PNR-SR: POLICY-NATIVE REWARD-POLICY CO-TRAINING FOR ONE-STEP REAL-WORLD IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step generative models have substantially improved the inference efficiency of real-world image super-resolution (Real-ISR), while reward-based post-training provides a promising paradigm for improving restoration quality beyond supervised learning. However, existing approaches typically rely on an offline-trained reward model that remains fixed during policy optimization. As the SR policy evolves, its output distribution can progressively diverge from the reward model's learned preference distribution, weakening the reward model's ability to reliably distinguish restoration quality on newly generated outputs. In this work, we formulate reward-based post-training for Real-ISR as a closed-loop optimization problem and propose **PNR-SR**, a **Policy-Native Reward-Policy Co-Training** framework for one-step Real-ISR. To this end, we introduce HR-exposure ordering to derive policy-native preference supervision from the current SR policy, enabling online reward model optimization without human annotations or fixed external evaluators. The reward model and SR policy are subsequently optimized in an online closed loop, where the reward model is continuously updated with policy-generated preferences to track the evolving SR policy distribution and provide dynamically updated reward guidance. To stabilize online reward model optimization, we further introduce dual-anchor calibration and a historical reward baseline to mitigate reward score drift. Extensive experiments demonstrate that our framework effectively improves one-step Real-ISR performance and outperforms existing state-of-the-art methods. The code is available at the anonymous link: https://anonymous.4open.science/r/PNR-SR-B550/.

1 INTRODUCTION

Real-World Image Super-Resolution (Real-ISR) aims to recover high-resolution (HR) images from low-resolution (LR) images with complex and unknown degradations. Recently, generative SR approaches Wang et al. (2024a); Yang et al. (2024); Wu et al. (2024b); Yu et al. (2024); Duan et al. (2025) based on diffusion models Ho et al. (2020) have made substantial progress in detail restoration and perceptual quality by leveraging the powerful generative priors of large-scale text-to-image (T2I) generative models. However, the high inference cost of multi-step generation processes limits their practical deployment. To improve efficiency, one-step generative SR methods Wu et al. (2024a); Wang et al. (2024b); Lin et al. (2025) compress the iterative generation process into a single forward step, but this compression also limits their ability to further exploit generative priors for detail recovery. Reward-based post-training Wu et al. (2025); Yi et al. (2026) has emerged as an effective paradigm for further improving restoration quality beyond supervised training by leveraging signals derived from reward models to optimize SR policies.

However, existing reward-based methods typically rely on fixed reward models during policy optimization, either from pre-trained task-specific reward models Song et al. (2026); Zhao et al. (2026) or directly from off-the-shelf evaluation metrics Wu et al. (2025); Han et al. (2026). Once established, these reward functions remain unchanged while the SR policy is further optimized using methods such as ReFL Xu et al. (2023) and DiffusionNFT Zheng et al. (2026). This paradigm assumes that an offline-trained reward model remains reliable throughout policy optimization. In practice, however, as the SR policy evolves, a fixed reward model may become less reliable at distin-

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.

guishing high- and low-quality samples generated by the evolving policy. Consequently, emerging policy-specific errors and artifacts may not be properly reflected in the reward signal, leading to misalignment between predicted rewards and actual visual quality. This issue is particularly critical for one-step SR, where unreliable reward guidance directly influences the final restoration quality without subsequent iterative refinement. These observations motivate us to move beyond fixed reward models and instead jointly optimize the reward model and SR policy in a closed loop.

We therefore propose **PNR-SR**, **a Policy-Native Reward-Policy Co-Training** framework that transforms reward-based post-training from a fixed-reward optimization paradigm into a closed-loop optimization process between the reward model and SR policy. In this framework, the reward model is continuously updated to track the evolving policy distribution, while the updated reward model guides policy optimization. The central challenge is then no longer how to design a fixed reward model, but how to obtain reliable and continuously refreshed supervision for online reward model optimization without additional human annotations or fixed external evaluators.

To address this challenge, we observe that controlling HR exposure in the starting state of one-step SR induces an intrinsic ordering among restoration outputs. For a given low-resolution (LR) image, the exposure level determines how much HR information from the paired image is injected into the starting state. Higher HR exposure provides stronger HR evidence to constrain the restoration, guiding the output toward the LR-conditioned HR solution space and reducing unconstrained hallucinations. This exposure-controlled relationship provides an intrinsic ordering of restoration outputs according to their plausibility, which we term HR-exposure ordering. Based on HR-exposure ordering, the SR policy can generate restoration pairs whose preference ordering is induced by different HR exposure levels. Since both the pairs and preference relations are derived from the policy, we regard them as policy-native preference supervision. With this supervision, reward model optimization enables the reward model to learn from the latest outputs of the policy, track the evolving policy distribution, and capture policy-specific restoration artifacts. Meanwhile, the updated reward model provides dynamic optimization signals for policy optimization. Together, these two processes form the proposed reward-policy co-training framework.

Online reward-policy co-training further introduces the issue of reward score drift. Continuous reward model optimization may cause the reward scale to shift over training, resulting in inconsistent optimization signals for policy optimization. To address this issue, we propose dual-anchor calibration, which uses stable score anchors to constrain reward scale drift. In addition, we maintain a historical reward baseline to provide a temporally smoothed reference for policy optimization, reducing its dependence on absolute reward values.

The main contributions are summarized as follows:

- We propose PNR-SR, a Policy-Native Reward-Policy Co-Training framework that reformulates reward-based post-training as a closed-loop optimization process.
- We present HR-exposure ordering, which leverages HR information exposure in one-step SR to construct policy-native preference supervision without human annotations or external evaluators.
- We introduce dual-anchor calibration together with a historical reward baseline to stabilize reward-policy co-training.
- Extensive experiments demonstrate that our method significantly improves one-step Real-ISR performance and outperforms existing state-of-the-art methods.

2 Related Work

Real-World Image Super-Resolution. Real-ISR involves complex and unknown degradations, limiting the effectiveness of PSNR-oriented methods trained with predefined degradations Dong et al. (2014); Kim et al. (2016). Generative approaches based on GANs Goodfellow et al. (2014), diffusion models Ho et al. (2020), and flow matching Lipman et al. (2022) have therefore been explored to improve perceptual quality. StableSR Wang et al. (2024a), DiffBIR Lin et al. (2024), PASD Yang et al. (2024), and SeeSR Wu et al. (2024b) leverage pretrained T2I priors to recover realistic details, but require costly multi-step sampling. This motivates the transfer of T2I priors to one-step SR for efficient inference while preserving visual quality. For example, OSEDiff Wu et al. (2024a), PiSA-SR Sun et al. (2025) and, HYPIR Lin et al. (2025) pursue this goal through

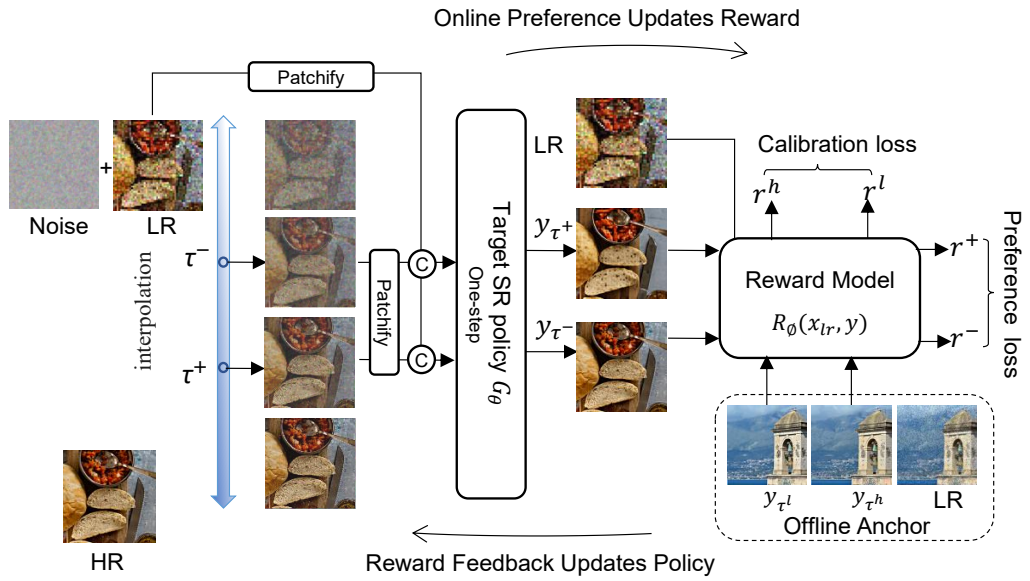


Figure 1: Overview of **PNR-SR**. The SR policy generates policy-native preference pairs to continually update the reward model, whose feedback in turn guides SR policy optimization.

variational score distillation (VSD), classifier score distillation (CSD), and diffusion-prior initialization followed by adversarial fine-tuning, respectively. Recent studies further explore reward-based post-training for both multi-step Wu et al. (2025); Liu et al. (2026) and one-step Han et al. (2026); Yi et al. (2026) SR models to improve perceptual quality.

Reward Models for Image Super-Resolution. In recent years, reward models have been applied to Real-ISR to provide optimization signals beyond conventional reconstruction losses. Some methods Wu et al. (2025); Han et al. (2026) directly adopt commonly used super-resolution evaluation metrics as reward functions, while others explicitly learn task-specific reward models for SR. FinPercep-RM Liu et al. (2026) learns fine-grained perceptual rewards from synthetic data. RefReward-SR Song et al. (2026) trains an LR-conditioned semantic reward model from large-scale human-annotated preference rankings and combines the learned reward with evaluation metrics to guide SR policy optimization. OARS Zhao et al. (2026) constructs a reward model based on fixed evaluation metrics to balance fidelity and perceptual quality. LucidNFT Fei et al. (2026) learns degradation-robust semantic representations through contrastive learning on LR–HR pairs, yielding an LR-conditioned faithfulness reward model. Despite their promising performance, these methods typically train the reward model offline and keep it fixed during policy optimization, which can lead to a mismatch between the reward model's learned preference distribution and the evolving policy distribution.

3 METHOD

Overview. We propose **PNR-SR**, a **Policy-Native Online Reward-Policy Co-Training** framework for one-step Real-ISR, where the reward model is continuously updated to track the evolving policy distribution and provide reward guidance for policy optimization. First, HR-exposure ordering enables the construction of policy-native preference supervision. Second, we use this supervision for reward model optimization and introduce dual-anchor calibration to mitigate reward score drift. Finally, the reward model and SR policy are alternately optimized in an online closed loop, with a historical reward baseline further stabilizing policy optimization.

Model Formulation. Let E and D denote the encoder and decoder of the pretrained variational autoencoder (VAE) Kingma & Welling (2013), respectively. Given a paired LR-HR sample $(x_{\mathrm{lr}}, x_{\mathrm{hr}})$, we obtain their latent representations as $z_{\mathrm{lr}} = E(x_{\mathrm{lr}})$ and $z_{\mathrm{hr}} = E(x_{\mathrm{hr}})$. Throughout the paper, x denotes images in RGB space, while z denotes latent representations or intermediate states in the

VAE latent space. We use G_θ to denote the one-step SR policy, consisting of the one-step denoising network followed by the VAE decoder D, and use y to denote the resulting restored RGB image.

3.1 One-Step SR with HR-Exposure Ordering

Within reward-policy co-training, reward model optimization requires continuously refreshed supervision to track the evolving policy distribution. However, repeatedly collecting human preferences is prohibitively expensive, while relying on external evaluators makes the supervision dependent on predefined evaluation criteria. We therefore seek to derive ordinal supervision directly from the SR policy itself.

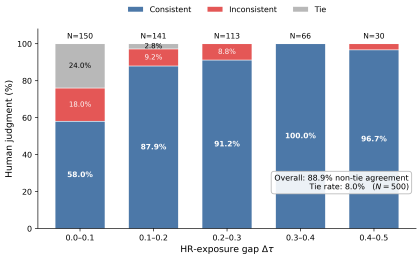


Figure 2: Human preferences across HR-exposure gaps $\Delta\tau$.

Our key observation is that the amount of HR information injected into the starting state of a one-step SR policy can be explicitly controlled. We therefore extend conventional fixed-starting-state one-step SR to an HR-exposure-controllable formulation, where the same SR policy restores from starting states with different strengths of real-image constraints. For the same LR input, increasing HR exposure provides stronger evidence from the paired HR image, constraining the restoration toward the LR-conditioned HR solution space. This controllable exposure naturally induces an ordinal restoration prior for reward model optimization.

We build our one-step SR policy upon Z-Image Cai et al. (2025), whose flow trajectory is parameterized such that $t = 0$ corresponds to the noise endpoint and $t = 1$ to the clean-image endpoint. Following the dual-input design of DNF-SR Han et al. (2026), we first construct the default starting state at an intermediate timestep t_m by mixing the LR latent z_{lr} with Gaussian noise:

$$z_{\mathrm{lr}}^{t_m} = t_m z_{\mathrm{lr}} + (1 - t_m)\boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{1}$$

The resulting LR-noise mixed latent serves as the primary input and is patchified into input tokens. Meanwhile, the LR latent z_{lr} is also patchified into condition tokens, which are concatenated with the input tokens before being fed into the one-step SR policy.

To construct HR-exposure-controlled starting states, we incorporate information from the paired HR image along the restoration trajectory. Treating the paired HR latent z_{hr} as the clean endpoint, we define the direction from the default starting state $z_{t_m}^{\mathrm{lr}}$ as $v = (z_{\mathrm{hr}} - z_{t_m}^{\mathrm{lr}})/(1 - t_m)$. We then define a discrete set of exposure timesteps $\mathcal{T} = \{\tau_1, \ldots, \tau_K\} \subseteq [t_m, t_{\mathrm{end}}]$, where $t_{\mathrm{end}} < 1$. For each $\tau \in \mathcal{T}$, we construct an HR-exposure-controlled starting state along this trajectory:

$$z_\tau^{\mathrm{in}} = z_{\mathrm{lr}}^{t_m} + (\tau - t_m)v. \tag{2}$$

The distance between the constructed starting state and the HR endpoint satisfies

$$\left\| z_\tau^{\mathrm{in}} - z_{\mathrm{hr}} \right\| = \frac{1 - \tau}{1 - t_m} \left\| z_{t_m}^{\mathrm{lr}} - z_{\mathrm{hr}} \right\|. \tag{3}$$

Thus, increasing τ moves the starting state monotonically toward the clean HR latent endpoint z_{hr}, exposing progressively more HR information to the one-step SR policy, while $\tau = t_m$ recovers the original LR–noise starting state.

Before reward-policy co-training, we first perform a supervised warm-up of G_θ, enabling the policy to reliably restore from starting states with different HR exposure levels. During this stage, we randomly sample $\tau \in \mathcal{T}$ and optimize the corresponding restoration output against the paired HR target using the following objective:

$$\mathcal{L}_{\mathrm{SR}} = \lambda_1 \mathcal{L}_{\mathrm{mse}} + \lambda_2 \mathcal{L}_{\mathrm{lpips}} + \lambda_3 \mathcal{L}_{\mathrm{FDL}}. \tag{4}$$

Here, $\mathcal{L}_{\text{lpips}}$ denotes the Learned Perceptual Image Patch Similarity (LPIPS) loss, while $\mathcal{L}_{\text{FDL}}$ denotes the Focal Frequency Distribution Loss (FDL) adopted in StrSR Wang et al. (2026). FDL matches the frequency distributions of the SR image and the paired HR target, thereby regularizing their spectral characteristics and mitigating grid-like periodic artifacts associated with high-frequency spectral leakage in one-step Diffusion Transformer (DiT)-based SR.

For an exposure level τ, the corresponding restored image is $y_\tau = G_\theta(z_\tau^{\text{in}}, z_{\text{lr}})$. We use the relative magnitude of τ to induce an ordinal relation between the resulting outputs:

$$\tau_i > \tau_j \quad \Longrightarrow \quad y_{\tau_i} \succ y_{\tau_j}. \tag{5}$$

We term this relation HR-exposure ordering and use it as ordinal supervision for reward model optimization. To assess whether this ordering aligns with human preference, we evaluate 500 restoration pairs. As shown in Fig. 2, 409 pairs agree with the exposure-based ordering, 51 disagree, and 40 are rated as ties, yielding 88.91% agreement among non-tied pairs. Agreement generally increases with the exposure gap $\Delta\tau = |\tau_i - \tau_j|$, while ties are concentrated among pairs with small $\Delta\tau$.

3.2 Policy-Native Reward Model Optimization with Dual-Anchor Calibration

The HR-exposure ordering established in Sec. 3.1 enables us to construct preference supervision directly from the current SR policy. We use these policy-native preferences to train an LR-conditioned reward model, together with dual-anchor calibration to maintain a stable reward reference.

Policy-Native Preference Construction. Given a paired LR–HR sample $(x_{\text{lr}}, x_{\text{hr}})$, we sample two HR-exposure levels satisfying $\tau^+ > \tau^-$. Following Sec. 3.1, the corresponding starting states $z_{\tau^+}^{\text{in}}$ and $z_{\tau^-}^{\text{in}}$ contain different amounts of HR information, and the current SR policy G_θ produces restoration outputs y_{τ^+} and y_{τ^-} from these states. According to the HR-exposure ordering, $y_{\tau^+} \succ y_{\tau^-}$. We refer to the resulting preference pair as **policy-native preference supervision**, since both the restoration samples and their ordinal relation are induced by the target SR policy and its HR-exposure-controlled restoration process, without requiring human annotation or external evaluators.

LR-Conditioned Reward Model. Given policy-native preference samples, we learn an LR-conditioned reward function $R_\phi(x_{\text{lr}}, y)$ to evaluate the restoration plausibility of an SR output y conditioned on the LR observation x_{lr}. A frozen visual encoder extracts token-level representations from the LR observation and the SR output. Using the SR tokens as the primary representation, we obtain LR context through cross-attention and incorporate it into the SR tokens via a learnable gate. A lightweight Transformer encoder refines the conditioned tokens, which are then aggregated into h_ϕ through attention pooling with a learnable query. A regression head produces the scalar reward $R_\phi(x_{\text{lr}}, y) = g_\phi(h_\phi)$. A higher reward indicates a more plausible high-resolution interpretation of the LR observation.

Policy-Native Reward Supervision with Dual-Anchor Calibration. Based on the policy-native preference samples, we optimize the reward model with a pairwise preference objective and a margin regularization term:

$$\mathcal{L}_{\text{pref}} = -\log\sigma\left(R_\phi(x_{\text{lr}}, y_{\tau^+}) - R_\phi(x_{\text{lr}}, y_{\tau^-})\right) + \lambda_m \,\text{ReLU}\left(\left|R_\phi(x_{\text{lr}}, y_{\tau^+}) - R_\phi(x_{\text{lr}}, y_{\tau^-})\right| - m_{\text{gap}}\right)^2. \tag{6}$$

where m_{gap} denotes a soft upper-bound threshold on the reward gap, and λ_m controls the contribution of the margin regularization. The pairwise objective trains the reward model to distinguish preferred from non-preferred restoration outputs according to the HR-exposure ordering, while the margin regularization prevents excessively large reward gaps between them.

Although the preference objective constrains relative reward ordering, it does not provide an absolute reference for the reward scale. Consequently, reward score drift may occur during reward-policy co-training, as empirically observed without dual-anchor calibration (see Fig. 5 in the supplementary material). Specifically, as the policy distribution evolves, restoration outputs with similar plausibility may receive inconsistent reward scores at different training steps, resulting in unstable optimization.

To address this issue, we propose a dual-anchor calibration strategy. Specifically, after the supervised warm-up stage, we construct a fixed calibration set using the fixed SR policy G_{θ_0}. For each paired LR-HR sample, we select the two extreme HR-exposure levels, $\tau_{\text{h}} = \max\mathcal{T}$ and $\tau_{\text{l}} = \min\mathcal{T}$, and

collect the corresponding restoration outputs y_{τ_h} and y_{τ_l} generated by the fixed policy G_{θ_0}. Since the policy is fixed during this stage, the collected restoration samples provide stable references for reward calibration. Based on the HR-exposure ordering, we assign fixed reward anchors to y_{τ_h} and y_{τ_l}, where the high-exposure output is assigned the higher reward anchor a_h, and the low-exposure output is assigned the lower reward anchor a_l. During reward model optimization, the fixed calibration samples are incorporated using the following calibration objective:

$$\mathcal{L}_{\mathrm{cal}} = \left(R_\phi(x_{\mathrm{lr}}, y_{\tau_\mathrm{h}}) - a_{\mathrm{h}}\right)^2 + \left(R_\phi(x_{\mathrm{lr}}, y_{\tau_\mathrm{l}}) - a_{\mathrm{l}}\right)^2. \tag{7}$$

The dual-anchor calibration provides stable reward references during online reward–policy co-training. The final reward model optimization objective is

$$\mathcal{L}_{\mathrm{RM}} = \mathcal{L}_{\mathrm{pref}} + \lambda_c \mathcal{L}_{\mathrm{cal}}. \tag{8}$$

where λ_c controls the contribution of the dual-anchor calibration term. With dual-anchor calibration, the reward model provides a stable LR-conditioned restoration plausibility estimate, which serves as the optimization signal for subsequent policy optimization during online reward–policy co-training.

3.3 Policy-Native Online Reward-Policy Co-Training

Following supervised warm-up and dual-anchor calibration set construction, we enter the online reward-policy co-training stage. We first warm up the reward model with the SR policy fixed, and then alternately optimize the reward model and SR policy.

During joint optimization, at each iteration k, the current policy G_{θ_k} generates new preference pairs following Sec. 3.2. The HR-exposure ordering of these samples provides supervision for reward model optimization, while the reward model R_{ϕ_k} provides reward signals for optimizing the SR policy. The updated policy and reward model then continue to generate and evaluate new preference samples, forming a continuously evolving closed-loop optimization process. Following the regularized RL formulation of DMDR Jiang et al. (2025), we directly optimize the SR policy using differentiable reward gradients as in ReFL Xu et al. (2023), while retaining the original supervised restoration objective as a regularization term. Specifically, the policy objective is defined as

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{SR}} + \lambda_r \mathcal{L}_{\mathrm{ReFL}}. \tag{9}$$

where $\mathcal{L}_{\mathrm{SR}}$ denotes the supervised restoration objective defined in Eq. 4, and λ_r controls the contribution of the reward optimization term $\mathcal{L}_{\mathrm{ReFL}}$.

Instead of directly optimizing the absolute reward score, we introduce a historical reward baseline to provide a temporally smoothed reference for policy optimization and reduce its dependence on absolute reward values. At iteration k, let $\bar{R}_k$ denote the average reward of the current policy outputs. We maintain an exponential moving average of historical rewards as

$$\mu_k = \rho \mu_{k-1} + (1-\rho)\bar{R}_k, \tag{10}$$

where ρ denotes the momentum coefficient. During policy optimization at iteration k, the previous historical baseline μ_{k-1} is detached from gradient propagation. Rather than maximizing the raw reward directly, we optimize the policy according to its reward improvement over this historical baseline. Specifically, the reward optimization objective is defined as

$$\mathcal{L}_{\mathrm{ReFL}} = \mathbb{E}\left[\mathrm{ReLU}\left(m_{\mathrm{imp}} - \left[R_{\phi_k}(x_{\mathrm{lr}}, y) - \mu_{k-1}\right]\right)\right]. \tag{11}$$

where m_{imp} is a fixed improvement margin. This objective penalizes restoration outputs whose LR-conditioned reward does not exceed the historical reward baseline by the required margin, thereby encouraging the current SR policy to achieve consistent reward improvement relative to its historical reference. The reward model R_{ϕ_k} is alternately updated using newly generated policy-native preference pairs together with the fixed dual-anchor samples according to Eq. 8.

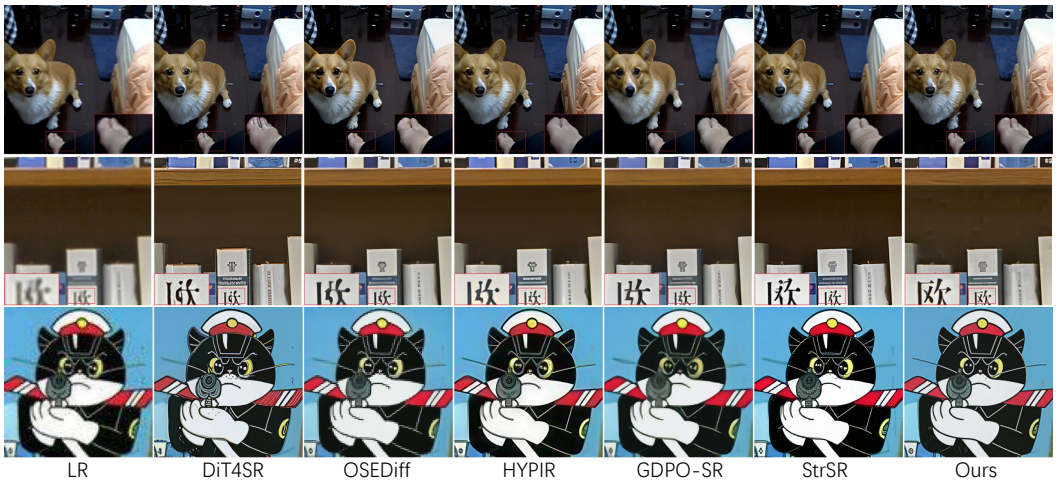


Figure 3: Qualitative comparison with state-of-the-art methods.

4 Experiments

4.1 Experimental Settings

Training Datasets. Following previous works Wu et al. (2024b); Dong et al. (2025), we construct the training dataset for our model using LSDIR Li et al. (2023) and the first 10,000 face images from the FFHQ Karras et al. (2019) dataset. The corresponding LR images are synthesized from the HR images using the Real-ESRGAN Wang et al. (2021) degradation pipeline.

Testing Datasets. To evaluate the effectiveness of the proposed method, we conduct experiments on five benchmarks: DIV2K-Val Agustsson & Timofte (2017), RealSR Cai et al. (2019), ScreenSR Wu et al. (2026), RealLQ250 Ai et al. (2024), and our newly introduced PhoneSR30. Specifically, we use the synthetic DIV2K-Val benchmark adopted in StableSR Wang et al. (2024a) and follow the SeeSR Wu et al. (2024b) evaluation setting for RealSR. To better reflect real-world telephoto imaging on smartphones, we further introduce PhoneSR30, a real-world benchmark comprising 30 images at a resolution of 512×512 images from recent smartphones. DIV2K-Val, RealSR, and RealLQ250 are evaluated at $\times 4$, PhoneSR30 at $\times 2$, and ScreenSR at its original resolution.

Evaluation Metrics. We employ both full-reference and no-reference metrics for evaluation. For reference-based evaluation, PSNR, LPIPS Zhang et al. (2018), and DISTS Ding et al. (2020) are used to measure pixel-level fidelity and perceptual similarity, while FID Heusel et al. (2017) is adopted to assess the distributional similarity between SR and HR images. For no-reference evaluation, we use MUSIQ Ke et al. (2021), CLIPIQA+ Wang et al. (2023), MANIQA Yang et al. (2022), Q-Align Wu et al. (2023), and TOPIQ Chen et al. (2024) to evaluate perceptual image quality.

Compared Methods. We compare PNR-SR against state-of-the-art Real-ISR methods, covering both multi-step and one-step approaches. The multi-step baselines include SeeSR Wu et al. (2024b), SUPIR Yu et al. (2024) and DiT4SR Duan et al. (2025), while the one-step baselines consist of HYPIR Lin et al. (2025), OSEDiff Wu et al. (2024a), SinSR Wang et al. (2024b), GDPO-SR Yi et al. (2026) and StrSR Wang et al. (2026).

4.2 Experimental Results

Quantitative Analysis. Tables 1 and 2 present quantitative comparisons between our method and state-of-the-art Real-ISR methods on four benchmark datasets. Our method achieves leading performance across multiple evaluation metrics. In particular, compared with StrSR, which adopts the same base model, our method obtains better results on multiple full-reference and no-reference metrics, further demonstrating its effectiveness in both restoration fidelity and perceptual quality.

Table 1: Quantitative comparison with SOTA Real-ISR methods on three datasets. Best and second-best results are highlighted in red and blue, respectively. ↑ / ↓ indicate higher / lower is better.

	Method	PSNR↑	LPIPS↓	DISTS↓	FID↓	MUSIQ↑	TOPIQ↑	MANIQA↑	CLIPIQA+↑	Q-Align↑
DIV2K-val	SeeSR	23.68	0.3194	0.1968	25.90	68.67	0.6857	0.6222	0.7043	3.977
	SUPIR	23.15	0.3637	0.2259	27.95	62.59	0.6126	0.5901	0.6629	3.778
	DiT4SR	21.80	0.3446	0.2102	30.68	68.02	0.6224	0.6419	0.6694	3.722
	SinSR	24.41	0.3239	0.2071	35.44	62.85	0.5726	0.5394	0.5581	3.512
	OSEDiff	23.73	0.2941	0.1959	26.08	67.82	0.6146	0.6108	0.6927	3.823
	HYPIR	22.22	0.3077	0.1951	26.86	66.25	0.6357	0.6244	0.6886	3.839
	GDPO-SR	23.92	0.2899	0.1965	26.28	68.84	0.6205	0.6393	0.7132	3.866
	StrSR	22.00	0.3224	0.2080	32.72	68.99	0.6455	0.6562	0.6740	3.776
	Ours	22.59	0.2905	0.1817	23.97	69.99	0.7014	0.6503	0.7077	4.105
RealSR	SeeSR	25.15	0.3007	0.2224	125.45	69.82	0.6891	0.6451	0.6908	3.718
	SUPIR	25.07	0.3728	0.2488	121.68	58.55	0.5316	0.5663	0.6050	3.391
	DiT4SR	23.40	0.3194	0.2237	121.20	67.74	0.5973	0.6570	0.6656	3.381
	SinSR	26.29	0.3198	0.2348	134.97	60.65	0.5162	0.5391	0.5416	3.147
	OSEDiff	25.15	0.2921	0.2128	123.48	69.09	0.6253	0.6335	0.6964	3.696
	HYPIR	22.90	0.3050	0.2272	124.96	66.21	0.5977	0.6513	0.6814	3.655
	GDPO-SR	25.48	0.2672	0.1978	119.06	69.19	0.6058	0.6614	0.7007	3.614
	StrSR	22.52	0.3182	0.2308	129.73	69.34	0.6621	0.6778	0.6954	3.700
	Ours	24.22	0.2943	0.2140	116.00	69.86	0.6902	0.6607	0.6937	3.830
ScreenSR	SeeSR	22.45	0.2233	0.1371	65.51	72.31	0.7363	0.6900	0.7454	4.344
	SUPIR	23.86	0.2001	0.1456	50.92	68.31	0.6599	0.6631	0.7059	4.058
	DiT4SR	20.73	0.2573	0.1600	79.08	71.44	0.6735	0.7098	0.7152	3.921
	SinSR	–	–	–	–	–	–	–	–	–
	OSEDiff	22.11	0.2301	0.1525	74.63	71.42	0.6271	0.6802	0.7310	4.012
	HYPIR	19.39	0.2515	0.1635	79.23	72.35	0.7031	0.7322	0.7505	4.087
	GDPO-SR	22.16	0.2251	0.1452	67.35	72.15	0.6409	0.7120	0.7463	4.062
	StrSR	22.09	0.1971	0.1373	65.56	72.00	0.7086	0.7396	0.7392	4.155
	Ours	21.21	0.2143	0.1339	67.91	72.97	0.7585	0.7182	0.7589	4.507

Table 2: Quantitative comparison on the RealLQ250 dataset.

Dataset	Metric	SeeSR	SUPIR	DiT4SR	SinSR	OSEDiff	HYPIR	GDPO-SR	StrSR	**Ours**
RealLQ250	MUSIQ↑	70.38	63.23	71.77	63.71	69.53	69.75	71.01	68.62	71.94
	TOPIQ↑	0.6637	0.5110	0.6420	0.5527	0.6073	0.6354	0.6098	0.6203	0.6859
	MANIQA↑	0.5927	0.5810	0.6308	0.5158	0.5787	0.6076	0.6154	0.5774	0.6253
	CLIPIQA+↑	0.7033	0.6253	0.7096	0.5923	0.6958	0.7197	0.7184	0.6873	0.7208
	Q-Align↑	4.143	4.154	4.226	3.745	4.244	4.226	4.361	4.261	4.258

Qualitative Analysis. As shown in Figure 3, we provide qualitative comparisons across diverse scenarios. In the first example, PNR-SR correctly reconstructs the human foot, whereas several competing methods mistakenly recover it as a dog paw. In the second example, our method most accurately restores the structure of the Chinese characters. In the third example involving an animated image, PNR-SR produces cleaner and sharper contours with fewer visible artifacts.

User Study. Following the user study protocol of DiT4SR Duan et al. (2025), we invited 18 volunteers to evaluate results on PhoneSR30. Participants were asked to compare the outputs in terms of image realism and fidelity. As shown in Table 3, the reported values denote the preference rate of our method over each competing method. Our method is consistently preferred over the other four competing methods.

Table 3: User study results.

Ours vs.	DiT4SR	HYPIR	GDPO-SR	StrSR
Realism	73.1%	59.6%	68.4%	62.7%
Fidelity	73.9%	58.1%	67.6%	64.9%

Comparison with Different One-Step Training Paradigms. We further view our method as a form of state-wise self-distillation, where preferences induced by different HR-exposure states are transferred toward the default one-step inference state. To validate the effectiveness of our method, we fix the base model as Z-Image and compare against three representative one-step training baselines: (1) CSD-based distillation; (2) GAN-based adversarial training following the HYPIR paradigm, using the HYPIR discriminator architecture, which is also adopted in StrSR; and (3) a strictly controlled GAN baseline, which is trained only at the default inference timestep t_m, while keeping the LR input

Table 4: Comparison of different one-step training paradigms under the same Z-Image base model on RealSR. CSD denotes CSD-based distillation, GAN-H denotes HYPIR-style adversarial training, and GAN-C denotes the strictly controlled GAN baseline.

Method	PSNR↑	LPIPS↓	DISTS↓	FID↓	MUSIQ↑	TOPIQ↑	MANIQA↑	CLIPIQA+↑	Q-Align↑
CSD	24.89	0.2898	0.2073	106.8	68.62	0.6636	0.6449	0.6662	3.702
GAN-H	23.88	0.3051	0.2188	122.0	68.33	0.6685	0.6560	0.6750	3.757
GAN-C	24.36	0.3031	0.2148	104.2	66.85	0.6350	0.6303	0.6476	3.739
Ours	24.22	0.2943	0.2140	116.0	69.86	0.6902	0.6607	0.6937	3.830

formulation and optimization hyperparameters identical to ours. Its discriminator adopts the same network architecture as our reward model, with the reward model optimization objective replaced by a standard adversarial loss. As shown in Table 4, on the RealSR test set, our method achieves the best performance across all no-reference metrics and ranks second on two full-reference metrics.

4.3 Ablation Studies

We further conduct component-wise ablation studies to assess the contribution of each key design. Specifically, we consider five variants: (1) **w/o reward-based post-training**: the SR policy is trained only with the original supervised objective; (2) **offline reward**: a two-stage training scheme is adopted, where the reward model is first trained offline using a fixed SR policy and is then frozen to optimize the SR policy; (3) **w/o LR conditioning**: the LR condition is removed from the reward model, which predicts rewards solely from the SR output; (4) **w/o dual-anchor calibration**: dual-anchor calibration is removed, leaving only preference ranking learning; and (5) **w/o historical reward baseline**: the historical reward baseline is removed, and the SR policy is optimized directly using the current reward scores.

Table 5: Ablation studies on RealSR set.

Variant	PSNR↑	LPIPS↓	CLIPIQA+↑	Q-Align↑
(1)	24.22	0.3095	0.6750	3.814
(2)	24.26	0.2959	0.6863	3.772
(3)	24.11	0.3151	0.6978	3.796
(4)	24.26	0.3170	0.6911	3.857
(5)	24.15	0.3130	0.6938	3.862
Full Model	24.22	0.2943	0.6937	3.830

As shown in Table 5, compared with the baseline without reward-based post-training, the full model improves or maintains both fidelity and perceptual quality metrics, indicating that reward-guided policy optimization can further enhance restoration quality beyond the original supervised objective. Notably, the offline reward variant also improves PSNR, LPIPS, and CLIPIQA+ over the baseline, showing that the reward model learned from policy-native preference supervision can already provide effective optimization signals for the SR policy; however, the full reward-policy co-training scheme achieves better overall performance than the offline variant, supporting the benefit of continuously updating the reward model as the policy evolves. Removing LR conditioning degrades both PSNR and LPIPS, indicating that LR conditioning is important for preserving restoration fidelity. In addition, removing either dual-anchor calibration or the historical reward baseline leads to degraded LPIPS, suggesting that both designs contribute to more stable online co-training. As shown in Figure 4, the full model generates textures with both high realism and strong fidelity to the input.

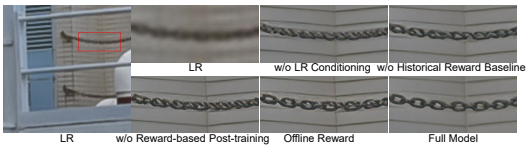


Figure 4: Ablation studies visualization.

5 Conclusion

In this work, we introduce PNR-SR, which reformulates reward-guided one-step Real-ISR as a closed-loop co-evolution process between the SR policy and reward model. By continuously deriving supervision from the policy itself, PNR-SR improves the reliability of reward-based post-training and consistently enhances restoration quality across diverse Real-ISR benchmarks.

Ethics Statement

This work adheres to the ICLR Code of Ethics. The study includes a user study involving human participants to compare the perceptual quality of restoration results produced by different methods. All datasets used in this work were obtained and used in accordance with their applicable usage guidelines and do not involve privacy violations. We also took care to minimize potential sources of bias or discriminatory outcomes throughout the research process. No personally identifiable information was collected or reported, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency, rigor, and research integrity throughout this work.

Reproducibility Statement

We have made every effort to ensure the reproducibility of the results reported in this paper. The code and relevant experimental materials are provided through an anonymous repository at https://anonymous.4open.science/r/PNR-SR-B550/. The paper also describes the experimental settings, training configurations, and methodological details required for reproduction.

All public training and testing datasets used in this work are publicly available. Our self-collected test dataset will be released upon acceptance of the paper. Since the dataset consists of real-world photographs that may contain location-related information, its release is deferred until after the review process.

We believe these measures will facilitate independent reproduction and further research based on our work.

References

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, pp. 126–135, 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. *NeurIPS*, 37:55443–55469, 2024.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, pp. 3086–3095, 2019.

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 TIP*, 33:2404–2418, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2020.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pp. 184–199. Springer, 2014.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *CVPR*, pp. 23174–23184, 2025.

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 *ICCV*, pp. 18948–18958, 2025.

Song Fei, Tian Ye, Sixiang Chen, Zhaohu Xing, Jianyu Lai, and Lei Zhu. Lucidnft: Lr-anchored multi-reward preference optimization for flow-based real-world super-resolution. *arXiv preprint arXiv:2603.05947*, 2026.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014.

Shuhao Han, Wenjie Liao, Hayden Vance, Hang Dong, Rui Zhang, Chun-Le Guo, and Chongyi Li. Dnf-sr: Dual-input and negative-aware feature fine-tuning for real-world image super-resolution. In *CVPR*, pp. 38176–38186, 2026.

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 *NeurIPS*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pp. 6840–6851, 2020.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Xin Jin, David Liu, Zhen Li, Mengmeng Wang, Peng Gao, and Harry Yang. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pp. 4401–4410, 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pp. 5148–5157, 2021.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pp. 1646–1654, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *CVPR*, pp. 1775–1787, 2023.

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 *ECCV*, pp. 430–448. Springer, 2024.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM TOG*, 44(6):1–21, 2025.

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

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 *CVPR*, pp. 4839–4849, 2026.

Yushuai Song, Weize Quan, Weining Wang, Jiahui Sun, Jing Liu, Pengbin Yu, Zhentao Chen, Meng Li, Wei Shen, Lunxi Yuan, et al. Refreward-sr: Lr-conditioned reward modeling for preference-aligned super-resolution. In *ECCV*, pp. 94–112. Springer, 2026.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *CVPR*, 2025.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, volume 37, pp. 2555–2563, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024a.

Jingkai Wang, Yixin Tang, Jue Gong, Jiatong Li, Shu Li, Libo Liu, Jianliang Lan, Yutong Liu, and Yulun Zhang. Spectral and trajectory regularization for diffusion transformer super-resolution. In *ECCV*, 2026.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pp. 1905–1914, 2021.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, pp. 25796–25805, 2024b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, pp. 25456–25467, 2024b.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Shihao Wang, Tianhe Wu, Qiaosi Yi, Shuai Li, and Lei Zhang. Dp^{2}o-sr: Direct perceptual preference optimization for real-world image super-resolution. *NeurIPS*, 38:172015–172041, 2025.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Xiangtao Kong, Jixin Zhao, Shihao Wang, and Lei Zhang. Vosr: A vision-only generative model for image super-resolution. In *CVPR*, 2026.

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. *NeurIPS*, 36:15903–15935, 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 *CVPR*, pp. 1191–1200, 2022.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, pp. 74–91. Springer, 2024.

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. *arXiv preprint arXiv:2603.16769*, 2026.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, pp. 25669–25680, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Shijie Zhao, Xuanyu Zhang, Bin Chen, Weiqi Li, Qunliang Xing, Kexin Zhang, Yan Wang, Junlin Li, Li Zhang, Jian Zhang, et al. Oars: Process-aware online alignment for generative real-world image super-resolution. *arXiv preprint arXiv:2603.12811*, 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 *ICLR*, volume 2026, pp. 134129–134150, 2026.

A Additional Methodological Details

For clarity, Algorithm 1 presents the training procedure of our method in detail.

Algorithm 1 Online Reward-Policy Co-Training

Require: Paired LR-HR dataset $\mathcal{D}$, HR-exposure timestep set $\mathcal{T}$, SR policy G_θ, reward model R_ϕ

Stage 1: Supervised Warm-up
Sample $\tau \in \mathcal{T}$ and optimize G_θ with $\mathcal{L}_{SR}$ under different HR-exposure levels.
Stage 2: Dual-Anchor Sample Collection
Freeze the policy G_{θ_0}.
Generate calibration samples using the highest and lowest exposure states $\tau_h, \tau_l \in \mathcal{T}$.
Stage 3: Online Reward-Policy Co-Training
Freeze G_{θ_0} and warm up R_ϕ with policy-native preference pairs and fixed dual-anchor samples.
Initialize the trainable policy $G_\theta \leftarrow G_{\theta_0}$.
Initialize the historical reward baseline μ_0 and set $\lambda_r > 0$.
for k = 1, 2, ... do
Sample LR-HR pairs and exposure levels $\tau^+, \tau^- \in \mathcal{T}$ with $\tau^+ > \tau^-$.
Generate policy-native restoration pair (y_{τ_+}, y_{τ_-}) using current G_θ.
Update G_θ with $\mathcal{L}_G = \mathcal{L}_{SR} + \lambda_r \mathcal{L}_{ReFL}$ using the current reward model and historical baseline μ_{k-1}.
Update R_ϕ with $\mathcal{L}_{RM}$ using the newly generated preference pairs and fixed dual-anchor samples.
Update the historical reward baseline μ_k.
end for
return G_θ and R_ϕ

B Additional Experiments

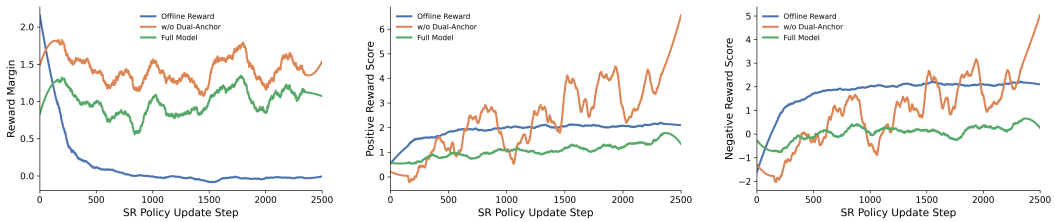
B.1 Experimental Settings

Implementation Details. For training, we apply LoRA to both the VAE and the base model, with the LoRA rank set to 64. We use the AdamW optimizer with a batch size of 12, and adopt a fixed text prompt, "high quality, sharp, detailed image", for all samples. The model is trained for a total of 43,000 optimization steps. Specifically, Stage 1 (Supervised Warm-up) is conducted for 40,000 steps with a learning rate of 5×10^{-5}. Stage 3 (Online Reward-Policy Co-Training) is then performed for an additional 3,000 steps, with the first 500 steps used for reward-model warm-up. During co-training, the learning rates of the SR model and reward model are both set to 2×10^{-5}. The hyperparameters λ_1, λ_2, λ_3, λ_c, λ_r, a_l, a_h, m_{gap}, m_{imp}, and ρ are set to 1, 2, 0.005, 0.25, 0.005, -0.98, 0.98, 2, 1 and 0.99, respectively. The HR-exposure timestep set is defined as $\mathcal{T} = \{0.50, 0.52, 0.54, \ldots, 0.98\}$. All experiments are conducted on three NVIDIA A100 GPUs.

PhoneSR30 and User Study. PhoneSR30 is a real-world super-resolution test set collected using the telephoto cameras of modern smartphones. It contains diverse indoor and outdoor scenes, covering a variety of content types, including plants, buildings, and both Chinese and English text. To further assess restoration quality under real-world conditions, we conduct a user study comparing our method with competing methods in terms of realism and fidelity. The study contains 120 comparison questions in total. Each participant is randomly assigned 30 questions, which helps reduce potential bias arising from a fixed subset or presentation order.

B.2 Experimental Results

Reward Dynamics Analysis. To further analyze the effectiveness of online reward-policy co-training and dual-anchor calibration, we visualize the evolution of positive and negative reward scores, together with their corresponding margins, during SR policy optimization in Fig. 5. Here, the positive and negative reward scores refer to the rewards assigned to the preferred sample y_{τ^+} and the non-preferred sample y_{τ^-}, respectively, where $\tau^+ > \tau^-$. As shown in Fig. 5, the offline reward strategy suffers from a rapid degradation in reward discrimination as the SR policy evolves.



(a) Reward Margin Evolution (b) Positive Reward Score (c) Negative Reward Score

Figure 5: Reward dynamics during SR policy optimization. (a) Reward margin evolution between positive and negative samples. (b) Positive reward score evolution. (c) Negative reward score evolution.

Specifically, the reward margin between positive and negative samples gradually collapses toward zero, indicating that the fixed reward model becomes increasingly unable to distinguish preferred from non-preferred restoration outputs under the evolving policy distribution. In contrast, removing dual-anchor calibration does not substantially reduce the reward margin, but leads to pronounced fluctuations in both positive and negative reward scores, revealing unstable reward-scale calibration during co-training. By providing stable reward references through dual anchors, the full model achieves smoother reward evolution while maintaining a consistent separation between preferred and non-preferred restoration outputs. These results show that online reward-policy co-training mitigates the distribution mismatch of a fixed reward model, while dual-anchor calibration further stabilizes the reward scale throughout policy optimization.

C Additional Visual Comparisons

We provide additional qualitative comparisons. (1) Figure 6 visualizes our preference data. As τ increases, both visual quality and fidelity improve, while preference pairs with different $\Delta\tau$ values remain visually distinguishable. (2) Figure 7 compares different training paradigms. Our method achieves better text restoration and artifact suppression. (3) Figure 8 presents additional visual comparisons between our method and other state-of-the-art methods. (4) Figure 9 presents comparisons on the PhoneSR30 dataset. Our method achieves better restoration quality across challenging content, including Chinese and English text, fur, and vegetation.

D AI use statement

Large Language Models (LLMs) were used as auxiliary tools during the preparation of this work. Their use included editorial assistance, such as language refinement, grammar checking, sentence rephrasing, and improving clarity and readability, as well as assistance with code debugging, refactoring, and implementation optimization.

LLMs were not used to determine the research objectives, formulate the core methodology, design the experiments, interpret the experimental results, or draw scientific conclusions. All research ideas, methodological decisions, experimental designs, and analyses were developed by the authors. Any LLM-assisted code suggestions were reviewed, tested, and validated by the authors before being incorporated into the experiments.

The authors take full responsibility for the manuscript, code, experimental results, and conclusions presented in this work.



Figure 6: Examples of preference data. A larger τ indicates that more information from the HR image is exposed. Here, we show preference pairs constructed with different $\Delta\tau$ values.

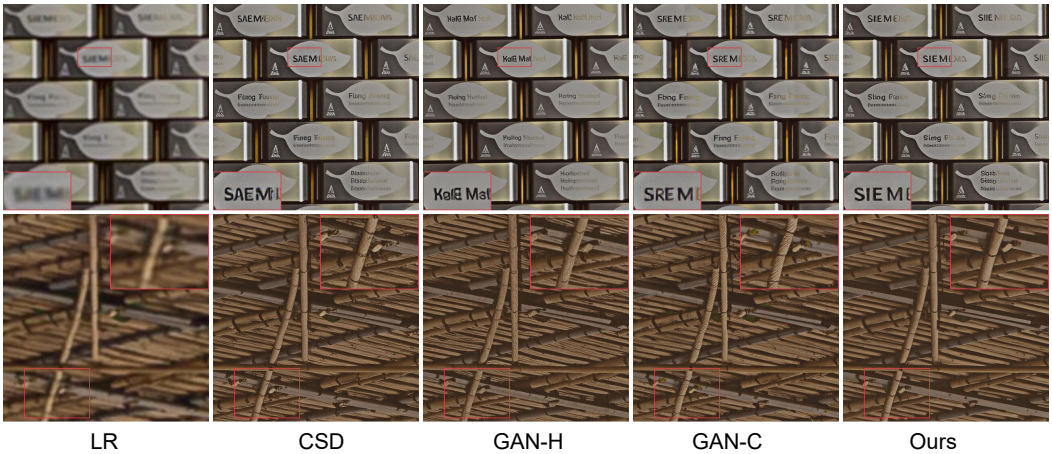


Figure 7: Comparison of different training paradigms using the same base model.

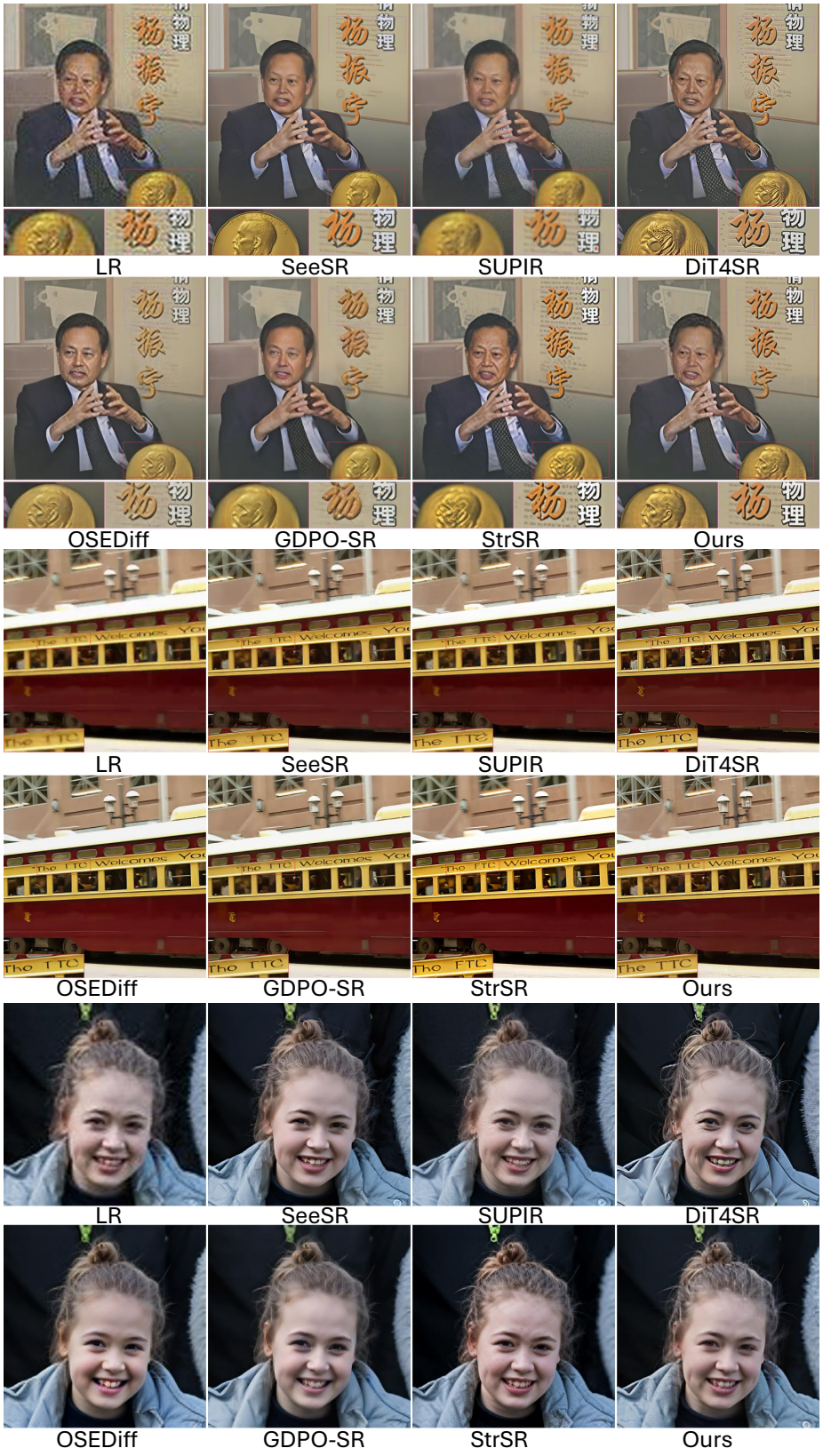


Figure 8: Additional qualitative comparison with state-of-the-art methods.

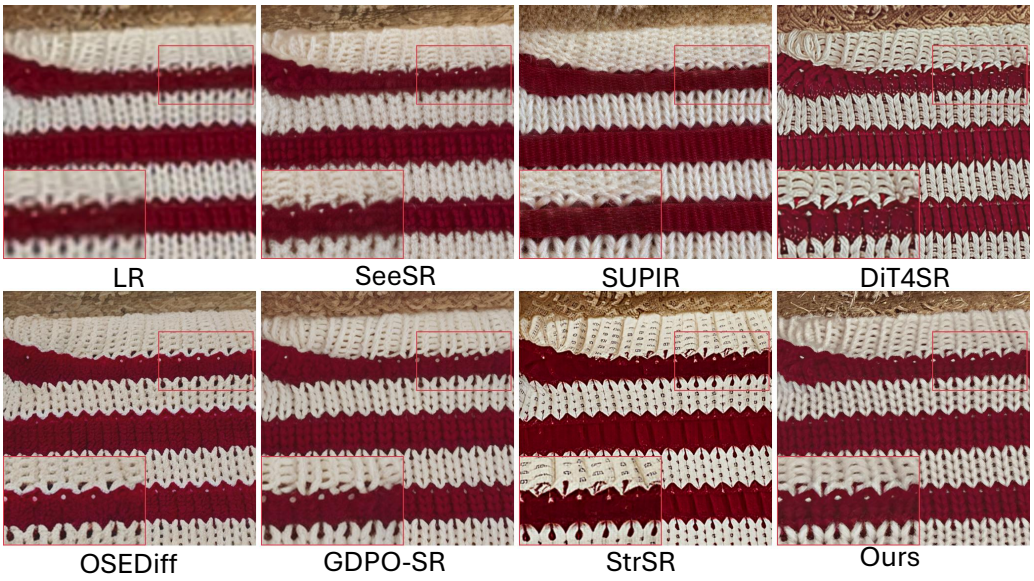


Figure 8: Additional qualitative comparison with state-of-the-art methods. (continued)



Figure 9: Qualitative comparison with state-of-the-art methods on PhoneSR30 set. The first example carries a risk of positional leakage, so we cropped the image.