OARS: Process-Aware Online Alignment for Generative Real-World Image Super-Resolution

Shijie Zhao[1*†], Xuanyu Zhang[1,2*], Bin Chen[1,2], Weiqi Li[1,2], Qunliang Xing[1], Kexin Zhang[1], Yan Wang[1], Junlin Li[1], Li Zhang[1], Jian Zhang[2], and Tianfan Xue[3]

[1] ByteDance Inc.
zhaoshijie.0526@bytedance.com
[2] Peking University
[3] The Chinese University of Hong Kong

Abstract. Aligning generative real-world image super-resolution models with human visual preference is challenging due to the perception–fidelity trade-off and diverse, unknown degradations. Prior approaches rely on offline preference optimization and static metric aggregation, which are often non-interpretable and prone to pseudo-diversity under strong conditioning. We propose **OARS**, a process-aware online alignment framework built on **COMPASS**, a MLLM-based reward that evaluates the $LR \rightarrow SR$ transition by jointly modeling fidelity preservation and perceptual gain with an input-quality-adaptive trade-off. To train COMPASS, we curate COMPASS-20K spanning synthetic and real degradations, and introduce a three-stage perceptual annotation pipeline that yields calibrated, fine-grained training labels. Guided by COMPASS, OARS performs progressive online alignment from cold-start flow matching to full-reference and finally reference-free RL via shallow LoRA optimization for on-policy exploration. Extensive experiments and user studies demonstrate consistent perceptual improvements while maintaining fidelity, achieving state-of-the-art performance on Real-ISR benchmarks.

Keywords: Super-Resolution · Reinforcement Learning

1 Introduction

Real-World Image Super-Resolution (Real-ISR) [4, 44] is a fundamental problem in low-level vision, aiming to restore high-fidelity and perceptually pleasing high-resolution (HR) images from low-resolution (LR) counterparts suffering from complex, unknown degradations. While early CNN-based methods [15, 28] focused on optimizing pixel-wise reconstruction errors, they often produce overly smooth textures that correlate poorly with human visual perception. The advent of Diffusion Models [43, 63] has shifted the paradigm from regression to generative distribution matching, significantly improving perceptual quality. However,

* Equal contribution.
† Corresponding author.

standard Supervised Fine-Tuning (SFT) on fixed LR-HR pairs faces inherent limitations: it struggles to generalize to unseen real-world degradations and lacks a direct optimization mechanism to align generated content with human aesthetic preferences, often leading to either hallucinations or over-smoothing artifacts.

Recently, Reinforcement Learning from Human Feedback (RLHF) [32, 52, 58, 72] has emerged as a powerful technique to align generative models in text and image tasks [17, 24, 34]. Despite its potential, applying RL to Real-ISR presents unique challenges, bottlenecked by the reward designs and the optimization mechanisms.

First, while leveraging Image Quality Assessment (IQA) models as RL reward signals is intuitive, directly applying existing metrics presents some hurdles. Full-Reference (FR) [14, 48] metrics are restricted by the absence of ground-truth (GT) in real-world scenarios. Meanwhile, previous No-Reference (NR) metrics [24, 42, 70] are primarily trained on natural images or generic synthetic degradations, severely lacking the fine-grained sensitivity required to distinguish subtly different generative SR outputs. Furthermore, how to effectively integrate FR and NR metrics into a unified reward remains under-explored. Simply combining these conflicting metrics via static linear summation imposes a rigid optimization objective that ignores the initial degradation severity. Without a dynamic scheme, such static aggregation can potentially lead to either insufficient enhancement or unnatural local over-sharpening.

Second, current offline RL frameworks suffer from pseudo-diversity and exploration collapse. Methods like DP^{2}O-SR [54] construct preference pairs by sampling outputs from a single SR model after SFT using varying noise seeds. However, under the rigid spatial constraints of the SR task, these noise-induced variations often degrade into mere random texture hallucination rather than genuine structural diversity. Optimizing over this narrow candidate pool forces the model to fit meaningless artifacts, drastically limiting preference alignment. To bridge these gaps, we argue that a robust post-training framework for Real-ISR requires two key innovations: *a process-aware, quality-adaptive reward model*, and *an online exploration strategy* that breaks the pseudo-diversity bottleneck.

In this paper, we propose **OARS** (**O**nline **A**lignment for **R**eal-world IS**R**), a progressive online RL framework driven by an MLLM-based reward, **COMPASS** (**COM**posite **P**rocess-**A**ware **S**R **S**core).

COMPASS adopts a process-oriented view of enhancement: instead of scoring outputs in isolation, it evaluates the $LR \rightarrow SR$ transition along two complementary factors, Fidelity for content and structure preservation, and Perceptual Gain for perceptual improvement relative to the input. To provide reliable supervision across diverse degradations, we curate **COMPASS-20K** with synthetic LR inputs paired with ground truth and real-world low-quality (LQ) inputs without references, where each input is associated with multiple SR outputs from diverse enhancement models. To ensure the perceptual scores are both globally comparable across diverse images and sensitive to subtle differences among SR outputs from the same LR, we further introduce a three-stage perceptual annotation pipeline that combines global anchor scoring, fine-grained intra-group

ranking, and rank-guided score calibration. Based on COMPASS-20K, we train a unified reward model. We then design an input-quality-adaptive scheme that gates the perceptual gain term by fidelity and adjusts its strictness according to the LR quality score, discouraging hallucinations on high-quality inputs while allowing larger improvements for severely degraded ones. Moreover, COMPASS achieves state-of-the-art pairwise preference accuracy on the SR quality assessment benchmark SRIQA-Bench [9].

On top of this reward, **OARS** performs progressive online alignment in three stages. It starts with a cold-start stage that learns basic SR capability via flow-matching SFT on paired LR–HR data. It then introduces a full-reference RL stage to stabilize early optimization with GT-based consistency rewards computed directly by distance metrics. Finally, it scales to unpaired real-world LQ in a non-reference RL stage guided solely by COMPASS. To avoid the exploration collapse of offline sampling from a rigid SR policy, OARS performs shallow policy optimization via LoRA on the base generative model, leveraging its higher stochasticity for on-policy exploration. Training is further stabilized by a negative-aware objective with group-wise reward normalization and variance-based filtering, which helps suppress reward hacking while maintaining fidelity. In inference, the learned LoRA is merged into the cold-start SR model to obtain the final aligned policy.

Our main contributions are summarized as follows:

- We curate **COMPASS-20K**, a comprehensive enhancement dataset covering diverse degradations, featuring a novel three-stage annotation pipeline to capture fine-grained intra-group variations.
- We propose **COMPASS**, a process-aware MLLM reward with an input-quality-adaptive mechanism to balance fidelity and perceptual gain, outperforming prior IQA-based rewards and achieving state-of-the-art accuracy on the SR preference benchmark
- We present **OARS**, an online RL framework that integrates Cold Start SFT, Full-Reference RL, and Non-Reference RL. By using base-model stochasticity to overcome pseudo-diversity and utilizing a negative-aware objective via LoRA shallow fine-tuning, OARS achieves state-of-the-art performance.

2 Related Work

Generative Real-ISR. Early deep learning approaches [5, 10, 15, 26, 28, 68] primarily rely on regression-based architectures optimized with pixel-wise losses like ℓ_1, ℓ_2, or PSNR-oriented objectives. GAN-based approaches [44,64] introduce adversarial learning to enhance realism. Recently, diffusion and flow-based methods have demonstrated superior generative capacity, and have emerged as the dominant paradigm for Real-ISR. Leveraging large-scale pretrained image generation priors [3, 38], approaches including StableSR [43], DiffBIR [29], SeeSR [55], and FluxSR [22] exploit SFT and distribution matching to synthesize details. These models enjoy higher robustness to real degradations [6, 7, 30, 39, 45, 46, 53, 62].

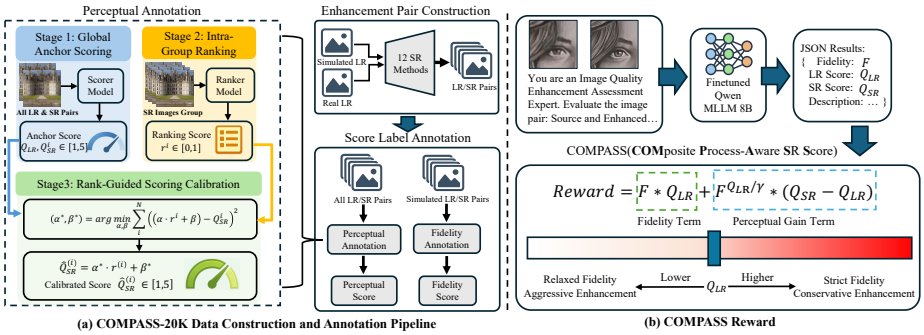

Fig. 1: Overview of the COMPASS-20K dataset construction and the COMPASS reward framework. (a) Data generation and score label annotation pipeline. (b) COMPASS uses an input-quality-adaptive mechanism to balance fidelity preservation and perceptual gain conditioned on input quality.

IQA. Traditional IQA can mainly be divided into FR methods, which evaluate degradation by comparing distorted images against pristine references [14,37,48], and NR approaches, which estimate perceptual quality blindly via learned data priors [19,20,42]. Recently, MLLMs have advanced this field by generating either numerical ratings [51,60], or interpretable textual descriptions [61,69]. However, these approaches rely heavily on extensive text annotations for SFT. To alleviate this and improve generalization, recent methods apply RL [24, 57, 67], enabling MLLMs to jointly deliver accurate quality scores and linguistic reasoning.

Preference Alignment. The alignment of generative models to human preferences has become an important research direction in MLLMs and image generation [12, 35, 50, 73]. RLHF [11] optimizes models using learned reward functions trained on human preference data. DPO [12,36] directly optimizes policies utilizing pairwise preference data under a Bradley–Terry formulation. Diff-DPO [41] and subsequent extensions [18,27,33] adapt DPO to diffusion models by defining preference losses at denoising steps. More recently, online RL [40] has been introduced to generative diffusion and flow-based models. Flow-GRPO [32] integrates policy-gradient RL into flow matching via ODE-to-SDE reformulation to enable on-policy explorations, while DiffusionNFT [72] performs negative-aware online fine-tuning directly on the forward process to improve training effectiveness and stability. These approaches demonstrate the promise of online RL for alignment.

3 COMPASS: COMposite Process-Aware SR Score

To provide a robust alignment signal for our subsequent online reinforcement learning, this section presents the proposed **COMPASS** reward framework. We detail the construction of the **COMPASS-20K** dataset and the hierarchical annotation pipeline, culminating in the final MLLM reward model.

3.1 Motivation

A practical RL reward for Real-ISR should reflect human preference: the output must look better and remain a plausible transformation of the input. However, existing IQA metrics present two fundamental limitations. First, the metrics themselves are inadequate: Full-Reference (FR) metrics require unavailable GT, while No-Reference (NR) metrics lack the fine-grained sensitivity for generative SR. Second, they are inherently *output-centric*—evaluating the restored image merely as a static outcome and entirely ignoring the starting LR condition.

To overcome these limitations, we must reconstruct the evaluation paradigm starting from the data level. A major bottleneck is the absence of comprehensive datasets featuring real-world degraded images, diverse generative SR outputs, and fine-grained preference annotations. Therefore, we first curate a dedicated enhancement dataset to fill this gap. To resolve the inherent drawbacks of existing FR and NR metrics, we design an annotation protocol that explicitly treats Real-ISR as a *process* ($LR \rightarrow SR$). Rather than scoring a static outcome, we evaluate this transition by decoupling it into **Fidelity** preservation and **Perceptual Gain**.

3.2 COMPASS-20K

To cover both synthetic and real-world degradation scenarios, we construct a diverse initial set of 2,400 images. Specifically, we apply Real-ESRGAN-style [44] random degradations to the DIV2K [1] dataset to generate 800 synthetic low-resolution (LR) images. Additionally, we collect 1,600 real-world low-quality (LQ) images from the Internet, encompassing a wide range of practical scenes. We then apply 12 mainstream image enhancement algorithms to these 2,400 inputs, resulting in 28,800 LR–SR image pairs. For notational simplicity, we denote both synthetic LR and real LQ inputs as x_{LR}.

For each $LR \rightarrow SR$ enhancement process, we annotate two dimensions—**Fidelity** and **Perceptual Gain**—and further provide textual descriptions for interpretability. Perceptual quality is annotated for all data, while fidelity is annotated only on the DIV2K subset, since real-world degraded images do not have ground-truth references.

Fidelity Annotation. Fidelity measures content consistency in the enhancement process. Following common practice, we use DISTS [13] to capture structural and textural similarity, and on the synthetic subset, we define fidelity using the SR–GT DISTS distance. For reward modeling, we normalize it to $[0, 1]$ as

$$F = 1 - \mathrm{DISTS}(x_{\mathrm{SR}}, x_{\mathrm{GT}}), \tag{1}$$

where $F = 1$ indicates the best fidelity and $F = 0$ the worst.

Perceptual Quality-Gain Annotation. A naive solution is to score LR and SR independently with an NR-IQA model and take the difference, but this is suboptimal for enhancement. Existing NR-IQA models are not optimized for modern enhancement outputs, especially generative ones. In addition, differences

among SR results from the same LR input are often subtle, and absolute cross-image scoring lacks sufficient sensitivity for such intra-group distinctions. Direct MOS labeling is also expensive and less reliable for fine-grained comparisons.

While the pioneering work DiffIQA [9] provides valuable fine-grained pairwise evaluations, it contains only partial intra-group comparisons that cannot yield a complete ranking score. To address this, we propose a **three-stage alignment annotation pipeline** that combines global comparability with comprehensive intra-group discriminability.

Stage 1: Global Anchor Scoring. We use Q-Insight [24] score model to score both LR and SR images, obtaining Q_{LR} and the anchor SR score Q_{SR}, both in $[1, 5]$. These scores provide a globally comparable quality scale.

Stage 2: Intra-Group Ranking. To capture the subtle quality differences among SR results derived from the same LR input, we establish a fine-grained intra-group ranking. We first build a dedicated pairwise comparison model upon the Q-Insight compare model and train it using the DiffIQA dataset as supervision. We then deploy this model to conduct exhaustive pairwise comparisons among all SR outputs within each group. Finally, we aggregate these pairwise outcomes into a continuous relative ranking score $r \in [0, 1]$, where 1 and 0 denote the best and worst SR results within that specific group, respectively.

Stage 3: Rank-Guided Scoring Calibration. To preserve the intra-group ranking while aligning it with the global quality scale, we perform a per-group linear calibration. For a group with N SR results, we estimate a scale α and bias β by

$$(\alpha^*, \beta^*) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \left(\alpha \cdot r^{(i)} + \beta - Q_{\mathrm{SR}}^{(i)} \right)^2 . \tag{2}$$

The calibrated SR quality score is then defined as

$$\hat{Q}_{\mathrm{SR}}^{(i)} = \alpha^* \cdot r^{(i)} + \beta^* . \tag{3}$$

This score preserves the fine-grained intra-group ranking while remaining aligned with the global scale, making it suitable for reward supervision. $\hat{Q}_{\mathrm{SR}}^{(i)}$ denotes the i-th calibrated SR quality score in the group.

Interpretability Annotation. We additionally generate concise, process-aware textual rationales for each LR–SR pair using Qwen3-VL-32B [2], describing fidelity-related changes and perceptual-gain effects. Importantly, these rationales are used for data quality control: we manually check the top/bottom 5% scored samples and discard pairs with obvious score–description conflicts, reducing corner cases and improving annotation reliability.
Detailed dataset construction procedures are provided in the **supplementary materials**.

3.3 COMPASS Reward

After constructing the COMPASS-20K dataset, we perform full-parameter SFT on QwenVL-8B [2]. For samples with fidelity annotations, the model is trained to

jointly predict fidelity and perceptual quality scores; for samples without fidelity annotations, only quality prediction is supervised. We use prompts to indicate the data type and task setting, allowing a unified model to handle different annotation configurations.

At inference time, the model generates both numerical scores and a textual description. The numerical outputs consist of the predicted fidelity score F for the enhancement process, the perceptual quality score of the input image Q_{LR}, and the perceptual quality score of the enhanced image Q_{SR}, where $F \in [0, 1]$ and $Q_{\mathrm{LR}}, Q_{\mathrm{SR}} \in [1, 5]$. To unify fidelity and perceptual quality change into a single reward signal, we propose an input-quality-adaptive mechanism that balances content preservation and quality improvement according to the input quality. Denoting the perceptual gain introduced by enhancement as $\Delta Q = Q_{\mathrm{SR}} - Q_{\mathrm{LR}}$, we define the reward as

$$R = F \cdot Q_{\mathrm{LR}} + F^{Q_{\mathrm{LR}}/\gamma} \cdot \Delta Q, \tag{4}$$

where γ is a hyper-parameter. We set $\gamma = 7$ based on ablation studies.

The first term, $F \cdot Q_{\mathrm{LR}}$, measures how much of the input image's original quality is preserved. The second term, $F^{Q_{\mathrm{LR}}/\gamma} \cdot \Delta Q$, captures the additional perceptual gain. Crucially, this gain is subjected to an input-quality-adaptive control via the exponent Q_{LR}/γ. When the input quality is high, this exponent becomes larger, making the reward highly sensitive to fidelity drops and thus encouraging conservative enhancement. Conversely, for low-quality inputs, this fidelity constraint relaxes, allowing for more aggressive perceptual improvements. Ultimately, this dynamic gating mechanism ensures that any perceptual enhancement is strictly bounded by content preservation, effectively balancing the perception-fidelity trade-off according to the initial degradation severity.

4 OARS: Online Alignment for Real-world ISR

After obtaining the COMPASS reward that reflects both fidelity and perceptual quality, we propose a progressive forward-process online RL framework. Through multi-stage continual learning, it can transform a general generative large model into an SR expert that aligns with human preferences. As shown in Fig. 2, our framework includes three stages: cold start, full-reference reinforcement learning, and non-reference reinforcement learning.

Cold Start Stage. To endow the model with initial super-resolution capability, we first warm up the model on large-scale LR–HR image pairs using a Flow Matching objective. This stage serves as a cold start that equips the model with basic perceptual reconstruction ability before reinforcement fine-tuning.

Let x_{HR}, x_{LR}, and c respectively denote high-resolution input, its corresponding low-resolution input, and the enhanced prompt. Given a Gaussian noised sample $\epsilon_{\mathrm{noise}} \sim \mathcal{N}(0, \mathbf{1})$, rectified flow defines an interpolated noisy sample at time $t \in [0, 1]$ as: $x_t = (1-t)\, x_{\mathrm{HR}} + t\, \epsilon_{\mathrm{noise}}$. Then, we train a conditional velocity field network $v_\theta(x_t, t \mid c, x_{\mathrm{LR}})$ to approximate the target velocity $v = \epsilon_{\mathrm{noise}} - x_{\mathrm{HR}}$

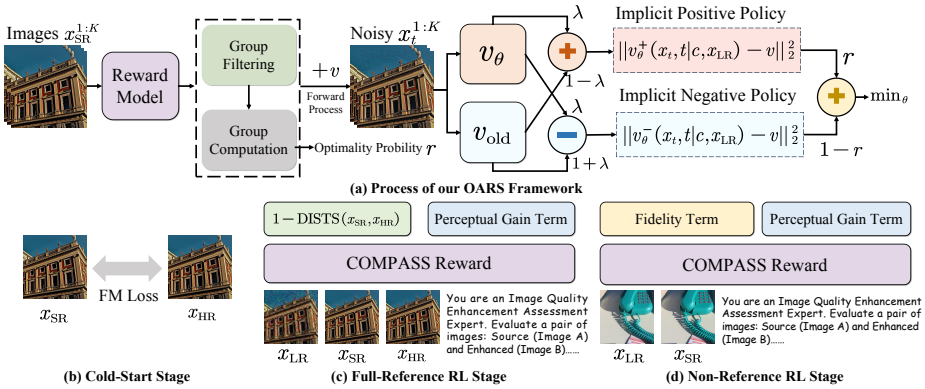

Fig. 2: Overview of the proposed OARS framework. (a) Policy optimization process of our OARS framework; (b) Cold-start stage with paired LR–HR supervision; (c) Full-reference RL stage; (d) Non-reference RL stage on unpaired data guided by our reward.

by minimizing the following loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),x_{\text{HR}},\epsilon_{\text{noise}}}\left[\|v - v_\theta(x_t, t \mid x_{\text{LR}}, c)\|_2^2\right]. \tag{5}$$

Through this cold start stage, the model learns a continuous conditional transport dynamics that maps samples from the noise distribution to the high-resolution image distribution conditioned on low-resolution inputs. It provides a strong initialization with reliable perceptual reconstruction capability.

Full-Reference RL Stage. After obtaining the SFT model $v_{\theta_{\text{SFT}}}$ from the cold start stage, a straightforward approach is to further perform RL on top of the SFT weights using additional low-resolution data. However, we observe two critical issues with this strategy: **1)** Directly optimizing both perceptual quality and structural consistency on out-of-domain data is highly challenging, leading to unstable training and poor convergence; **2)** RL initialized from $v_{\theta_{\text{SFT}}}$ tends to exploit the reward signal by over-optimizing perceptual scores while ignoring content and structural fidelity, resulting in severe reward hacking.

To address these issues, we introduce a reference-based RL stage as a buffer between SFT and reference-free optimization. In this stage, RL is still performed on LR inputs with available HR references. To reduce optimization difficulty and improve the reliability of consistency supervision, we directly compute the consistency reward using distance-based metrics (e.g., DISTS) between the generated image x_{SR} and the ground-truth image x_{HR}, instead of relying on the learned reward model to predict consistency shown in Eq. 4.

Inspired by shallow fine-tuning [16,31,56], we do not apply RL directly on the SFT weights. Instead, we perform policy optimization by applying LoRA-based updates on the base model weights $v_{\theta_{\text{base}}}$. During inference, the learned LoRA parameters are merged into the SFT model to obtain the final policy. This design is motivated by the following observations: 1) The parameter distributions of

$v_{\theta_{\text{base}}}$ and $v_{\theta_{\text{SFT}}}$ are closely aligned, enabling stable merging without introducing abnormal outputs; 2) Sampling rollouts from the base model exhibits higher stochasticity, which is beneficial for exploration during reinforcement learning; 3) Optimizing on the base model is less prone to reward hacking and better preserves fidelity and structural consistency.

Since SR can be treated as a strongly constrained generation task that does not heavily rely on step-wise trajectory optimization [32], and considering the optimization efficiency, we adopt a forward-process RL paradigm [72]. Given K samples $x_{\text{SR}}^{1:K}$ from the same LR input and enhanced prompt, we first evaluate each sample via our reward to obtain raw reward scores $s_{\text{raw}}^{1:K}$. Notably, the fidelity term in this stage is directly computed between the x_{HR} and the x_{SR}, rather than being predicted by our reward. To avoid ambiguous supervision caused by nearly identical samples, we discard groups whose raw rewards exhibit a high mean but low variance, as such samples provide limited discriminative learning signals. After filtering, we perform group computation on the remaining raw rewards, followed by clipping, to obtain the final optimal probability $r \in [0, 1]$. Formally, the optimal probability is computed as: $r = 0.5 + 0.5\ \text{clip}\big(\big(s_{\text{raw}} - \mathbb{E}_{\pi^{\text{old}}(\cdot|c)}[s_{\text{raw}}]\big)/\text{std}, -1, 1\big)$, where std > 0 is the standard deviation of raw rewards in the group and π^{old} denotes the reference policy.

Furthermore, we define a contrastive objective that steers the velocity predictor toward high-reward behaviors while suppressing low-reward ones. Formally, let v denote the target velocity field, and let v_{old} be the frozen reference policy. We define implicit positive and negative policies as linear combinations of the old policy and the current policy v_θ:

$$v_\theta^+(x_t, t \mid c, x_{\text{LR}}) = (1-\lambda)\, v_{\text{old}}(x_t, t \mid c, x_{\text{LR}}) + \lambda\, v_\theta(x_t, t \mid c, x_{\text{LR}}), \tag{6}$$

$$v_\theta^-(x_t, t \mid c, x_{\text{LR}}) = (1+\lambda)\, v_{\text{old}}(x_t, t \mid c, x_{\text{LR}}) - \lambda\, v_\theta(x_t, t \mid c, x_{\text{LR}}), \tag{7}$$

where λ=1.0. The final RL objective is defined as:

$$\mathcal{L}_{\text{RL}}(\theta) = \mathbb{E}\Big[r \left\|v_\theta^+(x_t, t \mid c, x_{\text{LR}}) - v\right\|_2^2 + (1-r) \left\|v_\theta^-(x_t, t \mid c, x_{\text{LR}}) - v\right\|_2^2\Big]. \tag{8}$$

By explicitly modeling the positive and negative policy directions, this negative-aware objective effectively balances perceptual quality enhancement and reference-based consistency preservation, getting the optimized lora weights Δ_{FR}.

Non-Reference RL Stage. In the final stage, we aim to further improve the perceptual quality while ensuring that fidelity degradation is minimized as much as possible. To this end, we perform RL on additional out-of-domain low-resolution data without ground-truth references in our COMPASS-20K, enabling the model to generalize beyond the training distribution. Building upon the learned LoRA Δ_{FR}, we continue training the model via LoRA finetuning. In the non-reference setting, no ground-truth high-resolution images are available. Therefore, the reward signal is entirely provided by the COMPASS reward model introduced in Eq. 4. Noted that, we still perform RL on the base model $v_{\theta_{\text{base}}}$. The policy optimization objective follows the same process as in the reference-based stage (Eq. 8). Thus, we can get the final lora weights Δ_{NR}. **During in-**

ference, we merge the learned LoRA parameters Δ_{NR} from the final stage into the backbone weights obtained from the cold start (SFT) stage $v_{\theta_{\mathrm{SFT}}}$.

5 Experiment

5.1 Experimental Setup

Reward Implementation and Evaluation. We train the reward on our COMPASS-20K, whose data generation and annotation procedure are described in Sec. 3, using 19,200 annotated pairs in total, with half from real-world LQ inputs and half from synthetic LR inputs. We adopt Qwen-3-VL-8B-Instruct [2] as the initialization and perform full-parameter supervised fine-tuning. Training is conducted on 8 GPUs (80 GB) with a global batch size of 64 for 8 epochs. We use AdamW for optimization, with an initial learning rate of 1e-5 and a cosine decay schedule.

To verify whether the learned reward aligns with human preference, we evaluate it on SRIQA-Bench [9], which contains 100 LR images and results produced by multiple SR methods. The benchmark provides pairwise comparisons between SR outputs for the same LR input, together with ground-truth preference annotations. Given a pair of SR candidates $(x^a_{\mathrm{SR}}, x^b_{\mathrm{SR}})$, we compute their reward scores and predict the preferred one by comparing the two scalar rewards. We then report **pairwise preference accuracy**, i.e., the percentage of comparisons where the reward-based ordering matches the annotated human preference. This evaluation directly measures the reward model's capability of ranking competing SR outputs in a fine-grained and perceptually consistent manner.

Online RL Implementation and Evaluation. We adopt Qwen-Image-Edit-2509 [50] as the base model and perform LoRA fine-tuning with rank=32 and alpha=64. The per-GPU batch size is set to 1, and the model is optimized using Adam with a learning rate of 3×10^{-4} and weight decay 1×10^{-4}. During training, the diffusion model uses 6 sampling steps for training and 40 steps for inference. For group filtering, the thresholds on the reward mean and variance are set to 0.9 and 0.05, respectively. For each x_{LR}, we sample $K = 24$ candidate groups. All RL experiments are conducted on a single computer with 8 GPUs. The reward server is deployed on a computer with 8 GPUs using vLLM for inference. The cold-start and full-reference RL stages are trained on the LSDIR dataset [25], while the non-reference RL stage is trained on our constructed COMPASS-20K, using a split that is strictly separated from the reward training set and contains 800 real-world LQ inputs. We evaluate the SR performance of our model on RealSR [4], DIV2K [1], and RealSet80 [63], with 512×512 resolution, performing $4\times$ SR. FR metrics include PSNR [48], SSIM [48], LPIPS, and DISTS, while NR metrics include LIQE [66], MUSIQ [20], MANIQA [59], Q-Insight [24], and TOPIQ [8].

5.2 Reward Experimental Results

To validate the effectiveness of our reward model, we compare it against two categories of visual quality assessment methods: (1) **GT-Ref** metrics (e.g., PSNR,

Table 1: Comparison of different IQA methods on SRIQA-Bench [9]. Throughout this paper, the best, second-best, and third-best results are highlighted in **bold red**, underlined blue, and *italic green*.

Type	Method	Reg-Acc	Gen-Acc	All-Acc
GT-Ref	PSNR	80.7	41.7	34.7
	SSIM [48]	83.0	45.3	37.4
	LPIPS [65]	84.7	63.8	72.2
	DISTS [14]	86.0	63.9	71.4
	AHIQ [21]	71.0	71.5	69.6
	TOPIQ [8]	78.3	*73.0*	77.7
	A-FINE [9]	83.3	**78.9**	82.4
GT-Free	CKDN [71]	76.7	64.3	59.1
	CLIP-IQA+ [42]	80.3	72.5	*79.5*
	Q-Insight [24]	**88.0**	71.0	78.8
	COMPASS (Ours)	*85.0*	78.3	**83.1**

LPIPS, A-FINE), which rely on GT references to compute quality scores; and (2) **GT-Free** metrics, which operate without GT. Within the GT-Free category, CKDN and our method use the LR input as a conditional reference, whereas Q-Insight and CLIP-IQA+ are reference-free. As shown in Table 1, our method achieves the highest overall accuracy (**83.1%**) across all 5500 test samples, outperforming all GT-Ref and GT-Free baselines. Beyond improved alignment with human perception, our approach is GT-Free and thus intrinsically suited for Real-ISR where HR references are unavailable.

5.3 Online RL on Super-Resolution Experimental Results

Quantitative Comparison. Following prior works [23, 53], we compare our OARS against diffusion-based methods: StableSR [43], DiffBIR [29], SeeSR [55], SinSR [47], OSEDiff [53], Qwen-SFT, the autoregressive method PURE [49], and the unified method UARE [23]. Note that Qwen-SFT denotes the Qwen-Image-Edit model after the cold-start stage. As reported on Tab. 2, our OARS achieves consistently top-ranked performance on NR metrics across RealSR, DIV2K, and RealSet80. Moreover, compared to perception-oriented methods such as PURE and UARE, our approach also attains significantly better FR metrics, showing that our reward design and RL strategy effectively balance fidelity preservation and perceptual enhancement. Notably, when compared to Qwen-SFT (our initial policy model), OARS yields consistent and pronounced improvements on all NR metrics, while the FR performance does not exhibit noticeable degradation.

Table 2: Quantitative comparison on three classical SR datasets, including RealSR, DIV2K, and RealSet80. ↑ / ↓ indicates higher/lower is better.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	LIQE↑	MUSIQ↑	MANIQA↑	Q-Insight↑	TOPIQ↑
RealSR	DiffBIR [29]	23.20	0.6346	0.3350	0.2162	3.5529	65.25	0.4620	3.530	0.6033
	OSEDiff [53]	23.07	*0.6850*	*0.2941*	0.2109	4.0681	*68.95*	0.4876	**3.712**	*0.6441*
	SeeSR [55]	**24.34**	**0.7187**	**0.2754**	*0.2134*	3.3938	65.53	0.4856	3.285	0.6246
	SinSR [47]	*23.68*	0.6649	0.3490	0.2445	3.2255	61.03	0.4230	3.264	0.5383
	StableSR [43]	23.73	0.6979	0.2792	**0.2023**	3.0532	61.65	0.3826	3.480	0.5201
	PURE [49]	21.31	0.5738	0.3859	0.2468	3.7881	66.57	0.4829	3.569	0.6301
	UARE [23]	21.38	0.6464	0.3095	0.2344	*4.0658*	69.67	0.5260	*3.664*	0.6796
	Qwen-SFT [50]	22.71	0.6462	0.3100	0.2203	3.8146	68.57	*0.4897*	3.545	0.6397
	OARS (Ours)	22.36	0.6481	0.3095	0.2244	**4.3045**	**71.41**	**0.5277**	3.701	**0.6803**
DIV2K	DiffBIR [29]	*18.94*	0.4332	0.4009	0.2238	3.8573	67.20	0.4574	3.577	0.6467
	OSEDiff [53]	18.86	*0.4563*	**0.3579**	0.2209	3.8877	67.83	0.4422	3.722	0.6269
	SeeSR [55]	19.11	0.4580	0.3769	0.2339	3.7445	66.31	0.4686	3.611	0.6330
	SinSR [47]	18.58	0.4059	0.4483	0.2455	3.4629	64.12	0.4483	3.224	0.5997
	StableSR [43]	**19.85**	**0.4940**	0.4796	0.2887	1.8466	43.25	0.2181	2.066	0.3276
	PURE [49]	16.71	0.3661	0.4449	0.2293	*4.2701*	70.06	**0.5201**	3.721	0.6621
	UARE [23]	16.59	0.3857	0.4074	*0.2138*	4.2627	*70.45*	*0.5028*	3.823	0.6864
	Qwen-SFT [50]	17.31	0.3988	*0.3811*	**0.1993**	4.3404	72.35	0.4875	*3.818*	*0.6745*
	OARS (Ours)	17.32	0.4072	0.3812	0.2036	**4.6668**	**74.07**	0.5052	**3.940**	**0.6960**
RealSet80	DiffBIR [29]	N/A				4.1113	68.10	**0.5527**	3.684	*0.6736*
	OSEDiff [53]					4.2251	68.88	0.4995	3.725	0.6062
	SeeSR [55]					4.3317	69.70	0.5362	3.774	0.6887
	SinSR [47]					3.6613	62.78	0.4483	3.382	0.5854
	StableSR [43]					3.9074	67.67	0.4682	3.562	0.6440
	PURE [49]					*4.2528*	69.55	0.5215	*3.744*	0.6647
	UARE [23]					4.1804	70.05	*0.5363*	3.508	0.6446
	Qwen-SFT [50]					4.1602	*69.79*	0.5171	3.701	0.6539
	OARS (Ours)					**4.5465**	**72.96**	0.5364	**3.823**	**0.6967**

Qualitative Comparison and User Study. Fig. 4 shows that OARS produces more natural restorations under complex real degradations, recovering fine details while preserving global structure. In comparison, SeeSR, UARE, DP^{2}O-SR, and OSEDiff more frequently exhibit over-smoothing or localized hallucinations/over-sharpening. We further conduct a user study on 12 LR inputs sampled from RealSR, DRealSR, and RealSet80, where 27 expert researchers select the most preferred restored result among the compared methods. OARS receives **47.62%** of the votes, higher than the strongest baseline (DP^{2}O-SR, 27.68%).

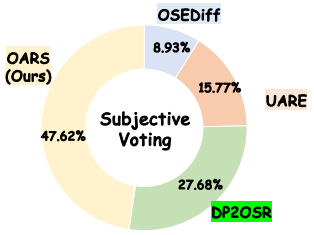

Fig. 3: Subjective user study results. OARS achieves the highest preference rate, receiving 47.62% of the total votes.

For more quantitative and qualitative comparison results, as well as the detailed user study setup, please refer to the **supplementary materials**.

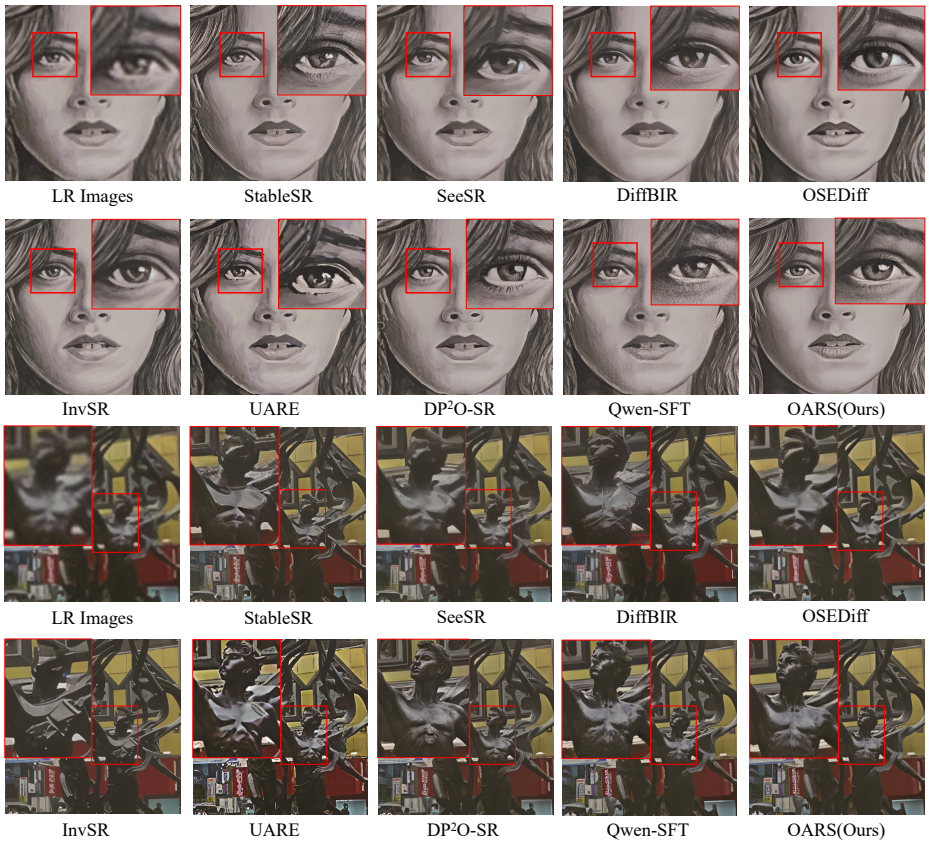

Fig. 4: Qualitative comparison of OARS against state-of-the-art Real-ISR methods under complex degradations.

5.4 Ablation Studies

Ablation on Reward Model. Table 3 evaluates key components of our reward formulation on COMPASS-20K. Introducing three-stage score calibration (Case 2) improves accuracy over naive absolute scoring (Case 1) by better aligning fine-grained intra-group rankings. Adding explicit fidelity modeling further boosts performance (Case 3), indicating that enforcing structural consistency is crucial. Finally, replacing a fixed linear combination with our Quality-Adaptive γ yields the best result, and we select $\gamma = 7$ based on the ablation (Case 4).
Ablation on Online RL. Tab. 5 analyzes the impact of RL stages and initialization. Starting from Qwen-SFT, applying our stage-1 RL on the **base** model (Case 1) yields clear perceptual gains across NR metrics while keeping FR metrics stable. Adding stage-2 reference-free RL (Case 2) further improves perceptual quality, giving the best overall NR performance. In contrast, performing RL on the **SFT** model (Cases 3–4) consistently degrades FR metrics, indicating

Table 3: Ablation study of dataset calibration, fidelity modeling, and adaptive γ.

Case	Score Calibration	Explicit Fidelity	Quality-Adaptive γ	Accuracy
1	✗	✗	✗	78.8
2	✓	✗	✗	81.5
3	✓	✓	✗	82.3
4	✓	✓	5	82.7
5	✓	✓	7	**83.1**
6	✓	✓	9	82.8

Table 4: Quantitative comparison between our OARS and DPO-based SR methods on RealSet80. Performance gains are reported in parentheses.

Method	LIQE↑	MUSIQ↑	MANIQA↑	Q-Insight↑	TOPIQ↑
Qwen-SFT	4.1602	69.7866	0.5171	3.701	0.6539
+COMPASS	4.5465 (+**0.39**)	72.9598 (+**3.17**)	0.5364 (+**0.02**)	3.823 (+**0.12**)	0.6967 (+**0.04**)
+COMPASS+TTT	4.5792 (+**0.03**)	73.2551 (+**0.30**)	0.5428 (+**0.01**)	3.841 (+**0.02**)	0.7025 (+**0.01**)
DP^{2}O-SR [54]	4.2159	67.9658	0.5703	3.865	0.6867
+COMPASS	4.5796 (+**0.36**)	72.8912 (+**4.93**)	0.6123 (+**0.04**)	3.949 (+**0.08**)	0.6982 (+**0.01**)

higher susceptibility to reward hacking and instability. These results validate our progressive RL design and the choice of shallow LoRA optimization on the base model for more effective on-policy exploration and robust alignment.

Ablation on Reward for RL. We further investigate how different reward signals influence the RL policy. Directly optimizing with naive no-reference metrics induces severe reward hacking, which clearly degrades fidelity and introduces over-generation artifacts. Additionally, dropping either the input-quality-adaptive gating or the fidelity term from our reward formulation results in a severe imbalance between objective fidelity and subjective visual quality. Conversely, employing the complete COMPASS reward achieves the optimal. For more experimental results, please refer to the **supplementary material**.

5.5 Comparison with DPO-based SR Methods

We further compare our online SR method with the offline RL approach DP^{2}O-SR [54], and also verify the generality of OARS on different backbones (e.g., Flux-ControlNet [3]). All experiments are conducted on RealSet80. Noted that DP^{2}O-SR can be viewed as most closely related to our stage-1 full-reference alignment; however, since its training data and protocol are not publicly available, we cannot reproduce an exactly matched stage-1 setting, and a fully controlled end-to-end comparison is therefore non-trivial. To ensure a consistent and well-defined training recipe, we adopt our stage-2 RL pipeline with the COMPASS reward.

As shown in Tab. 4, applying our two-stage training yields great improvements on RealSet80 across multiple NR metrics (LIQE, MUSIQ, TOPIQ), and additional test-time tuning (TTT) further enhances the results, where "TTT" denotes an extra RL stage performed on the RealSet80 test set after completing the non-reference RL stage. Moreover, when we directly apply our non-reference RL stage on top of the DP^{2}O-SR model, we observe consistent gains across all

Table 5: Ablation study on different RL stages and initialization settings. Stage-1 refers to the full-reference RL stage optimized with ground-truth references, whereas Stage-2 refers to the subsequent RL stage optimized with the COMPASS reward.

Method	RL Stage	SFT / Base	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	LIQE↑	MUSIQ↑	MANIQA↑	Q-Insight↑	TOPIQ↑
Qwen-SFT	-	-	**22.71**	*0.6462*	0.3100	**0.2203**	3.8146	68.5687	0.4897	3.545	0.6397
case (1)	stage1	base	22.52	**0.6494**	*0.3115*	0.2224	4.2345	71.0159	*0.5244*	3.685	0.6741
case (2)	stage1+stage2	base	*22.36*	0.6481	**0.3095**	*0.2244*	**4.3045**	**71.4136**	0.5277	**3.701**	**0.6803**
case (3)	stage1	sft	22.15	0.6266	0.3249	0.2272	4.0784	*70.6030*	0.5233	*3.617*	*0.6756*
case (4)	stage1+stage2	sft	21.31	0.5956	0.3326	0.2334	*4.0937*	70.5618	**0.5293**	3.667	0.6768

metrics. This suggests that our online, process-aware optimization provides an additional refinement signal beyond offline DPO-style alignment. More backbone results are provided in the **supplementary materials**.

6 Conclusion

We presented OARS, a process-aware online alignment framework for generative real-world super-resolution. We first propose COMPASS, an MLLM-based reward that scores the LR→SR transition by jointly modeling fidelity and perceptual gain with a quality-adaptive trade-off. To train it, we curate COMPASS-20K with diverse LR–SR pairs from multiple algorithms and produce fine-grained perceptual labels via a three-stage pipeline. Notably, COMPASS achieves state-of-the-art performance on SR preference benchmarks without ground-truth references. Guided by this reward, OARS performs progressive online alignment from cold-start flow matching to reference-based and finally reference-free RL, enabling on-policy exploration while reducing reward hacking and hallucinations. Experiments demonstrate that OARS provides an effective post-training paradigm for aligning generative SR models with human preferences.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 126–135 (2017)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Black Forest Labs: Flux. `https://github.com/black-forest-labs/flux` (2024), accessed: 2026-03-01
4. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3086–3095 (2019)
5. Cao, M., Mou, C., Yu, F., Wang, X., Zheng, Y., Zhang, J., Dong, C., Li, G., Shan, Y., Timofte, R., et al.: Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1731–1745 (2023)
6. Chen, B., Li, G., Wu, R., Zhang, X., Chen, J., Zhang, J., Zhang, L.: Adversarial diffusion compression for real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28208–28220 (2025)
7. Chen, B., Li, W., Zhao, S., Zhang, X., Li, J., Zhang, J., et al.: Improved adversarial diffusion compression for real-world video super-resolution. In: Proceedings of the International Conference on Learning Representations (ICLR) (2026)
8. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Topiq: A top-down approach from semantics to distortions for image quality assessment. IEEE Transactions on Image Processing (TIP) **33**, 2404–2418 (2024)
9. Chen, D., Wu, T., Ma, K., Zhang, L.: Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR). pp. 12742–12752 (2025)
10. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22367–22377 (2023)
11. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems (NeurIPS) pp. 4302–4310 (2017)
12. Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., Yang, Y.: Safe rlhf: Safe reinforcement learning from human feedback. arXiv preprint arXiv:2310.12773 (2023)
13. Ding, K., Liu, Y., Zou, X., Wang, S., Ma, K.: Locally adaptive structure and texture similarity for image quality assessment. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). pp. 2483–2491 (2021)

14. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **44**(5), 2567–2581 (2020)
15. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 184–199 (2014)
16. Fleshman, W., Van Durme, B.: Re-adapt: Reverse engineered adaptation of large language models. arXiv preprint arXiv:2405.15007 (2024)
17. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
18. Hu, Z., Zhang, F., Chen, L., Kuang, K., Li, J., Gao, K., Xiao, J., Wang, X., Zhu, W.: Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 23604–23614 (2025)
19. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1733–1740 (2014)
20. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5148–5157 (2021)
21. Lao, S., Gong, Y., Shi, S., Yang, S., Wu, T., Wang, J., Xia, W., Yang, Y.: Attentions help cnns see better: Attention-based hybrid image quality assessment network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1139–1148 (2022)
22. Li, J., Cao, J., Guo, Y., Li, W., Zhang, Y.: One diffusion step to real-world super-resolution via flow trajectory distillation. arXiv preprint arXiv:2502.01993 (2025)
23. Li, W., Zhang, X., Chen, B., Xie, J., Wang, Y., Zhang, K., Li, J., Zhang, L., Zhang, J., Zhao, S.: Uare: A unified vision-language model for image quality assessment, restoration, and enhancement. arXiv preprint arXiv:2512.06750 (2025)
24. Li, W., Zhang, X., Zhao, S., Zhang, Y., Li, J., Zhang, L., Zhang, J.: Q-insight: Understanding image quality via visual reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS) (2025)
25. Li, Y., Zhang, K., Liang, J., Cao, J., Liu, C., Gong, R., Zhang, Y., Tang, H., Liu, Y., Demandolx, D., et al.: Lsdir: A large scale dataset for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)
26. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1833–1844 (2021)
27. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Li, J., Zheng, L.: Step-aware preference optimization: Aligning preference with denoising performance at each step. arXiv preprint arXiv:2406.04314 (2024)
28. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 136–144 (2017)
29. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 430–448 (2024)

30. Lin, X., Yu, F., Hu, J., You, Z., Shi, W., Ren, J.S., Gu, J., Dong, C.: Harnessing diffusion-yielded score priors for image restoration. ACM Transactions on Graphics (TOG) **44**(6), 1–21 (2025)
31. Liu, A., Han, X., Wang, Y., Tsvetkov, Y., Choi, Y., Smith, N.A.: Tuning language models by proxy. arXiv preprint arXiv:2401.08565 (2024)
32. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
33. Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 8009–8019 (2025)
34. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025)
35. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (NeurIPS) pp. 27730–27744 (2022)
36. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems (NeurIPS) pp. 53728–53741 (2023)
37. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on Image Processing (TIP) **15**(2), 430–444 (2006)
38. Stability.ai: Stable diffusion. `https://stability.ai/stable-diffusion` (2023), accessed: 2026-03-01
39. Sun, L., Wu, R., Zhang, Z., Yong, H., Zhang, L.: Improving the stability of diffusion models for content consistent super-resolution. arXiv preprint arXiv:2401.00877 (2024)
40. Sutton, R.S., Barto, A.G., et al.: Reinforcement learning: An introduction, vol. 1. MIT press Cambridge (1998)
41. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8228–8238 (2024)
42. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 37, pp. 2555–2563 (2023)
43. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision (IJCV) **132**(12), 5929–5949 (2024)
44. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 1905–1914 (2021)
45. Wang, Y., Zhao, S., Li, J., Zhang, L.: Eliminating vae for fast and high-resolution generative detail restoration. In: Proceedings of the International Conference on Learning Representations (ICLR) (2026)
46. Wang, Y., Zhao, S., Zhang, K., Li, J., Zhang, L.: Gendr: Lightning generative detail restorator. In: Proceedings of the International Conference on Learning Representations (ICLR) (2026)

47. Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: Sinsr: diffusion-based image super-resolution in a single step. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25796–25805 (2024)
48. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (TIP) **13**(4), 600–612 (2004)
49. Wei, H., Liu, S., Yuan, C., Zhang, L.: Perceive, understand and restore: Real-world image super-resolution with autoregressive multimodal generative models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
50. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
51. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In: Proceedings of the International Conference on Machine Learning (ICML) (2024)
52. Wu, K., Jiang, S., Ku, M., Nie, P., Liu, M., Chen, W.: Editreward: A human-aligned reward model for instruction-guided image editing. arXiv preprint arXiv:2509.26346 (2025)
53. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 92529–92553 (2024)
54. Wu, R., Sun, L., Zhang, Z., Wang, S., Wu, T., Yi, Q., Li, S., Zhang, L.: Dp^{2}o-sr: Direct perceptual preference optimization for real-world image super-resolution. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
55. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25456–25467 (2024)
56. Wu, T., Yang, R., Li, J., Hu, P., Wu, Y.C., Wong, N., Yang, Y.: Shadow-ft: Tuning instruct model via training on paired base model. arXiv preprint arXiv:2505.12716 (2025)
57. Wu, T., Zou, J., Liang, J., Zhang, L., Ma, K.: Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
58. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: learning and evaluating human preferences for text-to-image generation. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 15903–15935 (2023)
59. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW). pp. 1191–1200 (2022)
60. You, Z., Cai, X., Gu, J., Xue, T., Dong, C.: Teaching large language models to regress accurate image quality scores using score distribution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14483–14494 (2025)
61. You, Z., Li, Z., Gu, J., Yin, Z., Xue, T., Dong, C.: Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In: Pro-

ceedings of the European Conference on Computer Vision (ECCV). pp. 259–276 (2024)
62. Yue, Z., Liao, K., Loy, C.C.: Arbitrary-steps image super-resolution via diffusion inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23153–23163 (2025)
63. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. Advances in Neural Information Processing Systems (NeurIPS) **36**, 13294–13307 (2023)
64. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4791–4800 (2021)
65. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)
66. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14071–14081 (2023)
67. Zhang, X., Li, W., Zhao, S., Li, J., Zhang, L., Zhang, J.: Vq-insight: Teaching vlms for ai-generated video quality understanding via progressive visual reinforcement learning. arXiv preprint arXiv:2506.18564 (2025)
68. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2018)
69. Zhang, Z., Kou, T., Wang, S., Li, C., Sun, W., Wang, W., Li, X., Wang, Z., Cao, X., Min, X., et al.: Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. arXiv preprint arXiv:2503.02357 (2025)
70. Zhao, S., Zhang, X., Li, W., Li, J., Zhang, L., Xue, T., Zhang, J.: Reasoning as representation: Rethinking visual reinforcement learning in image quality assessment. In: Proceedings of the International Conference on Learning Representations (ICLR) (2026)
71. Zheng, H., Yang, H., Fu, J., Zha, Z.J., Luo, J.: Learning conditional knowledge distillation for degraded-reference image quality assessment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10242–10251 (2021)
72. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. arXiv preprint arXiv:2509.16117 (2025)
73. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)