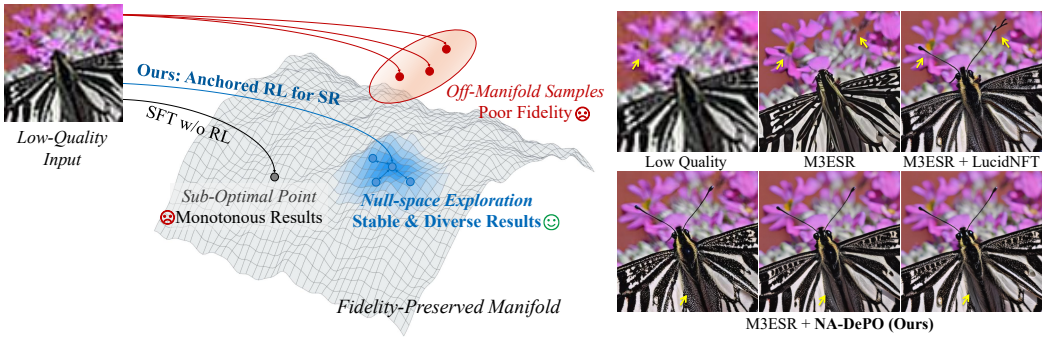
DECOUPLE THEN REVIVE: TOWARDS PERCEPTUAL SUPER-RESOLUTION VIA REINFORCEMENT LEARNING WITH ANCHORED GENERATIVE PRIORS

Anonymous authors
Paper under double-blind review



(a) Conceptual Illustration. **(b)** Qualitative Comparison.

Figure 1: We propose **NA-DePO**, a null-space anchored RL-for-SR method to improve balance between structural fidelity and perceptual detail quality.

ABSTRACT

Diffusion models advance RealSR through progressive distribution matching, yet existing training paradigms remain challenged: supervised finetuning on paired data tends to produce over-smoothed textures or misaligned semantics, while RL-based approaches introduce structural and textural inconsistency. In this paper, we propose NA-DePO (Null-space Anchored Dense Preference Optimization), which decomposes the SR signal into observation-constrained and unconstrained components via null-space decomposition, and restricts RL exploration exclusively to the unconstrained generative subspace to leverage generative priors without compromising structural consistency. We decouple the optimization process: we maintain the LR fidelity via a range-space constraint, and conduct negative-aware contrastive policy learning within the null-space. Furthermore, to provide informative and stable guidance, we design a dense reward mechanism incorporating VLM-driven local reward maps and timestep-aware scheduling, effectively reducing multi-reward conflicts. Extensive experiments demonstrate that NA-DePO significantly improves the balance between realistic detail generation and faithful structural restoration.

1 INTRODUCTION

Image Super-Resolution (SR) is a fundamental problem in low-level vision. Formally, this reduces to estimating the conditional distribution $p(\text{HR}|\text{LR})$, naturally framed as a MAP inference problem. Early methods train end-to-end networks (Dong et al., 2016a;b; Liang et al., 2021; Chen et al., 2023; Guo et al., 2025) under pixel-wise losses (e.g., ℓ_1 or ℓ_2), which implicitly impose a unimodal, pixel-wise prior over $p(\text{HR}|\text{LR})$. This inevitably produces unrealistic over-smoothed results with poor perceptuality, and limits generalization to the complex, heterogeneous degradations of real-world scenarios. The root cause is that such pixel-level objectives reduce the ill-posed, one-to-many mapping of Real-World SR (RealSR) to a deterministic one-to-one regression, which is known to suffer from regression-to-the-mean issue (Bruna et al., 2016).

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.

Generative models reframe RealSR as distribution alignment over $p(\text{HR}|\text{LR})$. GAN-based methods (Zhang et al., 2021; Wang et al., 2021) leverage adversarial training to match the HR distribution in a single forward pass, while **diffusion models** further model this distribution through iterative refinement, progressively recovering perceptually realistic high-frequency details step by step. Current diffusion-based SR methods typically train on synthetic paired data with multiple degradations, and design specific condition insertion architectures (Wang et al., 2024; Lin et al., 2024; Yi et al., 2025; Duan et al., 2025; Yu et al., 2024), reformulate noise-to-image trajectories with minimal architectural modifications via LR guidance (Zhang et al., 2025; Yue et al., 2025a;b; Sun et al., 2025), adversarial training (Lin et al., 2025), or incorporate degradation-robust high-level guidance, such as enriched prompts (Yang et al., 2024; Wu et al., 2024), object bounding boxes (He et al., 2025), segmentation maps (Xiao et al., 2024), and integrated multi-modal features (Mei et al., 2025; Luo et al., 2026) (*e.g.*, depth, edge, segmentation, uncertainty, and DINO features). Among these, multi-modal guidance provides degradation-robust representations that complement pixel-level supervision, offering a promising direction for RealSR.

Despite their advances, diffusion-based RealSR remains limited by an inaccurate characterization of the true conditional manifold: training on synthetic paired data forces the model to optimize against an oversimplified proxy of an inherently diverse target space. As a result, enforcing single-point fitting under such incomplete constraints inadvertently suppresses valid high-quality variations, causing the generator to **collapse toward a narrow conditional sub-manifold** rather than converge to the true distribution.

Reinforcement Learning (RL) (Christiano et al., 2017) offers a natural remedy: rather than passively fitting paired data, it enables *on-policy exploration* guided by IQA metrics, steering the generative trajectory toward a broader target distribution and alleviating mode collapse. Pioneering efforts integrate RL into diffusion-based RealSR, using reward-driven policy optimization to align generation with perceptual quality (Wu et al., 2025; Yi et al., 2026; Fei et al., 2026). However, existing RL-based RealSR methods **couple** faithful signal restoration with perceptual generation, leaving two key obstacles unresolved: **(1) Unbounded exploration.** RL's strong exploration capacity is a double-edged sword: without explicit constraints, it can distort underlying structures and push the generated distribution off the valid manifold. **(2) Reward design.** SR's inherent spatial diversity makes it ill-suited to global scalar rewards, weakening policy guidance. Meanwhile, discrepancies among IQA metrics cause *perceptual divergence* (*e.g.*, textures mixed with semantic hallucinations), introducing gradient variance that destabilizes training. Another line of work turns to IQA metrics as proxy estimators for the high-quality manifold. While IQA metrics (*e.g.*, LIQE (Zhang et al., 2023), CLIPIQA (Wang et al., 2023a), Q-Align (Wu et al., 2023)) serve as proxy estimators for the high-quality manifold, they characterize perceptual distance but do not provide optimization signals compatible with diffusion model training.

This raises a core question: *How can we decouple fidelity preservation from perception-driven RL exploration to enable more free-form generation, while providing more informative reward guidance?* To this end, we propose **NA-DePO**, a **N**ull-space **A**nchored **De**nse **P**reference **O**ptimization framework. Our approach alleviates the fidelity-exploration dilemma by decoupling both the policy optimization and reward formulation. As for the **RL policy**, we anchor the RL exploration within the **null-space**, which essentially ensures the intrinsic signal constancy during generative distribution alignment. Specifically, we decouple the optimization target into two parts: one for negative sample-aware contrastive fine-tuning in the re-composed latent space, and the other for the strict ℓ_2 constraint in the range-space basis. As for the **reward formulation**, we leverage Vision-Language Models (VLMs) to perform robust image quality and semantic consistency assessment, and propose a training-free attention modulation strategy to query fine-grained semantics from VLMs. To strengthen the informative supervision with multiple rewards, we factorize the reward spatially and temporally. We expand scalar rewards into spatialized reward maps, and assign each reward to different diffusion timesteps, alleviating potential multiple reward conflicts. For more stable optimization, we design a simple yet theoretically-sound sigmoid-normalization, which reduces the noisy gradient variance caused by biased IQA observations. By systematically decoupling both the exploration space and the rewards, our NA-DePO achieves a more stable and accurate distribution alignment, enabling faithful and realistic restoration.

Our contributions can be summarized as follows:

- **Null-Space Anchored RL Exploration.** We propose NA-DePO, which explicitly anchors the VLM generative exploration within the null-space, achieving significantly improved balance between fidelity and perceptual quality.
- **Spatial-Temporal Dense Guidance.** We expand scalar rewards to local reward maps, and assign each reward to different diffusion stages, enabling informative and stable guidance.
- **VLM-Guided Semantic Rewards.** We propose a semantically enriched reward via a training-free attention modulation scheme, which enhances semantic consistency.

2 Related Works

2.1 Diffusion Models for Real-world SR

RealSR aims to restore high-quality images from LR inputs suffering from unknown and complex degradations. Moving beyond traditional regression-based networks (Dong et al., 2016a; Liang et al., 2021) and GAN-based frameworks (Zhang et al., 2021; Wang et al., 2021), diffusion models (Ho et al., 2020; Lipman et al., 2023) have recently bloomed thanks to their exceptional distribution modeling capabilities. Current diffusion-based SR methods generally fall into two paradigms. The first paradigm introduces specific condition insertion architectures or reformulates the noise-to-image sampling trajectory to explicitly inject LR features into the denoising process (Wang et al., 2024; Lin et al., 2024; Yue et al., 2025b; Zhang et al., 2025). The second paradigm leverages informative high-level features as the degradation-robust representation, such as enriched textual prompts (Yang et al., 2024; Wu et al., 2024), bounding boxes (He et al., 2025), or pre-trained multi-modal features like depth, edge, or DINO (Mei et al., 2025; Luo et al., 2026). While these methods successfully harness the generative prior to synthesize realistic details, they universally rely on SFT with synthetic paired data, which inevitably forces the pre-trained generative prior to collapse onto a narrow, deterministic sub-manifold and compromises the optimal trade-off between fidelity and perceptuality.

2.2 Reinforcement Learning for Diffusion Models

RL has demonstrated remarkable success in aligning pre-trained diffusion models with complex human preferences or non-differentiable objectives. In the realm of Text-to-Image (T2I) generation, there are two RL paradigms, called reverse-process (GRPO-style) and forward-process (DPO-style) RL. GRPO-style RL (such as FlowGRPO (Liu et al., 2025a), DanceGRPO (Xue et al., 2025) and MixGRPO (Li et al., 2025)) adopts the main scheme from post-training of Large Language Models (LLMs), which treats the whole denoising path as a multi-step Markov Decision Process (MDP) and optimizes the model likelihood through reward guidance. However, these methods are limited to inaccurate likelihood estimation, costly rollout and manifold deviation (Li et al., 2025). On the contrary, forward-process RL directly performs policy optimization on the forward diffusion process through the flow matching objective on the split positive/negative samples, such as DiffusionDPO (Wallace et al., 2024; Liu et al., 2025b) and DiffusionNFT (Zheng et al., 2026). Recently, adapting RL for image restoration tasks like SR has garnered increasing attention. However, fundamental obstacles remain when adapting generative RL to RealSR.

3 NA-DePO

3.1 Preliminary

1) Flow Matching. Flow Matching (Lipman et al., 2023; Liu et al., 2023) models a generative process by learning a time-dependent vector field $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)$ that transports a noise distribution p_0 to a target data distribution p_1. Specifically, given noise $\boldsymbol{x}_0 \sim p_0$ and data $\boldsymbol{x}_1 \sim p_1$, they are bridged by a linear probability path $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$ for $t \in [0, 1]$. The model is trained to regress the constant ground-truth trajectory velocity, $\boldsymbol{x}_1 - \boldsymbol{x}_0$, using an MSE loss:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\boldsymbol{x}_0,\boldsymbol{x}_1}\left[||\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)||_2^2\right]. \tag{1}$$

Once trained, new samples are generated by solving the corresponding ODE along the learned velocity field starting from random noise.

2) DiffusionNFT: Diffusion Negative-aware Fine-tuning. Different from existing GRPO-style RL methods that optimize the whole denoising trajectory through reverse diffusion process, DiffusionNFT (Zheng et al., 2026) performs a smarter policy optimization within the forward process, exploring the generation space with a pair of contrastive directions and a reward-weighted velocity

loss. Formally, DiffusionNFT first performs interior and exterior interpolation between the original/current velocity $\boldsymbol{v}_{\text{old}}, \boldsymbol{v}_\theta$ to construct implicit positive and negative indicators $\boldsymbol{v}_\theta^+$ and $\boldsymbol{v}_\theta^-$, which represent the current contrastive exploration directions:

$$\boldsymbol{v}_\theta^+(\boldsymbol{x}_t, t, \boldsymbol{c}) = (1-\beta)\boldsymbol{v}_{old}(\boldsymbol{x}_t, t, \boldsymbol{c}) + \beta \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}), \tag{2}$$

$$\boldsymbol{v}_\theta^-(\boldsymbol{x}_t, t, \boldsymbol{c}) = (1+\beta)\boldsymbol{v}_{old}(\boldsymbol{x}_t, t, \boldsymbol{c}) - \beta \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}), \tag{3}$$

where the hyperparameter β controls the exploration strength, and $\boldsymbol{c}$ denotes the conditions according to the task. To align the generative model with external reward models, the training objective is formulated as a reward-weighted velocity matching loss:

$$\mathcal{L}_{\text{NFT-v}} = \mathbb{E}_{\boldsymbol{c}, \boldsymbol{x}_0, t}\left[r\|\boldsymbol{v}_\theta^+ - \boldsymbol{v}\|_2^2 + (1-r)\|\boldsymbol{v}_\theta^- - \boldsymbol{v}\|_2^2 \right], \tag{4}$$

where $\boldsymbol{v}$ denotes the velocity target pointing from the rollout samples $\boldsymbol{x}$ to Gaussian noises ϵ, *i.e.*, $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{x}$. In Eq. 4, high-reward samples ($r \to 1$) emphasize the positive direction term to reinforce the current exploration, while low-reward samples ($r \to 0$) increase the contribution of the negative term, effectively pushing the model away from undesired generation trajectories.

In practice, DiffusionNFT projects velocity to images[1] by one-step denoising along the velocity:

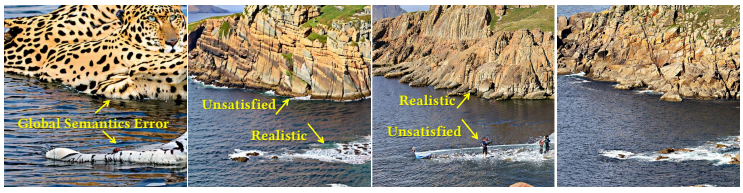
$$\hat{\boldsymbol{x}} = \boldsymbol{x}_t - t\boldsymbol{v}, \quad \hat{\boldsymbol{x}}^+ = \boldsymbol{x}_t - t\boldsymbol{v}_\theta^+, \quad \hat{\boldsymbol{x}}^- = \boldsymbol{x}_t - t\boldsymbol{v}_\theta^-. \tag{5}$$

Then, it performs image-level flow matching loss, which is equivalent to Eq. 4:

$$\mathcal{L}_{\text{NFT-x}} = \mathbb{E}_{\boldsymbol{c}, \hat{\boldsymbol{x}} \sim \pi_{\text{old}}, t}\left[r \circ \|\hat{\boldsymbol{x}}^+ - \hat{\boldsymbol{x}}\|_2^2 + (1-r) \circ \|\hat{\boldsymbol{x}}^- - \hat{\boldsymbol{x}}\|_2^2 \right]. \tag{6}$$

3.2 Motivation

Recent advances in RL-based post-training have substantially improved generative tasks, including text-to-image synthesis and image editing (Liu et al., 2025b;a; Zheng et al., 2026; Wu et al., 2026; 2025). Meanwhile, large generative models have demonstrated strong priors for real-world SR (Esser et al., 2024; Liu et al., 2023; Lin et al., 2024; Yang et al., 2024; Wang et al., 2024; Wu et al., 2024; Yu et al., 2024), motivating the exploration of RL to further enhance restoration quality.



(a) Seed 1 Result **(b)** Seed 2 Result **(c)** Seed 3 Result **(d)** Ground-Truth

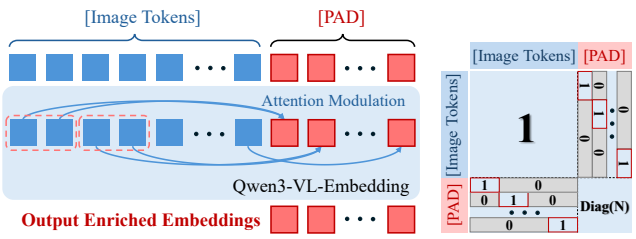
Figure 2: **Motivation of our Null-Space Anchored RL and Dense Rewards.** As different seeds leads to diverse SR results, we observe two primary failure modes: **(a)** some samples suffer from severe global semantic degradation; **(b&c)** others achieve satisfactory restoration in specific local regions but fail to maintain visual consistency elsewhere.

However, as shown in Fig. 1, 2, directly applying RL to SR remains challenging due to three key discrepancies between generation and restoration: **(1)** exploration may violate fidelity, **(2)** global scalar rewards overlook spatially localized quality variations, and **(3)** noisy SR metrics can yield unstable supervision. To address these issues, we propose **NA-DePO**, a null-space anchored RL framework with dense and robust reward guidance, trained on paired synthetic data.

3.3 VLM-based Dense Reward via Attention Modulation

Targeting a local reward with enriched semantic guidance, we propose **QwenVL-Local**, a novel reward built upon the QwenVL3-Embedding (Li et al., 2026) with a training-free attention modulation approach. A standard QwenVL-Embedding model takes a sequence of concatenated image tokens as input, and utilizes a single embedding token to query semantic information from image tokens. As shown in Fig. 3, we involve N additional embedding tokens to query local semantics and reformulate the attention allocation mechanism. The customized masking rules are simple yet effective: (1) **Global Image Context**: Image tokens maintain full visibility to each other to extract coherent global semantics. (2) **Embedding Isolation**: Embedding tokens are completely separated from one another. (3) **Spatial Restriction**: The visibility of each embedding token to image tokens is confined to their corresponding spatial assignments.

[1] We use latent diffusion models in practice and denote it in "images" here for simplicity.



(a) Overall Workflow of Enriched QwenVL-Embedding (b) Attention Masks

Figure 3: The proposed **attention modulation** mechanism for local QwenVL reward.

In summary, each embedding token is forced to independently query its assigned local patch, while the image tokens still maintain a global information flow, leading to a local semantically enriched representation. After that, we calculate the cosine similarity between corresponding local embeddings of the GT and policy-generated images. This naturally generates a dense, patch-level reward map $\mathbf{r}$, providing precise and discriminative supervision for fine-grained preference optimization.

Based on this, we replace the scalar reward r in Eq. 4 with a spatial reward map $\mathbf{r} \in [0,1]^{H\times W}$:

$$\mathcal{L}_{\text{NFT-Local}} = \mathbb{E}_{\boldsymbol{c},\,\hat{\boldsymbol{x}},\,t}\Big[\big\|\sqrt{\mathbf{r}}\circ(\hat{\boldsymbol{x}}^{+}-\hat{\boldsymbol{x}})\big\|_2^2 + \big\|\sqrt{1-\mathbf{r}}\circ(\hat{\boldsymbol{x}}^{-}-\hat{\boldsymbol{x}})\big\|_2^2\Big],$$

where $\circ$ denotes the Hadamard product.

3.4 Null-Space Anchored RL Exploration

When applying RL to SR, our objective is fundamentally dual: we seek to harness preference alignment to fully unleash the generative model's profound capacity for recovering semantics and textures, while preserving signal-level fidelity throughout the optimization process.

Our primary idea is to constrain the RL exploration in the decoupled space. According to *Range-Null Space Decomposition* (Wang et al., 2023b), any image $\boldsymbol{x}_0$ can be decomposed into two spaces with any linear operator $\boldsymbol{A}$ (*e.g.,* down-sample, grayish, *etc*) and its pseudo-inverse $\boldsymbol{A}^\dagger$:

$$\boldsymbol{x}_0 = \underbrace{\boldsymbol{A}^\dagger \boldsymbol{y}}_{\text{Range-space}} + \underbrace{(\boldsymbol{I}-\boldsymbol{A}^\dagger\boldsymbol{A})\boldsymbol{x}_0}_{\text{Null-space}}. \tag{7}$$

Motivated by this, we define $\boldsymbol{A}$ to be the $\times 4$ down-sample operator (average pooling in our implementation), decoupling original generation space into high-frequency details space as null-space and low-frequency basis space as range-space. By projecting RL into the residual null-space, and preserving the fundamental basis components that remain invariant across resolution changes, the RL exploration effectively maintains structural consistency and improves fidelity.

Formally, we decouple the optimization target in Eq. 6 to two parts: On one hand, we anchor RL exploration by **re-compositing** the null-space of the RL policy and range-space basis of the ground-truth images $\boldsymbol{x}_0$:

$$\tilde{\boldsymbol{x}}_0^{+} = \boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{x_0} + (\boldsymbol{I}-\boldsymbol{A}^\dagger\boldsymbol{A})\hat{\boldsymbol{x}}^{+}, \tag{8}$$

$$\tilde{\boldsymbol{x}}_0^{-} = \boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{x_0} + (\boldsymbol{I}-\boldsymbol{A}^\dagger\boldsymbol{A})\hat{\boldsymbol{x}}^{-}, \tag{9}$$

$$\mathcal{L}_{\text{anchor-RL}} = \|\sqrt{\mathbf{r}}\circ(\tilde{\boldsymbol{x}}_0^{+}-\hat{\boldsymbol{x}})\|_2^2 + \|\sqrt{1-\mathbf{r}}\circ(\tilde{\boldsymbol{x}}_0^{-}-\hat{\boldsymbol{x}})\|_2^2, \tag{10}$$

where $\hat{x}$ denotes the policy samples. On the other hand, we apply L_2 constraints on current velocity estimation v_θ and the basis components of GT images $\boldsymbol{A}\boldsymbol{x}_{\text{gt}}$ within range-space:

$$\mathcal{L}_{\text{constrain}} = \|\boldsymbol{A}\hat{\boldsymbol{x}}_\theta - \boldsymbol{A}\boldsymbol{x}_{\text{gt}}\|_2, \quad \hat{\boldsymbol{x}}_\theta = \boldsymbol{x}_t - t\boldsymbol{v}_\theta. \tag{11}$$

The final training objective can be formulated as:

$$\mathcal{L}_{\text{NA-DePO}} = \mathcal{L}_{\text{anchor-RL}} + \lambda\cdot\mathcal{L}_{\text{constrain}}, \tag{12}$$

where λ denotes the regularization strength and is set to 1.

Notably, the down-sample operator $\boldsymbol{A}$ serves as a surrogate observation operator: It defines an LR-observable component that is kept fixed during preference optimization, and a complementary null-space component where the generative prior is allowed to explore. $\boldsymbol{A}$ necessitates **no prior knowledge** of the RealSR degradation process.

3.5 Reward Sensitivity Control via Sigmoid Reweighting

Our policy optimization relies on empirical reward models, whose predictions may contain observation noise and scale variations. We theoretically prove that noisy empirical reward leads to unstable optimization, and propose a **sigmoid-reweighting strategy** to use it as a relative preference signal to provide informative RL guidance.

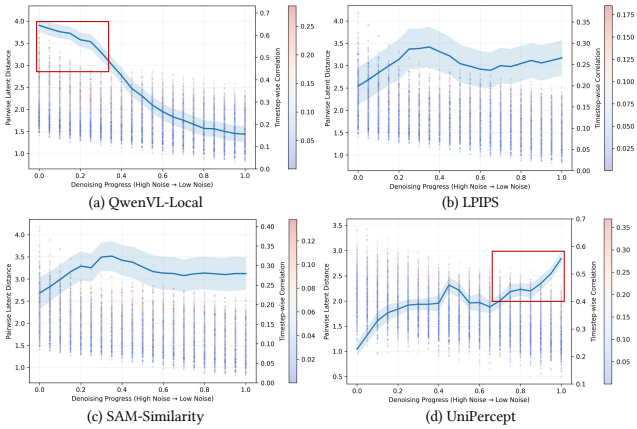


Figure 4: **Motivation experiment of reward scheduling**. We compute timestep-wise correlation between pairwise latent distances and final reward differences. Point colors indicate reward gaps, while the curve shows their Spearman correlation. We find that QwenVL-Local is sensitive during early denoising, while UniPercept emerges later.

1) Gradient Sensitivity to Reward Estimation Error. Eq. 4 is theoretically motivated when $r = r^* \equiv p(\mathbf{o} = 1|\boldsymbol{x}_0, \boldsymbol{c}) \in [0, 1]$ (Levine, 2018). In practice, we only have an empirical estimate $\hat{r} = r^* + \delta$, where δ denotes the reward estimation error. We define the baseline residual as $\boldsymbol{e} = \boldsymbol{v}^{old} - \boldsymbol{v}$ and the policy shift as $\boldsymbol{d}_\theta = \boldsymbol{v}_\theta - \boldsymbol{v}^{old}$. For a stochastic sample, the gradient is

$$\hat{\boldsymbol{g}} = 2\beta \left[\beta \boldsymbol{d}_\theta + (2\hat{r} - 1)\boldsymbol{e}\right] \nabla_\theta \boldsymbol{v}_\theta. \tag{13}$$

Let $\boldsymbol{g}^*$ denote the corresponding gradient obtained with the ideal reward r^*. Their difference is

$$\hat{\boldsymbol{g}} - \boldsymbol{g}^* = 4\beta\delta\, \boldsymbol{e} \nabla_\theta \boldsymbol{v}_\theta. \tag{14}$$

Therefore, conditioned on the sampled state and diffusion trajectory, the reward-noise-induced gradient variance is

$$\mathbb{E}_\delta \left[\|\hat{\boldsymbol{g}} - \mathbb{E}_\delta[\hat{\boldsymbol{g}}]\|_2^2 \right] = 16\beta^2 \operatorname{Var}(\delta) \left\| \boldsymbol{e} \nabla_\theta \boldsymbol{v}_\theta \right\|_2^2. \tag{15}$$

Since both r^* and $\hat{r}$ are bounded in $[0, 1]$, $\delta \in [-1, 1]$ and hence $\operatorname{Var}(\delta)$ is finite. Nevertheless, Eq. 15 shows that reward estimation noise is directly propagated to the policy gradient and can be amplified by a large baseline residual or policy Jacobian.

2) Variance Contraction via Sigmoid Reweighting. To reduce the sensitivity of policy optimization to reward perturbations, we reweight the empirical reward as

$$h(\hat{r}) := \sigma(\tau(2\hat{r} - 1)), \tag{16}$$

where $\sigma(x) = 1/(1 + e^{-x})$ and τ is set to 1 by default. The derivative of h satisfies

$$|h'(r)| = 2\tau\sigma'(\tau(2r - 1)) \leq \frac{\tau}{2}, \tag{17}$$

since $\sigma'(x) \leq 1/4$. Thus, h is globally $\tau/2$-Lipschitz. Defining $\delta_{\text{sig}} = h(\hat{r}) - h(r^*)$, we obtain the exact variance bound

$$\operatorname{Var}(\delta_{\text{sig}}) \leq \frac{\tau^2}{4} \operatorname{Var}(\delta). \tag{18}$$

Consequently, the reward-noise-induced gradient variance after sigmoid reweighting satisfies

$$\operatorname{Var}_\delta(\hat{\boldsymbol{g}}_{\text{sig}}) \leq \frac{\tau^2}{4} \operatorname{Var}_\delta(\hat{\boldsymbol{g}}), \tag{19}$$

where the variance denotes the conditional squared gradient deviation as in Eq. 15. For $\tau = 1$, sigmoid reweighting therefore reduces the corresponding variance upper bound by a factor of four.

3) Sensitivity Interpretation. Locally, a first-order expansion further gives

$$\operatorname{Var}(\delta_{\text{sig}}) \approx 4\tau^2 \left[\sigma'(\tau(2r^* - 1))\right]^2 \operatorname{Var}(\delta). \tag{20}$$

This sensitivity is maximal around the ambiguous region $r^* = 0.5$ and decreases toward $r^* = 0$ or 1, providing stronger attenuation for high-confidence reward predictions. Since $h(\cdot)$ is strictly monotonic, the transformation preserves reward ranking while trading calibrated reward magnitude for reduced optimization sensitivity.



Figure 5: **Qualitative Comparison.** Our NA-DePO achieves superior details generation and faithful structural restoration.

3.6 Multi-Reward Scheduling

We employ multiple reward models tailored to different quality dimensions, broadly categorized into two groups: (1) **Semantic Consistency**, including SAM3 (Carion et al., 2025) feature cosine similarity, LPIPS and the QwenVL-Local reward proposed in Sec. 3.3, and (2) **Texture Realism**, evaluating low-level details via UniPercept (Cao et al., 2025).

To ensemble these rewards with more informative guidance, we propose a timestep-decoupled multi-reward ensemble method. Recent studies (Esser et al., 2024; Meng et al., 2022) reveal that flow matching inherently decouples generation along the probability path, formulating global semantics during the early stages (high noise) while refining high-frequency textures in the later stages (low noise). Inspired by this, we apply semantic-oriented rewards to supervise the early steps (initial 20% for QwenVL-Local), and texture rewards in the later steps (last 70% for UniPercept) to enhance detail realism. We adopt LPIPS and SAM rewards at all steps.

We provide a toy experiment in Fig. 4 to verify our motivation. We sample multiple restoration trajectories for each LR image in a validation subset, save their intermediate latents, and compute the final rewards on generated HR images. For each denoising timestep, we compute the correlation between pairwise latent gaps and pairwise final reward gaps across sampled trajectories. This allows us to directly inspect when a given final reward dimension becomes aligned with the trajectory geometry. The results in Fig. 4 show that semantic reward differences are reflected in the trajectory geometry earlier, whereas texture-related differences emerge mainly during later refinement, motivating our timestep-decoupled reward schedule. The ablation in Tab. 2 provides verification.

4 Experiments

4.1 Implementation Details

1) Training Data. We utilize paired synthetic data for RL post-training using RealESRGAN (Wang et al., 2021) as degradation pipeline [2]. We sample a hard subset filtered by the largest IQA metrics dropping based on LSDIR (Li et al., 2023) and FFHQ (Karras et al., 2019) (1,358 images in total).

2) RL Policy. We choose M3ESR (Luo et al., 2026) as a pretrained backbone to employ our RL framework. We use classifier-free guidance (Ho & Salimans, 2021) with scale=8 during roll-out. The group size is set to 12, and the learning rate is set to 5e-5. Following DiffusionNFT (Zheng et al.,

[2] Although we utilize synthetic paired data, we allow diffusion models to explore a broader space of SR solutions rather than fit a point-to-point mapping, which differs fundamentally from existing SFT-based methods

Table 1: **Quantitative comparison results on DIV2K (Agustsson & Timofte, 2017) and RealLQ (Ai et al., 2024) datasets.** Best and second-best results are in **bold** and underlined. We additionally denote the improvement compared to the pretrained backbone without RL in magenta. † denotes those methods that we re-train on our data and backbone.

Method	**DIV2K** (Agustsson & Timofte, 2017)					
	LIQE↑	BRISQUE↓	Q-Align↑	WADIQAM_NR↑	CLIPIQA↑	CLIPIQA+↑
BSRGAN (Zhang et al., 2021)	4.0155	17.0244	3.4473	-0.1794	0.5924	0.6551
RealESRGAN (Wang et al., 2021)	4.1617	20.0862	3.5039	-0.1634	0.6539	0.6733
SeeSR (Wu et al., 2024)	4.6037	15.6099	4.3008	0.0609	0.7361	0.7594
DiffBIR (Lin et al., 2024)	4.5559	19.9810	4.1641	-0.0098	0.7368	0.7592
PiSA-SR (Sun et al., 2025)	4.5994	12.9555	4.2500	0.0397	0.7335	0.7635
InvSR (Yue et al., 2025a)	4.0425	16.8770	3.9980	0.0208	0.7357	0.7551
DiT4SR (Duan et al., 2025)	4.5917	17.8489	4.1250	0.0152	0.7176	0.7367
M3ESR (Luo et al., 2026)	4.8172	15.9677	4.3594	0.1146	0.7807	0.7576
M3ESR + DP2O-SR† (Wu et al., 2025)	4.7621	14.8379	4.2710	0.1056	0.7652	0.7229
M3ESR + LucidNFT† (Fei et al., 2026)	4.7507	13.0861	4.3001	0.1307	0.7799	0.7547
M3ESR + Ours	**4.8405**	**12.4988**	**4.4648**	**0.1377**	**0.8041**	**0.7636**
Performance Gain to M3ESR	↑ 0.48%	↑ 21.72%	↑ 2.42%	↑ 20.16%	↑ 3.00%	↑ 0.79%
Method	**RealLQ** (Ai et al., 2024)					
	LIQE↑	BRISQUE↓	Q-Align↑	WADIQAM_NR↑	CLIPIQA↑	CLIPIQA+↑
BSRGAN (Zhang et al., 2021)	3.9045	14.6321	3.5536	-0.2191	0.5988	0.6567
RealESRGAN (Wang et al., 2021)	4.0390	20.7046	3.4778	-0.2145	0.6371	0.6659
SeeSR (Wu et al., 2024)	4.3827	14.0860	4.1564	0.0020	0.7173	0.7446
DiffBIR (Lin et al., 2024)	4.4526	19.6524	4.0010	-0.0722	0.7233	0.7208
PiSA-SR (Sun et al., 2025)	4.0267	19.0132	4.2058	-0.0533	0.7053	0.7151
InvSR (Yue et al., 2025a)	4.4514	11.6125	4.0081	-0.0799	0.7262	0.7479
DiT4SR (Duan et al., 2025)	4.5987	13.3985	4.2104	0.0539	0.7513	0.7435
M3ESR (Luo et al., 2026)	4.6539	12.9490	4.2119	0.0857	0.7630	0.7508
M3ESR + DP2O-SR† (Wu et al., 2025)	4.7096	12.5802	4.2905	0.0872	0.7650	0.7503
M3ESR + LucidNFT† (Fei et al., 2026)	4.7021	12.0981	4.1023	0.0982	0.7692	0.7515
M3ESR + Ours	**4.7149**	**11.4443**	**4.3256**	**0.1102**	**0.7932**	**0.7561**
Performance Gain to M3ESR	↑ 1.31%	↑ 11.62%	↑ 2.70%	↑ 28.59%	↑ 3.96%	↑ 0.71%

2026), we use LoRA (Hu et al., 2022) to fine-tune the pretrained model, and use EMA ($\eta = 0.9$) to update the old policy model.

3) Evaluation Benchmark. We adopt DIV2K-val (Agustsson & Timofte, 2017) with 100 pairs of synthetic data, and RealLQ (Ai et al., 2024) with 250 unpaired in-the-wild LR as the evaluation data. We use CFG-Scale=6 by default, and also provide a curve illustration with diverse CFG-scale in Fig. 6. We adopt 8 evaluation metrics in total, consisting of 2 full-reference (FR) metrics including PSNR (Wang et al., 2004) and LPIPS (Zhang et al., 2018), with 6 no-reference (NR) metrics including LIQE (Zhang et al., 2023), BRISQUE (Mittal et al., 2012), Q-Align (Wu et al., 2023), WADIQAM-NR (Bosse et al., 2018), CLIPIQA, and CLIPIQA+ (Wang et al., 2023a) to extensively evaluate both fidelity and perceptuality.

4.2 Qualitative Comparison

Fig. 5 presents a qualitative comparison of our method and other diffusion-based RealSR methods. LucidNFT (Fei et al., 2026) produces over-generated textures, unrealistic artifacts and distorted details due to the lack of a stable anchor, while our method achieves notable superiority in both fidelity and realistic details. More comparisons between our method and other RL methods are shown in Fig. 7.

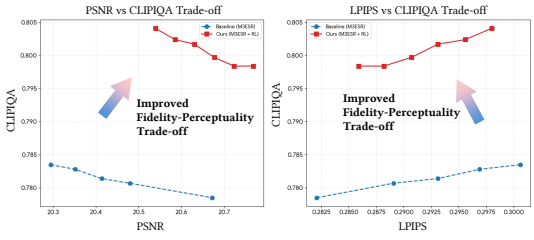


Figure 6: Our NA-DePO achieves an **improved balance between faithful restoration and realistic enhancement**.

4.3 Quantitative Comparison

We compare our method with two GAN-based methods (BSRGAN (Zhang et al., 2021), RealESRGAN (Wang et al., 2021)), six state-of-the-art diffusion-based methods (SeeSR (Wu et al., 2024), PiSA-SR (Sun et al., 2025), InvSR (Yue et al., 2025a), DiffBIR (Lin et al., 2024), DiT4SR (Duan et al., 2025), M3ESR (Luo et al., 2026)), and two state-of-the-art RL-for-SR methods (DP2O-SR (Wu et al., 2025), LucidNFT (Fei et al., 2026)). Tab. 1 shows quantitative results. **Our NA-DePO achieves the state-of-the-art performance on all perceptual metrics by a large margin.** We further present a comparison with the non-RL baseline in Fig. 6 to illustrate that our method improves the fidelity-perceptuality trade-off.

4.4 Ablation Study and Further Analysis

We conduct comprehensive ablation experiments to analyze each component of our framework.

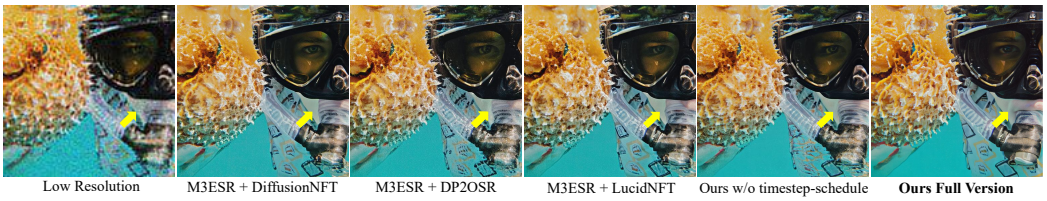


Figure 7: **Qualitative comparison of RL-for-SR variants and timestep reward scheduling.** Reward scheduling proposed in Sec. 3.6 improves semantic consistency.

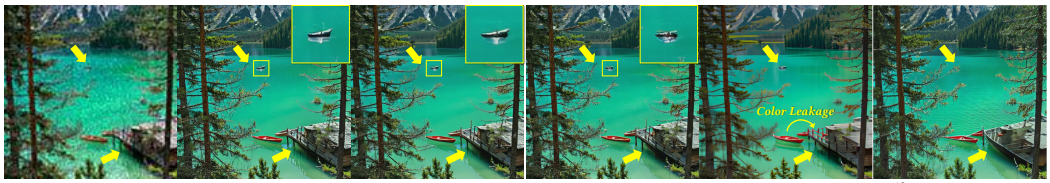


Figure 8: **Ablation of Null-space anchored RL. Compare (b&c).** We set coarser range-space scale ($\times 2$) to enable more diverse exploration in null-space. **Compare (b&d).** Simple SFT on synthetic paired data induces unrealistic artifacts, while our RL enables realistic detail generation. **Compare (b&e).** When we apply over-weighted constraints on range-space, it hampers effective RL exploration and causes over-smooth and color leakage. **Compare (b&f).** Abolishing range-space constraints leads to inconsistent structures and semantics, as pointed by the yellow arrows.

1) Effect of Null-Space Exploration and Range-Space Constraint. To verify the effectiveness of the proposed null-space anchored RL framework, we design four ablation experiments: (1) Replace the whole RL system with SFT (Fig. 8(d)), (2) Scale the downsample ratio in operator $\boldsymbol{A}$ (Fig. 8(c)), (3) Over-weight the range-space constraints (Fig. 8(e)), and (4) Remove the range-space constraints (Fig. 8(f)). Quantitative results are shown in Tab. 2.

Table 2: **Quantitative ablation experiment on DIV2K.**

Method	LIQE↑	Q-Align↑	CLIPIQA↑	PSNR↑	LPIPS↓
M3ESR w/o RL	4.8172	4.3594	0.7807	20.4136	0.3006
M3ESR w/ DiffusionNFT	4.7605	4.2177	0.7825	20.0892	0.3320
w/o Range Constraint	4.8152	4.4023	0.7889	20.0265	0.3118
Overweighted Range	4.8251	4.1768	0.7687	**21.3281**	0.3339
Null-Space Ratio $\times 2$	4.7489	4.2930	0.7735	20.5853	0.3037
w/o sigmoid reweight	4.8203	4.4042	0.7891	20.0962	0.3028
w/o reward scheduling	4.7717	4.4023	0.7984	20.4573	0.3053
Ours	**4.8405**	**4.4648**	**0.8041**	20.5393	**0.2980**

2) Effect of Reward Sigmoid-Reweighting. We find that without reward sigmoid-reweighting proposed in Eq. 16, both FR and NR metrics degrade in Tab. 2, verifying its effectiveness to bound gradient variance and stabilize training.

3) Effect of Reward Timestep Scheduling. We remove timestep scheduling and apply the same weight on all timesteps, observing degraded detail restoration and over-generated artifacts (Fig. 7).

4) Improved Fidelity-Perceptuality Trade-Off. As shown in Fig. 6, under different settings of CFG-scale, our NA-DePO consistently yields higher CLIPIQA scores at the same PSNR or LPIPS levels, demonstrating a notably improved trade-off between fidelity and perceptual quality.

We present **more ablations and analysis in the appendix**, including the effect of reward timestep scheduling, training stability, reduced artifacts or over-generated details, generalization on hard degradation, limitations and future works.

5 Conclusion

In this paper, we propose NA-DePO, a null-space anchored RL framework with dense reward for RealSR that improves the fidelity–perceptuality trade-off. We decouple the optimization process by anchoring the RL exploration within the null-space, which improves signal constancy while unleashing generative capabilities in the unconstrained sub-space. Furthermore, we propose QwenVL-Local, which formulates a dense reward mechanism. We theoretically analyze the gradient variance and propose the sigmoid reward mapping. We design reward scheduling for more informative supervision. We conduct extensive experiments to demonstrate that NA-DePO achieves consistent improvements across multiple perceptual metrics and FR-NR balance, and thoroughly verify the effectiveness of each component.

AI USE STATEMENT

In this work, we used generative AI tools only to polish the English writing and to assist in writing simple code for data preprocessing and visualization. These tools were not used to formulate research ideas, design the method or experiments, generate experimental data or results, or draw scientific conclusions. All AI-assisted text was manually reviewed and revised by the authors. All AI-assisted code was inspected, executed, and verified for correctness, and the resulting visualizations and preprocessed data were checked against the source data and expected outputs. The authors take full responsibility for the final content of this work, including all text, code, results, and claims.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 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 *Adv. Neural Inform. Process. Syst.*, 2024.

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.*, 27(1):206–219, 2018.

Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. In *Int. Conf. Learn. Represent.*, 2016.

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, Xiangyu Chen, Wenlong Zhang, Bin Fu, Yu Qiao, and Yihao Liu. UniPercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. *arXiv:2512.21675*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, et al. SAM 3: Segment anything with concepts. *arXiv: 2511.16719*, 2025.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Adv. Neural Inform. Process. Syst.*, 2017.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, 2016a.

Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Eur. Conf. Comput. Vis.*, 2016b.

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 *IEEE Int. Conf. Comput. Vis.*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Int. Conf. Mach. Learn.*, 2024.

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

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. MambaIRv2: Attentive state space restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Haodong He, Yancheng Bai, Rui Lan, Xu Duan, Lei Sun, Xiangxiang Chu, and Gui-Song Xia. RAGSR: Regional attention guided diffusion for image super-resolution. *arXiv Preprint arXiv:2508.16158*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Adv. Neural Inform. Process. Syst.*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *Int. Conf. Learn. Represent.*, 2022.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv:1805.00909*, 2018.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. MixGRPO: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv:2507.21802*, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv: 2601.04720*, 2026.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *IEEE Int. Conf. Comput. Vis. Worksh.*, 2021.

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 *Eur. Conf. Comput. Vis.*, 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 Trans. Graph.*, 44(6), 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Int. Conf. Mach. Learn.*, 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. *arXiv:2505.05470*, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv:2501.13918*, 2025b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Int. Conf. Mach. Learn.*, 2023.

Jinyi Luo, Minghao Liu, Yifan Li, Zejia Fan, and Jiaying Liu. Adaptive context matters: Towards provable multi-modality guidance for super-resolution. *arXiv: 2605.10470*, 2026.

Kangfu Mei, Hossein Talebi, Mojtaba Ardakani, Vishal M Patel, Peyman Milanfar, and Mauricio Delbracio. The power of context: How multimodality improves image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *Int. Conf. Learn. Represent.*, 2022.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence*, 2023a.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *Int. J. Comput. Vis.*, 132(12):5929–5949, 2024.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *IEEE Int. Conf. Comput. Vis. Worksh.*, 2021.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *Int. Conf. Learn. Represent.*, 2023b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

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:2312.17090*, 2023.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Shihao Wang, Tianhe Wu, Qiaosi Yi, Shuai Li, and Lei Zhang. DP²O-SR: Direct perceptual preference optimization for real-world image super-resolution. In *Adv. Neural Inform. Process. Syst.*, 2025.

Yuhui Wu, Chenxi Xie, Ruibin Li, Liyi Chen, Qiaosi Yi, and Lei Zhang. CoCoEdit: Content-consistent image editing via region regularized reinforcement learning. *arXiv:2602.14068*, 2026.

Jiahua Xiao, Jiawei Zhang, Dongqing Zou, Xiaodan Zhang, Jimmy Ren, and Xing Wei. Semantic segmentation prior for diffusion-based real-world super-resolution. *arXiv Preprint arXiv:2412.02960*, 2024.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. DanceGRPO: Unleashing grpo on visual generation. 2025.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *Eur. Conf. Comput. Vis.*, 2024.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Yuhui Wu, and Lei Zhang. Fine-structure preserved real-world image super-resolution via transfer VAE training. In *IEEE Int. Conf. Comput. Vis.*, 2025.

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: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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025a.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(1):116–130, 2025b.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE Int. Conf. Comput. Vis.*, 2021.

Leheng Zhang, Weiyi You, Kexuan Shi, and Shuhang Gu. Uncertainty-guided perturbation for image super-resolution diffusion model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

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 *Int. Conf. Learn. Represent.*, 2026.

A Additional Qualitative Results

In this section, we provide additional qualitative comparisons and analyses to further evaluate the perceptual quality, semantic consistency, robustness, and training stability of NA-DePO.

A.1 Visual Comparison with Ground Truth



Figure 9: **Which is Which? Comparison between our results and ground-truth images.** We hide the labels here. Just guess which are generated by our model? The answer is *First group: Down, down, down; Second group: Up, down, up*.

Fig. 9 compares our restored results with the corresponding ground-truth images. Despite the inherent ambiguity of super-resolution, NA-DePO produces visually plausible structures and high-frequency details that closely resemble real captured images in these examples.

A.2 Semantic and Structural Consistency

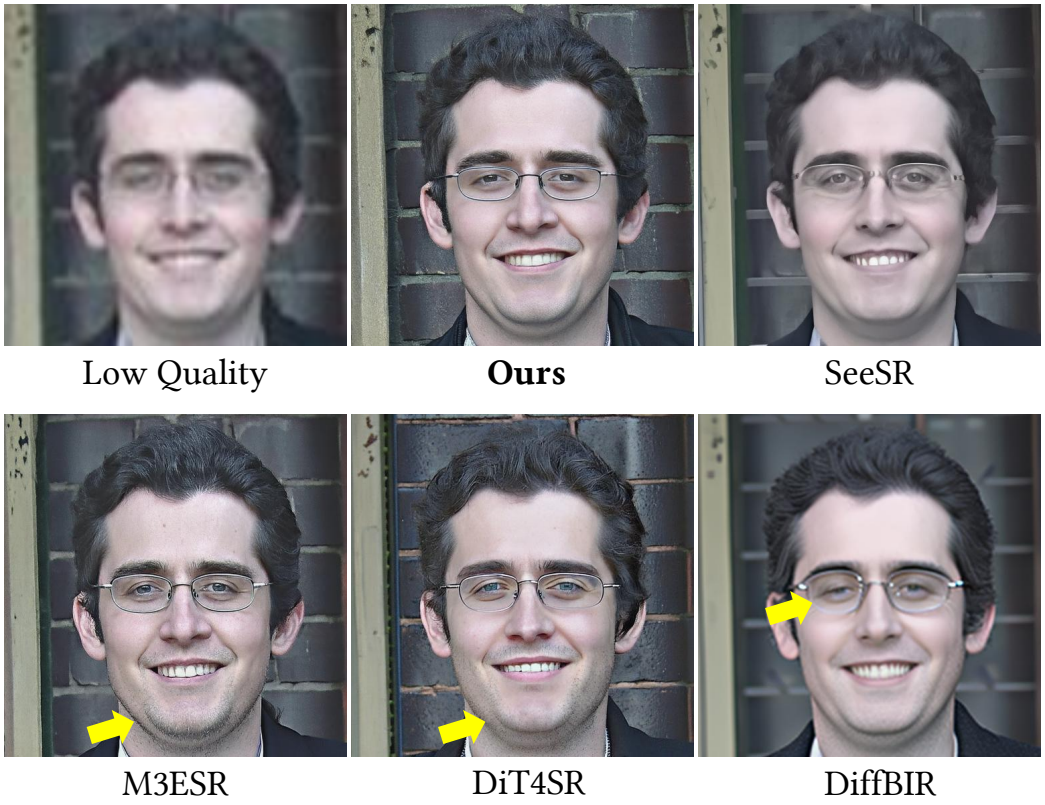


Figure 10: **Additional comparison on portrait restoration.** NA-DePO better preserves facial structures while generating realistic local details.

As shown in Figs. 10, competing methods may introduce distorted facial structures or inconsistent object details. In contrast, NA-DePO achieves more faithful semantic restoration while retaining perceptually realistic textures.

A.3 Reward Ablations

The ablations in Figs. 12 and 11 further illustrate the complementary roles of the reward components. QwenVL-Local mainly improves local semantic consistency, whereas UniPercept provides texture-oriented perceptual guidance. Together with SAM and LPIPS, these rewards improve multiple quality dimensions rather than optimizing a single perceptual criterion.

A.4 Artifact Suppression and Diverse Detail Restoration

As shown in Fig. 13, RL post-training with the proposed anchoring mechanism also reduces unrealistic artifacts and over-generated structures observed in the pretrained non-RL baseline. This leads to a more balanced restoration between perceptual enhancement and structural faithfulness.

Fig. 14 shows that NA-DePO can generate diverse yet plausible high-frequency details from the same low-resolution observation. The restored variations remain semantically consistent with the input while reflecting the one-to-many nature of the super-resolution problem.

A.5 Generalization to Severe Degradations

Figs. 17 and 22 present additional results under severe and irregular degradations. Even when substantial information is lost in the low-resolution input, NA-DePO can recover coherent structures and

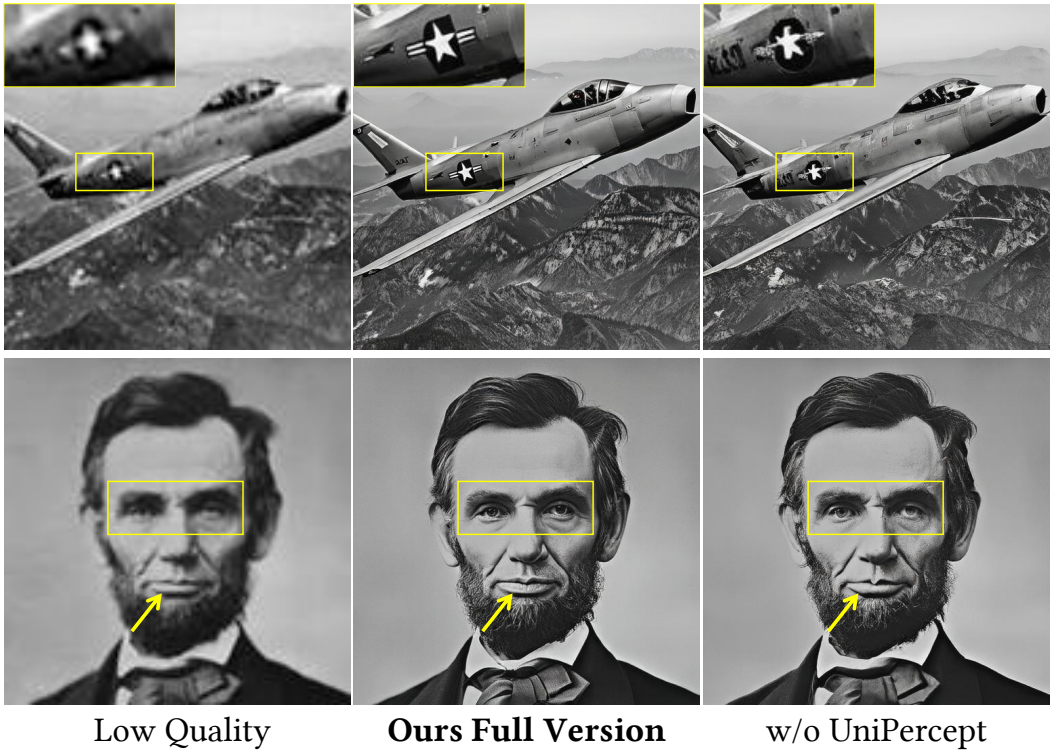


Figure 11: **Ablation of the UniPercept reward.** UniPercept provides complementary texture-level guidance and improves the realism of fine details.

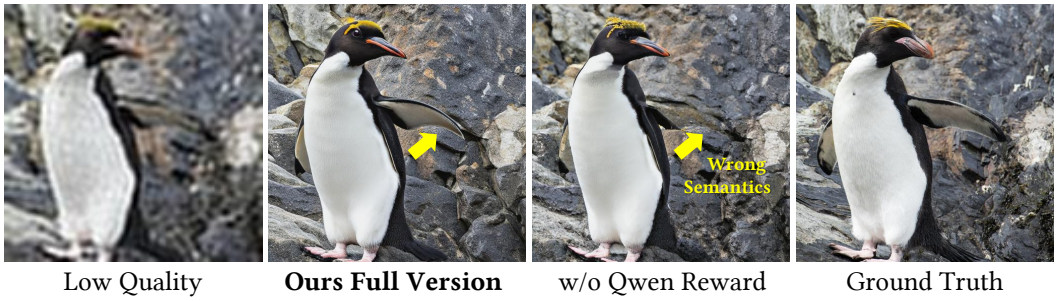


Figure 12: **Ablation of the QwenVL-Local reward.** Removing QwenVL-Local leads to less consistent local semantics, demonstrating its contribution to structure-aware restoration.

generate plausible local textures. Nevertheless, the exact ground-truth details remain fundamentally ambiguous in such highly ill-posed cases.

A.6 Training Efficiency and Stability

Our RL training is lightweight, which requires only 10 hours on one H20 GPU.

The RL post-training pipeline contains several stochastic factors, including policy rollout, flow-trajectory matching, synthetic degradation, and noisy reward estimation. Fig. 21 shows the evolution of CLIPIQA during training. Although short-term fluctuations are observed, the overall trend remains stable and improves progressively, indicating that the proposed training strategy does not compromise convergence.

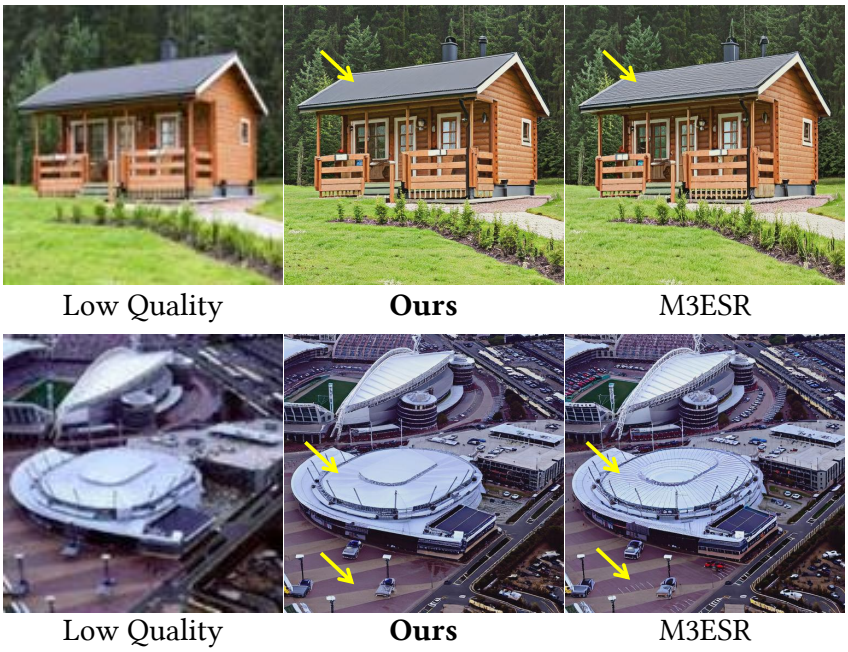


Figure 13: **Our RL post-training reduces over-generated details, achieving more faithful restoration.**



(a) Strong Detail Generation (b) Diverse Detail Generation

Figure 14: **Our method achieves diverse and powerful detail generation.**

We also present the reward curve during training in Fig. 19 and Fig. 20, demonstrating the proposed timestep reward scheduling and sigmoid reweighting scheme effectively stabilize the RL post-training.

B Further Interpretation of Null-Space Decoupling

In this section, we provide further analysis to verify that applying null-space decoupling in VAE latent space is reasonable and matches our claim.

We propose null-space anchored RL exploration in Sec. 3.4. In our implementation, we use `torch.nn.functional.adaptive_avg_pool2d` as the degradation operator $\boldsymbol{A}$ and `torch.nn.functional.interpolate` as the corresponding upsampling operator $\boldsymbol{A}^\dagger$. For simplicity, we use the notation $\boldsymbol{A}^\dagger$, although the interpolation operator is an approximate inverse rather than the exact Moore-Penrose pseudo-inverse.

In practice, we decompose the diffusion-VAE latent representation into

$$\boldsymbol{z}_{\mathcal{R}} = \boldsymbol{A}^\dagger \boldsymbol{A}\boldsymbol{z}, \qquad \boldsymbol{z}_{\mathcal{N}} = (\boldsymbol{I} - \boldsymbol{A}^\dagger \boldsymbol{A})\boldsymbol{z}, \tag{21}$$

where $\boldsymbol{z}_{\mathcal{R}}$ mainly retains LR-observation-constrained low-frequency information, while $\boldsymbol{z}_{\mathcal{N}}$ captures complementary high-frequency variations that are less constrained by the LR observation.

Strictly speaking, such an interpretation is exact only for a linear observation model in the image space. Since the VAE decoder is non-linear, we do not assume that the latent decomposition constitutes an exact mathematical null-space decomposition after decoding. **Instead**, we empirically

examine whether it approximately preserves the desired observation/detail decoupling, and whether its behavior is consistent with the corresponding decomposition performed in the image space.

Verification Experiment. Our goal here is not to assume an exact linear null space under the non-linear VAE decoder, but to examine whether applying the same spatial projection in the VAE latent produces a decomposition behavior similar to that in the image space.

Since VAE decoder is non-linear, rather than decoding the isolated component, we directly examine **whether applying the same spatial decomposition to the latent representation exhibits properties similar to those observed in the image space**.

Let $\boldsymbol{P} = \boldsymbol{A}^\dagger \boldsymbol{A}$. For each image $\boldsymbol{x}$, we construct

$$\boldsymbol{x}_{\mathcal{R}} = \boldsymbol{P}\boldsymbol{x}, \qquad \boldsymbol{x}_{\mathcal{N}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}, \tag{22}$$

$$\boldsymbol{z}_{\mathcal{R}} = \boldsymbol{P}\boldsymbol{z}, \qquad \boldsymbol{z}_{\mathcal{N}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{z}, \tag{23}$$

where $\boldsymbol{z} = E(\boldsymbol{x})$ is the VAE latent. We then compare the decomposition behavior in their respective native spaces. Specifically, we first measure the residual projection of the null component as

$$\eta_{\mathcal{N}}(\boldsymbol{u}) = \frac{\|\boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{u}\|_2}{\|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{u}\|_2 + \epsilon}, \tag{24}$$

and analogously measure the complementary leakage of the range component,

$$\eta_{\mathcal{R}}(\boldsymbol{u}) = \frac{\|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P}\boldsymbol{u}\|_2}{\|\boldsymbol{P}\boldsymbol{u}\|_2 + \epsilon}. \tag{25}$$

Because average pooling followed by interpolation is only an approximate projection, these quantities are not exactly zero. Importantly, we find that **their magnitudes in the VAE latent are comparable to those measured directly in image space** as shown in Fig. 15. We further analyze the *normalized radial power spectra* of the decomposed components in Fig. 16. In both image and latent spaces, the range component is dominated by low-frequency energy, whereas the null component contains a substantially larger proportion of high-frequency energy. This consistent behavior suggests that the operator $(\boldsymbol{I} - \boldsymbol{A}^\dagger \boldsymbol{A})$ retains a similar structural interpretation when directly applied to the diffusion-VAE latent, supporting its use as the null-space descriptor in our RL exploration.



Figure 15: **Verification of latent null-space decoupling.** The behavior between image-space and latent-space remains similar.

In conclusion, despite operating in different representation space, applying the same spatial projection to the VAE latent preserves the characteristic range–null frequency separation observed in image space, with the latent null component exhibiting an **even stronger high-frequency bias**.

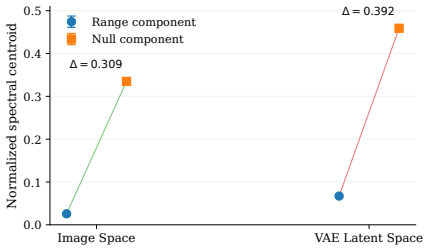


Figure 16: **Spectral centroid analysis between latent/image space range-null decoupling.** The same low-/high-frequency separation behavior is preserved.



Figure 17: **Generalization on Hard Degradations.** Our method achieves high-quality restoration and plausible detail generation even under irregular degradation, while PiSA-SR Sun et al. (2025) fail to remove the artifacts, and others methods struggle with over-generated textures.

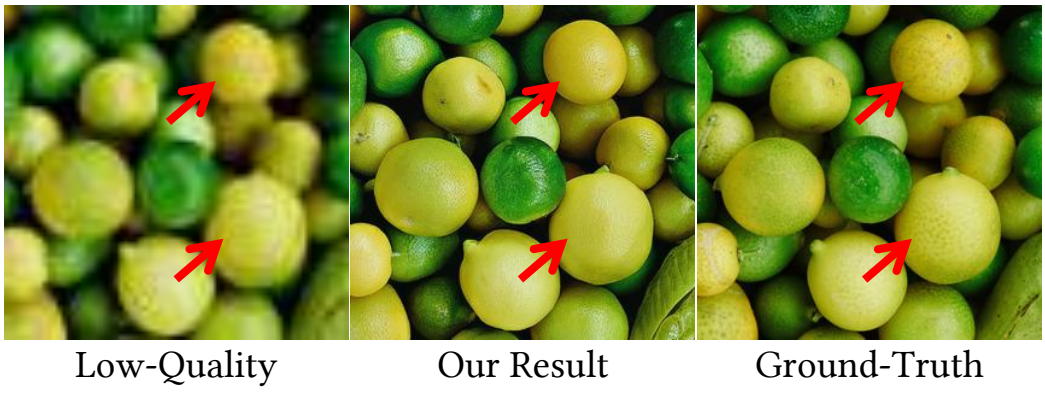


Figure 18: **Failure Case.** Although realistic restoration, our method struggles to generate textures that exactly aligned with the ground-truth if the essential information has been lost under the severe degradation.

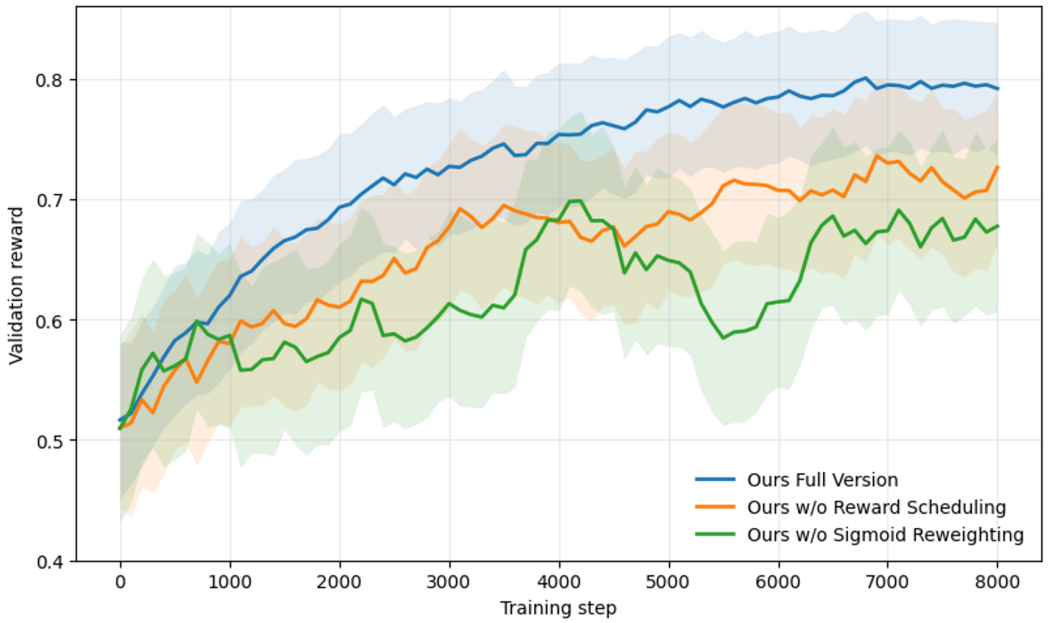


Figure 19: **Ablative Reward Curve.** We report the mean validation reward over 250 validation images with shaded regions indicating the standard deviation. Removing timestep-aware reward scheduling or sigmoid reweighting results in less stable optimization and lower final rewards.

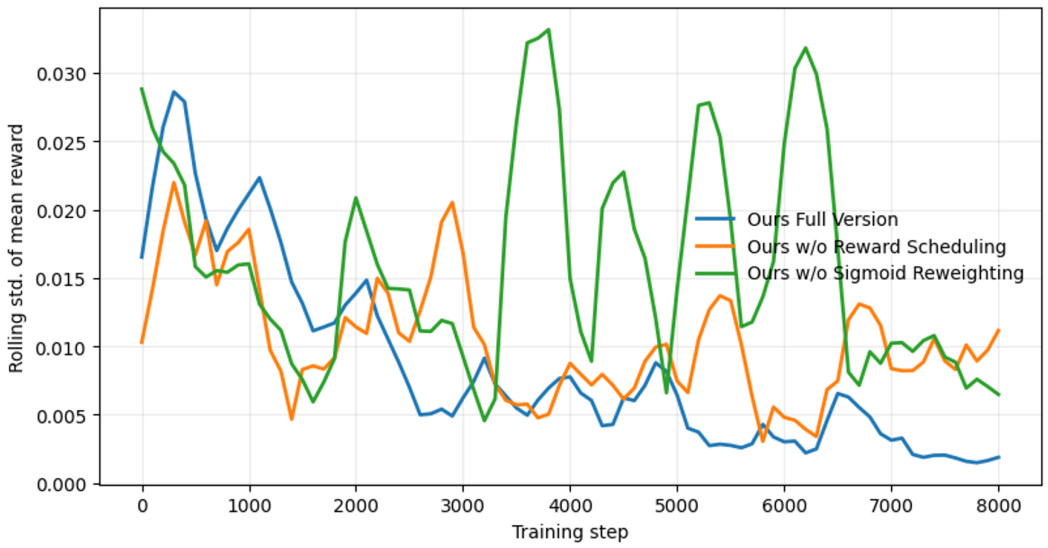


Figure 20: We report the mean validation reward standard deviation over 250 validation images. Removing timestep-aware reward scheduling or sigmoid reweighting results in less stable optimization.

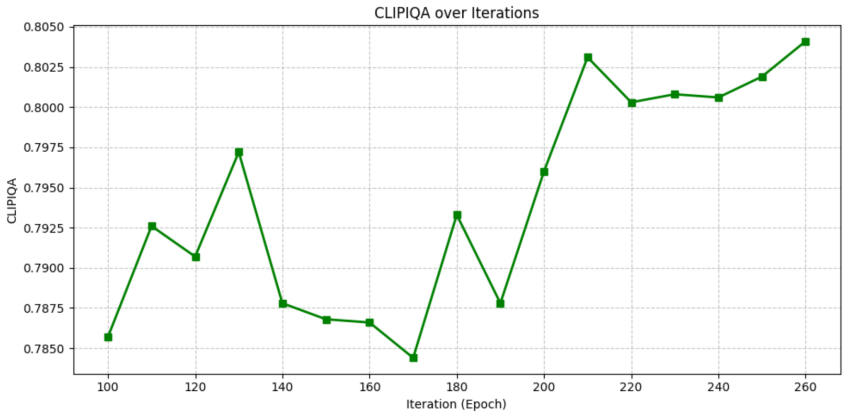


Figure 21: **Illustration of CLIPIQA improvement during training iterations.** The improvement is stable, verifying that our framework is robust to the stochastic nature of RL optimization.

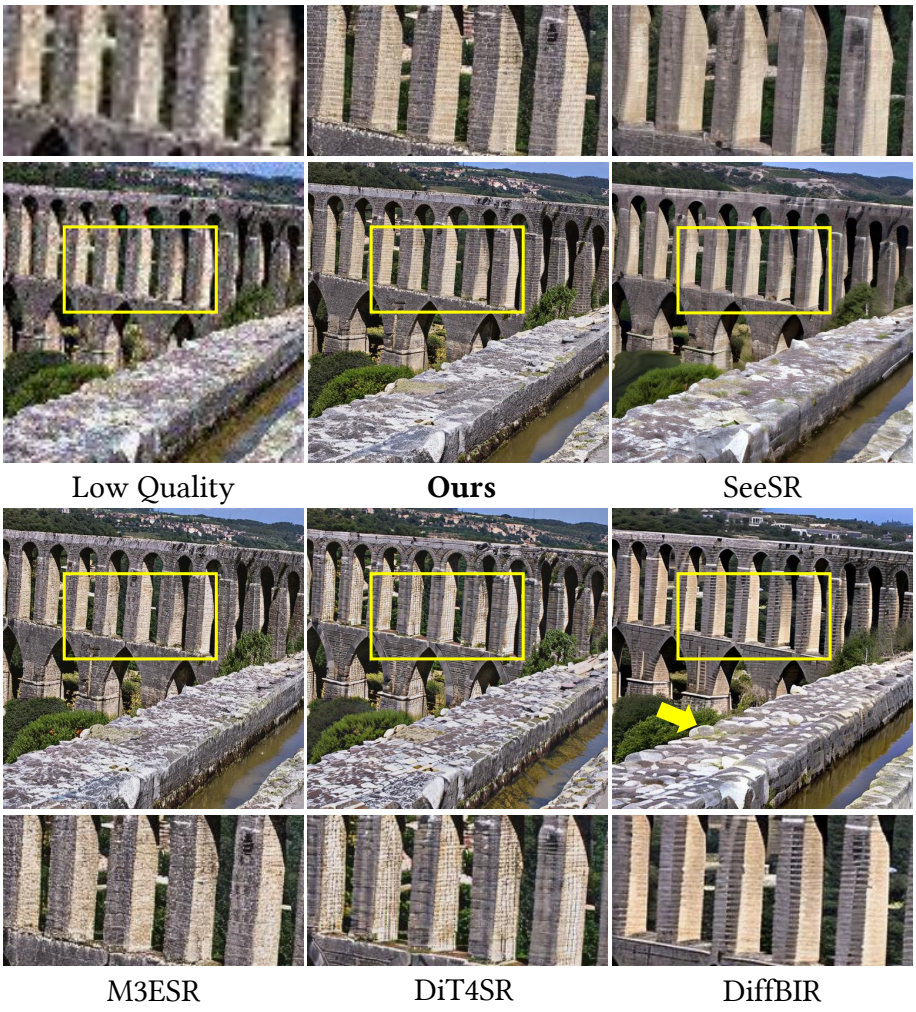


Figure 22: **Our method can restore and generate realistic and rich textures even under heavy distortion.**