer, and adaptive fidelity weighting are given in Appendix A. Training settings for every backbone, the data, reward, and evaluation protocols, and the scene-text protocol are documented in Appendices B.1, B.2, and B.5. The adaptations of the compared reward post-training methods and the ablation variants are specified in Appendices D.1 and D.2.

References

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. DreamClear: High-capacity real-world image restoration with privacy-safe dataset curation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 55443–55469, 2024.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4291–4301, 2019.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S. Ren, Chunle Guo, and Chongyi Li. DiT4SR: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18948–18958, 2025. doi: 10.1109/ICCV51701.2025.01761.

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, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 79858–79885, 2023.

Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S. Ren, and Chao Dong. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision (ECCV)*, 2020.

Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision (ECCV)*, volume 13678 of *Lecture Notes in Computer Science*, pp. 126–143. Springer, 2022. doi: 10.1007/978-3-031-19797-0_8.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, 2021. doi: 10.1109/ICCV48922.2021.00510.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18750–18759, 2022.

Harshat Kumar, Dionysios S. Kalogerias, George J. Pappas, and Alejandro Ribeiro. Zeroth-order deterministic policy gradient. *arXiv preprint arXiv:2006.07314*, 2020.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, volume 15117, pp. 430–448. Springer, 2024. doi: 10.1007/978-3-031-73202-7_25.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

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 (NeurIPS)*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Yidi Liu, Zihao Fan, Jie Huang, Jie Xiao, Dong Li, Wenlong Zhang, Lei Bai, Xueyang Fu, and Zheng-Jun Zha. FinPercep-RM: A fine-grained reward model and co-evolutionary curriculum for RL-based real-world super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4839–4849, 2026.

David McAllister, Miika Aittala, Tero Karras, Janne Hellsten, Angjoo Kanazawa, Timo Aila, and Samuli Laine. Finite difference flow optimization for RL post-training of text-to-image models. *arXiv preprint arXiv:2603.12893*, 2026a.

David McAllister, Songwei Ge, Brent Yi, Chung Min Kim, Ethan Weber, Hongsuk Choi, Haiwen Feng, and Angjoo Kanazawa. Flow matching policy gradients. In *International Conference on Learning Representations (ICLR)*, 2026b.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

Guy Ohayon, Tomer Michaeli, and Michael Elad. Posterior-mean rectified flow: Towards minimum MSE photo-realistic image restoration. In *International Conference on Learning Representations (ICLR)*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 53728–53741, 2023.

Christoph Schuhmann. LAION-Aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022.

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

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.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8802–8812, 2021.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 9460–9471, 2022.

Yushuai Song, Weize Quan, Weining Wang, Jiahui Sun, Jing Liu, Pengbin Yu, Zhentao Chen, Meng Li, Wei Shen, Lunxi Yuan, and Dong-Ming Yan. RefReward-SR: LR-conditioned reward modeling for preference-aligned super-resolution. In *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pp. 94–112. Springer, 2026. doi: 10.1007/978-3-032-37324-3_6.

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 (CVPR)*, pp. 8228–8238, 2024. doi: 10.1109/CVPR52733.2024.00786.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision (IJCV)*, 132(12):5929–5949, 2024. doi: 10.1007/s11263-024-02168-7.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905–1914, 2021. doi: 10.1109/ICCVW54120.2021.00217.

Bin Wu, Wei Wang, Yahui Liu, Zixiang Li, and Yao Zhao. DiffusionReward: Enhancing blind face restoration through reward feedback learning. *arXiv preprint arXiv:2505.17910*, 2025.

Bin Wu, Wei Wang, Yahui Liu, Chi Zhang, and Yao Zhao. Beyond the boundary: RL-driven solution space exploration for blind face restoration. In *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pp. 445–462. Springer, 2026. doi: 10.1007/978-3-032-37553-7_25.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029, 2024a.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25456–25467, 2024c.

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.

Zhengpeng Xie, Qiang Zhang, Fan Yang, Marco Hutter, and Renjing Xu. Simple policy optimization. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68813–68824, 2025.

Haoxuan Xu, Yi Liu, Tianfu Li, Ruolin Shen, Boyuan Jiang, Jinlong Peng, Donghao Luo, Xiaobin Hu, Shuicheng Yan, and Haoang Li. IRPO: Boosting image restoration via post-training GRPO. *arXiv preprint arXiv:2512.00814*, 2025a.

Jiaqi Xu, Wenbo Li, Haoze Sun, Fan Li, Zhixin Wang, Long Peng, Jingjing Ren, Haoran Yang, Xiaowei Hu, Renjing Pei, and Pheng-Ann Heng. Fast image super-resolution via consistency rectified flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11755–11765, 2025b.

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 (NeurIPS)*, 2023.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1190–1199, 2022. doi: 10.1109/CVPRW56347.2022.00126.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, and Lei Zhang. GDPO-SR: Group direct preference optimization for one-step generative image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2177–2187, 2026.

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23153–23163, 2025.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081, 2023.

Shijie Zhao, Xuanyu Zhang, Weiqi Li, Junlin Li, Li Zhang, Tianfan Xue, and Jian Zhang. Reasoning as representation: Rethinking visual reinforcement learning in image quality assessment. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

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 (ICLR)*, 2026.

Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, and Jiwen Lu. FlowIE: Efficient image enhancement via rectified flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13–22, 2024. doi: 10.1109/CVPR52733.2024.00010.

Yuanzhi Zhu, Ruiqing Wang, Shilin Lu, Hanshu Yan, Junnan Li, and Kai Zhang. One-step flow for image super-resolution with tunable fidelity-realism trade-offs. In *International Conference on Learning Representations (ICLR)*, 2026.

APPENDIX ROADMAP

The appendix covers method details, reproducibility protocols, empirical analyses, comparison protocols, and extended qualitative results, in this order.

A. Method Details
Derivations behind the APD-Flow update.

- **A.1** Matched comparisons from antithetic sources
- **A.2** Output-space direction from paired rewards
- **A.3** Proxy ratio for deterministic velocity fields
- **A.4** Common-mode regularization
- **A.5** Adaptive fidelity weighting
- **A.6** Complete training loop

B. Reproducibility
Training and evaluation protocols.

- **B.1** Essential training settings
- **B.2** Data, rewards, and evaluation
- **B.3** Additional intuition and diagnostics
- **B.4** Controller diagnostics
- **B.5** Scene-text restoration protocol

C. Empirical Evidence
Controlled analyses of the core design choices.

- **C.1** Source states versus trajectory noise
- **C.2** Closed-loop fidelity control
- **C.3** Alignment with the reward gradient
- **C.4** Sensitivity to the antithetic strength
- **C.5** Reward-model generality

D. Alternative Updates
Reward post-training baselines and variants.

- **D.1** ReFL, Flow-GRPO, DiffusionNFT, and FDFO
- **D.2** Mechanism-ablation variants
- **D.3** Direct backpropagation through the sampler
- **D.4** Human preference study protocol

E. Qualitative Results and Limitations
Where to find extended visual evidence and the scope of the method.

- **E.1** Additional qualitative comparisons
- **E.2** Limitations

A Additional Method Details

This section derives the source-state construction, the output-space APD, the proxy objective, and adaptive fidelity weighting, and ends with the complete training loop. Appendix C reports the corresponding empirical diagnostics.

A.1 Why Antithetic Source States Yield Matched Comparisons

Equation (4) changes how the two branches are coupled without changing the source distribution seen by either branch. Since ϵ and δ are independent standard Gaussians,

$$\mathbb{E}[z_0^{\pm}] = m, \qquad \mathrm{Var}(z_0^{\pm}) = \sigma^2 I, \qquad \mathrm{Cov}(z_0^{+}, z_0^{-}) = \sigma^2(1-2w^2)I. \tag{A.1}$$

Thus each branch remains marginally distributed as $\mathcal{N}(m, \sigma^2 I)$ (Figure A.1a). For the values used in our experiments, the branches remain correlated because they share the same base noise, while the offsets along δ are opposite.

Transporting the pair through the deterministic restorer exposes the effect of this coupling. For fixed input, annotation, and base noise, define

$$g(\delta) := \mathcal{G}_\theta\big(m + \sigma(\sqrt{1-w^2}\epsilon + w\delta);\, y\big), \qquad h(\delta) := R\big(g(\delta)\big). \tag{A.2}$$

Writing each response as the sum of its even and odd components,

$$g_e(\delta) = \tfrac{1}{2}\big(g(\delta) + g(-\delta)\big), \qquad g_o(\delta) = \tfrac{1}{2}\big(g(\delta) - g(-\delta)\big), \tag{A.3}$$

and analogously for h, subtraction gives

$$\Delta x = g(\delta) - g(-\delta) = 2g_o(\delta), \qquad \Delta R = h(\delta) - h(-\delta) = 2h_o(\delta). \tag{A.4}$$

Subtracting the two branches cancels the responses shared by the mirrored perturbations and isolates the change induced by the sign of δ. Swapping the branch labels changes the signs of both Δx and ΔR, leaving d_{APD} invariant. Independent source noises do not provide this matched cancellation. Section C.1 evaluates the resulting comparison empirically.

A.2 From Paired Rewards to an Output-Space Direction

Let $\hat{x}^{\pm} = \mathcal{G}_\theta(z_0^{\pm}; y) \in \mathbb{R}^N$, $\Delta x = \hat{x}^{+} - \hat{x}^{-}$, and $\Delta R = R(\hat{x}^{+}) - R(\hat{x}^{-})$. We work in the output space that contains these candidates and the targets of Eq. (6), with the dimension-normalized geometry

$$\langle a, b\rangle_{\mathrm{RMS}} = \frac{1}{N} a^\top b, \qquad \|a\|_{\mathrm{RMS}}^2 = \langle a, a\rangle_{\mathrm{RMS}}. \tag{A.5}$$

When $\varepsilon = 0$, the APD in Eq. (5) is the shortest vector whose projection onto the observed displacement reproduces the measured reward difference,

$$d_{\mathrm{APD}} = \operatorname*{arg\,min}_{g:\, \langle g, \Delta x\rangle_{\mathrm{RMS}} = \Delta R} \|g\|_{\mathrm{RMS}}^2. \tag{A.6}$$

To see this, write $g = \alpha \Delta x + g_\perp$ with $\langle g_\perp, \Delta x\rangle_{\mathrm{RMS}} = 0$. The constraint fixes $\alpha = \Delta R / \|\Delta x\|_{\mathrm{RMS}}^2$, and any nonzero $g_\perp$ only increases the norm. With the stabilizer used in the implementation, the APD becomes the regularized secant

$$d_{\mathrm{APD}} = \operatorname*{arg\,min}_{g} \left\{ \frac{1}{2}\|g\|_{\mathrm{RMS}}^2 + \frac{1}{2\varepsilon}\big(\langle g, \Delta x\rangle_{\mathrm{RMS}} - \Delta R\big)^2 \right\}. \tag{A.7}$$

The same decomposition yields $g = \Delta R \Delta x / (\|\Delta x\|_{\mathrm{RMS}}^2 + \varepsilon)$, exactly Eq. (5). For a differentiable reward, ΔR is the projection of the path-averaged reward gradient onto Δx, and APD retains this observed one-dimensional component (Figure A.1b). For a non-differentiable reward, APD remains a pairwise secant estimated directly from terminal rewards.

Relation to two-point and zeroth-order estimation. Zeroth-order Deterministic Policy Gradient obtains deterministic-policy updates from two-point reward evaluations (Kumar et al., 2020). APD also uses two evaluations, but it is not an unbiased estimate of the full parameter gradient. It is a pair-conditioned secant built from the realized restoration displacement rather than the injected source noise. Reward access is zeroth order, while the detached velocity targets are optimized by ordinary backpropagation.

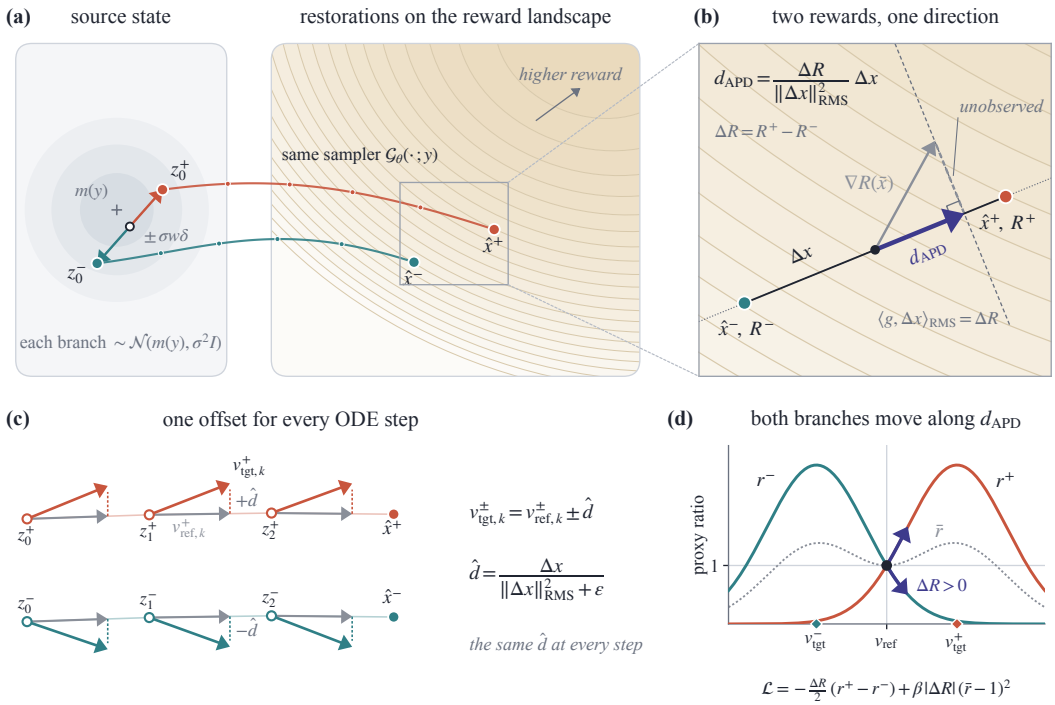

Figure A.1: Geometry of APD from the pair to the update. (a) The two branches share the base noise ϵ and mirror the perturbation δ, and the same deterministic sampler carries them to two restorations on the reward landscape. (b) The two rewards fix the component of the reward gradient along Δx, and d_{APD} keeps exactly this component. For the quadratic reward drawn here with $N = 1$ and $\varepsilon = 0$, it is the projection of $\nabla R(\bar{x})$ onto Δx. (c) The same offset $\hat{d}$ enters the velocity target of every ODE step, with opposite signs for the two branches. (d) Proxy ratios along $\hat{d}$ for $\|\hat{d}\|_{\mathrm{RMS}} = 1$. For $\Delta R > 0$, the update raises r^+ and lowers r^-, and both branches move along d_{APD} while the mean ratio $\bar{r}$ stays flat.

A.3 Proxy Ratio for a Deterministic Velocity Field

The proxy ratio compares the current and rollout-reference velocities through their energies at a shared detached target. It is used only for training and is not a policy likelihood.

For branch $b \in \{+1, -1\}$, write

$$\begin{aligned} v^b_k(\theta) &:= v_\theta(z^b_k, t_k;\, y), \\ v^b_{\mathrm{ref},k} &:= v_{\mathrm{ref}}(z^b_k, t_k;\, y), \\ v^b_{k,\mathrm{tgt}} &:= \big(v^b_{\mathrm{ref},k} + b\hat{d}\big)_{\mathrm{detach}}. \end{aligned} \tag{A.8}$$

The two targets shift the branch references by the same offset $\hat{d}$ with opposite signs at every ODE step (Figure A.1c). The two ratios are evaluated at different detached branch states, while the velocity network is shared.

Following the distance-based surrogates in FPO and FDFO (McAllister et al., 2026b;a), we define the velocity-field energy

$$\log q_\theta(u \mid z^b_k, t_k, y) = -\|u - v^b_k(\theta)\|^2_{\mathrm{RMS}} + \mathrm{const}. \tag{A.9}$$

The corresponding rollout-reference energy q_{ref} has the same form with $v^b_{\mathrm{ref},k}$. Evaluating the ratio at the detached target gives

$$\begin{aligned} r^b_k(\theta) &:= \frac{q_\theta(v^b_{k,\mathrm{tgt}} \mid z^b_k)}{q_{\mathrm{ref}}(v^b_{k,\mathrm{tgt}} \mid z^b_k)} \\ &= \exp\Big(- \|v^b_k(\theta) - v^b_{k,\mathrm{tgt}}\|^2_{\mathrm{RMS}} + \|v^b_{\mathrm{ref},k} - v^b_{k,\mathrm{tgt}}\|^2_{\mathrm{RMS}}\Big), \end{aligned} \tag{A.10}$$

which recovers Eq. (7). The ratio equals one at the rollout reference and increases as the current branch velocity approaches its detached target.

The rollout-dependent quantities z_k^b, $v_{\mathrm{ref},k}^b$, $v_{k,\mathrm{tgt}}^b$, Δx, and the rewards are all detached, and gradients flow only through $v_k^b(\theta)$. This stop-gradient construction makes the update compatible with deterministic samplers and non-differentiable rewards.

Local relation to target regression. At the rollout reference, the proxy-ratio objective reduces locally to signed target regression. Define

$$\mathcal{L}_{\mathrm{reg},k} := \frac{\Delta R}{2}\left(\|v_k^+(\theta) - v_{k,\mathrm{tgt}}^+\|_{\mathrm{RMS}}^2 - \|v_k^-(\theta) - v_{k,\mathrm{tgt}}^-\|_{\mathrm{RMS}}^2\right). \tag{A.11}$$

At $v_k^b(\theta) = v_{\mathrm{ref},k}^b$, the preference term $-\Delta R(r_k^+ - r_k^-)/2$ and Eq. (A.11) have the same value and first derivatives with respect to both branch-local velocities. Away from the reference, the ratio provides a nonlinear extension when multiple optimizer steps reuse the same rollout.

A.4 Common-Mode Regularization

The regularizer limits the joint drift of the pair from the rollout reference without suppressing the difference that the reward favors. Define the mean $\bar{r}_k = (r_k^+ + r_k^-)/2$ and half-difference $\delta_k = (r_k^+ - r_k^-)/2$. A per-branch quadratic penalty decomposes as

$$\frac{|\Delta R|}{4c}\left[(r_k^+ - 1)^2 + (r_k^- - 1)^2\right] = \frac{|\Delta R|}{2c}(\bar{r}_k - 1)^2 + \frac{|\Delta R|}{2c}\delta_k^2. \tag{A.12}$$

The first term controls joint drift, whereas the second would penalize the branch contrast that carries the preference signal. Following the quadratic stabilization of Xie et al. (2025), *APD-Flow* keeps only the first term, with $\beta = \lambda_c/(2c)$. This stabilizes the pair around $\bar{r}_k = 1$ while leaving its reward-driven contrast unpenalized.

The local relation to APD can be stated directly. Let $\mathcal{L}_k = -\Delta R\,\delta_k + \beta|\Delta R|(\bar{r}_k - 1)^2$. At the rollout reference,

$$\nabla_{v_k^+}\mathcal{L}_k = \nabla_{v_k^-}\mathcal{L}_k = -\frac{1}{N}\,d_{\mathrm{APD}}, \tag{A.13}$$

because $r_k^+ = r_k^- = 1$. At the rollout reference, the common-mode term has zero first derivative, and gradient descent moves both branch-local velocities along the APD direction (Figure A.1d). Parameter sharing couples these functional updates through the branch Jacobians. Table 6 evaluates the corresponding direct-regression variant.

A.5 Adaptive Fidelity Weighting

Adaptive weighting combines pairwise ranking agreement with closed-loop fidelity feedback. The agreement term measures whether the preference and fidelity scores favor the same branch, while the state μ_t tracks the deviation of realized batch fidelity $\bar{F}_t$ from its target τ. The resulting weight λ_t supplies the remaining fidelity correction. It increases when the rankings conflict or fidelity falls below the target and relaxes after fidelity recovers.

For a batch of B antithetic pairs, define

$$p_t = (\Delta P_1, \ldots, \Delta P_B)^\top, \qquad f_t = (\Delta F_1, \ldots, \Delta F_B)^\top, \qquad r_t(\lambda) = p_t + \lambda f_t. \tag{A.14}$$

The preference and fidelity components are normalized before these differences are formed. When $\|f_t\|_2^2 > 0$, the coefficient of the composite pairwise ranking $r_t(\lambda)$ along the fidelity-difference direction is

$$\gamma_F(\lambda) := \frac{f_t^\top r_t(\lambda)}{\|f_t\|_2^2} = \frac{A_t}{B_t} + \lambda, \qquad A_t := \sum_i \Delta P_i \Delta F_i, \quad B_t := \sum_i \Delta F_i^2. \tag{A.15}$$

The controller requires this score-space projection coefficient to be at least the feedback state $\mu_t \geq 0$, that is,

$$f_t^\top r_t(\lambda) = \sum_i \Delta F_i \Delta R_i(\lambda) \geq \mu_t \sum_i \Delta F_i^2. \tag{A.16}$$

Substituting $\Delta R_i(\lambda) = \Delta P_i + \lambda \Delta F_i$ gives

$$A_t + \lambda B_t \geq \mu_t B_t. \tag{A.17}$$

For $B_t > 0$, the exact smallest nonnegative solution of this score-space constraint is

$$\lambda_t^\star = \left[\frac{\mu_t B_t - A_t}{B_t}\right]_+ . \tag{A.18}$$

The implementation uses the damped numerical form reported with Eq. (10),

$$\lambda_t = \left[\frac{\mu_t B_t - A_t}{B_t + \varepsilon_\lambda}\right]_+ , \tag{A.19}$$

with a small $\varepsilon_\lambda > 0$. When $B_t = 0$, the batch contains no pairwise fidelity direction and $\lambda_t = 0$.

Here, A_t/B_t measures the fidelity alignment already contributed by the preference ranking, whereas μ_t specifies the alignment requested by the feedback controller. Their nonnegative residual determines λ_t, and the sign of A_t indicates whether preference and fidelity agree or conflict. The state μ_t is updated from the realized normalized batch fidelity $\bar{F}_t$ by

$$\mu_{t+1} = \Pi_{[0,\mu_{\max}]}\big(\mu_t + \eta_\mu(\tau - \bar{F}_t)\big). \tag{A.20}$$

Thus μ_t increases when batch fidelity falls below the target and decreases otherwise, which compensates for the mismatch between the score-space surrogate and the realized model response. The choice of P and F (Table B.2) changes the reward but not the *APD-Flow* update.

Equation (A.19) implements score-space feedback for the constrained objective rather than an exact KKT solution. Section C.2 evaluates its closed-loop behavior through the controller traces.

A.6 Training Loop

APD-Flow alternates frozen-reference rollout collection with optimization on cached states. Collection records paired deterministic trajectories and terminal rewards. Training re-evaluates the cached states to update the velocity field.

Algorithm 1 *APD-Flow* Reward Post-Training

Require: pretrained velocity field v_θ, training pool $\mathcal{D}$, rewards P and F, deterministic solver $\{(t_k, \Delta t_k)\}_{k=0}^{K-1}$

for each collection round $t = 1, 2, \ldots$ **do**
  freeze $v_{\text{ref}} \leftarrow v_\theta$ and initialize rollout buffer $\mathcal{B}$
  for each sampled (x_{gt}, y) **do**
    construct $z_0^\pm$ from shared noise by Eq. (4)
    for $k = 0, \ldots, K-1$ **do**
      roll out both branches with v_{ref} and cache detached $(z_k^\pm, v_{\text{ref},k}^\pm)$
    end for
    decode $\hat{x}^\pm$, evaluate terminal $P^\pm$ and $F^\pm$, and store Δx in $\mathcal{B}$
  end for
  compute λ_t, ΔR, and $\hat{d}$ by Eqs. (5) and (10)
  for pair minibatches in $\mathcal{B}$ and solver steps k **do**
    re-evaluate $v_\theta(z_k^\pm, t_k; y)$ and construct detached targets by Eq. (6)
    compute $r_k^\pm$ by Eq. (7) and update θ with Eq. (8)
  end for
end for
return v_θ; inference uses the original deterministic sampler, unchanged

The reference is refreshed at each collection round. Gradients flow only through v_θ during the update phase because all rollout quantities and terminal scores are detached. Both $P^\pm$ and $F^\pm$ are evaluated on decoded terminal outputs.

B Reproducibility and Evaluation Protocol

B.1 Essential Training Settings

We evaluate *APD-Flow* on five pretrained configurations spanning blind face restoration, blind image super-resolution, and scene-text restoration. Table B.1 lists the training settings. All runs use AdamW, one antithetic pair per input, and the original deterministic inference sampler. Training uses a single NVIDIA GeForce RTX 4090 (24 GB) with PyTorch 2.6.0 and CUDA 12.6. The optimizer batch size is two, except for scene-text restoration, where it is one.

Table B.1: Essential post-training settings.

Instance	ODE steps	Updated module	Peak LR	w
FlowIE-BFR	4	U-Net LoRA + ControlNet	2×10^{-6}	0.25
PMRF	10	Full flow model	2×10^{-6}	0.25
FlowIE-BSR	1	U-Net LoRA + ControlNet	2×10^{-6}	0.25
DiT4SR	15	Attention LoRA ($r = 16$)	2×10^{-5}	0.25
Text restoration	15	Attention LoRA ($r = 64$)	2×10^{-5}	0.50

B.2 Data, Rewards, and Evaluation

Face restoration follows the standard blind-face degradation pipeline on FFHQ (Wu et al., 2026). Blind super-resolution uses DIV2K with the two-stage Real-ESRGAN degradation pipeline (Wang et al., 2021). Scene-text training uses TextOCR with geometry-preserving blur, resampling, noise, and compression.

Table B.2 summarizes the signal families used to construct the preference component P and fidelity component F, including AdaFace (Kim et al., 2022) identity features and DINOv2 (Oquab et al., 2024) semantic features. Each signal is normalized before aggregation. The adaptive controller sets the balance between preference and fidelity for the general restoration tasks, whereas text restoration uses a fixed composite reward.

Table B.2: Signal families used to construct the preference and fidelity components.

Instance	Preference signals	Fidelity signals	Control
FlowIE-BFR	RALI, TOPIQ-Face	AdaFace, DISTS, DINOv2	Adaptive
PMRF	RALI, TOPIQ-Face	PSNR, AdaFace, DISTS, LPIPS, DINOv2	Adaptive
FlowIE-BSR	RALI, MANIQA	DISTS, DINOv2	Adaptive
DiT4SR	RALI, MANIQA	PSNR, DINOv2	Adaptive
Text restoration	OCR word similarity	DISTS	Fixed

The main results use the final checkpoints and complete paired benchmarks. A fixed 100-image CelebA-Test subset is used only for pilot ablations and diagnostics. Real-world benchmarks are held out for final evaluation. Matched comparisons share degraded inputs, source noise, input order, and evaluation procedure.

B.3 Additional Intuition and Diagnostics

Sections C.1 and C.3 evaluate source-state pairing, trajectory noise, and the alignment between APD and a differentiable reward gradient.

B.4 Controller Diagnostics

Figure C.5 reports the closed-loop behavior of adaptive fidelity weighting. Across PMRF and DiT4SR, decreases in batch fidelity are followed by larger adaptive weights, while the controller avoids sustained saturation. Section C.2 provides the complete traces and threshold-calibration analysis.

B.5 Scene-Text Restoration Protocol

TextOCR (Singh et al., 2021) is a large-scale collection of natural-scene images with dense, arbitrarily shaped text. It provides localized word instances with transcriptions, and the annotated regions define both the text condition and the word-level evaluation targets. We use its training split for post-training and its validation split for evaluation.

We apply *APD-Flow* to DiT4SR to test whether the same framework can optimize a non-differentiable, task-specific objective while retaining the pretrained deterministic sampler. The terminal reward measures word-level character agreement inside annotated text regions and includes a fidelity term to discourage unrelated visual changes.

The pretrained and adapted models receive the same annotation-derived text condition and share degraded inputs, source states, and word boxes. The null-text comparison in Table B.3 measures the gain from conditioning, while the matched DiT4SR–*APD-Flow* comparison isolates reward post-training. We evaluate valid Latin-alphanumeric words from the TextOCR validation set using exact match, ANLS, 1-NED, and exact match by character-height group.

Table B.3: TextOCR results without and with the text condition. Small, Medium, and Large give exact match by character height.

Method	Exact ↑	ANLS ↑	1-NED ↑	Small ↑	Medium ↑	Large ↑
DiT4SR (null text)	0.0461	0.1269	0.1070	0.0222	0.2392	0.3332
DiT4SR (annotation text)	0.0589	0.1418	0.1217	0.0326	0.2782	0.3523
+ *APD-Flow*	**0.0644**	**0.1517**	**0.1311**	**0.0373**	**0.2932**	**0.3567**

Adding the annotation-derived condition to the frozen DiT4SR increases exact match by 27.8% relative. Holding that condition fixed, *APD-Flow* further improves exact match, ANLS, and 1-NED by 9.3%, 7.0%, and 7.7%, respectively. The matched comparison contains 742 wrong-to-right and 303 right-to-wrong transitions (net +439), and all three character-height groups improve. Figures 5 and E.1 show the corresponding glyph corrections.

C Additional Empirical Evidence

C.1 Source-State Antithetic Exploration versus Trajectory Noise

Figure C.1 extends the pairing analysis of Sec. 5.4 to 100 CelebA-Test inputs. Antithetic source pairs concentrate at low candidate mismatch and high reward efficiency per unit displacement, whereas independent source pairs and trajectory-SDE candidates are more dispersed and less efficient.

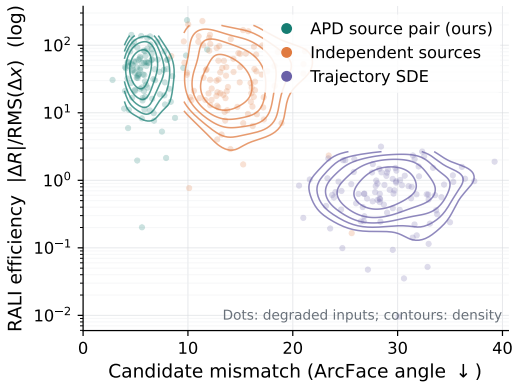

Figure C.1: Candidate mismatch (ArcFace angle) against reward efficiency $|\Delta R|/\mathrm{RMS}(\Delta x)$ for PMRF on 100 CelebA-Test inputs.

We compare source-state antithetic exploration with the marginal-preserving trajectory SDE of Flow-GRPO on pretrained PMRF before reward post-training. The SDE drift follows the official implementation and recovers the deterministic rollout at noise level 0. On CelebA-Test (3000), source-state exploration yields lower DISTS and NIQE than trajectory noise, and both metrics worsen as the noise level increases (Table C.1). The table evaluates the sampler output itself, whereas the trained Trajectory SDE variant in Table 6 uses noise level 0.1 and is evaluated with the deterministic sampler.

Table C.1: DISTS and NIQE of pretrained PMRF on CelebA-Test (3000) under source-state antithetic exploration and trajectory SDE, before reward post-training.

Exploration	DISTS ↓	NIQE ↓
Source-state antithetic (ours)	**0.156**	**5.14**
Trajectory SDE, noise level 0.05	0.489	14.40
Trajectory SDE, noise level 0.1	0.579	19.31

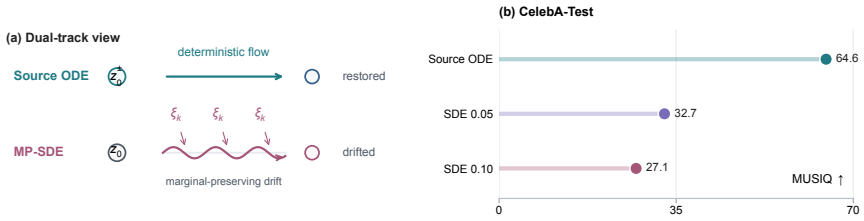

Figure C.2: Where stochasticity enters, before reward post-training. (a) Source-state antithetic exploration keeps the deterministic ODE trajectory, whereas the marginal-preserving SDE injects noise at intermediate states. (b) Aggregate MUSIQ on CelebA-Test decreases as trajectory noise increases.

Figure C.3 visualizes the corresponding exploration samples. Source-state perturbations remain natural restorations, whereas trajectory noise introduces visible corruption at noise level 0.05 and stronger artifacts at noise level 0.1.

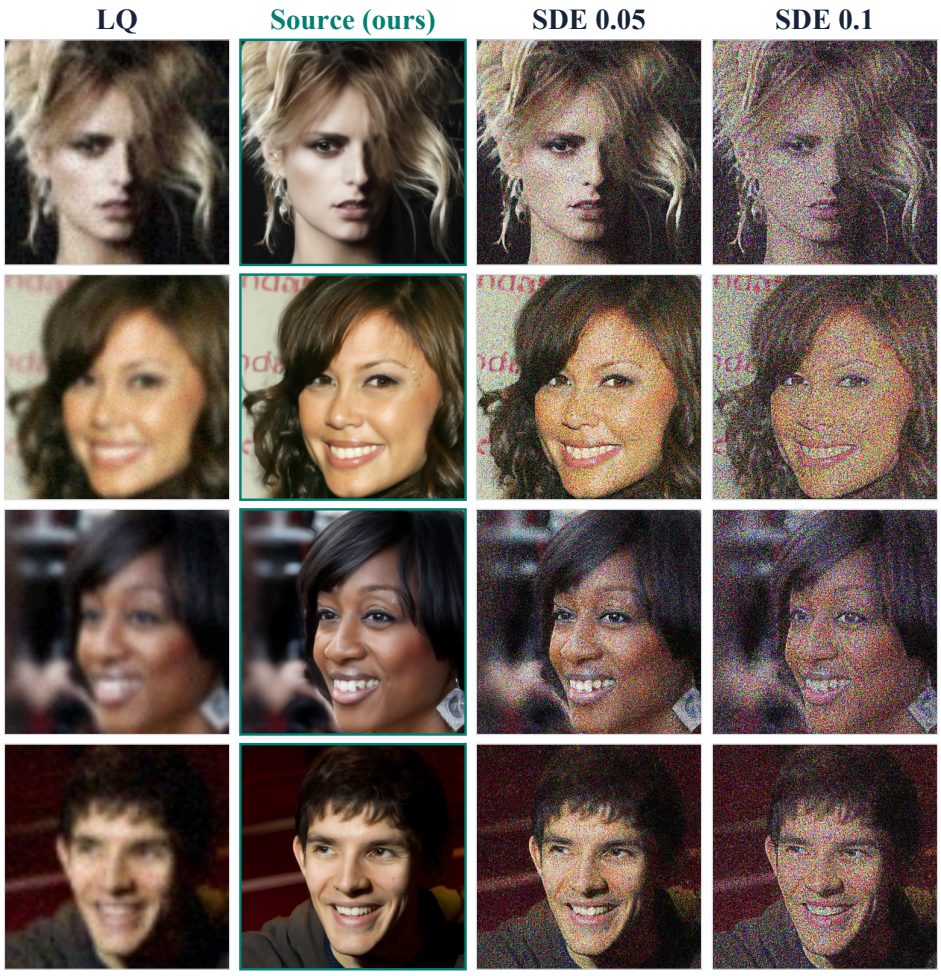

Figure C.3: Exploration samples of pretrained PMRF on CelebA-Test. The columns show the degraded input, source-state antithetic exploration, and trajectory SDE at noise levels 0.05 and 0.1.

C.2 Closed-Loop Fidelity Control

Figure C.4 shows normalized reward trajectories for all restorers. The identity retention of PMRF stays stable while its preference rewards increase.

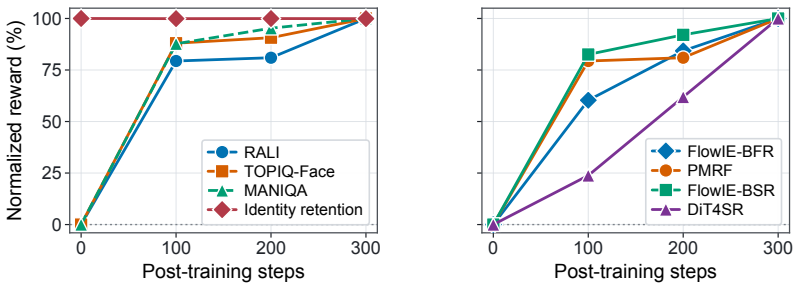

Figure C.4: Reward dynamics normalized per curve, with the pretrained value at 0 and the training peak at 100. The left panel shows RALI, TOPIQ-Face, held-out MANIQA, and identity retention for PMRF. The right panel shows RALI for FlowIE-BFR and PMRF and MANIQA for FlowIE-BSR and DiT4SR.

Figure C.5 shows the corresponding controller states λ_t and μ_t and the normalized batch fidelity $\bar{F}_t$ for PMRF and DiT4SR.

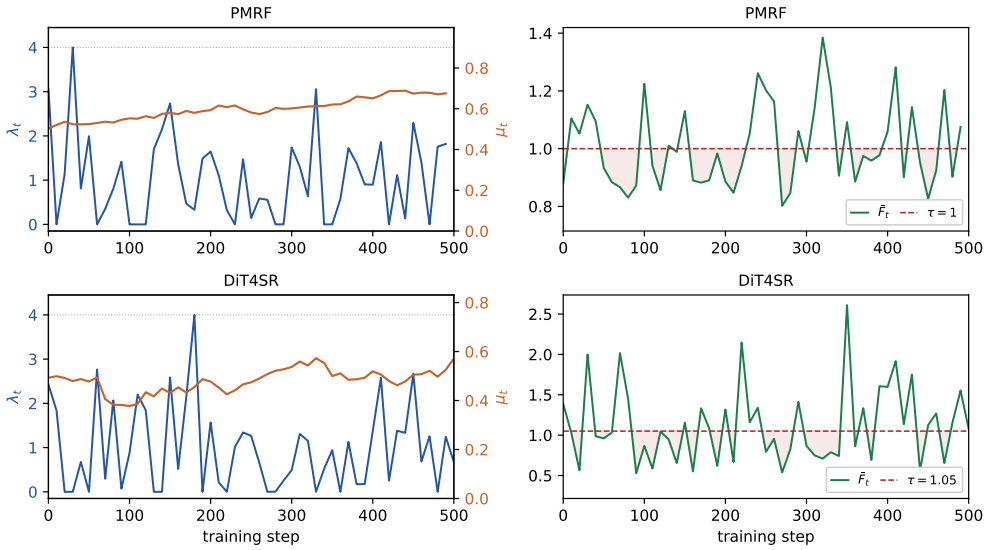

Figure C.5: Controller traces for PMRF and DiT4SR, logged every 10 optimizer steps. The left panel shows the adaptive weight λ_t and the feedback state μ_t. The right panel shows the normalized batch fidelity $\bar{F}_t$ against the threshold, $\tau = 1$ for PMRF and $\tau = 1.05$ for DiT4SR.

Neither trace shows sustained saturation. Decreases in $\bar{F}_t$ are followed by larger λ_t, and the final ten logged points average $\bar{F} = 1.03$ for PMRF and 1.19 for DiT4SR. The weight drops to zero on 24% of logged steps for PMRF and 20% for DiT4SR, and it reaches the cap $\lambda_t = 4$ on only 2%.

Threshold calibration. Each fidelity component is normalized by the validation statistics of the corresponding pretrained restorer, and $\bar{F}_t = 1$ therefore represents pretrained-level fidelity. The threshold is calibrated once at this operating point with a per-restorer margin of at most 5%, namely $\tau = 0.95$ for FlowIE-BFR, $\tau = 1$ for PMRF and FlowIE-BSR, and $\tau = 1.05$ for DiT4SR. A threshold below 1 therefore tolerates a bounded decrease in composite fidelity, at most 5% for FlowIE-BFR, whereas a threshold above 1 requires fidelity above the pretrained level. The controller uses $\mu_0 = 0.5$, step size $\eta_\mu = 0.02$, and $\lambda_t, \mu_t \leq 4$.

Table C.2: Adaptive versus fixed fidelity weighting on PMRF / CelebA-Test (3000). The column $\bar{F}_t - \tau$ is the gap between realized batch fidelity and the target.

Reward weighting	$\bar{F}_t - \tau$	**DISTS** ↓	**FID** ↓	**Deg.** ↓	**MUSIQ** ↑	**NIQE** ↓	**Q-Align** ↑
Adaptive (ours)	+0.03	0.154	10.59	30.37	71.35	4.39	4.317
Fixed λ_t=1.0	−0.02	0.158	10.76	30.41	70.80	4.32	4.310
Fixed λ_t=2.0	+0.06	0.153	11.18	30.34	69.90	4.51	4.190

The fixed rows replace the batch-adaptive weight with a constant and keep every other setting. Adaptive weighting meets the fidelity target and achieves higher MUSIQ and Q-Align than either fixed setting. Fixed λ_t=1.0 yields a slightly better NIQE but falls below the fidelity target. Fixed λ_t=2.0 exceeds the target and slightly improves DISTS and identity distance over adaptive weighting, but has worse FID and no-reference scores.

C.3 Alignment of the APD Direction with the Reward Gradient

To test the output-space geometry of Sec. A.2, we compare d_{APD} with the true gradient of a differentiable reward. On pretrained PMRF with the MANIQA reward, we evaluate four directions at each pair midpoint, namely d_{APD}, its counterpart built from the source perturbation δ, the unnormalized product $\Delta R\,\Delta x$, and a random direction.

Table C.3: Alignment with the true reward gradient on 300 CelebA-Test inputs (PMRF, MANIQA reward, $w = 0.25$), with 10^4-sample bootstrap intervals over inputs.

Direction	$\langle d, g\rangle$	**95% CI**	$\langle d, g\rangle{>}0$
APD d_{APD}	$+5.66$	$[+4.41, +7.11]$	**96%**
Source perturbation δ	$+4.8\times10^{-3}$	$[+3.7, +6.1]\times10^{-3}$	70%
Unnormalized $\Delta R\,\Delta x$	$+6.7\times10^{-4}$	$[+5.3, +8.3]\times10^{-4}$	96%
Random direction	$+1.4\times10^{-2}$	$[-1.9, +4.8]\times10^{-2}$	52%

Table C.3 shows that d_{APD} agrees in sign with the true reward gradient on 96% of inputs, and its bootstrap interval excludes zero. The random direction reaches 52%, and its interval crosses zero. An independent run on a disjoint set of 100 inputs gives sign agreements of 95%, 70%, 95%, and 49% for the four directions. The measured cosine of d_{APD}, 7.1×10^{-3}, also exceeds the random scale $O(D^{-1/2}) = 1.1\times10^{-3}$ for $D = 3{\cdot}512{\cdot}512$.

Replacing the realized displacement Δx with the injected perturbation δ reduces sign agreement to 70% and places the cosine below the random scale. This gap is consistent with nonlinear transport through the restorer and supports defining APD in output space.

RMS normalization preserves orientation while calibrating scale across pairs. The unnormalized $\Delta R\,\Delta x$ has the same cosine and sign agreement, but its magnitude grows with $\|\Delta x\|^2_{\mathrm{RMS}}$. This is why APD uses the normalized secant in Eq. (5).

C.4 Sensitivity to the Antithetic Strength w

The exploration strength w controls how far the two branches separate at the source state. Table C.4 sweeps it under the otherwise unchanged default setting, with every entry evaluated on the full CelebA-Test set.

Table C.4: Antithetic strength sweep on PMRF / CelebA-Test (3000). All other settings, the reward, and the update budget are unchanged.

w	**LPIPS** ↓	**DISTS** ↓	**Deg.** ↓	**MUSIQ** ↑	**NIQE** ↓	**CLIP-IQA** ↑
0.10	0.3459	0.1523	30.38	71.10	4.40	0.552
0.25 (default)	0.3489	0.1542	30.37	71.35	4.39	0.539
0.50	0.3443	0.1545	30.34	71.58	4.49	0.566

Performance varies little over the tested range of w, with no single value dominating every metric. We therefore use the intermediate setting $w = 0.25$ for face and general-image restoration.

C.5 Reward-Model Generality

To test whether the update depends on a particular preference model, we vary only the preference component P of PMRF over five choices, namely our RALI and TOPIQ-Face ensemble, each component alone, MANIQA, and HPS-v2.1 (Wu et al., 2023). All runs share the adaptive fidelity control, target τ, seed, data pool, and update count, and all are evaluated on the complete CelebA-Test benchmark. The four judge columns of Table C.5 are training rewards in at least one run, whereas the held-out metrics enter no training reward, with the exception of MANIQA in its own row.

Table C.5: Reward-model generality on CelebA-Test. Only the preference component P differs between rows.

	Trained judges				Held-out			Fidelity			
Preference P	RALI ↑	TOPIQ ↑	MANIQA ↑	HPS ↑	NIQE ↓	Q-Align ↑	LIQE ↑	LPIPS ↓	DISTS ↓	FID ↓	Deg. ↓
PMRF (pretrained)	3.657	0.778	0.402	0.2401	5.136	3.941	3.564	0.331	0.1556	13.38	30.47
RALI + TOPIQ-Face	3.837	0.848	0.504	0.2468	4.442	4.298	4.198	0.346	0.1536	10.82	30.37
RALI only	3.846	0.851	0.510	0.2473	4.404	4.321	4.222	0.346	0.1533	10.68	30.39
TOPIQ-Face only	3.794	0.836	0.481	0.2465	4.444	4.210	4.169	0.348	0.1536	10.96	30.37
MANIQA	3.809	0.844	0.519	0.2474	4.609	4.275	4.167	0.356	0.1594	10.86	30.34
HPS-v2.1	3.766	0.830	0.482	0.2466	4.595	4.198	4.216	0.351	0.1578	11.27	30.26

All five preference configurations improve the trained judges and the held-out NIQE, Q-Align, and LIQE metrics over pretrained PMRF. For two of the four judges, the best score comes from a run trained on another judge, which points to gains that transfer across judges rather than to overfitting of a single reward. The full-reference metrics show the same trade-off as the main results, with slightly higher LPIPS and lower FID than the pretrained model.

D Comparison Protocols for Alternative Updates

D.1 Adapting ReFL, Flow-GRPO, DiffusionNFT, and FDFO to Restoration

Table 5 compares alternative update rules on the same PMRF backbone, data pool, optimizer, and evaluation protocol. Each adaptation below keeps the defining assumption of the method and fixes everything else.

ReFL. ReFL differentiates a reward through the sampler and minimizes its negative value (Xu et al., 2023), and Wu et al. (2025) apply this design to blind face restoration. We follow the standard memory-efficient adaptation, which collects the trajectory without gradients and re-evaluates its final step with gradients. All training choices other than the update rule match those of *APD-Flow*.

Flow-GRPO. Flow-GRPO introduces a marginal-preserving SDE to obtain per-step Gaussian likelihood ratios and group-relative advantages (Liu et al., 2025). We use the official SDE construction with $G = 6$ candidates and noise level 0.1, and evaluate the adapted model with the deterministic sampler. This configuration requires three times as many rollouts and reward evaluations per update as *APD-Flow*. Table C.1 separately quantifies the effect of trajectory noise on restoration quality.

DiffusionNFT. DiffusionNFT performs reinforcement on the forward process. Rollouts come from any black-box sampler, and the velocity field is trained with a negative-aware flow-matching objective that contrasts implicit positive and negative policies built from the current and a slowly updated sampling policy (Zheng et al., 2026). We port the official implementation to PMRF with its default settings, including guidance strength $\beta = 1$, advantage clip 5, velocity KL weight 10^{-4}, the soft old-policy update, adaptive x_1-regression weighting, and EMA weights for evaluation. Rollouts use the deterministic PMRF sampler from $G = 6$ independent source samples per input, and the group-relative advantage is the one used for Flow-GRPO. Since PMRF transports its source $m(y)+\sigma\epsilon$ to the clean image, the forward process draws a fresh source sample and interpolates it with the clean rollout at the timesteps of the sampler.

FDFO. FDFO pairs two trajectories that diverge through an EDM-style stochastic sampler (Karras et al., 2022) and pulls every velocity toward their normalized, reward-weighted difference (McAllister et al., 2026a). We transfer its configuration to PMRF unchanged and train it with the same reward, data, and number of updates as the other methods.

Comparison scope. These experiments isolate the update rule on a common PMRF backbone. Systems that additionally replace the restoration backbone or train a separate reward model vary multiple components and are therefore outside this controlled comparison (Xu et al., 2025a; Liu et al., 2026; Song et al., 2026).

D.2 Variants in the Mechanism Ablation

The two alternatives in Table 6 each replace one component of *APD-Flow* while keeping the backbone, reward, controller, and update budget unchanged.

Trajectory SDE. This variant retains the update objective but replaces source-state perturbations with marginal-preserving trajectory noise, and is evaluated with the deterministic sampler after training. Its candidates exhibit intermediate-state corruption (Sec. C.1), and the resulting model provides only limited perceptual improvement over the pretrained restorer.

Direct regression. This variant uses the same antithetic source pairs and APD targets, but replaces the proxy ratio with one reward-weighted regression update per rollout. It improves over the pretrained restorer but trails the full objective in perceptual quality at comparable fidelity.

D.3 Comparison with Direct Backpropagation Through the Sampler

We also compare direct reward backpropagation on one-step FlowIE-BSR, where the full sampler contains a single network evaluation. Its objective combines MANIQA, DISTS, and DINOv2.

Table D.1: Direct reward backpropagation versus *APD-Flow* on FlowIE-BSR / DIV2K-Val (100) with a matched training budget.

Method	LPIPS↓	DISTS↓	FID↓	MUSIQ↑	CLIP-IQA↑	NIQE↓
FlowIE-BSR (pretrained)	0.457	0.232	144.1	58.47	0.476	5.13
+ *APD-Flow*	0.451	0.215	142.6	67.26	0.545	4.29
+ direct backpropagation	0.520	0.273	187.8	73.05	0.811	8.61

Direct backpropagation strongly raises its optimized no-reference metrics but degrades every reference-based metric and the held-out NIQE. Under the same protocol, *APD-Flow* improves DISTS, FID, and NIQE while retaining gains in MUSIQ and CLIP-IQA. The comparison illustrates why unconstrained reward gradients are insufficient for fidelity-aware restoration.

D.4 Human Preference Study Protocol

Twenty volunteers took part in the study of Table 3. We randomly sampled 100 CelebA-Test images and restored each with the pretrained restorer and its *APD-Flow* counterpart for both face backbones, PMRF and FlowIE-BFR.

Each trial presented one backbone under one criterion. For visual quality, the two restorations were shown side by side at their native 512×512 resolution, with left and right placement randomized and balanced across trials, and the participant chose the image with better overall quality in sharpness, natural texture, and freedom from artifacts, or declared a tie. For fidelity, the ground-truth image was also shown, and the participant chose the restoration more consistent with it in identity and content, again with a tie option. The ground truth appeared only in fidelity trials. Method identities were never revealed, viewing time was unlimited, and trial order was randomized independently for each participant.

Each participant judged 25 images per combination of backbone and criterion. Every image therefore received five independent votes, each combination accumulated 500 judgments, and Table 3 reports percentages over all of them.

E Qualitative Results and Limitations

E.1 Additional Qualitative Comparisons

Figure E.1 adds two matched scene-text examples in the format of Figure 5. The degraded input and the text condition are identical, and only the *APD-Flow* LoRA differs between the middle rows. The text-conditioned DiT4SR renders a closed "B" where the ground truth reads "Double" and turns the second "o" of "good" into a "b", and OCR reads its outputs as "Bouble" and "gobd". *APD-Flow* restores the correct glyphs, and the recognizer reads both words exactly.



Figure E.1: Additional matched TextOCR examples under the same text condition, extending Figure 5. Each group shows the ground-truth scene with the word marked in orange, followed by word crops of LQ, text-conditioned DiT4SR, *APD-Flow*, and GT. Orange boxes mark the corrected letter.

E.2 Limitations

The effect of reward post-training depends on the output space and conditioning of the pretrained restorer. Figure E.2 compares all four restorers, including cases where *APD-Flow* changes the pretrained output only modestly. In these examples, *APD-Flow* mainly refines sharpness and texture rather than content. Larger structural corrections may require stronger conditioning or a more capable pretrained model, as the text-conditioned setting of Sec. B.5 illustrates.

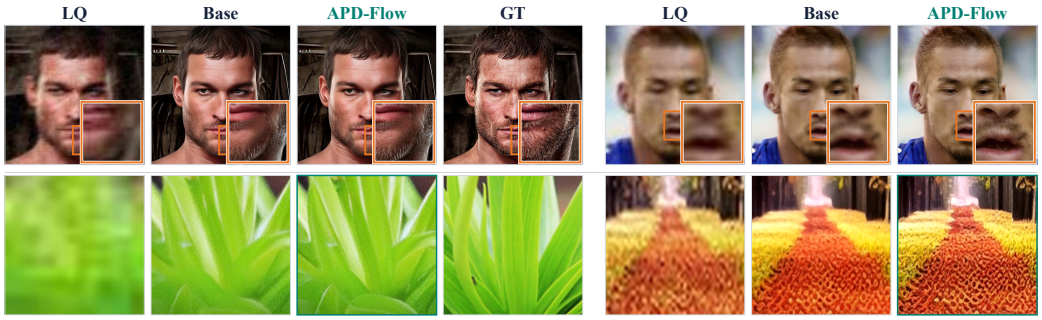

Figure E.2: Qualitative comparisons of LQ, Base, *APD-Flow*, and GT when available. The top row shows FlowIE-BFR on CelebA-Test and PMRF on LFW, and the bottom row shows DiT4SR on DIV2K-Val and FlowIE-BSR on RealLQ250.