RefReward-SR: LR-Conditioned Reward Modeling for Preference-Aligned Super-Resolution

Yushuai Song[1,2], Weize Quan[1,2⋆], Weining Wang[1,2], Jiahui Sun[1,2], Jing Liu[1,2], Pengbin Yu[3], Zhentao Chen[3], Meng Li[3], Wei Shen[3], Lunxi Yuan[3], and Dong-Ming Yan[1,2]

[1] Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] OPPO AI Center, OPPO Inc.

Abstract. Recent advances in generative super-resolution (SR) have greatly improved visual realism, yet existing evaluation and optimization frameworks remain misaligned with human perception. Full-Reference and No-Reference metrics often fail to reflect perceptual preference, either penalizing semantically plausible details due to pixel misalignment or favoring visually sharp but inconsistent artifacts. Moreover, most SR methods rely on ground-truth (GT)–dependent distribution matching, which does not necessarily correspond to human judgments. In this work, we propose RefReward-SR, a low-resolution (LR) reference-aware reward model for preference-aligned SR. Instead of relying on GT supervision or NR evaluation, RefReward-SR assesses high-resolution (HR) reconstructions conditioned on their LR inputs, treating the LR image as a semantic anchor. Leveraging the visual–linguistic priors of a Multimodal Large Language Model (MLLM), it evaluates semantic consistency and plausibility in a reasoning-aware manner. To support this paradigm, we construct RefSR-18K, the first large-scale LR-conditioned preference dataset for SR, providing pairwise rankings based on LR–HR consistency and HR naturalness. We fine-tune the MLLM with Group Relative Policy Optimization (GRPO) using LR-conditioned ranking rewards, and further integrate GRPO into SR model training with RefReward-SR as the core reward signal for preference-aligned generation. Extensive experiments show that our framework achieves substantially better alignment with human judgments, producing reconstructions that preserve semantic consistency while enhancing perceptual plausibility and visual naturalness.

Keywords: Image Super-Resolution · Multimodal Large Language Models · Preference Alignment · Group Relative Policy Optimization

⋆ Corresponding author: *qweizework@gmail.com*

1 Introduction

Image Super-Resolution (SR) [12, 22, 24, 46, 47, 53, 67] aims to reconstruct high-resolution (HR) images from low-resolution (LR) observations. As an inherently ill-posed inverse problem, SR has evolved from bicubic degradation settings to real-world scenarios with complex and unknown degradations [31]. The ultimate goal of real-world SR is to produce perceptually pleasing results that align with human visual aesthetics while preserving content fidelity. Early representative methods, such as BSRGAN [64] and Real-ESRGAN [46], formulated this task as a supervised distribution-matching problem, leveraging adversarial losses to synthesize plausible high-frequency details. With diffusion priors [5, 35], recent methods [2, 27, 37, 44, 52, 61] surpass GAN-based models, generating fine details even under severe degradation.

However, as generative models become increasingly powerful, existing evaluation and optimization frameworks for SR are encountering significant limitations. We summarize these challenges into two critical aspects:

1) Inadequacy of Existing Metrics for Semantic Alignment. Traditional Full-Reference (FR) metrics [48, 65] provide indispensable benchmarks for measuring pixel-level or structural fidelity. However, their reliance on strict spatial correspondence can be overly conservative for generative SR, where they may penalize "reasonable hallucinations"—details that are semantically sound and visually enriching but spatially misaligned with the ground-truth (GT) [6]. On the other hand, No-Reference (NR) metrics [9, 18, 43, 49, 54, 60, 62] excel at assessing low-level visual sharpness and textural naturalness without requiring a reference. However, due to the lack of reference, they inherently struggle to preserve content fidelity. This often allows models to "game the metric" by introducing sharp but unfaithful artifacts, or even altering the subject identity, while still receiving high scores. More importantly, both FR and NR metrics evaluate high-level semantic plausibility only passively through low-level cues, ignoring the top-down cognitive mechanism where human perception naturally prioritizes semantic consistency before scrutinizing low-level details. Thus, they fail to explicitly penalize obvious semantic violations or structural distortions that contradict world knowledge. Consequently, existing metrics struggle to fully align with true human preferences.

2) The Insufficiency of Exclusively GT-Driven Optimization. Most SR methods [27, 44, 52, 61] rely on supervised losses for fidelity, while advanced approaches [28, 56] incorporate adversarial losses. However, while GT supervision provides necessary structural bounds, relying exclusively on GT distribution matching restricts generative potential. Strictly fitting empirical GT is insufficient to capture higher-level human semantic preferences. Furthermore, real-world GT sometimes contains inherent degradations, causing overly GT-dependent models to reproduce imperfections rather than enhance quality. Unlike the Text-to-Image field, which actively aligns with human preferences via RLHF [4, 42, 58], using RLHF to complement GT constraints in SR remains limited.

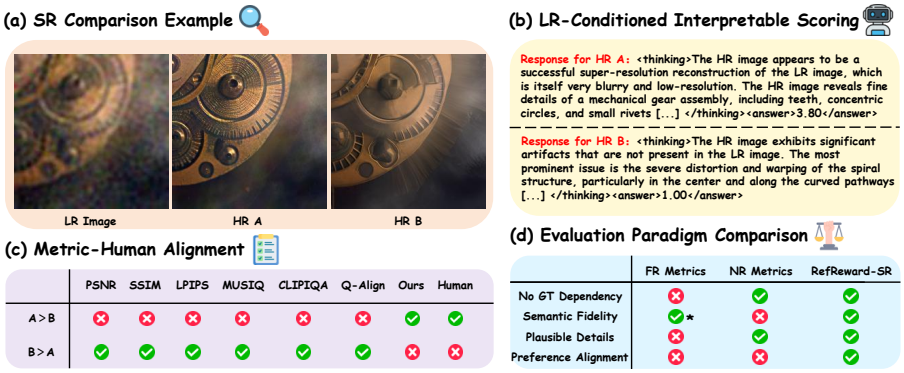

Fig. 1: Overview of our proposed evaluation paradigm. (a) HR A preserves structures, while HR B shows severe distortions. (b) RefReward-SR uses an MLLM for interpretable, LR-conditioned semantic scoring. (c) Existing metrics wrongly favor the distorted HR B, whereas ours aligns with human perception. (d) RefReward-SR uniquely ensures semantic fidelity and preference alignment without GT dependency. The asterisk(*) indicates that FR metrics evaluate high-level semantics only passively via low-level correspondence.

In this work, we propose **RefReward-SR**, an LR reference-aware semantic scoring model designed to complement existing evaluation paradigms. Addressing the ill-posed nature of super-resolution, RefReward-SR mimics human assessment by treating the LR image as a semantic anchor, leveraging the visual-linguistic priors of a Multimodal Large Language Model (MLLM) to jointly assess semantic consistency and plausibility from a high-level cognitive perspective. To this end, we construct **RefSR-18K**, the first large-scale LR-conditioned preference dataset focused on semantic-level annotations, establishing a critical foundation to shift SR evaluation and optimization from pixel-matching to human preference alignment. We fine-tune the MLLM using Group Relative Policy Optimization (GRPO) [11, 33] to align its judgments with human preferences, thereby capturing the underlying logical reasoning intrinsic to human perception. This enables RefReward-SR to serve as a key semantic complement to low-level metrics, establishing a more comprehensive evaluation and optimization framework for generative SR.

Building upon the learned reward models, we introduce GRPO into the super-resolution framework, employing a joint reward function with RefReward-SR as the core semantic signal to guide preference-aligned generation. Experimental results demonstrate that incorporating RefReward-SR alongside perceptual and quality constraints significantly enhances alignment with human judgments, leading to reconstructions with superior semantic consistency and plausibility.

Our contributions can be summarized as follows:

– We construct the first large-scale LR-conditioned preference dataset for SR. By treating the LR image as a semantic anchor, RefSR-18K provides 18,000

pairwise ranking annotations across diverse images, prioritizing high-level semantic consistency and plausibility.
- We propose RefReward-SR, an LR reference-aware semantic scoring model explicitly aligned with human perception. To address fine-grained discrimination, we further introduce a *global-local crop scoring* strategy to enhance sensitivity to localized semantic violations and artifacts.
- We introduce GRPO into the SR task with RefReward-SR as the reward signal, enabling preference-aligned optimization that improves semantic consistency and visual plausibility.

2 Related Work

2.1 Evaluation Metrics for Super-Resolution

Traditional FR metrics, including PSNR, SSIM [48], and LPIPS [65], quantify reconstruction quality by measuring deviations from the ground truth (GT). While indispensable for ensuring basic structural fidelity, operating within a strict distortion-minimization framework [6] causes them to systematically penalize spatially misaligned yet perceptually plausible details. Moreover, the frequent unavailability of high-quality GT in real-world scenarios further limits their practicality [10]. To remove GT dependency, NR metrics such as MUSIQ [18], MANIQA [60], and CLIPIQA [43] assess image quality without supervision. Recent IQA approaches built on Multimodal Large Language Models (MLLMs) [49,54,62] introduce partial interpretability and region awareness. Although effective at evaluating low-level visual aesthetics, their lack of a semantic anchor inherently causes them to struggle to preserve content consistency, often allowing models to inflate scores via excessive sharpening or unfaithful hallucinations. Furthermore, neither FR nor NR metrics can systematically evaluate semantic plausibility, which is the primary focus of human assessment for SR results. In contrast, we formulate SR evaluation as an LR-conditioned preference modeling problem. By jointly reasoning over the LR input and its HR reconstruction, RefReward-SR leverages the world knowledge of a Multimodal Large Language Model to provide semantically grounded, LR-anchored, and preference-aligned evaluation signals, effectively complementing the limitations of traditional FR/NR paradigms in semantic-level human preference assessment.

2.2 Generative Image Super-Resolution

Early SR methods optimized pixel-wise reconstruction losses under predefined degradations, achieving high PSNR but producing over-smoothed results [12, 19,21,25,41,67]. To improve visual quality, adversarial learning was introduced. Models such as ESRGAN [64] and Real-ESRGAN [46] framed SR as distribution matching, generating high-frequency details via GAN objectives. Later, diffusion-based approaches [27, 28, 45, 50, 52] leveraging pretrained models like Stable Diffusion [35] and FLUX [5] further advanced generative SR, demonstrating stronger semantic priors and improved perceptual quality under complex

degradations. Despite these advances, existing optimization paradigms remain focused on fitting the GT data distribution, rather than directly optimizing objectives aligned with human preferences. Moreover, they are fundamentally constrained by the quality of GT images. To address this limitation, inspired by reinforcement learning in text-to-image generation [4, 17, 29, 42, 58, 59], several works [8,32,51,57] have begun exploring reinforcement learning for SR. However, these methods often rely on existing evaluation metrics [8,51] with inherent limitations, or on MLLMs [32,57] that are not well-suited for low-level vision tasks, preventing them from fully unlocking the potential of reinforcement learning in SR.

3 Method

3.1 Construction of the RefSR-18K Dataset

Motivation. In real-world generative SR, models frequently exhibit similar failure modes, such as structural distortions, unreasonable textures, and localized content hallucinations that alter the semantics of the LR input. Traditional Image Quality Assessment (IQA) [15,26,34] and aesthetic datasets [20,55,58] fail to capture these generative anomalies, as they primarily focus on low-level distortions (e.g., noise) or global artistic appeal at the single-image level. Consequently, they lack corresponding LR references to evaluate semantic consistency and fail to provide explicit rankings for the semantic plausibility of the SR results. To address this gap, we construct RefSR-18K, a preference dataset specifically tailored for generative SR. Conditioned on the LR input, this dataset comprehensively evaluates the reconstructed HR images from two perspectives: semantic consistency and semantic plausibility. This provides crucial data support for training LR-conditioned, semantic-aware evaluation models.

Data Collection and Candidate Generation. To ensure diversity and broad scene coverage, we curated 5,130 high-quality images from the LSDIR dataset [23], encompassing a wide range of visual categories. Following the degradation pipeline proposed in SeeSR [52], we applied random cropping and complex degradations to synthesize the corresponding LR inputs. To ensure comprehensiveness and broad representativeness, the dataset must encompass diverse reconstruction styles and typical generative distortions produced by various models. To achieve this, we employed 8 diverse, state-of-the-art SR models to generate HR candidates: DiffBIR [27], SeeSR [52], OSEDiff [50], S3Diff [63], LucidFlux [16], DiT4SR [14], OMGSR-S [56], and HYPIR [28]. This model pool deliberately covers both single-step and multi-step generation paradigms, as well as distinct generative priors based on Stable Diffusion [35] and FLUX [5] architectures.

Annotation Protocol and Quality Control. Each annotation group consists of one LR reference image and four HR candidates randomly sampled from a pool of eight model outputs and the GT. Annotators are instructed to rank the four candidates jointly based on two primary criteria: (1) **LR–HR semantic consistency**, i.e., whether the generated details faithfully align with the structures and semantic cues present in the LR input; and (2) **semantic**

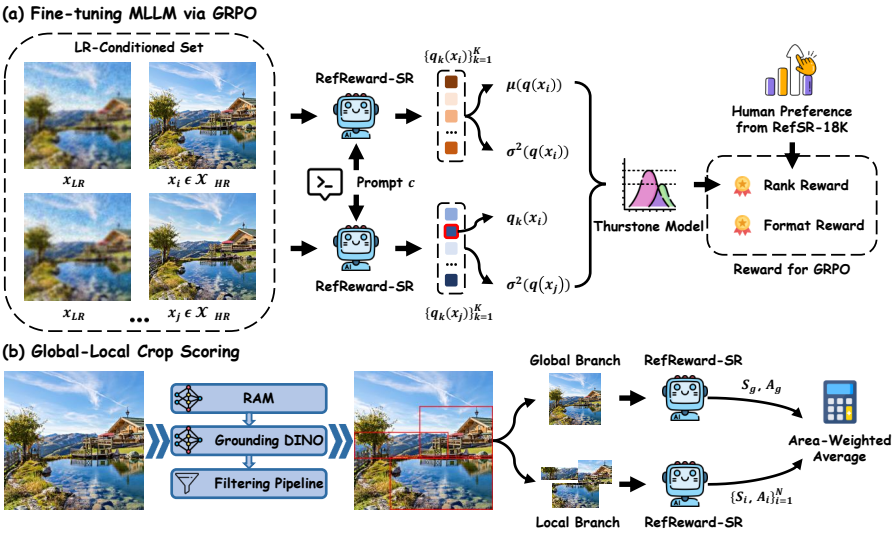

Fig. 2: The proposed RefReward-SR framework. (a) **MLLM Fine-tuning**: Fine-tuning via GRPO with format and LR-conditioned rank rewards from RefSR-18K for human preference alignment. (b) **Global-Local Crop Scoring**: Extracting representative local crops via RAM and Grounding DINO, then fusing multi-scale MLLM scores via area-weighted averaging for comprehensive evaluation.

plausibility, i.e., whether reconstructed objects are semantically natural and free from distortion-induced anomalies. When quality differences are indistinguishable, tied rankings are allowed. The resulting group-level rankings are then converted into pairwise preferences for training RefReward-SR and evaluating preference alignment. Detailed annotation guidelines and visual examples are provided in the *Suppl. Mater.*. Given the subjective and fine-grained nature of human preference annotation in image restoration, we adopt a strict quality control pipeline. Annotators are required to complete training and qualification tests prior to participation, and only those demonstrating high agreement with expert judgments are retained (approximately 50% are filtered out). Each group is independently annotated by at least three qualified annotators, and the final ranking is obtained via average rank aggregation. We further conduct expert review and post-hoc filtering to remove unreliable samples, including cases where candidate qualities are nearly indistinguishable, images lack sufficient semantic content for evaluation, or substantial disagreement exists among annotators. After this rigorous filtering process, 4,699 high-quality annotation groups are retained, resulting in 18,796 annotated HR reconstructions paired with LR references, forming the RefSR-18K dataset.

3.2 RefReward-SR: A Reference-Aware MLLM Evaluator

To build a human-aligned, reference-aware semantic evaluator, we fine-tune Qwen3-VL 8B [3] via GRPO [11, 33]. Given a low-resolution image x_{LR}, an evaluation prompt c specifying the scoring criteria and reasoning instructions, and a high-resolution candidate x_i from a candidate set $\mathcal{X}_{HR}$ (where $|\mathcal{X}_{HR}| = 4$) in the RefSR-18K dataset, the MLLM generates a detailed reasoning process followed by a structured score. The total reward for an output O_i comprises two core components:

Format Reward. To ensure the model functions as *an interpretable evaluator* rather than a "black-box" scorer, we enforce a structured output format. By inducing reasoning through reinforcement learning, we encourage the model to automatically explore reasonable reasoning paths guided by the evaluation prompt c. The model must first generate a reasoning chain within <thinking> ... </thinking> tags, explicitly articulating the issues present in the HR image and their specific locations conditioned on the LR reference. The final numerical rating is then enclosed in <answer>...</answer> tags. If any of these formats are incorrect, the format score is 0. The reward is defined as:

$$R_{\text{format}}(O_i) = \begin{cases} 1.0, & \text{if } O_i \text{ satisfies all format requirements} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

LR-Conditioned Rank Reward. The core of RefReward-SR is to learn the relative quality ordering of HR candidates conditioned on the same LR anchor x_{LR} and the evaluation prompt c. For a set of G HR candidates $\mathcal{X}_{HR} = \{x_1, \ldots, x_G\}$ corresponding to a specific LR input x_{LR}, we apply GRPO to generate K quality score predictions for each candidate x_i under the joint condition (x_{LR}, c), denoted as $\{q_k(x_i \mid x_{LR}, c)\}_{k=1}^{K}$. Here, $q_k(x_i \mid x_{LR}, c)$ denotes the scalar quality score produced by the k-th stochastic rollout of the policy $\pi_\theta(\cdot \mid x_i, x_{LR}, c)$. For notational simplicity, the conditioning on (x_{LR}, c) is implicit in the following formulations.

Based on the Thurstone model [40, 54], we compute the comparative probability of the k-th prediction of x_i against another candidate x_j within the same set by leveraging their sample means and variances:

$$p_k(x_i, x_j) = \Phi\left(\frac{q_k(x_i) - \mu(q(x_j))}{\sqrt{\sigma^2(q(x_i)) + \sigma^2(q(x_j)) + \gamma}}\right), \tag{2}$$

where $\mu(\cdot)$ and $\sigma^2(\cdot)$ are the sample mean and variance of the K predictions, Φ is the standard normal cumulative distribution, and γ is a small stabilizing constant. This explicitly accommodates predictive uncertainty for different images.

The true preference $p(x_i, x_j) \in \{1, 0.5, 0\}$ is derived from human annotations in RefSR-18K (representing win, tie, or loss). We define the reward $R_k(x_i)$ using a continuous fidelity measure averaged across the remaining $G - 1$ candidates:

$$R_k(x_i) = \frac{1}{G-1} \sum_{j \neq i} \left(\sqrt{p(x_i, x_j) p_k(x_i, x_j)} + \sqrt{(1 - p(x_i, x_j))(1 - p_k(x_i, x_j))} \right). \tag{3}$$

This continuous reward provides precise guidance by capturing subtle variations in quality ranking, penalizing outlier predictions.

Global-Local Crop Scoring. In practice, different SR results often exhibit consistent overall structures, with quality differences primarily manifesting in local details. Consequently, relying solely on global scoring can easily cause these critical nuances to be overlooked. During manual annotation, human evaluators typically observe the entire image first before comparing the primary objects individually. To mimic this behavior, we propose Global-Local Crop Scoring, which performs a holistic assessment followed by a fine-grained local comparison.

The strategy operates in two stages: (1) **Region Proposals**: We first leverage the RAM [66] to extract semantic tags, which then guide Grounding DINO [30] to generate candidate bounding boxes. To mimic human visual behavior that prioritizes texture-rich and salient objects, while balancing evaluation efficiency, we design a bounding box filtering pipeline to select representative regions with low overlap (details in *Suppl. Mater.*). (2) **Joint Evaluation**: The MLLM evaluates the global image and the selected crops separately. To integrate these multi-scale assessments, the final score S_{final} is computed as an area-weighted average:

$$S_{\text{final}} = \frac{A_g S_g + \sum_{i=1}^{N} A_i S_i}{A_g + \sum_{i=1}^{N} A_i}, \tag{4}$$

where S_g, A_g denote the score and area of the global image, and S_i, A_i represent those of the i-th local crop. This mechanism accounts for both global information and local details, leading to a more accurate assessment score.

3.3 Preference-Aligned Super-Resolution via GRPO

Having established RefReward-SR as a reliable human-aligned evaluator, we further integrate it into the SR generation stage to directly optimize reconstruction quality toward human preference alignment. In this work, we adopt the C-FLUX model from DP^{2}O-SR [51] as our base generative SR model. Specifically, we formulate the fine-tuning of this C-FLUX model as a reinforcement learning problem, which is optimized using GRPO [29, 59].

Let the pre-trained SR diffusion model be parameterized as a conditional policy π_θ, which generates a high-resolution reconstruction $y \sim \pi_\theta(\cdot \mid x_{LR})$ conditioned on a low-resolution input x_{LR}. For each input x_{LR}, the policy produces a group of stochastic reconstructions $\mathcal{Y} = \{y_1, y_2, \ldots, y_G\}$, and optimization is performed based on relative quality comparisons within each group. Since the learning signal is entirely driven by reward feedback, the design of the reward function becomes the key factor for achieving preference alignment. To comprehensively enhance SR reconstruction quality, we construct a composite reward consisting of three complementary components.

RefReward-SR Reward ($R_{\mathbf{Ref}}$). The primary reward signal is provided by RefReward-SR, our LR reference-aware evaluator described in Sec. 3.2. It jointly assesses semantic consistency between the generated image y_i and the LR observation x_{LR}, as well as the semantic plausibility of the reconstructed content.

While global-local crop scoring enhances fine-grained assessment during evaluation, we exclusively employ the global-only RefReward-SR score during the GRPO fine-tuning stage. This ensures high computational efficiency, while the global score already provides strong and sufficient semantic guidance to align the generated results with human preferences.

LPIPS Perceptual Reward ($R_{\mathbf{LPIPS}}$). To prevent excessive structural deviation and maintain perceptual fidelity with respect to the ground-truth image y_{GT}, we incorporate the LPIPS [65] metric as a perceptual constraint. Since lower LPIPS values indicate higher perceptual similarity, it is incorporated as a negative penalty term.

DEQA Quality Reward ($R_{\mathbf{DEQA}}$). We further incorporate DEQA [62] as an image quality assessment reward to enhance overall visual sharpness and strengthen the penalization of low-level visual degradations (e.g., blur and noise).

The overall reward for each candidate reconstruction y_i is defined as

$$R_{\mathrm{total}}(y_i) = \lambda_1 R_{\mathrm{Ref}}(y_i, x_{LR}) - \lambda_2 R_{\mathrm{LPIPS}}(y_i, y_{\mathrm{GT}}) + \lambda_3 R_{\mathrm{DEQA}}(y_i), \quad (5)$$

where λ_1, λ_2, and λ_3 are hyperparameters balancing semantic human preference alignment, perceptual fidelity, and artifact suppression, respectively.

4 Experiments

4.1 Effectiveness of RefReward-SR

To validate whether RefReward-SR can assess SR results at the semantic level in alignment with human judgment, and to investigate to what extent existing evaluation metrics capture human preferences, we conduct a comprehensive evaluation mimicking human selection behavior and compare our model against traditional NR/FR metrics and state-of-the-art MLLMs.

Experimental Setup. To evaluate performance and generalization, we establish two test sets. Both sets share the same 200 LR reference images but differ in the source of their HR candidates: (1) **In-domain Test Set**: We randomly sample 200 groups directly from the original RefSR-18K dataset. These groups contain HR candidates generated by the 8 models seen during training, along with GT. (2) **Out-of-domain Test Set**: Utilizing the identical 200 LR inputs from the In-domain set, we construct new comparison groups to assess generalization. In each group, 2 candidates are generated by entirely unseen models (TSD-SR [13] and PiSA-SR [36]), while the remaining 2 are sampled from the GT and the SR models used in training. The out-of-domain set is annotated following the exact same pipeline as RefSR-18K to ensure consistency.

We compare RefReward-SR with three categories of evaluators: (1) FR metrics, including PSNR, SSIM [48], and LPIPS [65]; (2) NR metrics, including MUSIQ [18], CLIPIQA [43], Q-Align [49], NIMA [38], VQ-R1 [54], and TOPIQ-IAA [9]; and (3) general-purpose MLLMs, including GPT-5.2 [1], Gemini 3

Table 1: Quantitative results (in %) on human preference alignment. We report agreement with human annotators, as well as Recall@1 and Filter@1 on both in-domain and out-of-domain test sets. "Annotators" denotes the average agreement among human annotators. All ± values indicate the maximum deviation from the mean.

Method	In-Domain			Out-of-Domain		
	Agreement	Recall@1	Filter@1	Agreement	Recall@1	Filter@1
Human Agreement						
Annotators	84.7 ± 1.4	-	-	81.8 ± 1.3	-	-
FR Metrics						
PSNR	41.8 ± 1.6	45.0	9.5	35.5 ± 2.3	25.5	8.5
SSIM [48]	51.8 ± 2.0	51.0	25.5	43.3 ± 1.8	26.0	19.0
LPIPS [65]	62.5 ± 0.9	49.5	46.0	58.8 ± 2.0	44.5	34.0
NR Metrics						
MUSIQ [18]	60.9 ± 1.4	40.5	42.0	57.6 ± 2.7	38.0	34.5
CLIPIQA [43]	54.0 ± 1.7	31.0	30.5	56.9 ± 2.5	32.0	37.5
NIMA [38]	54.5 ± 2.0	31.5	24.0	59.5 ± 2.3	41.0	40.0
TOPIQ-IAA [9]	54.1 ± 1.6	36.0	27.0	56.4 ± 1.7	36.0	33.5
Q-Align [49]	54.9 ± 1.7	35.0	29.5	54.3 ± 1.0	40.5	32.0
VQ-R1 [54]	47.8 ± 1.3	49.0	48.0	44.1 ± 2.8	50.5	39.0
MLLMs						
GPT-5.2 [1]	52.5 ± 1.0	30.7	46.2	49.7 ± 2.2	38.1	29.9
Gemini 3 Pro [39]	58.8 ± 1.5	54.0	52.0	47.9 ± 2.1	42.6	36.6
Qwen3-VL 8B [3]	31.8 ± 0.4	67.5	58.0	26.1 ± 0.9	59.0	57.0
Qwen3-VL 32B [3]	46.4 ± 1.3	39.5	43.0	40.7 ± 0.4	38.5	36.0
RefReward-SR	**85.0 ± 2.4**	**84.5**	**78.0**	**80.2 ± 2.2**	**77.0**	**73.0**

Pro [39], and the non-fine-tuned Qwen3-VL 8B [3]. To ensure a fair comparison, all MLLMs are evaluated using the same prompt design as RefReward-SR.

Following the evaluation protocol established by ImageReward [58], we assess the models based on agreement and their ability to successfully recall the best images or filter out the worst images. Specifically, *Agreement* assesses the likelihood of the model sharing consistent preferences with human annotators. *Recall@1* evaluates the model's ability to select the highly preferred images, measuring the frequency with which the human-annotated best image is ranked first by the model. Conversely, *Filter@1* evaluates the model's ability to identify and penalize the worst generations by measuring how often the human-annotated worst image is ranked last. More details are provided in the *Suppl. Mater.*.

Results and Analysis. Table 1 reports the quantitative results on semantic-level human preference alignment. RefReward-SR achieves performance comparable to human annotators on both in-domain and out-of-domain test sets, obtaining the best Recall@1 and Filter@1 while consistently outperforming all baselines. On the out-of-domain set, the performance of both human annotators and most metrics decreases, likely because the two unseen SR methods produce outputs with quality highly similar to other candidates, making preference dis-

Table 2: Quantitative comparison of state-of-the-art Real-ISR methods. To evaluate the effectiveness of our alignment strategy, the best results among the base model (C-FLUX), its DP^{2}O-fine-tuned variant (DP^{2}O-FLUX), and our proposed method are highlighted with a light red background. The best and second best results across all methods are highlighted in **red** and **blue**, respectively.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA↑	NIMA↑	TOPIQ-IAA↑	Q-Align↑	VQ-R1↑
RealSR	StableSR	24.54	0.7004	0.3068	65.92	0.6325	4.7880	4.5738	3.2930	3.7670
	DiffBIR	24.83	0.6501	0.3650	69.28	0.7053	4.9192	4.8973	3.7889	4.1030
	SeeSR	25.14	0.7211	0.3007	69.82	0.6705	4.9196	4.8548	3.7190	3.9830
	OSEDiff	**25.15**	**0.7341**	0.2921	69.09	0.6693	4.8953	4.7545	3.6962	4.0930
	PiSA-SR	**25.50**	**0.7418**	**0.2672**	70.15	0.6698	4.8954	4.7431	3.6354	4.0180
	TSD-SR	23.40	0.6938	**0.2823**	71.26	**0.7416**	5.0439	4.8958	3.8444	4.0800
	DIT4SR	23.52	0.6683	0.3153	67.70	0.6305	4.8737	4.5709	3.3875	3.9430
	LucidFlux	20.62	0.6016	0.3895	**71.95**	**0.7397**	**5.2414**	**5.1494**	3.7662	3.9010
	HYPIR	22.76	0.6669	0.3180	65.89	0.6369	4.9407	4.7330	3.6814	4.1530
	C-FLUX	24.42	0.6742	0.3376	68.70	0.6221	4.8799	4.6303	3.4943	3.9650
	DP^{2}O-FLUX	24.51	0.6774	0.3259	**72.42**	0.7156	4.9950	4.8873	**3.9757**	**4.2170**
	Ours	24.68	0.7101	0.3202	71.70	0.7007	**5.0567**	**4.9533**	**4.1596**	**4.2930**
DIV2K-Val	StableSR	23.26	0.5728	0.3112	65.78	0.6767	4.9243	4.7608	3.5213	3.8925
	DiffBIR	23.14	0.5441	0.3669	69.87	0.7298	5.2070	**5.1431**	4.1015	4.2014
	SeeSR	23.68	0.6043	0.3194	68.68	0.6936	5.0749	4.9928	3.9770	4.2013
	OSEDiff	**23.72**	**0.6109**	0.2942	67.97	0.6682	4.9664	4.8432	3.8358	4.1832
	PiSA-SR	**23.87**	**0.6058**	**0.2823**	69.68	0.6929	4.9584	4.8572	3.8804	4.2054
	TSD-SR	22.41	0.5644	**0.2710**	71.64	**0.7559**	5.1239	5.0116	4.0147	4.1529
	DIT4SR	21.79	0.5487	0.3451	67.96	0.6619	5.0888	4.8515	3.7135	4.1551
	LucidFlux	20.42	0.5106	0.3769	71.52	**0.7827**	**5.3984**	**5.3549**	4.0886	4.0852
	HYPIR	22.17	0.5669	0.3073	64.73	0.6421	5.0178	4.8767	3.7544	4.0757
	C-FLUX	22.71	0.5520	0.3349	69.70	0.6781	5.1084	4.9519	4.0573	4.2558
	DP^{2}O-FLUX	22.72	0.5536	0.3216	**72.98**	0.7536	**5.2482**	5.1347	**4.3983**	**4.4343**
	Ours	22.88	0.5786	0.3260	**71.95**	0.7302	5.1507	5.0695	**4.4605**	**4.4396**

crimination more challenging. Although existing NR/FR metrics mainly focus on low-level features, they can indirectly reflect semantic-level human preferences, allowing most to maintain a positive agreement (i.e., >50%). Notably, except for PSNR, all methods with agreement below 50% are MLLM-based. When SR candidates share similar structures, zero-shot MLLMs tend to be insensitive to subtle image details and often assign identical scores. Under our calculation protocol (detailed in the *Suppl. Mater.*), this behavior drives the agreement below 50%, while paradoxically causing an upward trend in Recall@1 and Filter@1 (as typically seen in Qwen3-VL 8B). This phenomenon highlights the limitation of zero-shot MLLMs for fine-grained SR evaluation and further motivates the construction of the RefSR-18K dataset for preference-aligned fine-tuning.

4.2 Performance of Preference-Aligned Super-Resolution

Training and Test Data. We construct our paired training data from the LSDIR dataset [23] using the Real-ESRGAN [46] degradation, setting the training resolution to 512×512 to align with our base model, C-FLUX [51]. For evaluation, we follow the benchmark protocol provided by StableSR [44], including both synthetic and real-world test sets. The synthetic set is based on DIV2K-Val [65], where 3,000 images are randomly cropped to 512×512, and LR inputs

are generated using the same Real-ESRGAN degradation process. We further evaluate real-world performance on the RealSR dataset [7], which contains 100 paired LR–HR images at resolutions of 128×128 and 512×512.

Experimental Setup. We compare our proposed method against a comprehensive suite of state-of-the-art generative SR approaches, including StableSR [44], DiffBIR [27], SeeSR [52], OSEDiff [50], PiSA-SR [36], TSD-SR [13], DIT4SR [14], LucidFlux [16], and HYPIR [28]. To explicitly demonstrate the performance gains achieved by our alignment strategy, we further include the base model C-FLUX and its DPO-fine-tuned [42] variant DP^{2}O-FLUX (derived from DP^{2}O-SR [51]) for comparison, evaluating all methods using their official default settings to ensure fairness. The performance is quantitatively assessed using a comprehensive set of FR metrics, including PSNR, SSIM [48], and LPIPS [65], as well as NR metrics such as MUSIQ [18], CLIPIQA [43], Q-Align [49], NIMA [38], VQ-R1 [54], and TOPIQ-IAA [9].

User Study. As demonstrated in Sec. 4.1, existing FR and NR metrics fail to fully capture human visual preferences in generative SR. Moreover, directly evaluating SR models using RefReward-SR may risk reward hacking [29, 59], since our method is optimized using this evaluator during reinforcement learning. To ensure an unbiased evaluation, we compare our method with the base model C-FLUX and its DPO-fine-tuned variant (DP^{2}O-FLUX) through a human preference study. Additionally, PiSA-SR and TSD-SR are included to examine RefReward-SR's generalization and alignment with human judgment on unseen methods. Specifically, we randomly sample 10 images from DIV2K-Val and 10 images from RealSR to ensure both synthetic and real-world scenarios are covered. A total of 15 volunteers with prior experience in image processing are invited to participate in the evaluation. During each trial, participants are presented with one LR reference image and two HR reconstructions generated by different models. Annotators are asked to select the better reconstruction according to the same criteria adopted in RefSR-18K. Ties are allowed and counted as 0.5 wins when computing overall win rates.

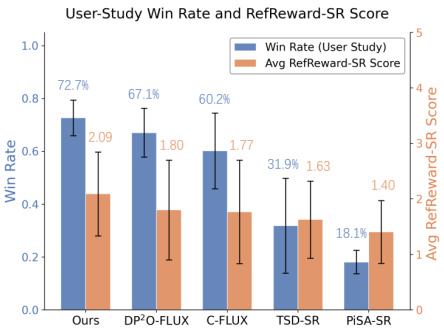

Fig. 3: The result of the user study win rate and average RefReward-SR score.

Results and Analysis. As shown in Table 2, our method consistently improves upon the base model (C-FLUX). Compared with DP^{2}O-FLUX, which is also fine-tuned using a reinforcement learning paradigm (DPO), our approach maintains a leading position on the majority of metrics. This demonstrates that our training strategy effectively enhances low-level fidelity while simultaneously boosting low-level visual naturalness. Qualitative comparisons in Fig. 4 further illustrate the advantages of our method. Compared with both the C-FLUX and

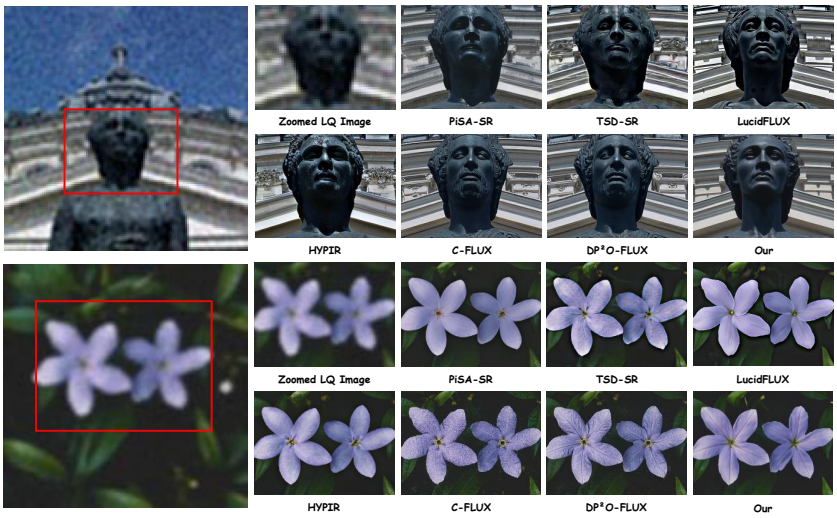

Fig. 4: Qualitative comparison of state-of-the-art generative SR methods. Please zoom in for a better view. Additional results are provided in the *Suppl. Mater.*.

the reinforcement learning baseline DP^{2}O-FLUX, our method produces visually cleaner, more realistic, and more semantically consistent high-resolution images, while remaining highly competitive with other recent SR methods.

As illustrated in Fig. 3, the user study shows that our method achieves the highest win rate of 72.7%, significantly outperforming DP^{2}O-FLUX (67.1%) and C-FLUX (60.2%). Meanwhile, the average RefReward-SR scores exhibit a strong positive correlation with the win rates, further validating the effectiveness of RefReward-SR as a reliable evaluation metric. Moreover, optimizing SR models with our proposed framework consistently drives the generation process toward better alignment with human semantic preferences. Interestingly, although C-FLUX performs worse than TSD-SR and PiSA-SR under conventional NR/FR metrics, it achieves a higher human preference win rate. This implies that existing metrics are insufficient for measuring alignment with human semantic preferences, which is consistent with the findings reported in Sec. 4.1.

4.3 Ablation Study

Ablation on the RefReward-SR Evaluator. As reported in Table 3, we assess our reward model's key design choices using human alignment metrics across in-domain and out-of-domain test sets. We compare the full model against variants trained on reduced data (1K and 2K groups) and without the global-local crop scoring mechanism. Scaling the training data yields consistent gains, demonstrating that larger preference datasets enhance human alignment and model generalization. Furthermore, removing crop scoring causes a performance drop, especially on the out-of-domain set, highlighting its effectiveness in capturing fine-grained semantic inconsistencies. Ultimately, the full configuration of

Table 3: Ablation study of the RefReward-SR evaluator on both in-domain and out-of-domain test sets. Agreement is reported in %.

Variant	In-Domain	Out-of-Domain
1K Data	81.2 ± 1.9	76.6 ± 3.2
2K Data	82.7 ± 1.9	77.4 ± 1.8
w/o G-L	83.6 ± 2.6	78.1 ± 2.3
Ours (Full)	$\mathbf{85.0 \pm 2.4}$	$\mathbf{80.2 \pm 2.2}$

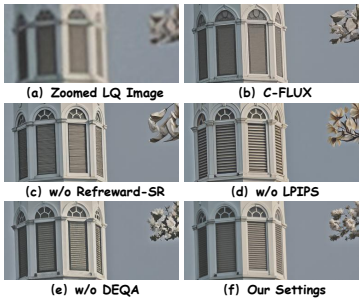

Fig. 5: Visual comparison of different reward components in the ablation study.

Table 4: Ablation study of reward components on the RealSR benchmark. The best and second best results are highlighted in **red** and **blue**, respectively.

Methods	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA↑	NIMA↑	TOPIQ-IAA↑	Q-Align↑	VQ-R1↑
C-FLUX (Base)	24.42	0.6742	0.3376	68.70	0.6221	4.8799	4.6303	3.4943	3.9650
w/o RefReward-SR	**23.89**	0.6920	0.3211	**73.48**	**0.7239**	**5.2534**	**5.0935**	**4.1656**	4.1655
w/o LPIPS	21.98	0.5798	0.4026	**74.55**	**0.7434**	**5.4195**	**5.2948**	**4.3521**	**4.3348**
w/o DEQA	23.87	**0.7099**	**0.2967**	69.02	0.6153	5.0189	4.6552	3.5313	4.0188
Ours (Full)	**24.68**	**0.7101**	**0.3202**	71.70	0.7007	5.0567	4.9533	4.1596	**4.2930**

RefReward-SR achieves the highest agreement, validating both data scaling and our global-local scoring design.

Ablation on Preference-Aligned SR Optimization. We investigate the contribution of each reward component defined in Eq. 5 during the GRPO fine-tuning of the C-FLUX model. As shown in Table 4 and Fig. 5, we iteratively remove each reward term and evaluate the resulting reconstruction quality on the RealSR benchmark. Analysis of the experimental results reveals that each reward component plays a distinct role. When RefReward-SR is removed, the FR metrics experience only a slight decline. This is because RefReward-SR primarily focuses on high-level semantics, whereas content fidelity is mainly maintained by the LPIPS reward. Although some NR metrics even show improvement, this absence still leads to semantic distortion and structural twisting in the generated results. Removing the LPIPS reward causes a significant drop in low-level fidelity metrics. While the model can still maintain a rough semantic correspondence, the generated details exhibit increased deviation from the LR input. Without the DEQA reward, the related NR metrics drop significantly compared to the full model. However, due to the enhanced semantic plausibility, these metrics still show a relative increase over the base model. Visually, the clarity of the results decreases, and the details become blurry. Overall, the full model achieves the most balanced performance across all metrics, yielding the most satisfactory generated results.

5 Conclusion

In this paper, we investigate the limitations of existing super-resolution (SR) evaluation metrics in capturing semantic understanding, which often leads to misalignment with human perceptual preferences. To address this issue, we introduce RefSR-18K, the first large-scale LR-conditioned preference dataset for generative SR with explicit annotations on semantic consistency and plausibility. Based on this dataset, we propose RefReward-SR, an MLLM-based evaluator designed to assess SR results from a semantic-level human preference perspective. We further incorporate RefReward-SR into the SR optimization pipeline as a core reward signal, enabling preference-aligned SR generation beyond conventional ground-truth distribution matching. Extensive experiments demonstrate that RefReward-SR provides a reliable proxy for human evaluation, and that the resulting preference-aligned SR model effectively suppresses generative artifacts while improving semantic fidelity and visual naturalness. We hope this work will inspire future research toward more human-centric evaluation and optimization paradigms for low-level vision tasks.

Acknowledgment. This work was partially supported by the National Natural Science Foundation of China (62572469, 62531026, 62437001), the Natural Science Foundation of Jiangsu Province under Grant BK20243051, and the OPPO Research Fund.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ai, Y., Zhou, X., Huang, H., Han, X., Chen, Z., You, Q., Yang, H.: Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. Advances in Neural Information Processing Systems **37**, 55443–55469 (2024)
3. 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)
4. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. In: International Conference on Learning Representations (2024)
5. Black Forest Labs: Flux. `https://github.com/black-forest-labs/flux` (2024)
6. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6228–6237 (2018)
7. 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: IEEE/CVF International Conference on Computer Vision. pp. 3086–3095 (2019)

8. Cai, M., Li, S., Li, W., Huang, X., Chen, H., Hu, J., Wang, Y.: Dspo: Direct semantic preference optimization for real-world image super-resolution. arXiv preprint arXiv:2504.15176 (2025)
9. 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 **33**, 2404–2418 (2024)
10. Chen, D., Wu, T., Ma, K., Zhang, L.: Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12742–12752 (2025)
11. DeepSeek-AI: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
12. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision. pp. 184–199. Springer (2014)
13. Dong, L., Fan, Q., Guo, Y., Wang, Z., Zhang, Q., Chen, J., Luo, Y., Zou, C.: Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23174–23184 (2025)
14. Duan, Z.P., Zhang, J., Jin, X., Zhang, Z., Xiong, Z., Zou, D., Ren, J.S., Guo, C., Li, C.: Dit4sr: Taming diffusion transformer for real-world image super-resolution. In: IEEE/CVF International Conference on Computer Vision. pp. 18948–18958 (2025)
15. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3677–3686 (2020)
16. Fei, S., Ye, T., Wang, L., Zhu, L.: Lucidflux: Caption-free universal image restoration via a large-scale diffusion transformer. arXiv preprint arXiv:2509.22414 (2025)
17. Guo, Y., Liu, J., Wang, Z., Chen, H., Sun, X., Zhao, Y., Wu, J., Yu, X., Liu, Z., Barsoum, E.: Imagedoctor: Diagnosing text-to-image generation via grounded image reasoning. arXiv preprint arXiv:2510.01010 (2025)
18. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
19. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1646–1654 (2016)
20. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 36652–36663 (2023)
21. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 624–632 (2017)
22. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4681–4690 (2017)
23. 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1775–1787 (2023)

24. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: IEEE/CVF International Conference on Computer Vision. pp. 1833–1844 (2021)
25. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 136–144 (2017)
26. Lin, H., Hosu, V., Saupe, D.: Kadid-10k: A large-scale artificially distorted iqa database. In: Eleventh International Conference on Quality of Multimedia Experience. pp. 1–3. IEEE (2019)
27. 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: European Conference on Computer Vision. pp. 430–448. Springer (2024)
28. 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 **44**(6), 1–21 (2025)
29. 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. Advances in Neural Information Processing Systems **38**, 40783–40818 (2025)
30. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
31. Moser, B.B., Shanbhag, A.S., Raue, F., Frolov, S., Palacio, S., Dengel, A.: Diffusion models, image super-resolution, and everything: A survey. IEEE Transactions on Neural Networks and Learning Systems (2024)
32. Qiao, J., Cai, M., Li, W., Liu, Y., Huang, X., He, G., Xie, J., Hu, J., Chen, X., Lin, S.: Realsr-r1: Reinforcement learning for real-world image super-resolution with vision-language chain-of-thought. arXiv preprint arXiv:2506.16796 (2025)
33. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
34. Sheikh, H.R., Sabir, M.F., Bovik, A.C., et al.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing **15**(11), 3440–3451 (2006)
35. Stability.ai: `https://stability.ai/stable-diffusion`
36. Sun, L., Wu, R., Ma, Z., Liu, S., Yi, Q., Zhang, L.: Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2333–2343 (2025)
37. Tai, Y., Xie, R., Zhao, C., Zhang, K., Zhang, Z., Zhou, J., Yang, J.: Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. Pattern Recognition p. 113012 (2026)
38. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE Transactions on Image Processing **27**(8), 3998–4011 (2018)
39. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
40. Thurstone, L.L.: A law of comparative judgment. In: Scaling, pp. 81–92. Routledge (2017)
41. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: IEEE/CVF International Conference on Computer Vision. pp. 4799–4807 (2017)

42. 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
43. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
44. 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 **132**(12), 5929–5949 (2024)
45. Wang, X., Yan, J.K., Cai, J.Y., Deng, J.H., Qin, Q., Cheng, Y.: Super-resolution reconstruction of single image for latent features. Computational Visual Media **10**(6), 1219–1239 (2024)
46. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: IEEE/CVF International Conference on Computer Vision. pp. 1905–1914 (2021)
47. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: European Conference on Computer Vision. pp. 63–79. Springer (2018)
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 **13**(4), 600–612 (2004)
49. 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: International Conference on Machine Learning. pp. 54015–54029. PMLR (2024)
50. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. Advances in Neural Information Processing Systems **37**, 92529–92553 (2024)
51. 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. Advances in Neural Information Processing Systems **38**, 172015–172041 (2025)
52. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25456–25467 (2024)
53. Wu, S., Dong, C., Qiao, Y.: Exploring contextual priors for real-world image super-resolution. Computational Visual Media **11**(1), 159–177 (2025)
54. Wu, T., Zou, J., Liang, J., Zhang, L., Ma, K.: VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. Advances in Neural Information Processing Systems **38**, 88167–88190 (2025)
55. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
56. Wu, Z., Sun, Z., Zhou, T., Fu, B., Cong, J., Dong, Y., Zhang, H., Tang, X., Chen, M., Wei, X.: Omgsr: You only need one mid-timestep guidance for real-world image super-resolution. arXiv preprint arXiv:2508.08227 (2025)
57. Xu, H., Liu, Y., Jiang, B., Peng, J., Luo, D., Hu, X., Yan, S., Li, H.: Irpo: Boosting image restoration via post-training grpo. arXiv preprint arXiv:2512.00814 (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. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)

59. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
60. 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1191–1200 (2022)
61. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: European Conference on Computer Vision. pp. 74–91. Springer (2024)
62. You, Z., Cai, X., Gu, J., Xue, T., Dong, C.: Teaching large language models to regress accurate image quality scores using score distribution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14483–14494 (2025)
63. Zhang, A., Yue, Z., Pei, R., Ren, W., Cao, X.: Degradation-guided one-step image super-resolution with diffusion priors. arXiv preprint arXiv:2409.17058 (2024)
64. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: IEEE/CVF International Conference on Computer Vision. 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
66. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize anything: A strong image tagging model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1724–1732 (2024)
67. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: European Conference on Computer Vision. pp. 286–301. Springer (2018)