GRASP: GROUNDED RESTORATION WITH AGENT-REFINED SEMANTIC PROMPTS FOR REAL-WORLD IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world image super-resolution (Real-ISR) requires preserving structures supported by a low-quality (LQ) input while synthesizing realistic fine details not directly observable in it. However, existing generative Real-ISR methods often leave the complementary roles of visual and semantic conditions implicit, limiting control over structural fidelity and detail realism. To address this challenge, we propose Grounded Restoration with Agent-Refined Semantic Prompts (GRASP), a Real-ISR framework that explicitly assigns complementary roles to local visual evidence, global image context, and semantic prompts. Specifically, we introduce a Decoupled Prior Injection mechanism that separates local visual evidence from global image context. Within this mechanism, a diptych provides spatially aligned local evidence to anchor observable structures and guide local detail synthesis, while a lightweight Vision-Context Bridge (VCB) supplies global image context to stabilize the scene-level layout. We further propose an inference-time agent-driven prompting strategy that iteratively refines semantic prompts to guide realistic detail synthesis in ambiguous regions. Experiments show that GRASP achieves strong perceptual performance across multiple Real-ISR benchmarks.

1 INTRODUCTION

Real-world image super-resolution (Real-ISR) aims to recover high-resolution (HR) images from low-quality (LQ) inputs with unknown and compound degradations (Wang et al., 2021; 2024; Wu et al., 2024b). With the emergence of large-scale generative models, super-resolution (SR) has shifted from deterministic reconstruction toward conditional generation under severe information uncertainty. In this new regime, the goal is no longer limited to recovering observable structures, but also to synthesizing plausible details that are not explicitly present in the degraded input. This raises a fundamental question: How can incomplete visual evidence and strong generative priors be coordinated to synthesize plausible details without sacrificing structural fidelity?

Generative Real-ISR methods differ in how they inject LQ information into the generative backbone. As illustrated in Figure 1(a–c), prior methods incorporate LQ-derived representations through latent-level conditioning (Wu et al., 2024a; Chen et al., 2025b; Kong et al., 2025), inject extracted visual features through auxiliary visual-control branches (Wang et al., 2024; Lin et al., 2024; Yang et al., 2024; Wu et al., 2024b; Ai et al., 2024), or enable interactions between visual and text tokens through joint token-level multimodal conditioning (Duan et al., 2025; Fang et al., 2026). Although these paradigms provide useful constraints, the complementary roles of visual and semantic conditions often remain implicit, making structural fidelity and detail realism difficult to coordinate under complex degradations. Moreover, semantic prompts that primarily summarize observable content may offer limited explicit guidance for synthesizing plausible fine details not directly observable in the LQ input. These limitations motivate a conditioning design that explicitly organizes heterogeneous cues according to their complementary roles in the SR process. Further discussion is provided in Appendix B.

To address this problem, we propose Grounded Restoration with Agent-Refined Semantic Prompts (GRASP), a Diffusion Transformer (DiT)-based Real-ISR framework that explicitly assigns complementary roles to local visual evidence, global image context, and semantic prompts, as illustrated in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

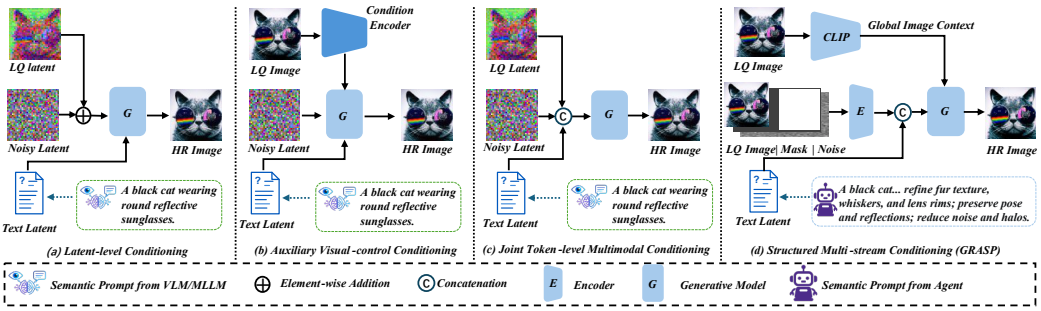

Figure 1: Schematic comparison of conditioning paradigms in generative Real-ISR. Prior methods typically inject low-quality observations and semantic prompts through latent-level conditioning, auxiliary visual-control branches, or joint token-level multimodal conditioning. Panel (a) illustrates an additive implementation of latent-level conditioning. GRASP organizes local visual evidence, global image context, and agent-refined semantic prompts as distinct condition streams with complementary functions for the SR objective.

Figure 1(d). Specifically, we formulate Real-ISR as visual in-context completion, with a Decoupled Prior Injection mechanism that separates local visual evidence from global image context. Here, the upsampled LQ input is placed beside a masked SR target region to form a diptych, providing spatially aligned local visual evidence to anchor observable structures and guide local detail synthesis. To complement this local reference, a lightweight Vision-Context Bridge (VCB) adapts Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) visual features to provide global image context, stabilizing scene-level layout and low-frequency content. For fine details not directly observable in the LQ input, we further propose an inference-time agent-driven prompting strategy that enriches semantic prompts with plausible detail descriptions. The agent iteratively refines these prompts based on the LQ input and intermediate SR results, guiding realistic detail synthesis while keeping visual conditions and model parameters fixed.

By assigning complementary roles to visual and semantic conditions, GRASP coordinates incomplete visual evidence with strong generative priors to maintain structural fidelity while synthesizing realistic details. Extensive experiments demonstrate that GRASP achieves strong perceptual performance across multiple Real-ISR benchmarks while using approximately 122M trainable adaptation parameters.

The main contributions of this paper are summarized as follows:

- We propose GRASP, a DiT-based framework that formulates Real-ISR as visual in-context completion and explicitly assigns complementary roles to visual and semantic conditions.
- We introduce a Decoupled Prior Injection mechanism that organizes visual conditions into two complementary streams: local visual evidence and global image context.
- We propose an inference-time agent-driven prompting strategy that iteratively refines semantic prompts to incorporate plausible descriptions of fine details not directly observable in the LQ input, guiding realistic detail synthesis.
- Extensive experiments demonstrate that GRASP achieves state-of-the-art performance across multiple Real-ISR benchmarks.

2 Related Work

Real-world Image Super-Resolution. Real-ISR aims to recover HR images from LQ inputs with complex and unknown degradations. Reconstruction-oriented SR methods (Dong et al., 2016; Lim et al., 2017; Zhang et al., 2018b; Liang et al., 2021; Conde et al., 2022) primarily emphasize reconstruction fidelity. To better handle real-world degradations, training with real-world data and realistic degradation models (Cai et al., 2019; Zhang et al., 2021; Wang et al., 2021; Wei et al., 2021) improves robustness, while GAN-based objectives (Ledig et al., 2017; Sajjadi et al., 2017; Wang et al., 2018a;b; Zhang et al., 2019; Fritsche et al., 2019; Maeda, 2020) extend the focus beyond reconstruction fidelity to perceptual realism. More recently, diffusion-based methods (Wang et al., 2024; Lin et al., 2024; Yang et al., 2024; Chen et al., 2025b; Yu et al., 2024; Wu et al., 2024a;

Yi et al., 2025) leverage pretrained diffusion or text-to-image priors to synthesize plausible details from severely degraded inputs. Recent DiT-based frameworks (Ai et al., 2024; Kong et al., 2025; Duan et al., 2025; Fang et al., 2026) extend this generative paradigm to transformer backbones for real-world image restoration and super-resolution. Despite these advances, coordinating visual and semantic conditions for structural fidelity and detail realism remains challenging.

Visual In-Context Generation and Restoration. Recent visual in-context systems built on diffusion transformers use images as visual context for generation and editing. FLUX.1 Fill (Black Forest Labs, 2024) provides a rectified-flow inpainting interface, and FLUX.1 Kontext (Black Forest Labs et al., 2025) unifies image and text inputs via sequence concatenation. In-Context LoRA (Huang et al., 2024), which uses low-rank adaptation (LoRA), demonstrates cross-image contextual modeling on a shared canvas. Diptych Prompting (Shin et al., 2025) formulates subject-driven generation as inpainting an incomplete diptych with a visible reference panel and a masked target panel. ICEdit (Zhang et al., 2025) uses in-context generation for instruction-based image editing, while Insert Anything (Song et al., 2025), In-Context Brush (Xu et al., 2025), and IC-Custom (Li et al., 2026) extend in-context generation to reference-based insertion and customization. VisualCloze (Li et al., 2025) generalizes this formulation to a unified visual in-context completion framework. Building on these in-context formulations, GRASP casts Real-ISR as diptych completion, using the upsampled LQ input as a visible reference panel alongside a masked SR target panel. The diptych supplies spatially aligned local visual evidence without a separate encoder dedicated to local conditioning.

Semantic Prompting for Image Super-Resolution. Severe degradations can obscure local visual evidence, motivating the use of language to disambiguate scene and object semantics. PASD (Yang et al., 2024) couples pixel-aware attention with classification, detection, and captioning signals. SeeSR (Wu et al., 2024b) trains a degradation-aware extractor producing hard tags and soft prompts, while SUPIR (Yu et al., 2024) uses prompts generated by multimodal large language models (MLLMs) and provided by users as a control interface. Recent works further spatialize or decouple textual priors: DTPSR (Jiang et al., 2026) separates textual priors by spatial hierarchy and frequency, while SRSR (Chen et al., 2025a) grounds cross-attention with masks. However, semantic prompts that primarily summarize observable content may offer limited explicit guidance for synthesizing plausible fine details not directly observable in the LQ input. To address this limitation, GRASP employs an inference-time agent that iteratively refines semantic prompts to guide detail synthesis in ambiguous regions.

3 METHOD

3.1 OVERALL FRAMEWORK

GRASP is a DiT-based Real-ISR framework that formulates Real-ISR as visual in-context completion and explicitly assigns complementary roles to local visual evidence, global image context, and semantic prompts. Given an LQ input x_{lq}, GRASP generates K SR candidates using fixed visual conditions and iteratively refined semantic prompts. The final SR output is expressed as

$$\hat{y} = \mathcal{S}\left(\left(\mathcal{G}_\theta\left(c_v^{\text{local}}, c_v^{\text{global}}, c_t^{(k)}; \epsilon\right)\right)_{k=1}^{K}\right), \tag{1}$$

where $\mathcal{G}_\theta$ denotes the complete generation process for a single SR candidate based on FLUX.1 Fill (Black Forest Labs, 2024), and $\mathcal{S}$ denotes the final Best-of-K selection operator. The local visual condition c_v^{local} is constructed from x_{lq} through the diptych, the global image context vector c_v^{global} is obtained through the Vision-Context Bridge (VCB), and $c_t^{(k)}$ is the textual condition encoded from the semantic prompt for the k-th candidate. Here, $c_t^{(1)}$ is encoded from the initial caption generated by the Large Language-and-Vision Assistant (LLaVA), while $c_t^{(k)}$ for $k \geq 2$ is encoded from agent-refined prompts; $\mathcal{S}$ performs per-image no-reference Best-of-K selection. The generator parameters θ, both visual conditions, and the initial latent noise ϵ remain fixed across rounds.

In this formulation, the fixed local and global visual conditions provide structural and scene-level constraints, while agent-refined semantic prompts supply plausible descriptions of details not directly observable in the LQ input. Figure 2 illustrates this condition organization in the overall GRASP pipeline, while Figure 3 details how the three condition streams are injected into the FLUX transformer

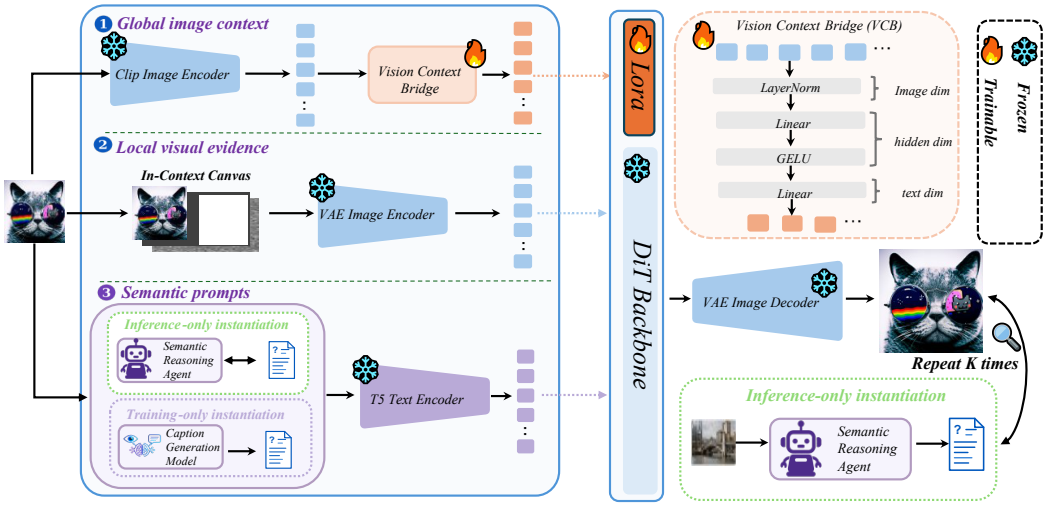

Figure 2: Overview of GRASP. Real-ISR is cast as visual in-context completion. Local visual evidence comes from the diptych, global image context from the Vision-Context Bridge over CLIP, and semantic prompts from LLaVA captions during training and agent refinement during inference.

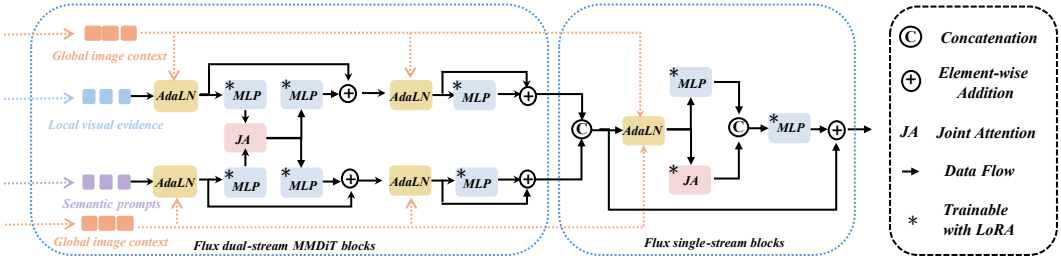

Figure 3: Detailed condition injection in GRASP. Global image context, local visual evidence, and semantic prompts are routed through the dual-stream and single-stream blocks of the FLUX multimodal diffusion transformer (MMDiT), while lightweight LoRA adaptation updates the trainable modulation and projection layers.

blocks. The following subsections describe the construction of each condition and the training and inference procedures.

3.2 Local visual evidence via diptych completion

To preserve pixel-wise spatial correspondence between the upsampled LQ input and the SR output, GRASP provides local visual evidence through diptych completion without an additional encoder dedicated to local conditioning. Let U_{bic} denote bicubic upsampling and $\text{cat}_w(\cdot,\cdot)$ denote concatenation along the width dimension. The upsampled input $x^{\text{up}} = U_{\text{bic}}(x_{lq})$ occupies the left panel as visible context, while an equally sized right panel defines the SR target region. During training, we construct the target diptych, base diptych, and image-space Fill mask as

$$d_{\text{tgt}} = \text{cat}_w(x^{\text{up}}, y), \qquad d_{\text{base}} = \text{cat}_w(x^{\text{up}}, x^{\text{up}}), \qquad m = \text{cat}_w(\mathbf{0}, \mathbf{1}), \tag{2}$$

where y is the HR target, $\mathbf{0}$ and $\mathbf{1}$ denote panel-sized all-zero and all-one mask tensors, and the mask m is zero on the visible left panel and one on the right target panel. After image and mask preprocessing, the masked diptych is formed as $x_{\text{mi}} = d_{\text{base}} \odot (1 - m)$. The target diptych d_{tgt} is used only during training; at inference, d_{base} and m are constructed from x_{lq} in the same way to form x_{mi}.

Following the FLUX.1 Fill interface (Black Forest Labs, 2024), the masked diptych x_{mi} is encoded by the variational autoencoder (VAE) and packed into masked-image tokens. These tokens are concatenated with the Fill-mask tokens along the feature dimension to form the local visual condition

c_v^{local}. The target diptych d_{tgt} is separately mapped to the training latent z_0 through VAE posterior sampling, latent scaling, and patch packing. The packed loss mask m_z corresponds to the same SR target panel as the image-space Fill mask m, but is constructed in the packed latent layout for loss computation.

This construction provides local structures, contours, and textures as spatially aligned in-context evidence to anchor observable structures and guide local detail synthesis.

3.3 Global image context via the Vision-Context Bridge

While the diptych provides spatially aligned local visual evidence, it does not explicitly summarize the input as a compact image-level descriptor. To complement this local reference, we introduce a lightweight Vision-Context Bridge (VCB) that provides global image context to stabilize scene-level layout and low-frequency content.

We first extract an image-level visual descriptor from x_{lq} using a frozen CLIP image encoder (Radford et al., 2021):

$$g_v = \text{Pool}\left(\Phi_{\text{CLIP}}(x_{lq})\right), \tag{3}$$

where Φ_{CLIP} denotes the CLIP image encoder and $\text{Pool}(\cdot)$ denotes the pooled output returned by the CLIP encoder. No additional patch mean pooling or L_2 normalization is applied. The VCB maps this descriptor into the conditioning space of the FLUX MMDiT backbone:

$$c_v^{\text{global}} = \text{VCB}(g_v) = s\left[W_2\,\sigma\left(W_1\,\text{LN}(g_v) + b_1\right) + b_2\right], \tag{4}$$

where $\text{LN}(\cdot)$ is LayerNorm, $\sigma(\cdot)$ is GELU, W_1 and W_2 are learnable projection matrices, b_1 and b_2 are biases, and s is a learnable scale.

The resulting c_v^{global} is provided to FLUX as an image-level conditioning vector, rather than as spatially aligned tokens. It provides global guidance through the modulation pathway illustrated in Figure 3, complementing the local visual evidence supplied by the diptych.

3.4 Inference-time agent-driven prompt refinement

When severe degradations leave fine details ambiguous, GRASP uses semantic prompts to provide plausible descriptions of details not directly observable in the LQ input. An inference-time agent refines these prompts to guide realistic detail synthesis while respecting the available visual evidence.

At inference time, a LLaVA-generated caption initializes the semantic prompt $p^{(1)}$, and the corresponding SR result $\hat{y}^{(1)}$ serves as the static-prompt baseline. Here, K denotes the total number of candidates, including the initial candidate and $K-1$ agent-refined candidates.

For each subsequent candidate $k = 2, \ldots, K$, the agent compares the previous raw SR result with the LQ input and conservatively revises the previous prompt. It is instructed to retain the main scene description while refining cues for ambiguous regions and visible artifacts. The prompt update is

$$p^{(k)} = \mathcal{A}\Big(x_{lq}, \hat{y}^{(k-1)}, p^{(k-1)}, \mathbf{q}^{(k-1)}\Big), \qquad k = 2, \ldots, K. \tag{5}$$

where $\mathcal{A}$ denotes the prompt-refinement agent, and $\mathbf{q}^{(k-1)}$ contains the four no-reference quality scores of the previous raw SR result as feedback. For all $k = 1, \ldots, K$, the text encoder $\mathcal{E}_t$ produces $c_t^{(k)} = \mathcal{E}_t(p^{(k)})$, where $p^{(1)}$ is initialized by LLaVA, and the corresponding SR candidate is generated as

$$\hat{y}^{(k)} = \mathcal{G}_\theta\Big(c_v^{\text{local}}, c_v^{\text{global}}, c_t^{(k)}; \epsilon\Big), \qquad k = 1, \ldots, K. \tag{6}$$

The generator parameters, visual conditions, and initial latent noise remain fixed throughout this process.

After all K candidates are generated, the selection operator $\mathcal{S}$ chooses the final output for each image without using HR references. Candidates are ranked separately by the same four no-reference metrics, with rank 1 assigned to the best candidate under each metric, and the candidate with the smallest rank sum is selected. Ties are resolved by the higher mean of the four scores after per-metric min–max normalization within the candidate set; any remaining tie is resolved in favor of the earlier candidate.

3.5 TRAINING AND INFERENCE STRATEGY

We jointly train the LoRA adapters and the Vision-Context Bridge while keeping the pretrained FLUX backbone frozen. For the main configuration, we use LLaVA-generated captions as textual supervision (Liu et al., 2023). Training uses a masked flow-matching objective, where v_θ denotes the velocity predictor used in $\mathcal{G}_\theta$:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{z_0,\epsilon,u}\left[\frac{\left\| m_z \odot \left(v_\theta(z_t, t, c_v^{\mathrm{local}}, c_v^{\mathrm{global}}, c_t) - (\epsilon - z_0)\right)\right\|_2^2}{\|m_z\|_1}\right], \tag{7}$$
$$u \sim \mathcal{N}(0,1), \quad t = \mathrm{sigmoid}(u), \quad \epsilon \sim \mathcal{N}(0, I), \quad z_t = (1-t)z_0 + t\epsilon.$$

where z_0 is the target diptych latent defined in Section 3.2, and m_z is the binary packed loss mask broadcast to the prediction shape, with ones on the SR target panel and zeros on the visible context panel. The denominator counts the supervised latent elements, including the feature dimension. Only the SR target panel contributes directly to the loss.

At inference, we follow the prompt-refinement and candidate-selection procedure in Section 3.4, using GPT-5.4 as the agent and $K = 10$ total candidates with Best-of-K selection unless otherwise specified. Model parameters and visual conditions remain fixed, and the same per-image denoising seed is used across refinement rounds. The instruction template and prompt examples are provided in Appendix E.3.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Training datasets. Following the training protocol of SeeSR (Wu et al., 2024b), we train GRASP on LSDIR (Li et al., 2023) and the first 10K face images from FFHQ (Karras et al., 2019). Low-resolution/high-resolution (LR-HR) training pairs are synthesized with the Real-ESRGAN degradation pipeline (Wang et al., 2021) under the $4\times$ SR setting.

Testing datasets. We evaluate GRASP on three standard $4\times$ SR benchmarks preprocessed by StableSR (Wang et al., 2024): DIV2K (Agustsson & Timofte, 2017), RealSR (Cai et al., 2019), and DRealSR (Wei et al., 2020), using the LR/HR pairs provided by StableSR. Among them, DIV2K is a synthetic benchmark, while RealSR and DRealSR are real-world SR benchmarks. To further assess generalization on real images without reference targets, we additionally report no-reference results on RealLR200 (Wu et al., 2024b) and RealLQ250 (Wang et al., 2024).

Implementation details. GRASP uses FLUX.1-Fill-dev (Black Forest Labs, 2024) as the generative backbone and is fine-tuned with LoRA (Hu et al., 2022) using rank 32. Global image context is extracted with CLIP ViT-L/14 (Radford et al., 2021) and adapted through the Vision-Context Bridge. Training uses 512×512 HR crops, a batch size of 16, and the Prodigy optimizer (Mishchenko & Defazio, 2023) with bias correction, safeguard warmup, and a weight decay of 0.01. The model is trained on 8 NVIDIA H800 GPUs for 15K iterations, taking approximately 24 hours. During inference, each SR candidate is generated with FlowMatchEulerDiscreteScheduler (von Platen et al., 2022) using 30 sampling steps and a guidance scale of 30. On an NVIDIA H800, static GRASP and the full $K = 10$ pipeline take 7.718 and 750.700 s/image, respectively, on 50 RealSR images at 512×512, excluding model loading and warm-up.

Evaluation metrics. Consistent with DiT4SR and prior studies (Blau & Michaeli, 2018; Gu et al., 2020; Yu et al., 2024; Duan et al., 2025), we use LPIPS (Zhang et al., 2018a) as the primary full-reference perceptual metric and report PSNR and SSIM (Wang et al., 2004) as supplementary fidelity indicators, since these distortion-oriented scores often fail to reflect the visual quality of generative SR results. On paired benchmarks, we additionally report DISTS and FID (Ding et al., 2022; Heusel et al., 2017), with PSNR/SSIM results provided in Appendix E.1. We report LIQE (Zhang et al., 2023), MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), and CLIPIQA (Wang et al., 2023) on both paired and no-reference benchmarks. All metrics use the same pyiqa backend (Chen & Mo, 2022); protocol details are in Appendix C.

Compared methods. We compare GRASP with representative recent generative Real-ISR methods, including StableSR (Wang et al., 2024), SeeSR (Wu et al., 2024b), SUPIR (Yu et al., 2024),

Table 1: Representative quantitative comparison on RealSR, DIV2K-Val, and DRealSR under the unified pyiqa backend. Red and blue denote the best and second-best results, respectively. GRASP denotes the full inference pipeline with agent-driven prompt refinement. GRASP w/o Agent Ref. uses the initial LLaVA prompt and a single output.

Dataset	Method	Metric						
		LPIPS↓	DISTS↓	FID↓	LIQE↑	MUSIQ↑	MANIQA↑	CLIPIQA↑
RealSR	StableSR	0.3272	0.2454	134.89	3.5055	61.47	0.4634	0.5653
	SeeSR	0.3007	0.2224	125.51	4.1365	69.82	0.5437	0.6704
	DiffBIR	0.4012	0.2603	151.33	4.2846	67.94	0.4865	0.6776
	FaithDiff	**0.2887**	**0.2114**	108.89	3.9926	68.85	0.4680	0.6204
	DiT4SR	0.3201	0.2258	120.12	3.9017	67.63	0.4582	0.6294
	FluxControlNet	0.3555	0.2438	126.46	3.6057	64.29	0.4554	0.6394
	GRASP w/o Agent Ref.	0.3323	0.2359	**108.24**	4.2739	69.71	0.6070	0.7007
	GRASP	0.3383	0.2383	110.46	**4.3833**	**70.52**	**0.6222**	**0.7155**
DIV2K-Val	StableSR	0.3456	0.2142	31.04	4.1023	65.07	0.5393	0.7124
	SeeSR	0.3190	0.1963	26.02	4.2642	68.62	0.5030	0.6926
	DiffBIR	0.3741	0.2291	42.09	4.3054	67.47	0.4685	0.6868
	FaithDiff	0.3124	0.1992	25.88	4.2441	69.32	0.4311	0.6467
	DiT4SR	0.3427	0.2123	36.60	4.1329	67.39	0.4392	0.6602
	FluxControlNet	0.3580	0.2137	26.51	3.7580	62.35	0.4321	0.6164
	GRASP w/o Agent Ref.	**0.3109**	**0.1955**	**23.82**	4.5183	70.12	0.6103	0.7283
	GRASP	0.3179	0.1989	24.93	**4.6074**	**70.98**	**0.6251**	**0.7431**
DRealSR	StableSR	0.3792	0.2683	169.83	2.9169	53.03	0.3943	0.5328
	SeeSR	**0.3174**	**0.2315**	147.46	4.1273	65.09	0.5129	**0.6910**
	DiffBIR	0.5095	0.2959	175.92	3.7407	59.83	0.4508	0.6486
	FaithDiff	0.3501	0.2416	153.34	4.0124	66.35	0.4512	0.6332
	DiT4SR	0.3721	0.2445	156.99	3.9953	65.29	0.4439	0.6660
	FluxControlNet	0.3972	0.2521	156.29	3.1211	56.01	0.4069	0.5959
	GRASP w/o Agent Ref.	0.3662	0.2456	**138.92**	3.9471	64.14	0.5163	0.6597
	GRASP	0.3863	0.2532	146.34	**4.1324**	**66.39**	**0.5484**	0.6874

PASD (Yang et al., 2024), DiffBIR (Lin et al., 2024), OSEDiff (Wu et al., 2024a), FaithDiff (Chen et al., 2025b), PiSA-SR (Sun et al., 2025), DiT4SR (Duan et al., 2025), and FluxControlNet. We use official checkpoints and configurations whenever available, and re-evaluate all methods under the unified pyiqa backend for fair comparison.

4.2 Comparison with existing state-of-the-art methods

4.2.1 Quantitative analysis

Table 1 reports representative paired-benchmark comparisons, while the representative no-reference comparison on RealLR200 and RealLQ250 is provided in Appendix D (Table 5). Full comparisons with all evaluated baselines are provided in Appendix D (Tables 4 and 6). As shown in Table 1, GRASP shows different strengths with and without agent-driven prompt refinement. GRASP w/o Agent Ref. achieves the lowest FID on all three paired benchmarks and the best LPIPS and DISTS results on DIV2K-Val. Relative to this static-prompt baseline, the full GRASP pipeline improves all four no-reference perceptual metrics on all three paired benchmarks. Across RealSR, DIV2K-Val, and DRealSR, these results demonstrate strong visual realism and perceptual quality.

On RealSR and DRealSR, GRASP does not lead on the two full-reference metrics, LPIPS and DISTS. Visual inspection suggests that some HR targets in these benchmarks contain relatively blurry fine details, consistent with their lower scores on all four no-reference quality metrics than those of DIV2K-Val (Appendix E.2, Table 8). In such cases, the sharper and more visually plausible details synthesized by GRASP may differ from the target textures and therefore receive less favorable LPIPS and DISTS scores, even when the outputs appear clearer and more natural. This behavior is also

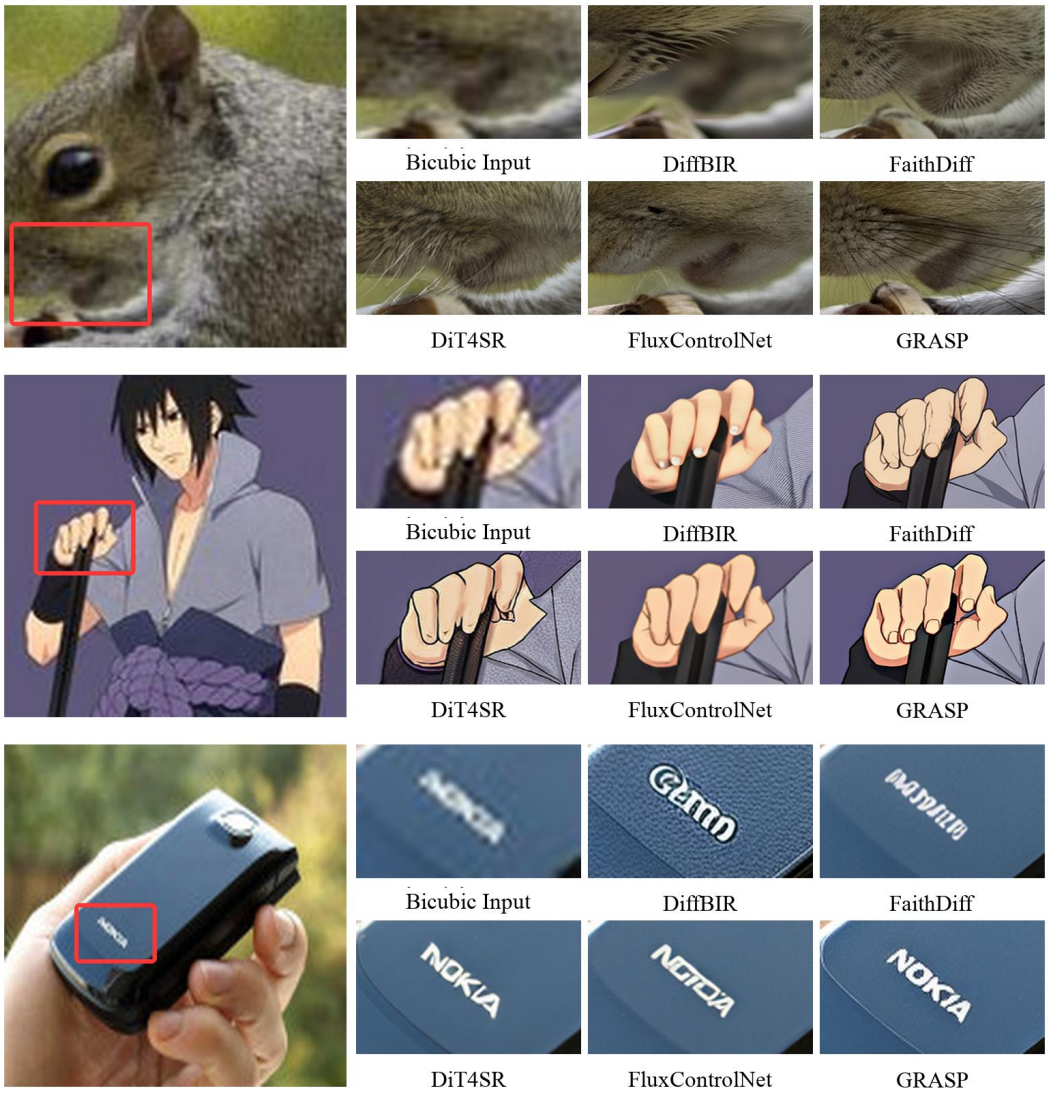

Figure 4: Qualitative comparison between GRASP and representative Real-ISR methods on challenging low-quality images. Red boxes indicate enlarged regions. Existing methods often oversmooth textures or synthesize sharp but spatially inconsistent details under severe degradations. GRASP coordinates local visual evidence, global image context, and agent-refined semantic prompts, restoring realistic fine details while preserving input structures and reducing hallucinated or misaligned content.

consistent with agent-driven prompt refinement: more specific and realistic synthesized details can improve no-reference perceptual scores, but these details may not exactly match the HR targets.

The no-reference results in Appendix D (Table 5) further support the generalization ability of GRASP on real images without reference targets. On RealLR200, GRASP achieves the best MANIQA and CLIPIQA scores and ranks second on LIQE and MUSIQ. On RealLQ250, GRASP achieves the best MUSIQ, MANIQA, and CLIPIQA scores, while ranking second on LIQE. Overall, these results demonstrate strong perceptual performance on the two additional real-world datasets.

4.2.2 Qualitative Analysis

Qualitative comparison. Figure 4 illustrates the challenge of balancing perceptual sharpness with structural fidelity under severe degradations. Some baselines oversmooth fine textures, while others synthesize sharper but misaligned details, such as distorted animal textures, deformed hand contours, or incorrect logo/text patterns. In contrast, GRASP better follows the LQ structure while restoring

Table 2: Condition-organization ablation on the RealLR200 benchmark. All variants use the same static LLaVA caption to isolate the effect of visual condition organization; A2 adapts the DiT4SR-style token injection strategy to Flux Fill.

		Condition organization				RealLR200			
Variant	Name	Local evidence	Global context	VCB	Token inj.	MANIQA↑	MUSIQ↑	CLIPIQA↑	LIQE↑
A1	Local-only GRASP	✓	×	×	×	0.5077	68.1042	0.6542	3.9563
A2	DiT4SR-style token injection	×	×	×	✓	0.4067	61.1938	0.5678	3.1952
A3	Global-only GRASP	×	✓	✓	×	0.5299	**70.8224**	0.6978	**4.5649**
A4	GRASP w/o VCB	✓	✓	×	×	0.4944	61.1463	0.6330	1.5596
A5	Full visual GRASP	✓	✓	✓	×	**0.5651**	69.5311	**0.7090**	4.2210

realistic local details in these examples. More qualitative comparisons on diverse real-world inputs and enlarged local regions are provided in Appendix F.

Human preference study. To complement the qualitative comparisons, we conducted a blinded human preference study with 50 participants on 10 anonymized comparison panels, each comparing GRASP with six representative Real-ISR methods. GRASP was selected as the most perceptually natural output in 202 of 500 participant–panel judgments (40.4%) and as best preserving visible structure and content in 174 of 500 participant–panel judgments (34.8%). These results provide complementary human evidence for perceptual naturalness and the preservation of visible structure and content on the evaluated images. Detailed study materials are provided in Appendix F.

4.3 Ablation Studies and Analysis

Beyond the main comparisons, we analyze GRASP from four aspects: visual condition organization, agent-driven prompt refinement, text supervision, and parameter efficiency.

Role of visual condition organization. Table 2 shows that the local-only variant already forms a meaningful SR baseline, while the DiT4SR-style token injection variant performs substantially worse under the evaluated FLUX Fill configuration. The global-only variant obtains favorable no-reference scores but may drift from the LQ structure without local visual evidence, as visualized in Appendix E (Figure 9). The performance drop when removing VCB further supports the effectiveness of adapting global image context through the VCB. The full visual design improves all four metrics over the local-only variant and achieves the highest MANIQA and CLIPIQA scores, supporting the complementary use of local visual evidence and global image context.

Effect of agent-driven prompt refinement. Appendix E.3 evaluates agent-driven prompt refinement on RealLR200 using raw round-wise outputs and Best-of-K selected outputs. Later raw outputs eventually outperform the static LLaVA baseline on all four no-reference metrics, supporting the effectiveness of inference-time prompt refinement without post-selection or model parameter updates. The Best-of-K results show an overall upward trend as K increases from 1 to 10, reflecting the combined effect of prompt refinement and candidate selection. To complement these metric-based evaluations, a blinded human fidelity audit assesses unsupported changes in the static, raw round-10, and Best-of-10 outputs.

Effect of text supervision. Text supervision mainly changes the fidelity–perception tendency of GRASP. Appendix E.1 compares DAPE tags and LLaVA-generated image descriptions. DAPE tags provide weaker semantic guidance and favor ground-truth (GT)-aligned, fidelity-oriented SR, whereas LLaVA captions provide richer scene context and favor perceptual SR under no-reference metrics. We use LLaVA captions for the main GRASP model.

Parameter efficiency. Appendix C (Table 3) shows that GRASP uses only approximately 122M trainable adaptation parameters, fewer than the listed generative Real-ISR baselines, while still achieving strong perceptual performance.

5 Conclusion

This paper presents GRASP, a visual in-context framework for real-world image super-resolution (Real-ISR) that assigns complementary roles to local visual evidence, global image context, and semantic prompts. Limitations and future work are discussed in Appendix G.

AI USE STATEMENT

This work uses generative AI tools as part of the proposed method: LLaVA generates the initial semantic captions, and GPT-5.4 performs inference-time prompt refinement as described in Sections 3.4 and 3.5. We also used generative AI tools to assist with manuscript language editing and organization. All AI-assisted outputs, experimental procedures, quantitative results, and final claims were reviewed by the authors. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

The human preference study was voluntary, anonymous, and minimal-risk. Participants provided consent before proceeding; no names, contact information, or other direct identifiers were collected. The study asked only for perceptual judgments of anonymized super-resolution outputs.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 126–135, 2017.

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

Black Forest Labs. FLUX.1 Fill [dev]: Inpainting and outpainting with rectified flow transformers. `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev`, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Muller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3086–3095, 2019.

Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. `https://github.com/chaofengc/IQA-PyTorch`, 2022.

Chen Chen, Majid Abdolshah, Violetta Shevchenko, Hongdong Li, Chang Xu, and Pulak Purkait. SRSR: Enhancing semantic accuracy in real-world image super-resolution with spatially re-focused text-conditioning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. FaithDiff: Unleashing diffusion priors for faithful image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28188–28197, 2025b.

Zheng Chen, Yulun Zhang, Jinjin Gu, Xin Yuan, Linghe Kong, Guihai Chen, and Xiaokang Yang. Image super-resolution with text prompt diffusion, 2023.

Marcos V. Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration. In *Proceedings of the European Conference on Computer Vision Workshops*, pp. 669–687, 2022. doi: 10.1007/978-3-031-25063-7_42.

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

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307, 2016.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12606–12633, 2024.

Yushun Fang, Yuxiang Chen, Shibo Yin, Qiang Hu, Jiangchao Yao, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. One-step diffusion transformer for controllable real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23440–23450, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fang_One-Step_Diffusion_Transformer_for_Controllable_Real-World_Image_Super-Resolution_CVPR_2026_paper.html.

Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3599–3608, 2019.

Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S. Ren, and Chao Dong. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *Proceedings of the European Conference on Computer Vision*, pp. 633–651, 2020.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Qiming Hu, Linlong Fan, Yiyan Luo, Yuhang Yu, Xiaojie Guo, and Qingnan Fan. Text-aware real-world image super-resolution via diffusion model with joint segmentation decoders. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context LoRA for diffusion transformers, 2024.

Aiwen Jiang, Zhi Wei, Long Peng, Feiqiang Liu, Wenbo Li, and Mingwen Wang. DaLPSR: Leverage degradation-aligned language prompt for real-world image super-resolution, 2024.

Lei Jiang, Xin Liu, Xinze Tong, Zhiliang Li, Jie Liu, Jie Tang, and Gangshan Wu. Disentangled textual priors for diffusion-based image super-resolution, 2026.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

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

Bryan Sangwoo Kim, Jeongsol Kim, and Jong Chul Ye. Chain-of-zoom: Extreme super-resolution via scale autoregression and preference alignment. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12809–12819, 2025.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017.

Yaowei Li, Xiaoyu Li, Zhaoyang Zhang, Yuxuan Bian, Gan Liu, Xinyuan Li, Jiale Xu, Wenbo Hu, Yating Liu, Lingen Li, Jing Cai, Yuexian Zou, Yancheng He, and Ying Shan. IC-Custom: Diverse image customization via in-context learning. In *International Conference on Learning Representations*, 2026.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. LSDIR: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1775–1787, 2023.

Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. VisualCloze: A universal image generation framework via visual in-context learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18969–18979, 2025.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144, 2017.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Proceedings of the European Conference on Computer Vision*, pp. 430–448, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 291–300, 2020.

Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Mehdi S. M. Sajjadi, Bernhard Scholkopf, and Michael Hirsch. EnhanceNet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4501–4510, 2017.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7986–7996, 2025.

Wensong Song, Hong Jiang, Zongxing Yang, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in DiT, 2025.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-LoRA approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2333–2343, 2025.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

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

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

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 606–615, 2018a.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops*, pp. 63–79, 2018b. doi: 10.1007/978-3-030-11021-5_5.

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

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

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pp. 101–117, 2020.

Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13385–13394, 2021.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

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

Yu Xu, Fan Tang, You Wu, Lin Gao, Oliver Deussen, Hongbin Yan, Jintao Li, Juan Cao, and Tong-Yee Lee. In-context brush: Zero-shot customized subject insertion with context-aware latent space manipulation, 2025.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Yabin Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *Proceedings of the European Conference on Computer Vision*, pp. 74–91, 2024.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Yuhui Wu, and Lei Zhang. Fine-structure preserved real-world image super-resolution via transfer VAE training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12415–12426, 2025.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25669–25680, 2024.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4791–4800, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018a.

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

Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. RankSRGAN: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3096–3105, 2019.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*, pp. 286–301, 2018b.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

A Supplementary Material Overview

This appendix is organized by function rather than by the order in which materials were added. Appendix B expands the motivation for condition organization in generative Real-ISR. Appendix C details the implementation, evaluation protocol, and parameter efficiency. Appendix D reports full quantitative comparisons. Appendix E provides additional ablations and analysis. Appendix F presents additional qualitative examples. Appendices G and H discuss limitations, future work, and broader impacts.

B Additional Discussion on Condition Organization

B.1 Condition-injection paradigms

Figure 1 summarizes three representative conditioning paradigms according to how LQ information is incorporated into the generative backbone. Latent-level conditioning incorporates LQ-derived representations into the latent generation process (Wu et al., 2024a; Chen et al., 2025b; Kong et al., 2025). Auxiliary visual-control branches inject extracted visual features through additional control pathways (Wang et al., 2024; Lin et al., 2024; Yang et al., 2024; Wu et al., 2024b; Ai et al., 2024). Joint token-level multimodal conditioning enables interactions between visual and text tokens (Duan et al., 2025; Fang et al., 2026). These paradigms characterize how conditions are injected, but the injection interface alone does not ensure effective coordination among cues with different roles in the SR process. GRASP explicitly organizes these roles: local visual evidence anchors observable structures and guides local detail synthesis, global image context stabilizes scene-level layout, and semantic prompts guide plausible detail synthesis in ambiguous regions.

B.2 DiT-based condition organization

Many earlier diffusion-based Real-ISR pipelines are built on U-Net-based latent diffusion architectures (Rombach et al., 2022; Wang et al., 2024; Lin et al., 2024; Yang et al., 2024; Wu et al., 2024b). Since these models generate in the compressed latent space of pre-trained autoencoders (Rombach et al., 2022), limited information bandwidth and the reconstruction ceiling may weaken fine-grained structures and textures (Yi et al., 2025). DiT-based generative models provide a different interface: heterogeneous conditions can interact at the token level through attention mechanisms (Peebles & Xie, 2023; Esser et al., 2024). This makes DiT a natural backbone for organizing local visual evidence, global image context, and semantic prompts as distinct condition streams. DiT4SR (Duan et al., 2025) and ODTSR (Fang et al., 2026) adopt joint token-level multimodal conditioning with distinct streams. Our design focuses on assigning complementary roles to local visual evidence, global image context, and semantic prompts.

B.3 Semantic prompting under severe degradations

Complex real-world degradations can make semantic interpretation from the LQ input uncertain. When local structures are severely blurred, corrupted, or missing, the low-quality input may provide insufficient evidence for recognizing scene semantics, object categories, and fine-grained visual attributes. Recent prompt-based SR methods introduce text or semantic prompts to complement such missing information (Chen et al., 2023; Yang et al., 2024; Wu et al., 2024b; Jiang et al., 2024; Chen et al., 2025a; Hu et al., 2025; Kim et al., 2025). However, semantic prompts that primarily summarize observable content may offer limited explicit guidance for synthesizing plausible fine details not directly observable in the LQ input. In GRASP, semantic prompts provide plausible detail descriptions to guide synthesis in ambiguous regions, complementing local visual evidence and global image context rather than replacing them. An inference-time agent iteratively refines these prompts while keeping the visual conditions fixed.

Table 3: Comparison of trainable parameters among representative generative Real-ISR methods.

Method	StableSR	SeeSR	SUPIR	PASD	DiffBIR	FaithDiff	DiT4SR	**GRASP**
Params↓	149.91M	749.9M	1.3B	609.52M	364M	2.61B	350.88M	**122M**

C Implementation and Evaluation Protocol

C.1 Evaluation protocol

All full-reference benchmarks follow the $4\times$ SR protocol and use the LR/HR pairs preprocessed by StableSR (Wang et al., 2024). For DIV2K, RealSR, and DRealSR, full-reference perceptual metrics are computed against the provided HR targets, while no-reference metrics are computed directly on SR outputs. RealLR200 and RealLQ250 do not provide reference HR targets, so we report only no-reference image quality metrics on these two datasets.

C.2 Training and inference settings

Section 4.1 summarizes the main training and inference hyperparameters. Here we emphasize that the same configuration is used for all main benchmarks and ablation variants unless explicitly stated. We do not tune the sampling steps, guidance scale, or training schedule separately for individual datasets, which keeps the paired and no-reference evaluations controlled under a unified protocol.

The training configuration uses the Prodigy optimizer with a learning rate of 1.0, bias correction, safeguard warmup, and a weight decay of 0.01. We use bfloat16 mixed precision, a per-process batch size of 16, one gradient-accumulation step, random seed 231, and eight data-loading workers. Training uses 512×512 HR crops for 15K iterations and an offline paired training manifest containing 949,910 rows. Text conditions are read from pre-generated prompt files.

C.3 LoRA adaptation and Vision-Context Bridge

GRASP adapts FLUX.1-Fill-dev with rank-32 LoRA adapters while keeping the base generative backbone frozen. The adapters are applied to conditioning-, attention-, and multilayer perceptron (MLP)-related projection layers in the FLUX transformer. The Vision-Context Bridge maps CLIP ViT-L/14 visual features into the FLUX conditioning space and is trained jointly with the LoRA adapters. Exact module patterns and configuration files are provided in the released code.

C.4 Trainable parameter comparison

In addition to SR quality, we compare the number of trainable parameters to evaluate trainable-parameter efficiency. Table 3 shows that GRASP uses only approximately 122M trainable adaptation parameters, which is the smallest among the listed generative Real-ISR baselines. Despite this compact adaptation, GRASP still delivers strong perceptual performance in the quantitative comparisons of Section 4.2. This result suggests that high-quality Real-ISR does not necessarily require large additional control branches or expensive full-model fine-tuning. Instead, GRASP benefits from a parameter-efficient conditioning design: its local visual evidence is provided directly through the diptych, avoiding a separate encoder for local LQ features, while the lightweight Vision-Context Bridge together with LoRA adapts global image context into the DiT conditioning space with few trainable parameters. Overall, GRASP achieves a favorable balance between Real-ISR quality and trainable adaptation cost.

For the complete $K = 10$ configuration, GRASP generates ten FLUX outputs and performs nine GPT-5.4 refinement calls per image. On an NVIDIA H800 GPU, processing 50 RealSR images at 512×512, the full pipeline takes 750.7 s per image and reaches a peak memory usage of 50,869 MiB, excluding model loading. The system loads 24.674B local parameters, and the recorded token-based API cost is estimated at \$0.4672 per image.

Table 4: Full quantitative comparison on RealSR, DIV2K-Val, and DRealSR under the unified pyiqa backend. Red and blue denote the best and second-best results, respectively. GRASP denotes the full inference pipeline with agent-driven prompt refinement. GRASP w/o Agent Ref. uses the initial LLaVA prompt and a single output.

Dataset	Method	Metric						
		LPIPS↓	DISTS↓	FID↓	LIQE↑	MUSIQ↑	MANIQA↑	CLIPIQA↑
RealSR	StableSR	0.3272	0.2454	134.89	3.5055	61.47	0.4634	0.5653
	SeeSR	0.3007	0.2224	125.51	4.1365	69.82	0.5437	0.6704
	SUPIR	0.4757	0.2657	136.05	3.2000	60.95	0.4319	0.5714
	PASD	**0.2859**	**0.2085**	127.95	2.8736	58.20	0.3767	0.4479
	DiffBIR	0.4012	0.2603	151.33	4.2846	67.94	0.4865	0.6776
	OSEDiff	0.3173	0.2363	126.05	3.9402	67.53	0.4853	0.6735
	FaithDiff	0.2887	0.2114	108.89	3.9926	68.85	0.4680	0.6204
	PiSA-SR	0.3130	0.2267	141.92	1.6308	47.13	0.3049	0.3593
	DiT4SR	0.3201	0.2258	120.12	3.9017	67.63	0.4582	0.6294
	FluxControlNet	0.3555	0.2438	126.46	3.6057	64.29	0.4554	0.6394
	GRASP w/o Agent Ref.	0.3323	0.2359	**108.24**	4.2739	69.71	0.6070	0.7007
	GRASP	0.3383	0.2383	110.46	**4.3833**	**70.52**	**0.6222**	**0.7155**
DIV2K-Val	StableSR	0.3456	0.2142	31.04	4.1023	65.07	0.5393	0.7124
	SeeSR	0.3190	0.1963	26.02	4.2642	68.62	0.5030	0.6926
	SUPIR	0.3887	0.2133	26.44	4.0395	65.83	0.4451	0.6455
	PASD	0.3843	0.2287	35.28	3.5030	60.53	0.3891	0.5380
	DiffBIR	0.3741	0.2291	42.09	4.3054	67.47	0.4685	0.6868
	OSEDiff	**0.3046**	0.2129	26.81	3.7510	65.56	0.4425	0.6589
	FaithDiff	0.3124	0.1992	25.88	4.2441	69.32	0.4311	0.6467
	PiSA-SR	0.4891	0.2693	46.02	1.4901	40.50	0.2560	0.3389
	DiT4SR	0.3427	0.2123	36.60	4.1329	67.39	0.4392	0.6602
	FluxControlNet	0.3580	0.2137	26.51	3.7580	62.35	0.4321	0.6164
	GRASP w/o Agent Ref.	0.3109	**0.1955**	**23.82**	4.5183	70.12	0.6103	0.7283
	GRASP	0.3179	0.1989	24.93	**4.6074**	**70.98**	**0.6251**	**0.7431**
DRealSR	StableSR	0.3792	0.2683	169.83	2.9169	53.03	0.3943	0.5328
	SeeSR	0.3174	0.2315	147.46	4.1273	65.09	0.5129	0.6910
	SUPIR	0.5662	0.2931	158.32	3.3557	58.95	0.4436	0.5962
	PASD	0.3224	**0.2273**	164.80	2.4823	49.89	0.3508	0.4704
	DiffBIR	0.5095	0.2959	175.92	3.7407	59.83	0.4508	0.6486
	OSEDiff	0.3177	0.2365	142.18	3.8654	63.55	0.4790	**0.7060**
	FaithDiff	0.3501	0.2416	153.34	4.0124	66.35	0.4512	0.6332
	PiSA-SR	**0.3054**	0.2285	156.64	1.5760	43.53	0.3038	0.3680
	DiT4SR	0.3721	0.2445	156.99	3.9953	65.29	0.4439	0.6660
	FluxControlNet	0.3972	0.2521	156.29	3.1211	56.01	0.4069	0.5959
	GRASP w/o Agent Ref.	0.3662	0.2456	**138.92**	3.9471	64.14	0.5163	0.6597
	GRASP	0.3863	0.2532	146.34	**4.1324**	**66.39**	**0.5484**	0.6874

D Additional Quantitative Results

D.1 Full paired-benchmark comparison

Table 4 reports the full paired-benchmark comparison on DIV2K-Val, RealSR, and DRealSR under the unified pyiqa backend. As the complete version of Table 1, it confirms the same overall trend: GRASP w/o Agent Ref. achieves strong FID and full-reference scores, while the full GRASP pipeline obtains leading results on most no-reference metrics across the three paired benchmarks. The lower LPIPS/DISTS rankings on the two real-world paired benchmarks are consistent with the analysis in Section 4.2.

Table 5: Representative no-reference quantitative comparison on RealLR200 and RealLQ250 under the unified pyiqa backend. Red and blue denote the best and second-best results, respectively.

Method	RealLR200				RealLQ250			
	LIQE↑	MUSIQ↑	MANIQA↑	CLIPIQA↑	LIQE↑	MUSIQ↑	MANIQA↑	CLIPIQA↑
StableSR	3.6087	64.53	0.4610	0.6540	3.5437	64.23	0.4279	0.6647
SeeSR	4.0390	69.60	0.4977	0.6834	3.9787	70.38	0.4895	0.7061
DiffBIR	3.8747	66.95	0.4576	0.6684	3.8728	66.68	0.4406	0.6703
FaithDiff	3.9225	68.99	0.4285	0.6592	3.8442	69.70	0.3984	0.6652
DiT4SR	4.1422	69.17	0.4552	0.6896	4.2885	71.06	0.4593	0.7120
FluxControlNet	**4.4060**	**71.32**	0.5186	0.7038	**4.3680**	70.58	0.5085	0.7081
GRASP	4.2898	70.2334	**0.5770**	**0.7251**	4.3016	**71.0707**	**0.5449**	**0.7268**

Table 6: Full no-reference quantitative comparison on RealLR200 and RealLQ250 under the unified pyiqa backend. Red and blue denote the best and second-best results, respectively.

Method	RealLR200				RealLQ250			
	LIQE↑	MUSIQ↑	MANIQA↑	CLIPIQA↑	LIQE↑	MUSIQ↑	MANIQA↑	CLIPIQA↑
StableSR	3.6087	64.53	0.4610	0.6540	3.5437	64.23	0.4279	0.6647
SeeSR	4.0390	69.60	0.4977	0.6834	3.9787	70.38	0.4895	0.7061
SUPIR	3.5202	64.15	0.3992	0.5573	3.6044	65.68	0.3825	0.5750
PASD	3.1836	61.95	0.3767	0.5536	3.0007	60.62	0.3516	0.5397
DiffBIR	3.8747	66.95	0.4576	0.6684	3.8728	66.68	0.4406	0.6703
OSEDiff	3.7757	68.05	0.4452	0.6889	3.6201	67.59	0.4307	0.6911
FaithDiff	3.9225	68.99	0.4285	0.6592	3.8442	69.70	0.3984	0.6652
PiSA-SR	1.7138	47.86	0.2873	0.4444	1.3716	45.90	0.2583	0.3805
DiT4SR	4.1422	69.17	0.4552	0.6896	4.2885	71.06	0.4593	0.7120
FluxControlNet	**4.4060**	**71.32**	0.5186	0.7038	**4.3680**	70.58	0.5085	0.7081
GRASP	4.2898	70.2334	**0.5770**	**0.7251**	4.3016	**71.0707**	**0.5449**	**0.7268**

D.2 Representative no-reference comparison

Table 5 reports the representative no-reference comparison on RealLR200 and RealLQ250 discussed in the main paper. GRASP obtains the best MANIQA and CLIPIQA scores on RealLR200 and the best MUSIQ, MANIQA, and CLIPIQA scores on RealLQ250, showing strong perceptual quality on real images without reference targets.

D.3 Full no-reference comparison

Table 6 reports the full no-reference comparison on RealLR200 and RealLQ250. Including all evaluated baselines does not change the main conclusion: GRASP obtains the best MANIQA and CLIPIQA scores on RealLR200 and the best MUSIQ, MANIQA, and CLIPIQA scores on RealLQ250, while ranking second on LIQE on both datasets.

E Additional Ablations and Analysis

E.1 Text supervision ablation

In this paper, we examine two types of text supervision sources: one is DAPE tags following SeeSR (Wu et al., 2024b), and the other is natural language descriptions generated by LLaVA (Liu et al., 2023) for the input low-quality images. Although both are used as text conditions for the generative model, they provide substantially different amounts of semantic information, which in turn affects how the model balances textual and visual conditions during SR generation and ultimately leads to different SR tendencies.

Table 7: Ablation of text supervision on RealSR, DIV2K-Val, and DRealSR. Red and blue denote the best and second-best results, respectively.

Dataset	Method	Metric								
		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	LIQE↑	MUSIQ↑	MANIQA↑	CLIPIQA↑
RealSR	GRASP-Faithful	**24.30**	**0.7213**	**0.2577**	**0.2025**	**101.48**	3.4044	62.78	0.4365	0.5046
	GRASP-Perceptual	22.41	0.6397	0.3323	0.2359	108.24	4.2739	69.71	0.6070	0.7007
	GRASP	22.35	0.6376	0.3383	0.2383	110.46	**4.3833**	**70.52**	**0.6222**	**0.7155**
DIV2K-Val	GRASP-Faithful	**22.75**	**0.5748**	**0.2837**	**0.1827**	**21.39**	4.0781	65.70	0.5089	0.6134
	GRASP-Perceptual	21.75	0.5315	0.3109	0.1955	23.82	4.5183	70.12	0.6103	0.7283
	GRASP	21.60	0.5274	0.3179	0.1989	24.93	**4.6074**	**70.98**	**0.6251**	**0.7431**
DRealSR	GRASP-Faithful	**27.04**	**0.7463**	**0.2999**	**0.2148**	**129.19**	3.0211	56.24	0.3698	0.4629
	GRASP-Perceptual	25.20	0.6724	0.3662	0.2456	138.92	3.9471	64.14	0.5163	0.6597
	GRASP	24.72	0.6488	0.3863	0.2532	146.34	**4.1324**	**66.39**	**0.5484**	**0.6874**

Table 8: No-reference quality scores of HR targets on the paired evaluation benchmarks. Higher values indicate better perceptual quality.

Dataset	LIQE↑	MUSIQ↑	MANIQA↑	CLIPIQA↑
RealSR	2.6720	57.46	0.3295	0.4591
DIV2K-Val	3.8829	63.76	0.4060	0.6014
DRealSR	2.1154	50.56	0.3152	0.4750

As shown in Table 7, GRASP-Faithful trained with DAPE tags is overall more advantageous on full-reference metrics such as PSNR, SSIM, LPIPS, DISTS, and FID, whereas GRASP-Perceptual trained with LLaVA captions performs better on no-reference metrics such as LIQE, MUSIQ, MANIQA, and CLIPIQA. A key reason is that DAPE tags provide much less semantic information than sentence-level captions. Under such weaker textual conditioning, the model is forced to rely more heavily on the image condition from the input LQ image itself, and the resulting outputs are therefore more likely to remain close to the reference GT. As a result, this setting tends to obtain better full-reference scores, which directly measure similarity to GT. In contrast, LLaVA captions provide richer scene context, object relations, and attribute information, and thus more strongly activate the high-level semantic and natural-image priors in the pretrained generative model. This makes the generated SR results more natural, sharper, and visually appealing, leading to stronger no-reference performance.

E.2 Benchmark-reference quality analysis

The GT scores in Table 8 further reveal an important property of the RealSR and DRealSR benchmarks. On these two datasets, the GT images themselves receive relatively low no-reference scores on LIQE, MUSIQ, MANIQA, and CLIPIQA. Visual inspection further suggests that some GT images in these datasets are themselves relatively blurry and do not fully align with the high-naturalness distribution preferred by no-reference image quality assessment (NR-IQA) models. Therefore, on RealSR and DRealSR, a result that is more similar to GT may naturally obtain lower no-reference scores, while a result that is visually sharper and more natural may deviate from the relatively blurry GT and thus fail to achieve the best full-reference scores. In this sense, the observed fidelity–perception discrepancy is not caused only by the form of text supervision, but is also closely related to the properties of the benchmark GT itself.

This interpretation also explains why the conflict is weaker on DIV2K-Val. Since DIV2K-Val is constructed by synthesizing degradations from high-quality images, its GT images are themselves cleaner and closer to the natural image distribution preferred by generative models and NR-IQA metrics. Under this setting, full-reference and no-reference metrics are less contradictory than on RealSR and DRealSR. As a result, it becomes easier to obtain favorable performance on both types of metrics simultaneously. Overall, these results suggest that DAPE tags are better suited to fidelity-oriented SR because they provide less semantic information and force the model to rely more on the LQ input, whereas LLaVA captions are better suited to perceptual SR because they more strongly activate generative priors toward natural and visually pleasing outputs.

E.3 AGENT PROMPT REFINEMENT ABLATION

For the GPT-5.4 experiments, we use a temperature of 1.0, top-p of 1.0, medium reasoning effort, and a maximum output length of 2,400 tokens. Each request has a 180-second timeout, no request-level seed, and at most one retry after a failed logical request. The full $K = 10$ configuration performs nine logical GPT refinement calls per image.

Agent instruction template. We summarize the fixed GPT-5.4 instruction template used for LLaVA-anchored closed-loop prompt refinement below.

> **Inputs:** LQ image; initial LLaVA caption; previous-round raw SR result and prompt; previous-round LIQE, MUSIQ, MANIQA, and CLIPIQA scores.
> **Initialization:** Use the LLaVA-generated caption as the initial semantic prompt.
> **Refinement:** Conservatively edit the previous prompt while preserving the main scene description supported by the LQ image. Make only one or two targeted cue additions or revisions to provide plausible fine-detail descriptions for ambiguous regions or address artifacts observed in the previous raw SR result.
> **State rule:** Use the immediately preceding raw SR output as the SR image input for the next refinement step. Do not substitute best-of-previous-rounds, selected, oracle, or fallback outputs.
> **Constraints:** Keep detail descriptions consistent with the available visual evidence. Do not invent object identities, readable text, or instance-specific attributes unsupported by the LQ image. Treat unresolved fine details as plausible synthesis guidance rather than confirmed observations.
> **Output:** Return the complete revised semantic prompt, not only the edited phrases.

Example agent refinement. The following example shows that the agent makes a conservative full-prompt edit rather than a free-form rewrite.

> **Initial prompt:** *The image features a tree with no leaves, standing tall against a cloudy sky. The tree appears to be bare, with no foliage visible. The sky is overcast, creating a moody atmosphere in the scene.*
>
> **Previous prompt:** *A cropped view of mostly bare tree branches extends upward against a pale overcast sky. Preserve the existing framing, the separated branch clusters, and the main branch directions, especially the prominent left-leaning branch and the finer rising branches near the center and right. The previous SR result is already clearer, but some twig tips and lower branch clusters still appear slightly too dark and wiry, with occasional duplicate branch strokes, faint halos, crunchy sharpening, and small speckle near the upper branches. Refine gently: keep the sky smooth, cool, and low-contrast; preserve the open gaps between branch groups; and keep the finest twigs slender and naturally tapered rather than dense or uniformly thick.*
>
> **Next-round prompt:** *A cropped view of mostly bare tree branches extends upward against a pale overcast sky. Preserve the current framing, the separated branch clusters, the main branch directions, and the open sky gaps. The current SR result is clearer, but some fine twigs remain a little too black and etched, especially around twig tips and lower branch clusters, and there are still mild duplicate strokes, halos, oversharpening, and faint speckle around upper branches. Refine conservatively: keep the sky smooth, cool, and low-contrast; maintain delicate, naturally tapered twig structure within the existing branch regions; reduce dark wiry outlines and repeated branching strokes; and avoid adding dense foliage, new major branches, or dramatic cloud texture.*

Table 9 and Figure 5 report the adopted static LLaVA initialization and subsequent raw refinement rounds without Best-of-K selection. We also include the agent-initialized output in the table for comparison; it is weaker than the static LLaVA initialization, which motivates using the latter as the adopted starting point. The trend is not strictly monotonic in early rounds, but later raw refinement outputs overtake the adopted initialization on all four no-reference metrics.

Table 10 reports the practical Best-of-K variant starting from the adopted static LLaVA initialization. As K increases, the candidate pool expands, and the aggregate no-reference scores of the selected outputs show an overall upward trend. Figure 6 visualizes this Best-of-K selection trend. These results support the effectiveness of using inference-time visual feedback to refine semantic prompts for Real-ISR. The selection rule is fixed before evaluation and applied automatically to every image; no result is manually selected. For $K = 1$, the selected output is the adopted static LLaVA initialization. For $K \geq 2$, only one output is chosen from the first K generated candidates for each image, and the final metrics are recomputed on this selected output set. Therefore, the Best-of-K results should be read together with the raw round-wise results in Table 9.

Table 9: Raw round-wise ablation of inference-time agent-driven prompt refinement on RealLR200. Static LLaVA initialization is the adopted round-1 input because it performs better than agent initialization, which is reported for comparison. Later rows report raw outputs from subsequent refinement rounds without Best-of-K selection.

Prompt setting	MANIQA↑	MUSIQ↑	CLIPIQA↑	LIQE↑
Static LLaVA initialization (round 1)	0.5651	69.5311	0.7090	4.2210
Agent initialization (not adopted)	0.5646	69.6605	0.7121	4.2012
Agent refinement, round 2	0.5679	69.7336	0.7144	4.2326
Agent refinement, round 3	0.5682	69.7769	0.7152	4.2362
Agent refinement, round 4	0.5684	69.7803	0.7166	4.2383
Agent refinement, round 5	0.5689	69.8128	0.7172	4.2403
Agent refinement, round 6	0.5698	69.8177	0.7172	4.2421
Agent refinement, round 7	0.5706	69.8331	0.7182	4.2436
Agent refinement, round 8	0.5716	69.8828	**0.7194**	4.2490
Agent refinement, round 9	0.5716	69.9195	0.7189	4.2520
Agent refinement, round 10	**0.5726**	**69.9498**	0.7190	**4.2586**

Table 10: Ablation of inference-time agent-driven prompt refinement on RealLR200. Static LLaVA initialization is used as $K = 1$ because it performs better than agent initialization, which is reported for comparison. For $K \geq 2$, Best-of-K selects one output per image using the rank-sum rule over LIQE, MUSIQ, MANIQA, and CLIPIQA, then recomputes aggregate metrics on the selected set.

Prompt setting	MANIQA↑	MUSIQ↑	CLIPIQA↑	LIQE↑
Static LLaVA initialization, $K = 1$	0.5651	69.5311	0.7090	4.2210
Agent initialization (not adopted)	0.5646	69.6605	0.7121	4.2012
Agent refinement, $K = 2$	0.5715	69.9638	0.7178	4.2591
Agent refinement, $K = 3$	0.5727	70.0562	0.7194	4.2629
Agent refinement, $K = 4$	0.5732	70.0989	0.7206	4.2657
Agent refinement, $K = 5$	0.5737	70.1191	0.7213	4.2675
Agent refinement, $K = 6$	0.5743	70.1445	0.7218	4.2728
Agent refinement, $K = 7$	0.5749	70.1604	0.7230	4.2767
Agent refinement, $K = 8$	0.5760	70.2035	0.7245	4.2816
Agent refinement, $K = 9$	0.5766	70.2015	0.7249	4.2857
Agent refinement, $K = 10$	**0.5770**	**70.2334**	**0.7251**	**4.2898**

Open-weight backend check. We additionally run the complete $K = 10$ refinement protocol on RealLR200 with the open-weight Qwen2.5-VL-7B-Instruct model. The SR model, state transition, sampling configuration, seed policy, selector, and evaluation protocol remain unchanged. Qwen Best-of-10 improves the shared static LLaVA candidate on all four aggregate no-reference metrics, showing that the refinement protocol can also be executed with an open-weight VLM backend. The resulting raw-round and Best-of-10 scores are summarized in Table 11.

Figure 7 provides qualitative examples consistent with the quantitative trend.

Human fidelity-audit questionnaire. The agent-specific blind audit presents the LQ input, an HR reference when available, and three anonymized outputs A–C corresponding to the static output, raw round 10, and Best-of-10 under a balanced codebook. Participants rate each output for unsupported changes using five levels: no finding, minor, obvious, severe, or unable to judge. Figure 8 shows a representative severity-rating matrix.

Human audit results. The audit uses 30 fixed images (10 per paired benchmark) and 50 valid participants; each participant rates 10 images, yielding 500 raw judgments per output setting.

Table 11: Open-weight prompt-refinement backend check on RealLR200. The static row uses the canonical RealLR200 baseline adopted throughout the paper; the Qwen rows come from the complete ten-candidate run.

Setting	LIQE↑	MUSIQ↑	MANIQA↑	CLIPIQA↑
Static LLaVA	4.2210	69.5311	0.5651	0.7090
Qwen raw round 10	4.2333	69.7212	0.5680	0.7126
Qwen Best-of-10	**4.2469**	**69.7897**	**0.5695**	**0.7146**

Table 12: Aggregate results of the blinded human fidelity audit on 30 fixed images. Naturalness percentages are computed over 440 non-tie judgments per setting; unsupported-change rates exclude unable-to-judge responses.

Setting	Naturalness (%)	Unsupported changes (%)	Marked-or-severe (%)
Static	23.0	12.1 (58/480)	3.8
Raw round 10	31.1	12.4 (60/482)	4.1
Best-of-10	45.9	11.0 (53/481)	3.3

Table 12 summarizes the naturalness preferences and unsupported-change rates for the static, raw round-10, and Best-of-10 outputs. Relative to Static, the Raw-minus-Static and Best-minus-Static unsupported-change-rate differences are $+0.4$ percentage points (95% CI $[-3.8, +4.6]$) and -1.1 points (95% CI $[-5.1, +2.9]$), respectively; both upper bounds are below the prespecified $+5$-point non-inferiority margin. Ordinal Krippendorff's α is 0.71, and no evident A/B/C position effect is observed.

E.4 Qualitative analysis of the global-only variant

Figure 9 further illustrates why favorable no-reference scores from global context alone should not be interpreted as faithful SR behavior.

F Additional Qualitative Results

F.1 Human preference questionnaire

The cross-method blind questionnaire presents the LQ input and seven anonymized outputs A–G, with output positions balanced across panels. Participants select the output with the highest perceptual naturalness and the output that best preserves visible structure and content; paired RealSR and DRealSR panels additionally ask about unsupported changes. Figure 10 shows a representative naturalness question. On the seven paired RealSR and DRealSR panels, unsupported changes were flagged in 37 of 350 participant–panel judgments (10.6%).

F.2 Teaser comparison

Figure 11 provides an additional teaser-style qualitative comparison.

F.3 Diverse real-world examples

Figure 12 shows additional GRASP outputs on diverse low-quality inputs.

F.4 Additional zoom-in comparisons

Figure 13 provides additional zoom-in comparisons that complement the main qualitative figure. Across architectural details, animal textures, fingers, faces, printed chip characters, and object parts,

GRASP more consistently restores recognizable local structures, while several baselines oversmooth details or synthesize sharper but misaligned patterns.

G Limitations and Future Work

GRASP is currently limited by its fixed-resolution visual in-context completion formulation. In our implementation, the model is trained with a diptych composed of two 512×512 images, where the left panel provides the upsampled LQ visible context and the right panel defines the SR target region to be completed. As a result, the model mainly learns context-to-target correspondence under a fixed spatial scale and a fixed left-right layout. For real images with arbitrary resolutions or non-standard aspect ratios, inference therefore requires resizing, cropping, or tiled processing, which may alter the original spatial structure or reduce global context modeling across image regions.

When an input image is divided into multiple tiles, global context can be fragmented across tiles, leading to structural misalignment, texture discontinuities, or low-frequency color inconsistency near tile boundaries. This issue can be more pronounced for large objects, long-range repetitive textures, and semantic regions that extend across several tiles, because each local tile may be insufficient to recover the complete scene relationship independently.

Future work will explore more flexible arbitrary-resolution visual in-context completion mechanisms for Real-ISR, such as multi-scale diptych construction, overlapping tiled inference, and consistency constraints across image tiles. In addition, inspired by spatially focused prompting strategies such as SRSR (Chen et al., 2025a), we plan to investigate region-aware textual conditioning in DiT-based Real-ISR. This direction may clarify how localized text guidance interacts with regional DiT tokens, how it should be aligned with visual evidence, and whether it can alleviate the context fragmentation caused by fixed-window or tiled inference.

H Broader Impacts

GRASP may benefit applications that require improving the perceptual quality and readability of degraded images, such as image archival, content creation, and visual inspection under low-quality capture conditions. However, as a generative Real-ISR method, GRASP may synthesize plausible but not necessarily authentic details in severely degraded regions. Such outputs should therefore not be used as evidence in high-stakes settings such as forensics, medical diagnosis, identity verification, or surveillance without task-specific validation and clear disclosure that the images have been enhanced. These risks motivate cautious deployment and transparent communication of the generative nature and limitations of SR outputs.

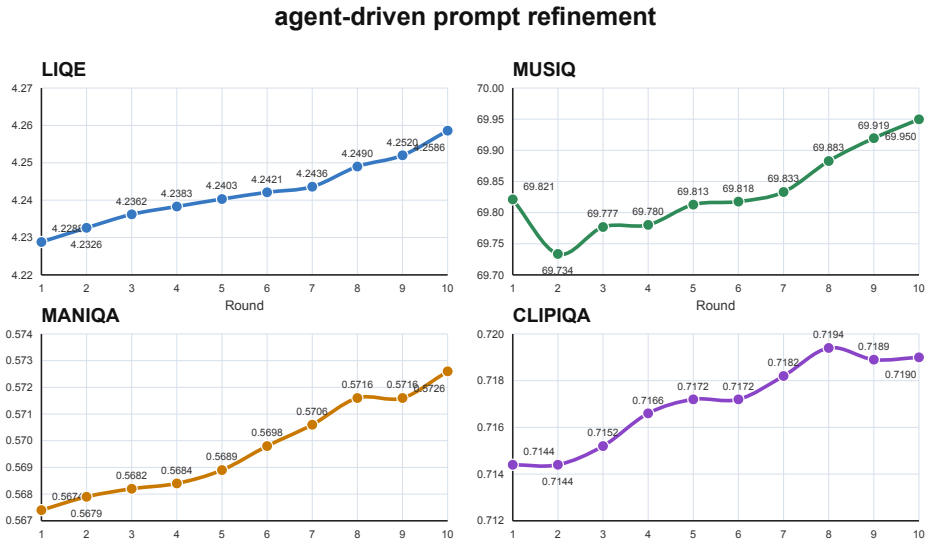

Figure 5: No-reference metric trends for raw outputs at each refinement round on RealLR200. Round 1 is the adopted static LLaVA initialization; no Best-of-K selection is applied to later raw outputs.

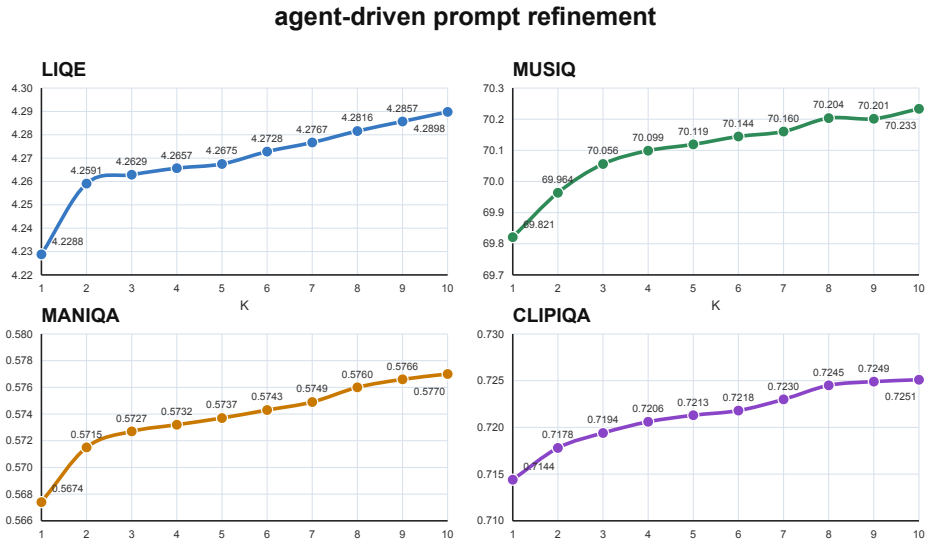

Figure 6: No-reference metric trends of the Best-of-K selected outputs on RealLR200. $K = 1$ is the adopted static LLaVA initialization; for $K \geq 2$, one output is selected per image from the first K generated candidates using the rank-sum rule over LIQE, MUSIQ, MANIQA, and CLIPIQA.

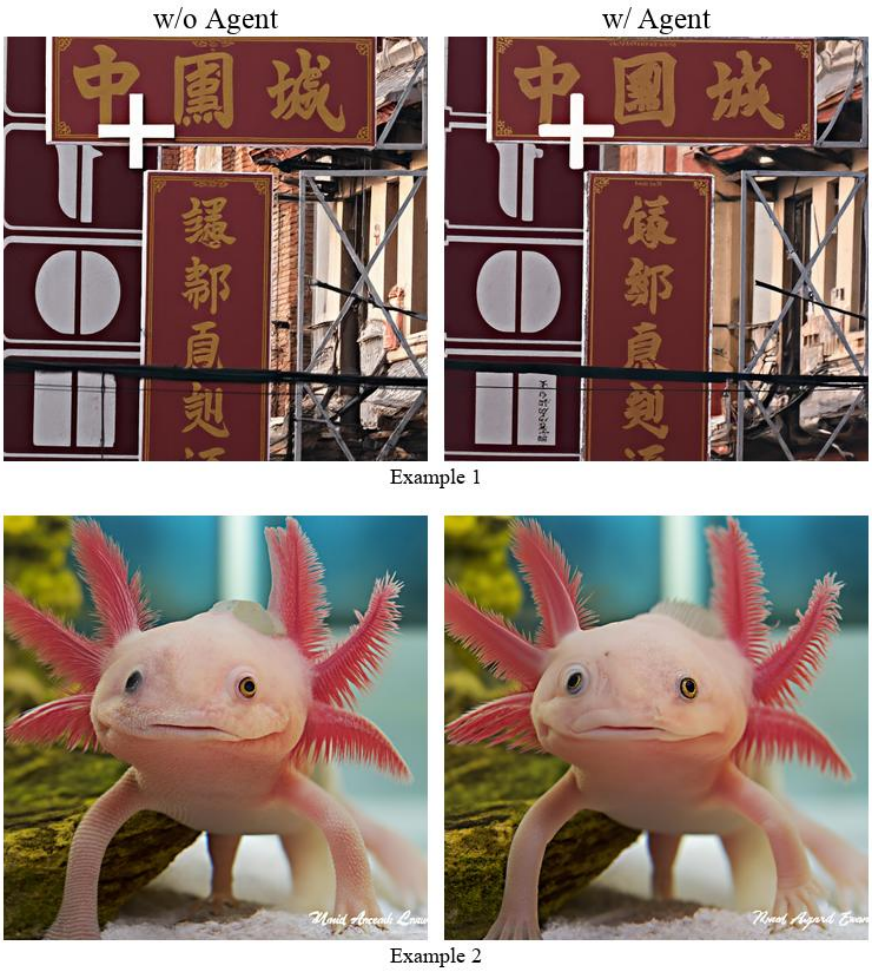

Figure 7: Qualitative examples showing the effect of agent-refined prompting. Each row compares the output without agent refinement and the output with agent refinement. The agent-refined setting produces more coherent semantic details and reduces locally inconsistent structures, supporting the effectiveness of inference-time visual feedback for prompt refinement.

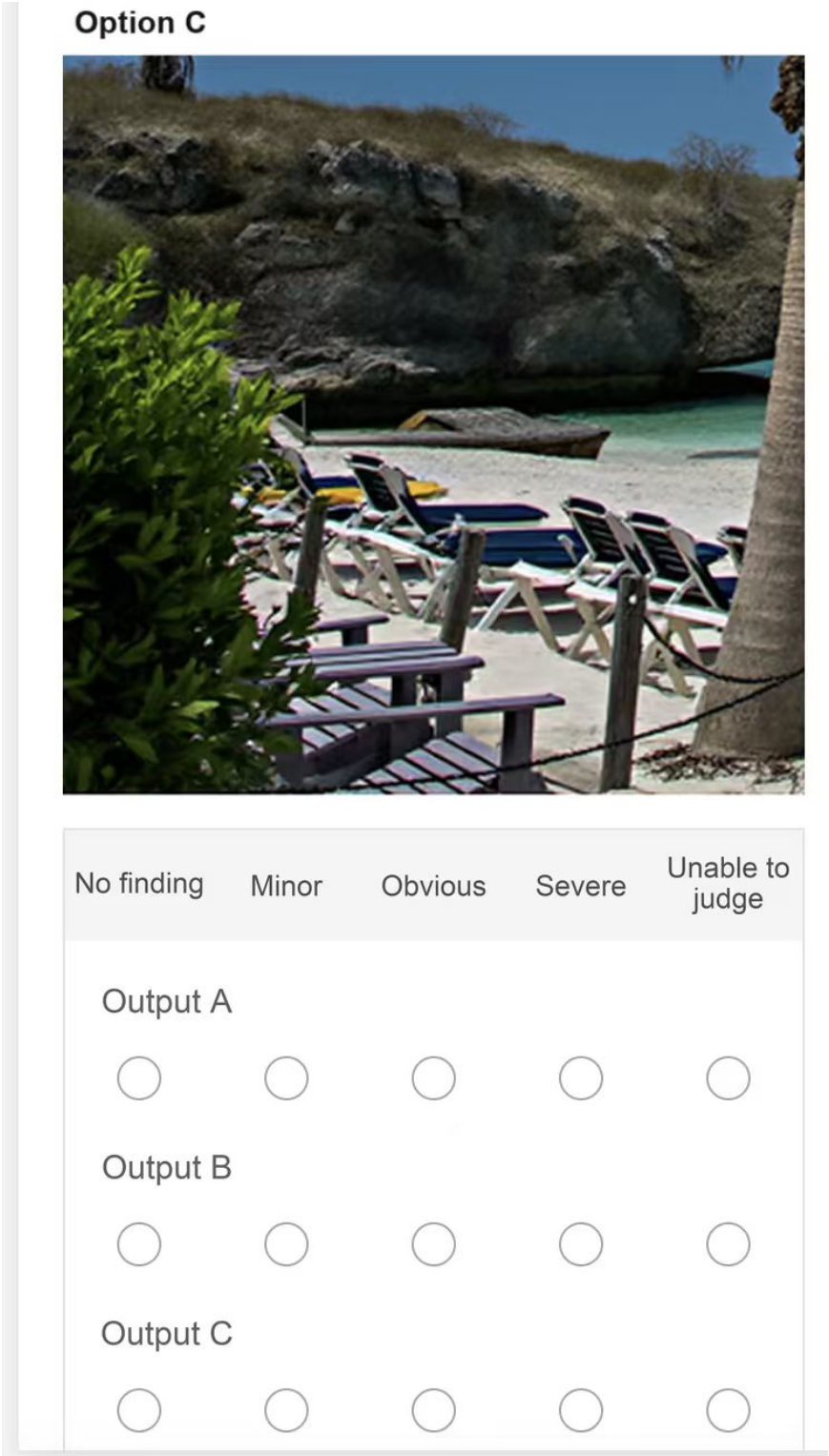

Figure 8: Representative severity-rating matrix from the agent-specific fidelity audit. Participants independently rate the three anonymized outputs for unsupported changes using five severity levels.

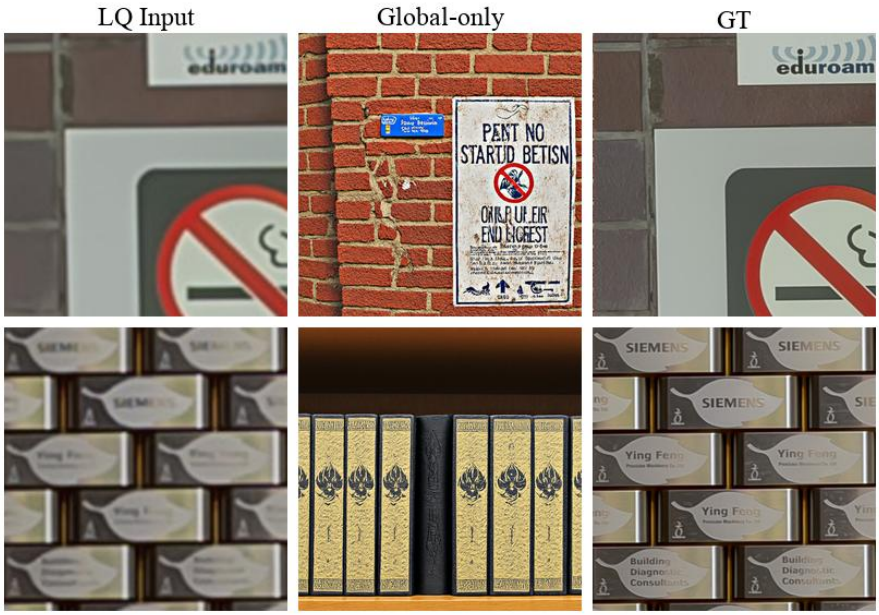

Figure 9: Qualitative examples of the global-only variant. Each row shows the LQ input, the global-only output, and the GT image. Although the global-only variant can produce perceptually plausible images and obtain favorable no-reference scores, it is weakly constrained by the input condition and may synthesize content that is inconsistent with the LQ image and GT.

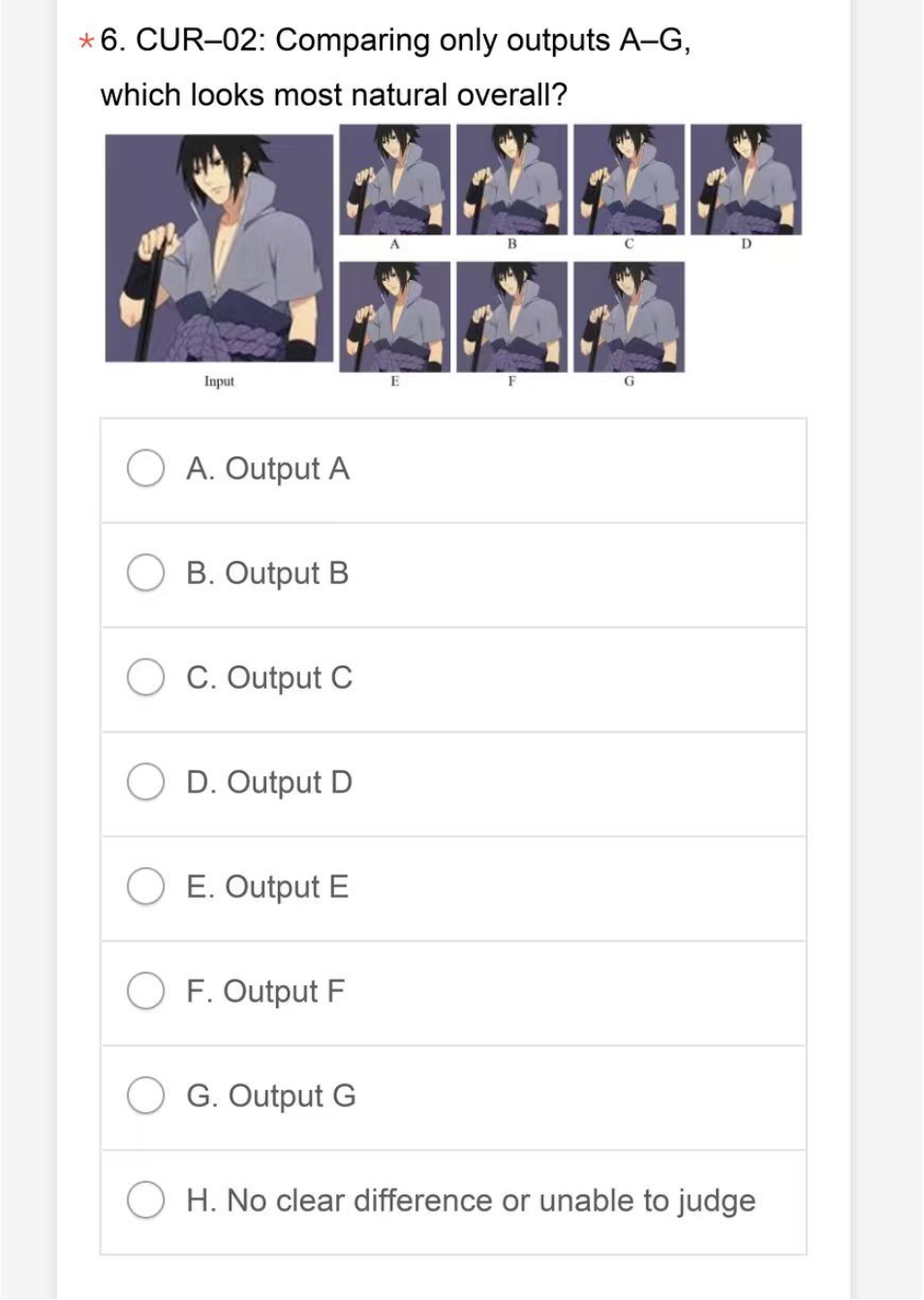

Figure 10: Representative naturalness question from the blinded seven-method human preference study. The LQ input and seven anonymized outputs A–G are shown, with an additional option for no clear difference or unable to judge.

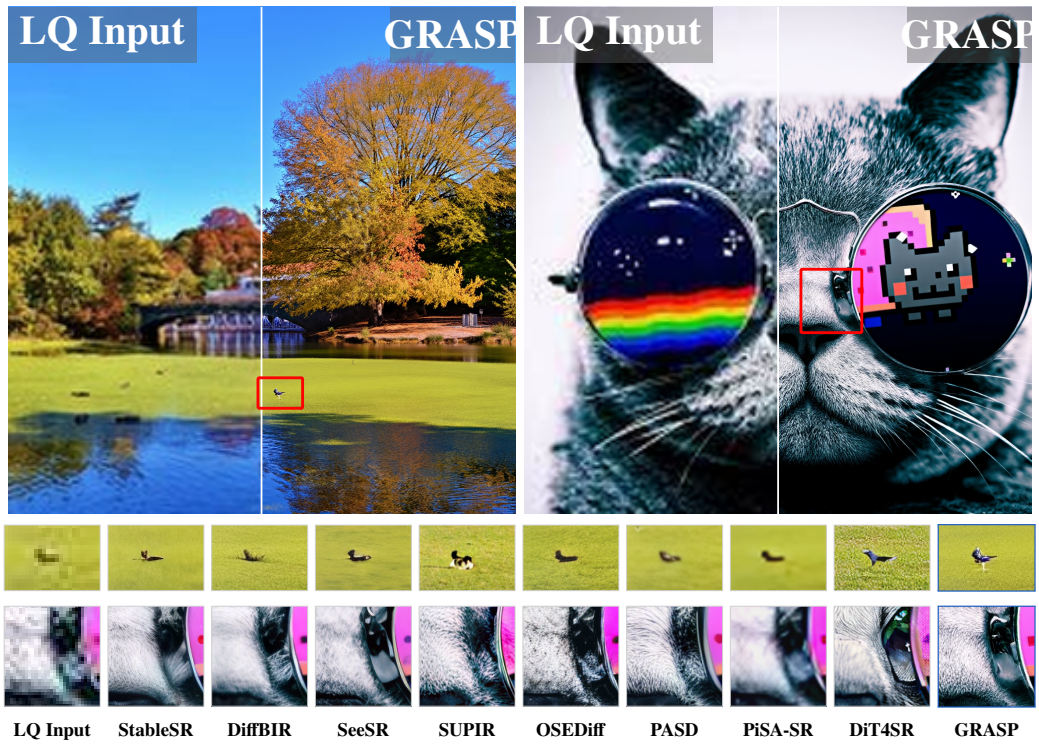

Figure 11: Performance comparison between GRASP and other state-of-the-art Real-ISR methods on two challenging real-world LQ images. With a DiT-based generative prior, structured visual conditioning, and semantic prompts refined at inference time by the agent for these examples, GRASP restores realistic fine details while remaining faithful to the input, avoiding hallucinated or spatially inconsistent content.

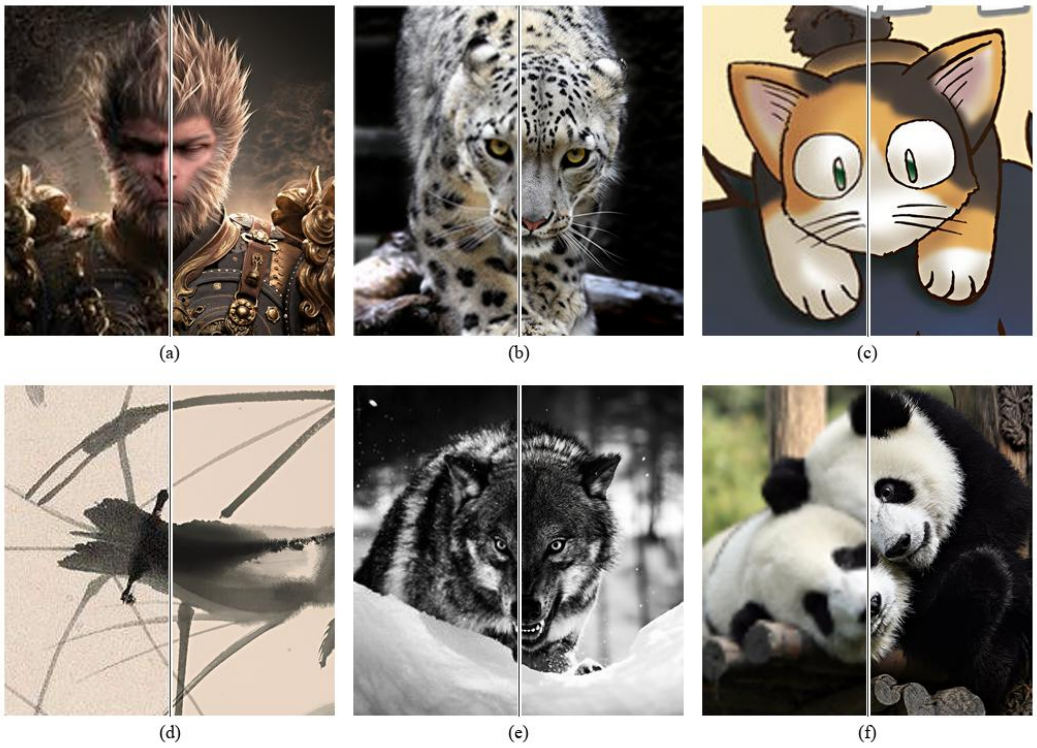

Figure 12: Additional qualitative examples of GRASP on diverse low-quality inputs. In each split view, the left half shows the low-quality input upsampled to the output resolution, and the right half shows the GRASP output.

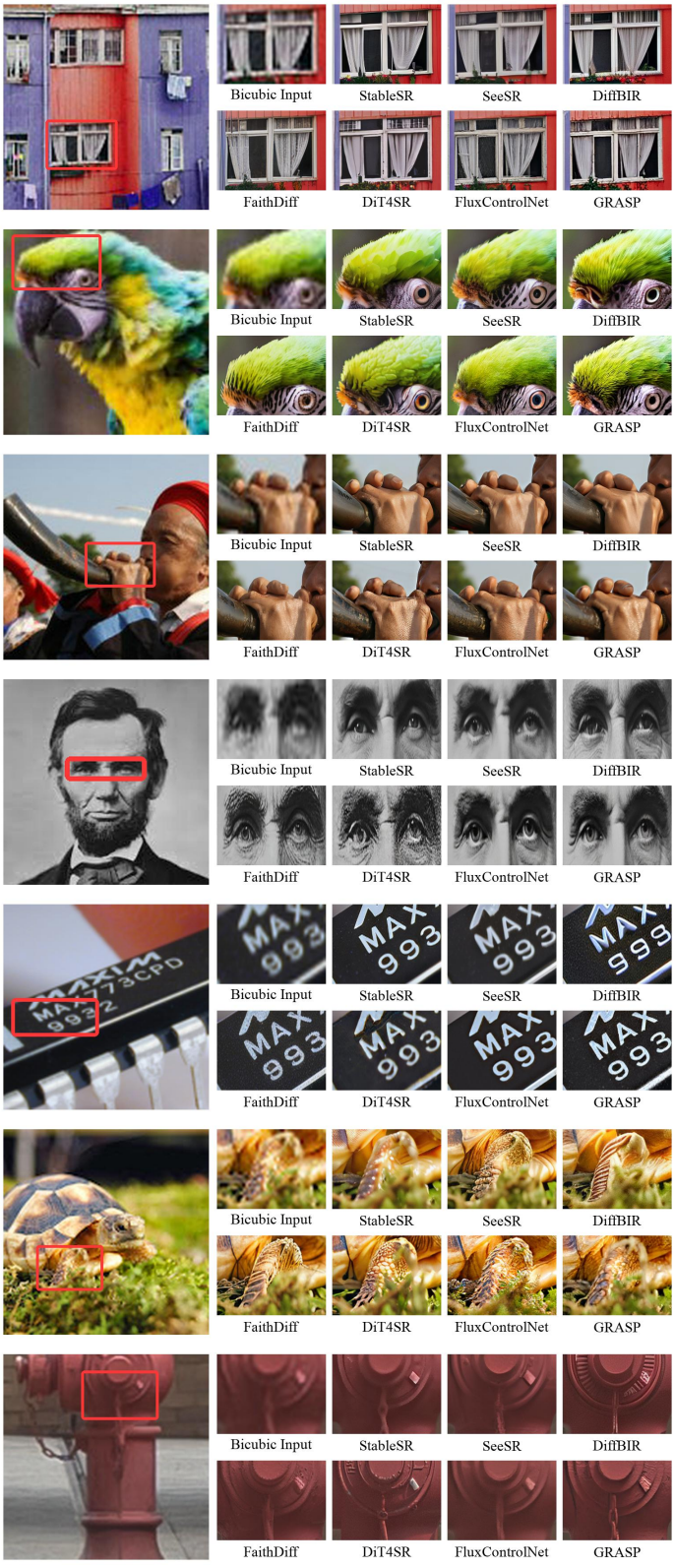

Figure 13: Additional qualitative zoom-in comparisons on challenging low-quality inputs. The examples complement Figure 4 by showing more cases where GRASP preserves local structures and realistic details under diverse degradations.