TOWARDS AUTONOMOUS IMAGE RESTORATION WITH FINE-GRAINED AGENTIC CONTROL AND HISTORY-GROUNDED RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most generative image restoration methods perform restoration in a single round, without explicitly reassessing residual degradations, newly introduced artifacts, or reconstruction errors in the restored results. Recent agentic approaches introduce iterative assessment and replanning, but their control largely remains at the level of tool selection, with visual history primarily used for decision-making rather than restoration execution. To build an autonomous restoration framework, we propose **AutoRestore**, which tightly integrates VLM-based assessment with generative restoration through fine-grained agentic control and history-grounded restoration. At each round, AutoRestore uses a VLM to identify restoration requirements and generate region-specific instructions. These instructions are further translated into spatially varying uncertainty and uncertainty-driven modulation, enabling fine-grained restoration control. Meanwhile, we introduce history-grounded restoration, which preserves reference representations from the original input and intermediate results as persistent visual memory across rounds. Extensive experiments on multiple image restoration benchmarks demonstrate that AutoRestore achieves superior restoration quality and a better trade-off between fidelity and perceptual quality.

1 INTRODUCTION

Image restoration (IR) aims to recover high-quality (HQ) images from low-quality (LQ) observations affected by degradations such as resolution loss (Dong et al., 2015b; Liang et al., 2021; Chen et al., 2023; Zhang et al., 2024), noise (Zhang et al., 2017; 2019; Zamir et al., 2022; Zhang et al., 2023), blur (Cho et al., 2021; Zamir et al., 2021), and compression artifacts (Dong et al., 2015a; Jiang et al., 2021). Early methods typically learn task-specific LQ-to-HQ mappings under predefined degradation assumptions, whereas real-world images contain combinations of unknown degradations. Generative methods address this challenge using diffusion priors (Ho et al., 2020; Song et al., 2020) by formulating restoration as conditional generation (Saharia et al., 2022; Li et al., 2022) or constructing diffusion processes that bridge LQ and HQ image distributions (Yue et al., 2023; Wang et al., 2024b; Zhang et al., 2025b). More recent methods further adapt large-scale pretrained text-to-image (T2I) models, including Stable Diffusion (Podell et al., 2024; Esser et al., 2024) and FLUX (Black Forest Labs, 2024), to leverage their rich natural-image priors for photo-realistic restoration (Wu et al., 2024c; Yu et al., 2024; Wu et al., 2024b; Zhu et al., 2024). In parallel, advances in vision-language models (VLMs) (Liu et al., 2023; OpenAI, 2023) and tool-using multimodal agents (Yao et al., 2022; Wu et al., 2023; Shen et al., 2023) have enabled agentic IR systems (Chen et al., 2024b; Zhu et al., 2025; Jiang et al., 2026), which diagnose degradations, coordinate specialized restoration tools, and assess intermediate results for reflection and replanning.

These advances are converging toward a broader goal: an autonomous restoration framework that can assess and improve its own output across rounds. Realizing this goal, however, requires answering three open questions. ***First, how can restoration become a closed loop with self-assessment and correction?*** A truly autonomous restoration system should continuously evaluate and refine its own outputs, rather than merely generating a plausible HQ image in a single pass. Existing generative IR methods, however, typically perform restoration in one round, with conditioning derived once from the initial LQ input. Without feedback to reassess intermediate results and trigger further correction,

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.

residual degradations, reconstruction errors, and newly introduced artifacts may persist in the final output. ***Second, how can high-level visual reasoning be translated into low-level restoration control?*** Visual reasoning should not only identify degradations, but also specify what to restore, where to restore, and how much correction is required, directly guiding the restoration process. Agentic IR systems introduce iterative assessment and replanning, but VLM reasoning largely remains at the orchestration level. Their assessments typically identify image-level degradations and select corresponding tools, while the actual restoration is still performed globally with limited spatial control. Such degradation-specific decomposition may be insufficient for complex real-world images where multiple degradations are spatially coupled, leaving semantic reasoning disconnected from fine-grained restoration. ***Lastly, how can restoration preserve and reuse visual evidence across iterations?*** Intermediate restoration results contain valuable visual evidence, including recovered structures, refined textures, and preserved content information. Rather than treating them merely as records for reflection, replanning, or rollback, an autonomous restoration system should retain and reuse this evidence to guide subsequent restoration. Existing agentic systems typically provide restoration tools only with the latest output, causing useful information from the original input and previous iterations to be discarded and potentially allowing errors to propagate across rounds.

To realize such an autonomous restoration paradigm, we propose **AutoRestore**, a closed-loop framework that integrates iterative VLM reasoning with generative restoration through fine-grained control and history-grounded refinement. At each restoration round, AutoRestore employs a VLM to analyze the current restoration result and generate region-specific restoration instructions. Rather than only identifying degradations or selecting restoration operations, the VLM specifies where corrections are needed, what visual elements should be recovered, and how individual regions should be refined. These instructions are translated into fine-grained restoration control through instruction-guided uncertainty modeling. Specifically, an uncertainty predictor leverages the restoration instructions to estimate the spatially varying scale of input-to-target latent residuals, indicating the required correction magnitude at different locations. The resulting uncertainty field provides spatially adaptive modulation for the generative restorer, bridging high-level visual reasoning and restoration execution. After each restoration round, the VLM reassesses the updated result and determines whether further refinement is required, enabling continuous self-correction. Furthermore, AutoRestore introduces history-grounded restoration by incorporating information accumulated from previous restoration rounds. We maintain reference representations from the original input and intermediate restoration results as a visual memory, allowing subsequent restoration to jointly utilize current and historical visual evidence. By coupling iterative reassessment, fine-grained agentic control, and history-grounded restoration, AutoRestore establishes a unified framework that bridges visual reasoning with restoration process.

Our main contributions are summarized as follows:

- We propose **AutoRestore**, an autonomous image restoration framework that tightly integrates iterative VLM assessment with generative restoration in a closed loop.
- We introduce instruction-guided uncertainty modeling to translate VLM reasoning into spatially adaptive restoration control, enabling fine-grained correction of different regions.
- We propose a history-grounded restoration mechanism that incorporates information from the original input and intermediate restoration results, allowing subsequent restoration rounds to leverage accumulated visual evidence.

2 Related Works

2.1 Generative Modeling and Image Restoration

Denoising diffusion probabilistic models (Ho et al., 2020; Nichol & Dhariwal, 2021) and score-based generative models (Song & Ermon, 2019; Song et al., 2020) formulated through stochastic differential equations model image distributions via a learned progressive denoising process. Building on this generative formulation, early IR methods (Saharia et al., 2022; Li et al., 2022) recast restoration as a conditional denoising process that progressively generates HQ images from noise under the guidance of LQ inputs. Subsequent work (Yue et al., 2023) simplifies distribution modeling through a residual-based generative trajectory between LQ and HQ images. Further advances incorporate distillation (Wang et al., 2024b), uncertainty modeling (Zhang et al., 2025b), and tra-

jectory matching (You et al., 2025) to improve restoration quality and sampling efficiency. With the emergence of large-scale pretrained text-to-image (T2I) models (Esser et al., 2024; Black Forest Labs, 2024), recent IR methods (Wang et al., 2024a; Yang et al., 2024; Wu et al., 2024c;b; Zhu et al., 2024) have shifted from learning restoration distributions from scratch to adapting pretrained image priors. They leverage strong generative priors and incorporate semantic guidance to recover more photorealistic image structures and textures.

2.2 Degradation Semantics and Agentic Image Restoration

Beyond content semantics, another line of research uses degradation understanding to guide image restoration. Early methods encode degradation characteristics through prompts or descriptors (Potlapalli et al., 2023; Guo et al., 2024), or adapt CLIP-based representations to identify image degradations (Luo et al., 2024; Jiang et al., 2024). Natural-language instructions (Conde et al., 2024) further provide explicit guidance on restoration requirements. With the rapid development of vision-language models (VLMs), such as LLaVA (Liu et al., 2023) and GPT-4V (OpenAI, 2023), recent studies have explored their capabilities in fine-grained image understanding and multimodal reasoning. These advances have also benefited low-level vision tasks, enabling more detailed analysis of image distortions and perceptual quality (You et al., 2024; Wu et al., 2024a). Building on these capabilities, agentic image restoration methods leverage high-level visual reasoning to analyze degradation patterns, decompose complex restoration problems into sub-tasks, and plan restoration workflows. RestoreAgent (Chen et al., 2024b) and AgenticIR (Zhu et al., 2025) further introduce iterative decision-making mechanisms, where restoration strategies are adjusted according to intermediate results. Subsequent works enhance restoration planning through stronger reasoning mechanisms, such as chain-of-thought reasoning (Wei et al., 2022; Zhou et al., 2025) and tool portrait banks (Wang et al., 2026), enabling more effective tool selection and restoration scheduling.

3 Methodology

In this section, we present AutoRestore, as illustrated in Fig. 1. We first outline the autonomous restoration loop in Sec. 3.1. We then describe how restoration instructions and instruction-guided uncertainty jointly translate VLM reasoning into fine-grained agentic control in Sec. 3.2. Finally, Sec. 3.3 presents history-grounded restoration, which uses a historical cache as visual memory to jointly leverage current and historical references during restoration.

3.1 Autonomous Restoration Framework

We formulate autonomous image restoration as a closed-loop process that continuously assesses the current result and executes the required corrections. To this end, we introduce **AutoRestore**, which couples a VLM with a generative restorer through fine-grained agentic control and history-grounded restoration. Rather than restricting VLM reasoning to tool selection, AutoRestore translates visual assessment into complementary linguistic and spatial conditions that directly guide region-specific restoration. Furthermore, restoration is grounded in visual references from both the current input and previous restoration rounds, allowing accumulated evidence to directly participate in restoration. We use Qwen3.8 (Qwen Team, 2026) as the VLM and build the restorer on FLUX.2 (Black Forest Labs, 2026), whose joint processing of text and image tokens supports this unified integration. Fig. 1(a) illustrates the overall framework.

We initialize the framework with a degraded image $\boldsymbol{y}^{(0)} \in \mathbb{R}^{3\times H\times W}$. At each round i, the VLM assesses the current image $\boldsymbol{y}^{(i-1)}$ and generates a restoration instruction $\boldsymbol{p}_r^{(i)}$, specifying the required local corrections and desired restoration outcomes. Conditioned on these instructions, an uncertainty predictor estimates the spatially varying residual scale $\boldsymbol{u}^{(i)}$, which provides adaptive modulation during restoration. To retain useful visual evidence from previous restoration stages, we maintain a visual memory initialized as $\boldsymbol{h}^{(0)} = \varnothing$. Before generating the next result, reference representations of the current image are integrated with accumulated historical information to update the visual memory $\boldsymbol{h}^{(i)}$. The restorer then produces $\boldsymbol{y}^{(i)}$ by jointly leveraging the restoration instruction, uncertainty-driven modulation, and the visual references stored in $\boldsymbol{h}^{(i)}$. The output is fed back to the VLM for reassessment and becomes the new visual reference for subsequent rounds. This closed loop progressively refines restoration decisions while grounding restoration in the orig-

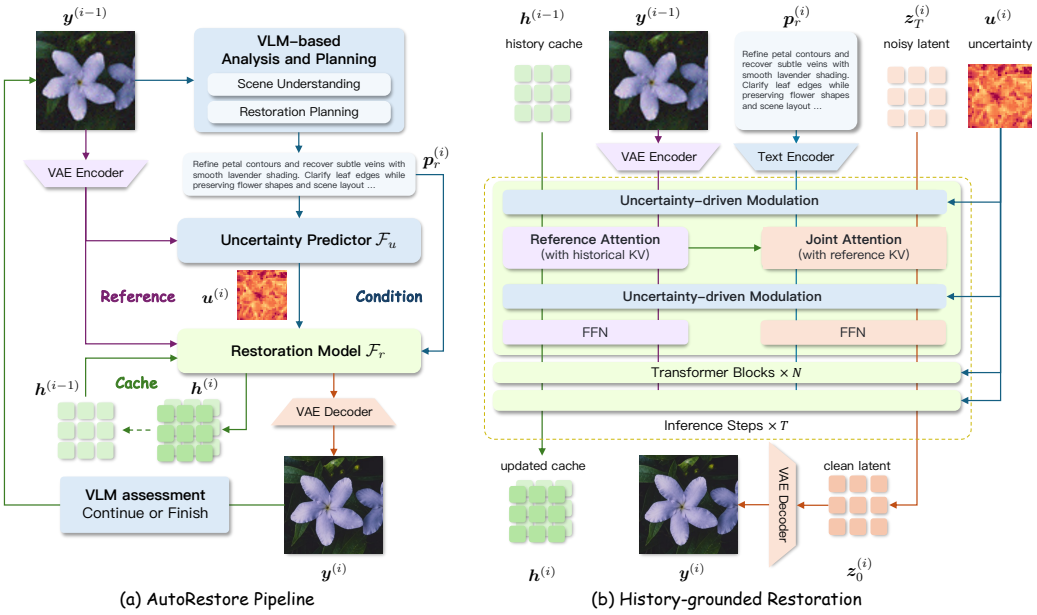

Figure 1: **Overview of AutoRestore. (a) Autonomous restoration loop.** At each round, the VLM analyzes the current image to generate region-specific restoration instructions, and $\mathcal{F}_u$ predicts uncertainty conditioned on the image and instructions. The instructions and uncertainty jointly provide fine-grained agentic control over the restorer $\mathcal{F}_r$. The VLM reassesses the output to determine whether further restoration is needed and, if so, updates the instructions for the next round. **(b) History-grounded restoration.** The original input and intermediate results are integrated as reference, which is then supplied to joint text–flow attention, making the restoration process grounded by both current and historical visual evidence.

inal input and intermediate results. We elaborate on fine-grained agentic control in Sec. 3.2 and history-grounded restoration in Sec. 3.3.

3.2 Fine-grained Agentic Control

Existing generative restoration methods typically rely on scene descriptions (Wu et al., 2024c) or image-level restoration requirements (Potlapalli et al., 2023; Conde et al., 2024) as guidance. However, such coarse-grained semantic conditions provide limited control over the type and extent of corrections required by different image regions (Yin et al., 2026; Ye et al., 2026). We therefore aim to translate VLM reasoning into fine-grained conditions that explicitly characterize local restoration requirements and enable spatially adaptive restoration. For notational simplicity, we omit round indices throughout this subsection and use $\boldsymbol{y}$ to denote the input of the current round.

Residual-aware Instruction Generation. We first leverage the VLM to generate restoration instructions that specify what and how each region or object should be corrected. To better ground these instructions in the underlying scene content, we decouple scene understanding from restoration planning through a two-step reasoning process. Specifically, the VLM first infers a scene description $\boldsymbol{p}_c \sim \pi_\theta^{\mathrm{VLM}}(\cdot \mid \boldsymbol{y}, \boldsymbol{q}_c)$. Using $\boldsymbol{p}_c$ as semantic context, it then analyzes the restoration requirements and desired outcomes, producing $\boldsymbol{p}_r \sim \pi_\theta^{\mathrm{VLM}}(\cdot \mid \boldsymbol{y}, \boldsymbol{p}_c, \boldsymbol{q}_r)$. $\boldsymbol{q}_c$ and $\boldsymbol{q}_r$ denote task prompts for scene understanding and restoration planning, respectively (see Appendix. A.6 for details). The generated restoration instruction $\boldsymbol{p}_r$ contains a brief description of the overall scene, followed by region-specific correction requirements. Each local instruction identifies the target region or object, the required correction, and the desired structure, texture, or appearance. Therefore, $\boldsymbol{p}_r$ establishes an explicit correspondence between image content and restoration objectives. Moreover, since instructions are regenerated from the latest restoration results at each round, the required corrections can be dynamically updated throughout the restoration process.

Instruction-guided Uncertainty Estimation. Although $\boldsymbol{p}_r$ provides semantic guidance on what should be restored, it does not explicitly determine the spatial magnitude of required corrections. We therefore introduce instruction-guided uncertainty estimation to translate VLM reasoning into an explicit spatial correction prior. Inspired by heteroscedastic uncertainty modeling (Kendall & Gal, 2017; Zhang et al., 2025b), we estimate a latent uncertainty field that characterizes the residual scale at different spatial locations. For paired LQ-HQ images $\boldsymbol{y}$ and $\boldsymbol{x}$, we obtain their latent representations $\boldsymbol{z}_y$ and $\boldsymbol{z}_x$ using the VAE encoder. The latent residual is defined as $\boldsymbol{z}_r := \boldsymbol{z}_x - \boldsymbol{z}_y \in \mathbb{R}^{N_l \times d_l}$, with N_l spatial locations and d_l channels. This residual is further modeled using a factorized, zero-centered Gaussian distribution:

$$p(\boldsymbol{z}_r \mid \boldsymbol{z}_y, \boldsymbol{p}_r) = \prod_{j=1}^{N_l} \mathcal{N}(\boldsymbol{z}_{r,j}; \mathbf{0}, \mathrm{Diag}(\boldsymbol{u}_j^2)), \tag{1}$$

where $\boldsymbol{u}_j \in \mathbb{R}^{d_l}_{>0}$ represents the channel-wise residual scale at location j, which we refer to as restoration uncertainty. The full uncertainty field is predicted as $\boldsymbol{u} = \mathcal{F}_u(\boldsymbol{z}_y, \boldsymbol{p}_r) \in \mathbb{R}^{N_l \times d_l}_{>0}$. We train the predictor by minimizing the corresponding negative log-likelihood:

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{2N_l} \sum_{j=1}^{N_l} \sum_{c=1}^{d_l} (z_{r,j,c}^2 / u_{j,c}^2 + \log u_{j,c}^2), \tag{2}$$

where $z_{r,j,c}$ and $u_{j,c}$ are the residual and predicted standard deviation at location j and channel c, respectively. This objective ties the predicted uncertainty to local residual magnitude, allowing $\boldsymbol{u}$ to reflect spatial variations in the scale of the required corrections. Through $\boldsymbol{p}_r$, the predictor draws on the VLM's understanding of the desired local structures and appearances to assess correction needs that may not be evident from image cues alone.

Uncertainty-driven Spatial Modulation. The predicted uncertainty provides a spatially varying prior that characterizes the required restoration strength at different locations. To inject this prior into the restoration process, we use it to modulate the timestep conditioning of the pretrained backbone. Specifically, we transform the globally shared timestep embedding into token-specific conditioning, enabling each region to receive restoration signals according to its estimated correction requirements. Let $\boldsymbol{e}_{t_s} \in \mathbb{R}^{1 \times d_t}$ denote the timestep embedding for image stream $s \in \{\mathrm{flow}, \mathrm{ref}\}$, with embedding dimension d_t. The flow stream uses the current sampling timestep $t_{\mathrm{flow}} = t$, while the reference stream uses a fixed timestep $t_{\mathrm{ref}} = 0$. Given the corresponding uncertainty features $\boldsymbol{u}_s \in \mathbb{R}^{N_s \times d_l}$ for N_s tokens, we construct spatially varying timestep embeddings through an affine transformation:

$$(\boldsymbol{\alpha}_s, \boldsymbol{\beta}_s) = \mathcal{P}_s(\boldsymbol{u}_s), \quad \widetilde{\boldsymbol{e}}_s = (1 + \boldsymbol{\alpha}_s) \odot \boldsymbol{e}_{t_s} + \boldsymbol{\beta}_s, \tag{3}$$

where $\mathcal{P}_s$ is a learnable projection that predicts token-specific scale and shift parameters $\boldsymbol{\alpha}_s, \boldsymbol{\beta}_s \in \mathbb{R}^{N_s \times d_t}$. The timestep embedding is broadcast along the token dimension, yielding $\widetilde{\boldsymbol{e}}_s \in \mathbb{R}^{N_s \times d_t}$, where N_s is the number of tokens in stream s. These embeddings are fed to the corresponding modulation heads to generate token-specific AdaLN (Peebles & Xie, 2023) parameters for the flow and reference streams. By progressively converting VLM reasoning into restoration instructions, spatial uncertainty, and token-level modulation, AutoRestore achieves fine-grained agentic control over both where and how strongly restoration is performed.

3.3 History-grounded Restoration

While local correction requirements are reassessed at each round, effective restoration also requires preserving reliable visual evidence throughout the iterative process. Relying only on the latest restoration result may discard useful information from earlier stages and allow error accumulation. We therefore maintain visual references from both the current restoration state and previous rounds, enabling subsequent restoration to be grounded in accumulated visual evidence.

To preserve this visual evidence, we maintain a historical cache $\boldsymbol{h}$ as persistent visual memory within the restorer. The cache stores layer-wise reference KV representations and is initialized as $\boldsymbol{h}^{(0)} = \varnothing$. At round i, we encode the current image $\boldsymbol{y}^{(i-1)}$ into reference tokens $\boldsymbol{z}_{\mathrm{ref}}^{(i)}$. As illustrated in Fig. 2, during training, reference attention is applied causally, where each reference token only attends to tokens from current and preceding rounds. During inference, the retained KV cache is introduced

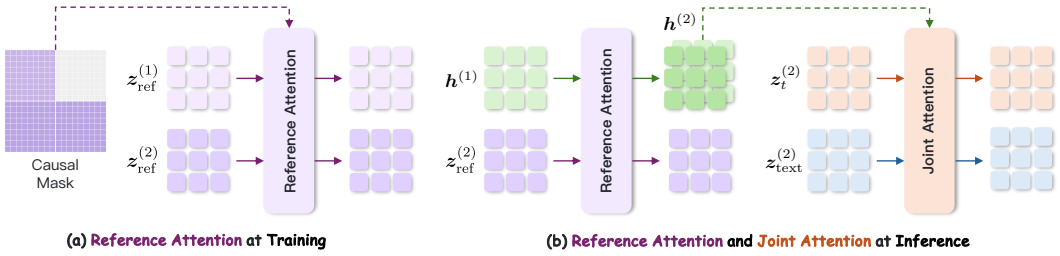

Figure 2: Illustration of reference attention (Eq. 4) and joint attention (Eq. 6) at round $i = 2$. During training, a block-level causal mask allows reference tokens to attend to tokens from the same and preceding rounds, but not future rounds. During inference, the current reference is processed once per restoration round to cache its keys and values. The updated cache is kept fixed and reused by joint attention throughout the flow sampling steps of the restoration round.

as additional reference context, enabling the current reference tokens to incorporate historical visual evidence:

$$\boldsymbol{A}_r^{(i)} = \mathrm{Attn}\left(\boldsymbol{Q}_r^{(i)}, [\boldsymbol{K}_r^{(i)}, \boldsymbol{K}_h^{(i-1)}], [\boldsymbol{V}_r^{(i)}, \boldsymbol{V}_h^{(i-1)}]\right), \tag{4}$$

where $(\boldsymbol{K}_h^{(i-1)}, \boldsymbol{V}_h^{(i-1)})$ denotes the historical KV cache accumulated from previous rounds. $[\cdot, \cdot]$ denotes concatenation along the token dimension, and layer indices are omitted for clarity. This interaction allows the current reference representations to incorporate historical visual evidence. The current reference KV representations are then appended to the retained historical memory:

$$\boldsymbol{K}_h^{(i)} = [\boldsymbol{K}_h^{(i-1)}, \boldsymbol{K}_r^{(i)}], \quad \boldsymbol{V}_h^{(i)} = [\boldsymbol{V}_h^{(i-1)}, \boldsymbol{V}_r^{(i)}], \quad \boldsymbol{h}^{(i)} = (\boldsymbol{K}_h^{(i)}, \boldsymbol{V}_h^{(i)}). \tag{5}$$

Through this recurrent update, the cache progressively accumulates visual evidence from the original input and intermediate restoration results, forming a persistent memory for subsequent rounds. The updated cache $\boldsymbol{h}^{(i)}$ is then kept fixed during the subsequent flow inference steps within the same restoration round and serves as visual memory for both text and flow streams. The restoration instructions $\boldsymbol{p}_r^{(i)}$ are encoded into text embeddings $\boldsymbol{z}_{\text{text}}^{(i)}$, while the noisy latent $\boldsymbol{z}_t^{(i)}$ is projected into flow tokens. At each attention layer, text and flow tokens jointly attend to their own representations and the cached visual memory:

$$\boldsymbol{A}_{tf}^{(i)} = \mathrm{Attn}\left(\boldsymbol{Q}_{tf}^{(i)}, [\boldsymbol{K}_{tf}^{(i)}, \boldsymbol{K}_h^{(i)}], [\boldsymbol{V}_{tf}^{(i)}, \boldsymbol{V}_h^{(i)}]\right), \tag{6}$$

where the subscript tf denotes concatenated text and flow tokens. Through this attention, historical visual evidence directly participates in restoration rather than serving only as context for assessment or replanning. Together with the restoration instructions and uncertainty-driven modulation described in Sec. 3.2, these visual references provide additional evidence for the restoration process:

$$\boldsymbol{y}^{(i)} = \mathcal{D}\left(\mathcal{F}_r\left(\boldsymbol{z}_T^{(i)}; \boldsymbol{z}_{\text{text}}^{(i)}, \boldsymbol{h}^{(i)}, \boldsymbol{u}^{(i)}\right)\right), \tag{7}$$

where $\mathcal{F}_r$ denotes the conditional latent sampling process implemented by the restorer and $\mathcal{D}$ is the VAE decoder. The resulting image $\boldsymbol{y}^{(i)}$ is fed back to the VLM, which determines whether further restoration is required. If further restoration is needed, the restoration instructions are revised and the instruction-guided uncertainty is re-estimated for the next round. The image then serves as the next visual reference which is appended to the cache to form $\boldsymbol{h}^{(i+1)}$. AutoRestore thus closes the loop between visual assessment and restoration execution, adaptively revising local correction requirements while grounding restoration in current and historical visual evidence.

4 Experiments

4.1 Experimental Setup

We build both the uncertainty predictor $\mathcal{F}_u$ and restoration model $\mathcal{F}_r$ upon the FLUX.2-Klein-base-4B backbone (Black Forest Labs, 2026). Our training data are collected from LSDIR (Li et al.,

Table 1: **Ablation study of AutoRestore.** (a) Effects of key restoration components. (b) Effects of instruction guidance for uncertainty estimation.

(a) Restoration components

Variant	PSNR ↑	LPIPS ↓	CLIPIQA+ ↑	MUSIQ ↑	A-FINE-NR ↓
baseline	23.09	0.2323	0.7324	71.48	-1.067
+ uncertainty modulation	23.07	0.2316	0.7375	72.07	-1.088
+ multi-round inference	22.26	0.2727	0.7505	72.71	-1.116
+ historical cache	23.06	0.2361	0.7556	72.53	-1.129

(b) Instruction guidance

Instruction	w/o	w/
NLL ↓	-0.468	-0.475
SRCC ↑	0.982	0.990
AUPRC ↑	0.891	0.922

2023), 4KLSDB (Zhu et al., 2026), UHR100K (Zhao et al., 2026), and Aesthetic-Train-V2 (Zhang et al., 2025a). We apply image filtering and preprocessing procedures to curate high-quality training data, and further generate restoration instructions for each sample. All models are trained using the AdamW optimizer (Loshchilov & Hutter, 2017) on eight NVIDIA RTX 5090 GPUs. We employ mixed-precision training (Micikevicius et al., 2017) and gradient checkpointing (Chen et al., 2016) for memory-efficient training. Training is conducted in two stages. In the first stage, we train the model on single-round restoration to establish its basic capability of recovering high-quality images from degraded inputs. Starting from this initialization, the second stage introduces multi-round training, enabling the model to adapt to iterative restoration, where restoration instructions and visual references are progressively updated across rounds. In each stage, we sequentially train the uncertainty predictor and restoration model. Details of the model architecture, data curation pipeline, training configurations and strategies are provided in the appendix.

4.2 Ablation Study

We conduct ablation studies to evaluate the contributions of the key components in AutoRestore, including uncertainty-driven spatial modulation, history-grounded restoration, and instruction-guided uncertainty estimation.

As shown in Table 1(a), we evaluate different variants on the real-world NTIRE2020-Track1 dataset (Lugmayr et al., 2020). The proposed uncertainty-driven spatial modulation consistently improves perceptual quality while maintaining comparable restoration fidelity, indicating that uncertainty guidance enables the restorer to adapt its behavior to spatially varying correction demands without sacrificing content preservation. We further investigate the effect of iterative restoration. A naive multi-round strategy that repeatedly restores the latest output improves no-reference perceptual metrics, but substantially decreases PSNR by 0.81 dB and degrades LPIPS by 0.0411. This suggests that relying only on the previous restoration result may accumulate restoration errors and lead to progressive content drift. In contrast, history-grounded restoration preserves visual evidence from both the original input and intermediate results, maintaining content fidelity while retaining the perceptual benefits of iterative refinement. Table 1(b) further evaluates the effect of restoration instructions on uncertainty estimation. We use negative log-likelihood (NLL) to measure the probabilistic fit of the predicted residual distribution, Spearman's rank correlation coefficient (SRCC) to evaluate the correlation between predicted uncertainty and actual residual magnitude, and area under the precision–recall curve (AUPRC) to measure the ability to identify samples with large reconstruction residuals. All metrics are computed on the validation split of our training dataset. Instruction guidance reduces NLL and improves SRCC, demonstrating more accurate uncertainty estimation and better ranking of residual magnitude. It also improves AUPRC from 0.891 to 0.922, indicating that VLM-based visual reasoning provides additional semantic cues for identifying challenging restoration cases with large residual errors.

4.3 Comparison with Existing Methods

We compare AutoRestore with two groups of state-of-the-art image restoration methods. Single-round diffusion-based baselines include SeeSR (Wu et al., 2024c), OSEDiff (Wu et al., 2024b), SUPIR (Yu et al., 2024), DreamClear (Ai et al., 2024), FaithDiff (Chen et al., 2025b), and Lucid-FLUX (Fei et al., 2026). We also compare with recent multi-round agentic restoration methods, including AutoDIR (Jiang et al., 2024), AgenticIR (Zhu et al., 2025), and RAR (Chen et al., 2026).

Table 2: Quantitative comparison with state-of-the-art methods on real-world (NTIRE2020-Track1 and RealSR) and synthetic (DIV2K-Val and LSDIR-Val) benchmarks.

Dataset		Metric	SeeSR	OSEDiff	SUPIR	Dreamclear	FaithDiff	LucidFlux	AutoDIR	AgenticIR	RAR	AutoRestore
Real-world	*NTIRE2020-Track1*	PSNR	22.77	22.44	22.75	21.86	22.85	20.36	22.20	22.54	22.46	23.06
		SSIM	0.5932	0.5964	0.5679	0.5540	0.5872	0.5359	0.5774	0.5697	0.5828	0.6053
		LPIPS	0.2633	0.2603	0.3172	0.2359	0.2435	0.2964	0.2820	0.3474	0.3227	0.2361
		TOPIQ-FR	0.5122	0.5129	0.4526	0.5692	0.5415	0.4648	0.4907	0.4295	0.4313	0.5696
		CLIPIQA+	0.7407	0.7329	0.6784	0.7148	0.7063	0.7887	0.6432	0.6732	0.6264	0.7556
		MUSIQ	71.47	70.88	65.47	69.03	70.93	73.95	64.69	67.87	64.49	72.53
		NIQE	4.319	4.342	4.551	5.806	4.221	4.136	4.247	4.878	3.704	3.832
		A-FINE-NR	-1.024	-1.044	-0.863	-0.896	-1.048	-1.153	-0.863	-1.024	-0.710	-1.129
	RealSR	PSNR	25.34	25.17	25.20	24.97	25.30	21.34	23.96	22.78	22.74	24.10
		SSIM	0.7253	0.7346	0.6987	0.6961	0.7048	0.6277	0.6903	0.6625	0.6565	0.6816
		LPIPS	0.2968	0.2921	0.3394	0.3389	0.2898	0.3446	0.3004	0.3598	0.3749	0.3018
		TOPIQ-FR	0.4918	0.5060	0.4373	0.4852	0.5101	0.4273	0.4633	0.4558	0.4186	0.4977
		CLIPIQA+	0.6905	0.6965	0.6049	0.5955	0.6595	0.7306	0.5503	0.6303	0.5675	0.7338
		MUSIQ	69.57	69.08	57.52	58.05	68.79	71.85	58.55	63.32	65.51	72.15
		NIQE	5.386	5.646	5.547	6.984	5.377	5.709	6.187	5.360	4.729	5.311
		A-FINE-NR	-1.034	-1.049	-0.852	-0.747	-1.051	-1.155	-0.821	-0.969	-0.822	-1.148
Synthetic	*DIV2K-Val*	PSNR	23.92	23.39	26.37	25.45	24.10	22.89	28.67	27.50	26.19	28.51
		SSIM	0.6305	0.6266	0.7321	0.7074	0.6421	0.6588	0.8050	0.8189	0.7560	0.8217
		LPIPS	0.2457	0.2625	0.1875	0.1660	0.2270	0.2472	0.1802	0.2682	0.1782	0.1343
		TOPIQ-FR	0.5376	0.5194	0.6600	0.7154	0.5746	0.5569	0.7046	0.6149	0.7137	0.7817
		CLIPIQA+	0.7551	0.7266	0.7262	0.7429	0.7129	0.7864	0.6294	0.5295	0.6836	0.7392
		MUSIQ	72.23	70.94	70.23	70.85	71.04	74.03	65.79	60.87	70.53	71.86
		NIQE	4.468	4.672	4.469	5.965	4.186	4.137	4.656	6.065	4.314	4.052
		A-FINE-NR	-1.046	-1.072	-0.956	-0.997	-1.050	-1.167	-0.908	-0.895	-0.987	-1.086
	LSDIR-Val	PSNR	21.61	21.28	23.46	22.97	21.67	21.61	27.13	25.93	24.60	26.55
		SSIM	0.5670	0.5631	0.6739	0.6676	0.5755	0.6468	0.7868	0.7890	0.7391	0.8051
		LPIPS	0.2495	0.2634	0.1867	0.1505	0.2272	0.2180	0.1649	0.2518	0.1449	0.1139
		TOPIQ-FR	0.5111	0.4939	0.6347	0.7116	0.5523	0.5872	0.6994	0.6152	0.7305	0.7991
		CLIPIQA+	0.7708	0.7370	0.7657	0.7856	0.7314	0.8003	0.6682	0.5761	0.7017	0.7731
		MUSIQ	73.45	72.11	72.15	73.01	72.47	74.75	69.34	63.84	72.47	73.76
		NIQE	4.207	4.306	4.621	6.217	3.918	4.250	4.280	5.553	3.930	3.786
		A-FINE-NR	-1.133	-1.172	-1.061	-1.084	-1.149	-1.257	-1.020	-1.020	-1.072	-1.218

Quantitative Comparison. We evaluate all methods on two real-world benchmarks, NTIRE2020-Track1 (Lugmayr et al., 2020) and RealSR (Cai et al., 2019), and two synthetic benchmarks, DIV2K-Val (Agustsson & Timofte, 2017) and LSDIR-Val (Li et al., 2023). For the synthetic benchmarks, we apply degradations including blur, noise, and JPEG compression to the corresponding high-quality validation images. All methods are evaluated at a resolution of 512×512. We assess restoration quality in terms of both fidelity and perceptual quality. PSNR, SSIM, LPIPS (Zhang et al., 2018), and TOPIQ-FR (Chen et al., 2024a) measure fidelity against ground-truth references, while CLIPIQA+ (Wang et al., 2023), MUSIQ (Ke et al., 2021), NIQE (Mittal et al., 2012), and A-FINE-NR (Chen et al., 2025a) evaluate no-reference perceptual quality.

As shown in Table 2, AutoRestore consistently achieves better balance between fidelity and perceptual quality across benchmarks. On NTIRE2020-Track1, it achieves competitive or superior perceptual quality compared with existing methods. Although LucidFlux performs slightly better on several perceptual metrics, AutoRestore provides substantially higher fidelity, improving PSNR by 2.70 dB and LPIPS by 0.0603. These results indicate that AutoRestore better preserves ground-truth structures while reducing visually plausible but inaccurate details introduced by generative priors. Compared with existing multi-round agentic methods, AutoRestore achieves a more favorable overall balance between restoration fidelity and perceptual quality. The gains are particularly evident on real-world benchmarks, where degradations are more complex and less predictable. This demonstrates the advantage of a unified restoration framework for handling diverse degradation patterns while maintaining both content fidelity and perceptual quality.

Visual Comparison. Fig. 3 presents qualitative comparisons with representative IR methods. AutoRestore produces visually realistic results while maintaining better consistency with the ground truth. In these examples, AutoRestore better preserves the reflective textures of the peppers and plastic bag and recovers text details that are more consistent with the ground truth. Additional visual comparisons are provided in the appendix. We further compare the first-round and multi-round results of AutoRestore in Fig. 4. Subsequent rounds further correct residual artifacts and refine lo-



Figure 3: Visual comparison on NTIRE2020-track1 and LSDIR-Val. *Zoom in for a better view.*

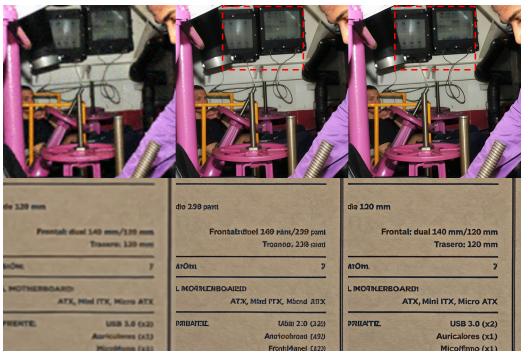

Figure 4: Visual comparison with and without multi-round inference.

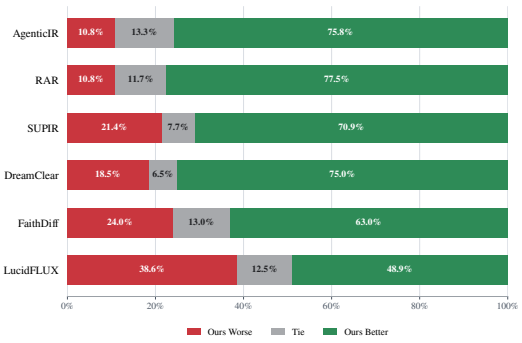

Figure 5: User study between AutoRestore and representative IR methods.

cal structures and text details that remain imperfect after the first round. These results highlight the benefit of framework in improving restoration quality beyond the first-round output through iterative reassessment and targeted correction.

User Study We conduct a blind pairwise user study with 12 participants to compare AutoRestore with representative single-round and multi-round restoration methods. Participants evaluate randomly ordered image pairs and select the preferred result or indicate a tie. As shown in Fig. 5, AutoRestore is preferred more often than each competing method, highlighting its perceptual advantages over both single-round and multi-round baselines.

5 Conclusion

In this work, we introduce **AutoRestore**, an autonomous image restoration framework that establishes a closed loop between visual reasoning and restoration execution. Unlike existing approaches that either perform restoration in a single round or employ agents mainly for tool-level decision making, AutoRestore enables iterative assessment and adaptive correction through VLM-guided restoration instructions. By analyzing restoration requirements at the region level, the VLM provides detailed semantic guidance on where corrections are needed and how they should be performed. We further introduce instruction-guided uncertainty modeling to translate high-level visual reasoning into fine-grained restoration control, allowing the restoration backbone to adjust correction strength according to local residual complexity. Moreover, AutoRestore maintains persistent visual memory to preserve and reuse evidence from the original input and intermediate restoration results, improving fidelity and reducing error accumulation throughout the iterative process. Extensive experiments demonstrate that AutoRestore achieves superior performance across diverse image restoration benchmarks, providing a better balance between perceptual quality and content fidelity.

AI USE STATEMENT

In this work, we used generative AI tools to assist with several research and writing tasks. Specifically, we used AI tools for generating parts of the synthetic datasets, including restoration instructions and related prompt data, as well as for translation assistance and improving the clarity and readability of the manuscript. All AI-assisted outputs were carefully reviewed, verified, and refined by the authors. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1122–1131. IEEE, 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. *Advances in Neural Information Processing Systems*, 37:55443–55469, 2024.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Black Forest Labs. FLUX.2. `https://github.com/black-forest-labs/flux2`, 2026.

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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3086–3095. IEEE, 2019.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024a.

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12742–12752. IEEE, 2025a.

Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. *Advances in Neural Information Processing Systems*, 37:110643–110666, 2024b.

I Chen, Isma Hadji, Enrique Sanchez, Adrian Bulat, Sy-Yen Kuo, Radu Timofte, Georgios Tzimiropoulos, Brais Martinez, et al. Restore, assess, repeat: A unified framework for iterative image restoration. *arXiv preprint arXiv:2603.26385*, 2026.

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

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22367–22377. IEEE, 2023.

Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4621–4630. IEEE, 2021.

Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, pp. 1–21. Springer, 2024.

Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE international conference on computer vision*, pp. 576–584, 2015a.

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, 2015b.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Song Fei, Tian Ye, Lujia Wang, and Lei Zhu. Lucidflux: Caption-free universal image restoration via a large-scale diffusion transformer. In *International Conference on Learning Representations*, volume 2026, pp. 86233–86260, 2026.

Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. In *European conference on computer vision*, pp. 255–272. Springer, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Jiaxi Jiang, Kai Zhang, and Radu Timofte. Towards flexible blind jpeg artifacts removal. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4977–4986. IEEE, 2021.

Xu Jiang, Gehui Li, Bin Chen, and Jian Zhang. Multi-agent image restoration. *International Journal of Computer Vision*, 134(5):205, 2026.

Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. In *European Conference on Computer Vision*, pp. 340–359. Springer, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1775–1787. IEEE, 2023.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF international conference on computer vision workshops (ICCVW)*, pp. 1833–1844. IEEE, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Andreas Lugmayr, Martin Danelljan, Radu Timofte, Namhyuk Ahn, Dongwoon Bai, Jie Cai, Yun Cao, Junyang Chen, Kaihua Cheng, SeYoung Chun, et al. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2058–2076. IEEE, 2020.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, volume 2024, pp. 16226–16246, 2024.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PmLR, 2021.

OpenAI. GPT-4V(ision) System Card. `https://openai.com/index/gpt-4v-system-card/`, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in neural information processing systems*, 36:71275–71293, 2023.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Jianyi Wang, Kelvin CK 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 CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024a.

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

Yijian Wang, Qingsen Yan, Jiantao Zhou, Duwei Dai, and Wei Dong. Paagent: Portrait-aware image restoration agent via subjective-objective reinforcement learning. *arXiv preprint arXiv:2603.17055*, 2026.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pp. 360–377. Springer, 2024a.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553, 2024b.

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

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Junyan Ye, Jun He, Zilong Huang, Dongzhi Jiang, Xuan Yang, Rui Chen, and Weijia Li. Genclaw: Code-driven agentic image generation. *arXiv preprint arXiv:2605.30248*, 2026.

Zijin Yin, Tiankai Hang, Yiji Cheng, Shiyi Zhang, Runze He, Yu Xu, Chunyu Wang, Bing Li, Zheng Chang, Kongming Liang, et al. Generative visual chain-of-thought for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4657–4667, 2026.

Weiyi You, Mingyang Zhang, Leheng Zhang, Xingyu Zhou, Kexuan Shi, and Shuhang Gu. Consistency trajectory matching for one-step generative super-resolution. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12747–12756. IEEE, 2025.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, pp. 259–276. Springer, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25669–25680. IEEE, 2024.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in neural information processing systems*, 36: 13294–13307, 2023.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14816–14826. IEEE, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Ultra-high-resolution image synthesis: Data, method and evaluation. *arXiv preprint arXiv:2506.01331*, 2025a.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26 (7):3142–3155, 2017.

Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023.

Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2856–2865. IEEE, 2024.

Leheng Zhang, Weiyi You, Kexuan Shi, and Shuhang Gu. Uncertainty-guided perturbation for image super-resolution diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17980–17989. IEEE, 2025b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019.

Chen Zhao, En Ci, Yunzhe Xu, Tiehan Fan, Shanyan Guan, Yanhao Ge, Jian Yang, and Ying Tai. Ultrahr-100k: Enhancing uhr image synthesis with a large-scale high-quality dataset. *Advances in Neural Information Processing Systems*, 38:3373–3393, 2026.

Yingjie Zhou, Jiezhang Cao, Farong Wen, Zicheng Zhang, Yu Zhou, Yue Shi, Xiaohong Liu, Radu Timofte, Luc Van Gool, and Guangtao Zhai. Q-agent: Quality-driven chain-of-thought image restoration agent through robust multimodal large language model. *arXiv preprint arXiv:2504.07148*, 2025.

Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for complex image restoration problems. In *International Conference on Learning Representations*, volume 2025, pp. 57985–58013, 2025.

Yuanzhi Zhu, Ruiqing Wang, Shilin Lu, Junnan Li, Hanshu Yan, and Kai Zhang. Oftsr: One-step flow for image super-resolution with tunable fidelity-realism trade-offs. *arXiv preprint arXiv:2412.09465*, 2024.

Zihao Zhu, Kuan-Ru Huang, Zhaoming Xu, Renjie Li, Bo Wu, Ruizheng Bai, Mingyang Wu, Sayak Paul, and Zhengzhong Tu. 4klsdb: A large-scale dataset for 4k image restoration and generation. *arXiv preprint arXiv:2605.24762*, 2026.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

We use a frozen Qwen3.8-27B (Qwen Team, 2026) VLM for scene description, restoration instruction generation, and image assessment during training and inference. At inference time, the VLM also reassesses each restored image to determine whether another restoration round is required. The uncertainty predictor $\mathcal{F}_u$ and the restorer $\mathcal{F}_r$ are built on FLUX.2-klein-base-4B (Black Forest Labs, 2026), with frozen pretrained backbone weights and separate LoRA adapters (Hu et al., 2021). We apply LoRA to the projection layers in the attention and feed-forward modules with the rank set to 64. For $\mathcal{F}_u$, we re-train the output head from scratch alongside its LoRA adapter, to predict the element-wise log variance of latent restoration residuals as $\boldsymbol{s} = \log \boldsymbol{u}^2$. For $\mathcal{F}_r$, we additionally introduce separate zero-initialized MLPs for the flow and reference streams to map uncertainty features to affine modulation parameters. These parameters modulate the timestep embeddings used for input, block, and output conditioning. Block-level modulation is applied in both double-stream and single-stream transformer blocks. The uncertainty predictor remains frozen while the restorer is optimized.

A.2 ADDITIONAL EXPERIMENTS

Model Size and Inference Efficiency. Table 3 compares model size and inference efficiency at 1024×1024 resolution under the same hardware and timing conditions. Parameter counts and inference times are reported separately for captioning/VLM processing and restoration, with timing covering the complete pipeline, including all restoration rounds. For AgenticIR, which relies on the closed-source GPT-4, we replace this component with Qwen3.8-27B for end-to-end timing, as indicated by the asterisk. Since RAR is trained at 256×256 resolution, we evaluate it using tiled inference and report the total time required to process a full 1024×1024 image.

These results indicate that AutoRestore delivers improved restoration quality while maintaining competitive inference speed despite its multi-round design. Its overall latency is lower than that of DreamClear, LucidFlux, AgenticIR, and RAR. These comparisons show that iterative assessment and refinement can be incorporated without sacrificing runtime competitiveness. Moreover, processing independent images concurrently reduces the amortized VLM reasoning time from 16.6 s to 3.0 s per image, further improving throughput.

Table 3: **Model size and inference efficiency.** Parameter counts and inference times are reported separately for captioning/VLM processing and restoration. The gray value denotes amortized prompt-generation time per image under concurrent processing.

Method	Model size (B)		Latency (s)	
	Captioner/VLM	Restorer	Captioner/VLM	Restorer
SUPIR	13.35	3.86	5.1	9.2
DreamClear	13.35	2.21	6.9	27.2
LucidFlux	-	13.53	-	36.9
AutoDIR	-	1.09	-	12.3
AgenticIR	7.11+27.00*	0.0∼1.72	11.1*	78.9
RAR	6.92	2.41	1.1	31.1
AutoRestore	27.00	4.10	16.6 (3.0)	12.0

Quantitative Comparison at 1024^2 Resolution. Since SUPIR, DreamClear, and LucidFlux are trained using images at 1024^2 resolution, we perform inference at 1024^2 resolution, after which their outputs are downsampled to 512^2 for metric computation and comparison. To complement this evaluation, we provide additional comparisons with these methods on NTIRE2020-Track1 at 1024^2 output resolution. Tables 4 reports the corresponding results. These results again verify the capability of the proposed AutoRestore in achieving better trade-off between fidelity and perceptual quality.

Table 4: Quantitative Comparison at 1024^2 resolution on NTIRE2020-Track1.

Method	PSNR	SSIM	LPIPS	TOPIQ-FR	CLIPIQA+	MUSIQ	NIQE	A-FINE-NR
SUPIR	24.88	0.6404	0.2471	0.5512	0.6708	68.02	**3.253**	-0.883
DreamClear	23.50	0.6178	0.2458	0.5405	0.7059	70.70	3.888	-0.894
LucidFlux	20.54	0.5647	0.3218	0.4134	**0.7749**	**74.88**	3.683	-1.019
AutoRestore	**25.78**	**0.7003**	**0.2111**	**0.6018**	0.7189	73.00	3.490	**-1.020**

Detailed Evaluation Results on Synthetic Datasets. We provide detailed quantitative comparisons on synthetic datasets in Tab. 5 and 6. The evaluations cover a diverse range of degradation types, including single degradations (e.g., blur, noise, and JPEG compression) and complex compound degradations. Compared with representative image restoration methods, AutoRestore consistently achieves a better trade-off between restoration fidelity and perceptual quality across different degradation settings. Notably, even for degradation-specific scenarios where agentic methods such as AgenticIR can invoke specialized restoration tools, AutoRestore still achieves competitive or superior performance, demonstrating the effectiveness and generality of our unified restoration framework. These results indicate that adaptive visual reasoning and iterative refinement provide stronger restoration capability than relying solely on predefined degradation-specific processing.

Table 5: Detailed quantitative comparison on synthetic dataset DIV2K-Val.

Degradation	**Method**	**PSNR**	**SSIM**	**LPIPS**	**TOPIQ-FR**	**CLIPIQA+**	**MUSIQ**	**NIQE**	**A-FINE-NR**
SR$\times$4	SUPIR	23.41	0.6127	0.3001	0.4373	0.7023	66.96	4.488	-0.834
	LucidFlux	21.05	0.5699	0.3036	0.4228	0.7758	72.81	4.004	-1.085
	AutoDIR	22.83	0.6244	0.2598	0.4714	0.6739	65.73	3.895	-0.809
	AgenticIR	21.54	0.5886	0.3218	0.4314	0.6663	67.66	4.349	-0.949
	RAR	22.42	0.6196	0.3198	0.4055	0.6254	65.01	4.490	-0.690
	AutoRestore	24.16	0.6620	0.2232	0.5761	0.7502	71.69	3.929	-1.061
Blur $\sigma = 2$	SUPIR	25.80	0.7137	0.2040	0.6368	0.7346	70.59	4.591	-0.970
	LucidFlux	21.83	0.6123	0.2696	0.5119	0.7915	73.94	4.282	-1.183
	AutoDIR	24.53	0.6745	0.2222	0.5884	0.6715	67.88	4.402	-0.940
	AgenticIR	25.11	0.7550	0.3453	0.5105	0.4856	47.77	6.600	-0.682
	RAR	25.48	0.7567	0.1935	0.6758	0.6905	71.72	4.103	-1.051
	AutoRestore	27.43	0.7971	0.1493	0.7562	0.7414	71.18	4.003	-1.071
JPEG $q = 20$	SUPIR	26.82	0.7517	0.1760	0.6805	0.7147	69.58	4.358	-0.934
	LucidFlux	23.52	0.6928	0.2346	0.5881	0.7818	74.11	4.100	-1.149
	AutoDIR	31.01	0.8698	0.1761	0.7452	0.5653	60.71	5.414	-0.844
	AgenticIR	30.07	0.8845	0.1795	0.7748	0.5872	69.46	5.105	-0.969
	RAR	26.66	0.7842	0.1615	0.7355	0.6962	71.18	3.672	-0.987
	AutoRestore	29.18	0.8408	0.1256	0.8031	0.7368	72.45	4.110	-1.093
Noise $\sigma = 30$	SUPIR	26.50	0.7310	0.1826	0.6627	0.7294	70.51	4.459	-0.965
	LucidFlux	23.31	0.6713	0.2374	0.5707	0.7861	74.06	4.029	-1.167
	AutoDIR	30.48	0.8707	0.1425	0.7803	0.6515	68.77	4.151	-0.940
	AgenticIR	27.31	0.8172	0.2797	0.5594	0.5158	65.39	6.491	-1.034
	RAR	26.41	0.7270	0.1796	0.7298	0.6640	68.71	5.167	-0.923
	AutoRestore	28.91	0.8272	0.1280	0.7859	0.7394	71.96	4.042	-1.094
Blur $\sigma = 2$ + Noise $\sigma = 30$ + JPEG $q = 40$	SUPIR	24.62	0.6172	0.2653	0.5399	0.7122	67.94	5.776	-0.909
	LucidFlux	20.75	0.5568	0.3053	0.4483	0.7975	74.21	4.386	-1.148
	AutoDIR	22.94	0.4106	0.7077	0.3217	0.3038	23.80	5.444	-0.275
	AgenticIR	21.32	0.5858	0.5369	0.3673	0.3423	37.23	8.452	-0.634
	RAR	20.95	0.4906	0.4322	0.4186	0.7064	70.38	4.789	-0.789
	AutoRestore	24.38	0.6291	0.2525	0.5413	0.7651	73.59	3.942	-1.140

A.3 Training Data Construction

Image Collection and Filtering. We collect HQ images from LSDIR, 4KLSDB, UHR100K, and Aesthetic-Train-V2 to construct training pairs at 512×512 resolution. To reduce uninformative crops while retaining diverse scene content, we adopt a two-stage selection procedure. For each image from 4KLSDB, UHR100K, and Aesthetic-Train-V2, we sample 12 square crops with side lengths drawn from the intervals $[512, \min(H, W)]$. For LSDIR, we sample five 512×512 crops per image. Crops larger than 512×512 are resized using area interpolation.

Table 6: Detailed quantitative comparison on synthetic dataset LSDIR-Val.

Degradation	Method	PSNR	SSIM	LPIPS	TOPIQ-FR	CLIPIQA+	MUSIQ	NIQE	A-FINE-NR
SR×4	SUPIR	20.60	0.5251	0.3290	0.3958	0.7118	66.34	4.507	-0.880
	LucidFlux	19.14	0.5178	0.2844	0.4300	0.7840	73.75	4.146	-1.141
	AutoDIR	20.87	0.5531	0.2848	0.4339	0.6822	67.83	3.758	-0.880
	AgenticIR	19.78	0.5136	0.3360	0.4053	0.6773	69.05	4.089	-1.020
	RAR	20.24	0.5532	0.3167	0.3781	0.6118	65.30	4.375	-0.720
	AutoRestore	21.51	0.5856	0.2271	0.5424	0.7725	72.73	3.663	-1.157
Blur $\sigma = 2$	SUPIR	22.87	0.6472	0.2067	0.5998	0.7682	72.00	4.761	-1.060
	LucidFlux	20.43	0.5885	0.2436	0.5320	0.8032	74.63	4.487	-1.264
	AutoDIR	22.60	0.6153	0.2511	0.5457	0.6877	68.55	4.384	-0.996
	AgenticIR	23.18	0.6821	0.3750	0.4631	0.4938	47.43	6.251	-0.764
	RAR	23.76	0.7140	0.1768	0.6721	0.6942	72.36	3.738	-1.106
	AutoRestore	24.93	0.7531	0.1469	0.7478	0.7706	72.76	3.795	-1.188
JPEG $q = 20$	SUPIR	23.79	0.6957	0.1756	0.6523	0.7599	72.10	4.490	-1.049
	LucidFlux	22.32	0.6855	0.2022	0.6250	0.7965	74.85	4.208	-1.245
	AutoDIR	28.71	0.8594	0.1457	0.7354	0.6303	66.80	4.739	-0.991
	AgenticIR	28.16	0.8739	0.1447	0.7958	0.6478	73.23	4.566	-1.106
	RAR	25.11	0.7729	0.1236	0.7628	0.7249	73.20	3.425	-1.079
	AutoRestore	27.39	0.8357	0.0982	0.8300	0.7736	74.32	3.810	-1.225
Noise $\sigma = 30$	SUPIR	23.72	0.6788	0.1779	0.6519	0.7690	72.34	4.612	-1.074
	LucidFlux	22.08	0.6664	0.2081	0.6047	0.8012	74.78	4.057	-1.262
	AutoDIR	30.08	0.8855	0.0979	0.8172	0.6867	72.67	3.716	-1.073
	AgenticIR	26.45	0.8111	0.2355	0.5866	0.5867	70.86	5.843	-1.190
	RAR	24.92	0.7304	0.1344	0.7566	0.6860	71.85	4.627	-1.030
	AutoRestore	27.33	0.8264	0.0965	0.8197	0.7751	74.21	3.753	-1.240
Blur $\sigma = 2$ + Noise $\sigma = 30$ + JPEG $q = 40$	SUPIR	21.94	0.5535	0.2570	0.5184	0.7585	71.17	5.300	-1.020
	LucidFlux	19.42	0.5311	0.2810	0.4689	0.8129	74.88	4.439	-1.229
	AutoDIR	21.36	0.3848	0.6775	0.3002	0.2993	23.97	5.270	-0.310
	AgenticIR	20.03	0.4949	0.5663	0.3398	0.3522	35.51	7.678	-0.642
	RAR	20.03	0.4846	0.3747	0.4245	0.7008	71.26	4.541	-0.856
	AutoRestore	22.22	0.5795	0.2429	0.5383	0.7844	74.47	3.716	-1.250

We evaluate each candidate using Laplacian variance, Shannon entropy, a GLCM-based texture statistic (Zhang et al., 2025a), CLIPIQA+, and MUSIQ. Within each source image, we separately average the normalized ranks of the first three metrics and those of the two perceptual metrics. The two group scores are then combined with equal weights, and the top-2 candidates are retained. We subsequently pool the retained candidates and rank them by the mean global percentile rank of Shannon entropy, the GLCM-based statistic, CLIPIQA+, and MUSIQ. The selected HQ crops are used to construct 330,000 LQ/HQ pairs before instruction generation. All crops from the same source image are assigned to the same data split.

Degradation Synthesis. We construct degraded images with two types of degradation patterns: single degradations and compound degradations. Single degradations independently apply individual distortion types, including noise (Gaussian or Poisson), blur (Gaussian, motion, or lens), and JPEG compression. Compound degradations follow the two-stage degradation pipeline of Real-ESRGAN (Wang et al., 2021), combining blur, random resizing, noise, and JPEG compression. To increase degradation diversity, the second degradation stage is randomly dropped. Single and compound degradations are sampled with probabilities of 0.25 and 0.75, respectively, where the two-stage compound setting is sampled with a higher probability than the single-stage variant.

To simulate spatially localized restoration scenarios, global and local degradation regions are sampled with probabilities of 0.75 and 0.25, respectively. For local degradations, the degraded image is blended with its HQ counterpart using a feathered random rectangular or circular mask, leaving regions outside the mask unchanged.

Text Data Construction. We generate restoration instructions through scene description followed by restoration planning. For training data, the VLM first describes the HQ target and then examines the aligned LQ/HQ pair, using the description as an auxiliary cue to generate the restoration instruction $\boldsymbol{p}_r$. The HQ image defines the desired appearance, while differences between the paired images indicate the required corrections. Instructions for globally degraded samples combine a brief scene description with region-specific restoration targets, whereas those for locally degraded sam-

ples contain only regional targets. Each target identifies a location or object and describes its desired structure, texture, material, or color in positive or neutral language.

At inference time, both the scene description and the initial restoration instruction are generated from the LQ image alone. For subsequent rounds, the VLM reviews the intermediate restoration against the HQ target during training and against the original LQ image during inference. The resulting instructions focus on remaining degradations and reconstruction errors, including incorrect structures, textures, and object categories. All instructions use concise English clauses separated by semicolons and begin with the prefix "Restore this image to high quality."

A.4 Training Details

Stage 1 Training. We first train the uncertainty predictor on LQ/HQ pairs using the Gaussian negative log-likelihood objective, updating its LoRA adapter and newly initialized output head. We then freeze the uncertainty predictor and optimize the restorer's LoRA adapter and uncertainty-driven modulation modules. The restoration objective combines latent flow matching loss, pixel-space MSE and LPIPS. Training uses AdamW with BF16 mixed precision. The uncertainty predictor uses a peak learning rate of 10^{-4}, while the restorer uses 7.5×10^{-5} after warm-up. Their effective batch sizes are 48 and 64, respectively.

Subsequent Training Stages. In subsequent stages, we train the model to further refine intermediate restoration results using historical visual evidence. To this end, we use intermediate results generated offline from the LQ images as training inputs. Offline generation is guided by restoration instructions $\boldsymbol{p}_r$ inferred from the LQ images alone. We apply wavelet-based color correction using the original LQ images as color references. These LQ-only instructions are used exclusively for offline inference. For training, the VLM compares each intermediate result with its corresponding HQ target to generate a new instruction addressing remaining degradations and reconstruction errors. The intermediate inputs and training instructions remain fixed within each stage.

A.5 Additional Visual Comparisons

We provide additional visual comparisons between AutoRestore and existing competing methods in Fig. 6, 7, 8, and 9 across various benchmark datasets. These examples complement the quantitative evaluation by offering a detailed examination of restoration fidelity, including structural recovery and texture reconstruction.

A.6 VLM Task Prompts

Figs. 10, 11, and 12 present the task prompts used for restoration-instruction generation and quality assessment. During training, the VLM is provided with both the degraded LQ image and the corresponding GT reference. By comparing the LQ image with the GT image, the VLM analyzes the missing or distorted visual information and generates restoration instructions that describe the required corrections, including degraded regions, target structures, and texture details. During inference, the VLM only receives the original LQ image and reasons about the most likely high-quality scene content based on the observed visual evidence. It then produces restoration instructions specifying the regions requiring refinement and the corresponding restoration operations. For refinement in subsequent rounds, the VLM additionally considers the current restoration result to identify remaining artifacts and determine further corrections. For iterative restoration, we further use a quality comparison prompt to evaluate two generated results. The VLM compares their visual fidelity and remaining degradation artifacts, providing a decision on whether additional restoration iterations are necessary.

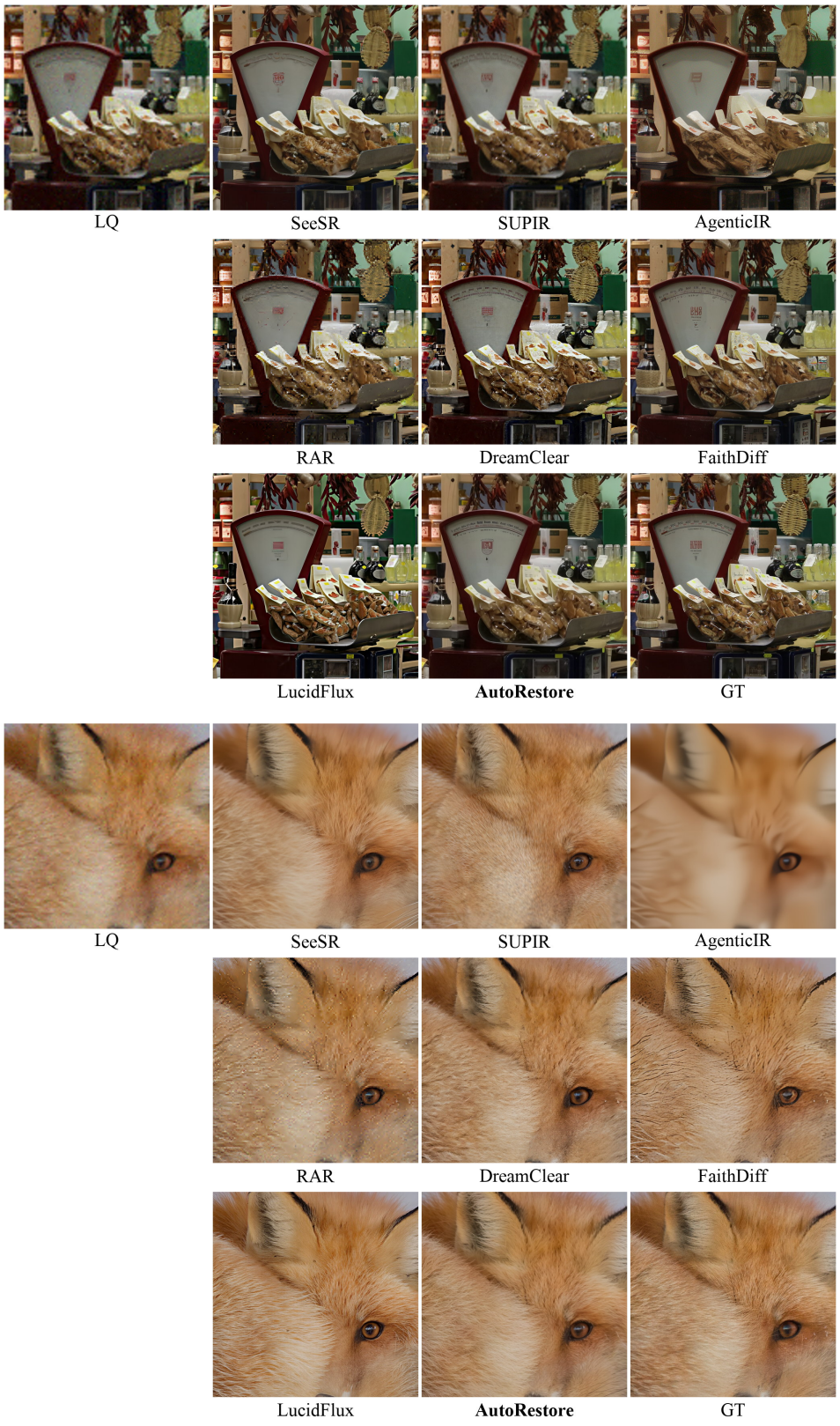

Figure 6: Additional visual comparison with existing IR methods on NTIRE2020-Track1.

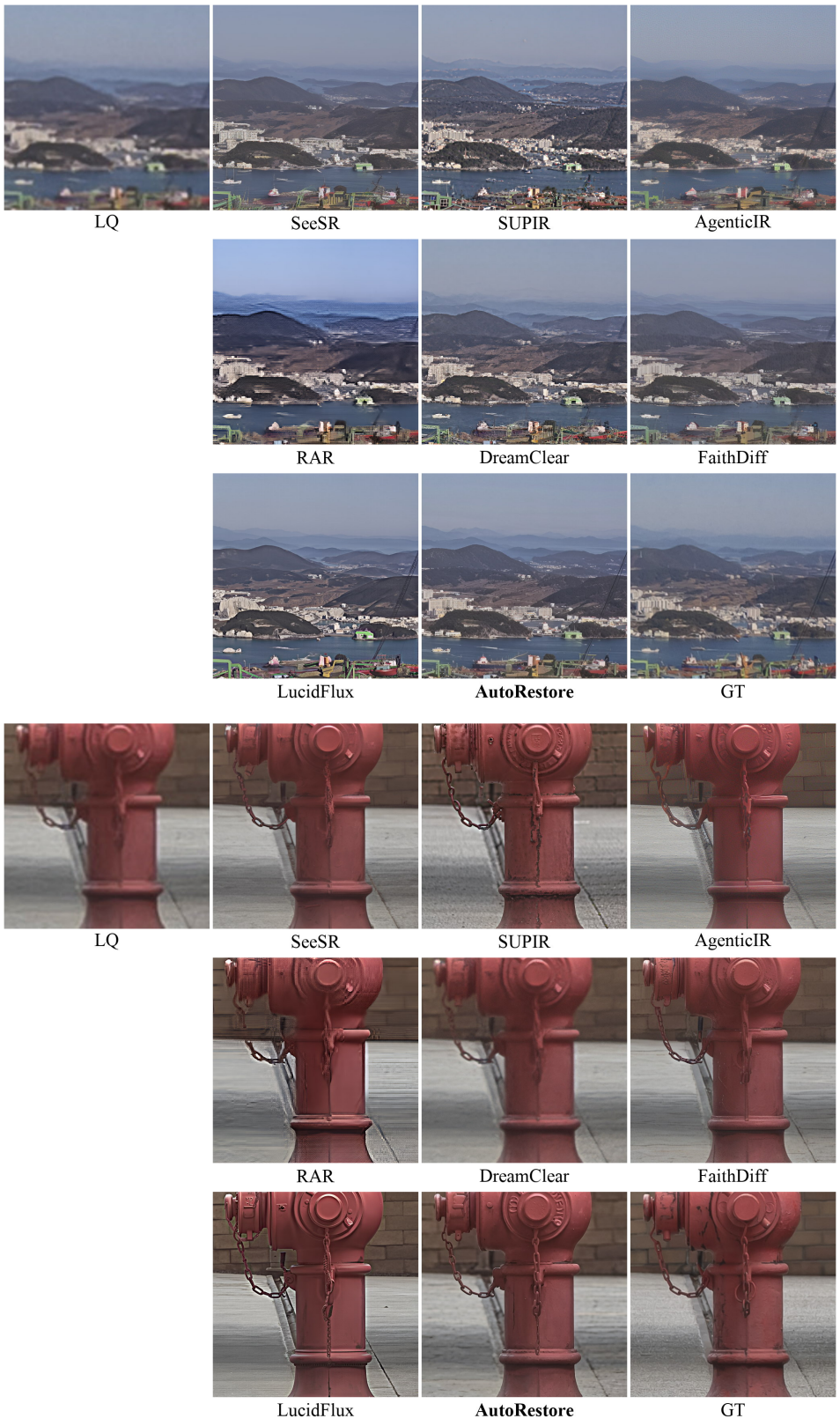

Figure 7: Additional visual comparison with existing IR methods on RealSR.

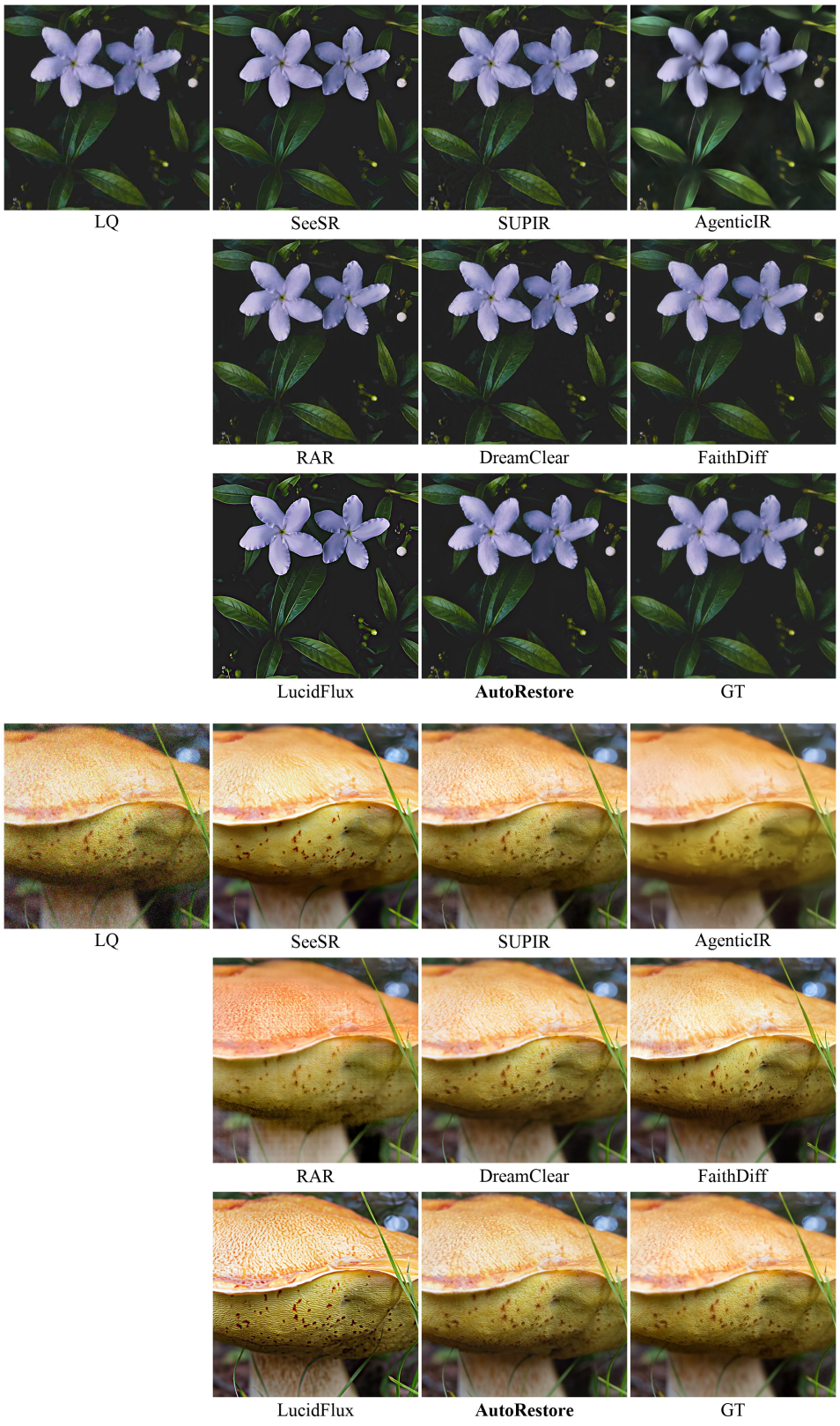

Figure 8: Additional visual comparison with existing IR methods on DIV2K-Val.

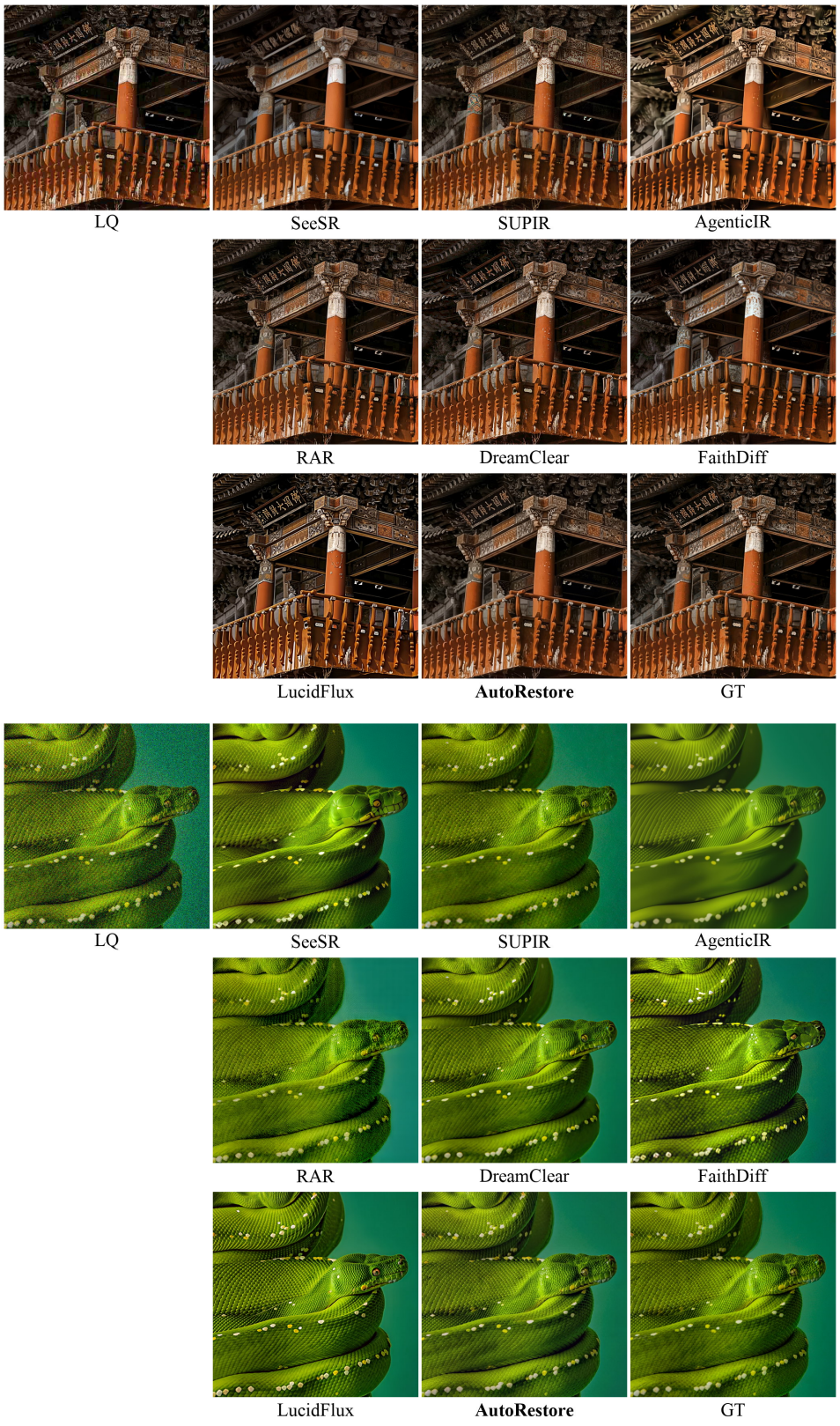

Figure 9: Additional visual comparison with existing IR methods on LSDIR-Val.

GT content description (round 1)
Inputs: [Image 1] Clean GT image.
Describe the visible scene in this image as a complete, grounded account of its content. Begin with the main subjects and the overall setting, then cover the important secondary objects and how they are arranged. Include useful details about number, position, pose, shape, color, material, surface pattern, lighting, and spatial relationships when those details can be seen. Keep the description coherent across the whole frame: connect an object to a person or another object only when the visual connection is apparent. Distinguish foreground, background, and areas of intentional focus where that helps explain the scene. Describe characteristic texture at the level of a visible region rather than inventing individual hairs, fibers, markings, or tiny components. If an element is small, distant, obscured, or ambiguous, describe its visible outline, color, and location instead of assigning an unsupported identity or function. Use familiar object knowledge only to interpret features that the image actually shows; do not complete hidden parts from expectation. Quote brief lettering in straight double quotation marks only when its characters are clearly readable and useful to the scene description. Otherwise describe the sign or lettering area by its appearance and placement without guessing missing words. Preserve visible relationships and distinctions between similar objects, and avoid merging separate subjects into one. Write one natural English paragraph of roughly 40 to 140 words that conveys the scene to someone who has not seen it. Use concrete visual language. Do not include a heading, list, alternative interpretations, a preamble, or commentary about how you reached the answer.

GT restoration instruction (round 1)
Inputs: [Image 1] Degraded input; [Image 2] Aligned clean counterpart.
Text context: [GT content description (round 1)], generated from Image 2.
Image 1 is the degraded input and Image 2 is its aligned clean counterpart. Compare the same regions in both images and use Image 2 to establish the appearance to recover. The supplied scene description is a helpful hint, but check its object identities and relationships against the visible pair. Start with one short global clause naming the main subjects and their arrangement; then add one to four local clauses for the most important restoration needs. Order those local clauses from the dominant subject or region toward secondary details. Each local clause should make clear where to act, what needs to be recovered or corrected, and the intended visible appearance. Combine related changes on the same object instead of producing a checklist of isolated defects. Describe supported geometry, boundaries, regional texture, color, material, and spatial relationships in concrete terms. Keep characteristic patterns at the scale of the object or surface; avoid spelling out individual hairs, fibers, or other microscopic details that are not important to the scene. Preserve natural low-texture areas, subtle shading, the scene lighting, and any intentional depth of field. Do not sharpen an intentionally distant or out-of-focus region merely because another region is crisp. When text is useful, quote its characters in straight double quotation marks only if clearly readable in Image 2; otherwise describe the text area and layout. Prefer positive, specific descriptions of the desired result over repeated defect labels or generic quality adjectives. Return a single-line English restoration instruction with clauses separated by semicolons, at most 180 words. Do not include image labels, analysis, a heading, or a preamble.

GT restoration instruction (Subsequent rounds)
Inputs: [Image 1] Most recent intermediate restoration result; [Image 2] Aligned clean counterpart.
Text context: [GT CONTENT DESCRIPTION], generated from Image 2.
Image 1 is the current intermediate result and Image 2 is its aligned clean counterpart. Review the whole result before deciding what still needs to change. Compare corresponding regions to find incomplete recovery, excessive edge contrast, invented texture, color shifts, or structural mistakes that matter in the scene. The supplied scene description is a hint, but the visible pair takes priority when it differs from that description. Write one to four local restoration clauses for the most important remaining corrections, and omit a global reconstruction clause. Each clause should identify the region or object, state the correction, and describe the appearance supported by Image 2. Correct mistaken identities, attributes, or relationships explicitly when the paired images establish them. Retain good parts of the intermediate result; do not request changes to already adequate regions just to fill a clause. Describe characteristic texture at the scale of its material and location, including spacing, variation, and edge definition when useful. Avoid turning natural low-texture surfaces into artificial detail or exaggerating fine patterns. Preserve supported colors, lighting, shape, and intentional depth of field. A smooth wall, sky, or out-of-focus background can remain visually calm when Image 2 supports it. Quote short text in straight double quotation marks only when clearly readable in Image 2; otherwise describe its placement and arrangement without completing uncertain characters. Use concrete, positive descriptions of the intended result rather than repeatedly naming defects. Return one single-line English instruction with semicolon-separated clauses, at most 180 words, and no image labels, analysis, heading, or preamble.

Figure 10: Task prompt for generating $\boldsymbol{p}_c$ and $\boldsymbol{p}_r$ with GT reference.

LQ content description (round 1)
Inputs: [Image 1] LQ input image.
Describe the visible scene in this image as one coherent account of its content. Identify the main subjects and setting, then describe important secondary objects and how they are arranged across the frame. Include the number, position, pose, broad shape, color, material, lighting, and spatial relationships of visible elements when the image supports those details. Distinguish foreground from background and preserve clear differences between separate subjects. When some detail is unclear, use visible silhouettes, larger shapes, color patches, nearby parts, and scene context to choose a conservative, internally consistent description. A familiar object category can be used when the visible cues support it, but a familiar name alone is not evidence for hidden parts, exact identities, or specific fine textures. Describe an uncertain small or distant element by its appearance and location at a broader level instead of listing guesses. Do not attach a loose shape to a person, vehicle, or building unless the physical relationship is visible. Mention characteristic surface patterns only at a scale that can be observed; do not invent individual hairs, fibers, printed marks, or tiny components. Quote brief lettering in straight double quotation marks only when the characters are readable. If they are not, describe the lettering area, its color, and placement without completing a word from context or a familiar brand. Write one natural English paragraph of roughly 40 to 140 words that covers the whole scene. Use concrete visual language, with no heading, list, competing interpretations, preamble, or commentary about your process.

LQ restoration instruction (round 1)
Inputs: [Image 1] LQ input image.
Text context: [LQ CONTENT DESCRIPTION (round 1)].
Inspect the degraded input image and use the supplied scene description as a fallible guide. Recheck its subjects and relationships against visible shapes, colors, layout, and context before choosing a plausible clean appearance. Start with one short global clause about the main subjects and their arrangement. Add one to four local clauses for the most important restoration needs, ordered from primary to secondary regions. Each local clause should identify where to act, what to recover or correct, and the intended appearance or subject. Combine related changes for the same region in natural wording rather than listing isolated defects. Describe concrete structure, geometry, color, and characteristic regional texture without inventing a precise identity, hidden component, or microtexture that the input cannot support. When a detail is uncertain, ask for a broad coherent form rather than a guessed object or a mixture of alternatives. Preserve naturally even surfaces, gentle tonal transitions, scene lighting, and any supported depth of field. Blur from degradation alone does not prove that the scene was intentionally out of focus. Do not force excessive sharpness or dense texture into a sky, wall, skin area, or distant background. Quote short text in straight double quotation marks only when the visible glyphs support its characters; otherwise describe the lettering region and its arrangement without inventing words. Favor positive and specific desired properties over a repetitive list of defects or generic quality labels. Return one single-line English restoration instruction with clauses separated by semicolons, at most 180 words, without image roles, analysis, or a preamble.

Figure 11: Task prompt for generating $\boldsymbol{p}_c$ and $\boldsymbol{p}_r$ with LQ input only.

LQ content description (Subsequent rounds)
Inputs: [Image 1] Original LQ input; [Image 2] Most recent intermediate restoration result.
Text context: [PREVIOUS LQ CONTENT DESCRIPTION].
Image 1 is the original input and Image 2 is the current intermediate result of the same scene. Reinspect the whole scene as if writing a fresh description. Use Image 1 to anchor the number of subjects, their identities at a supported level, broad shapes, colors, positions, and spatial relationships. Use clearer details from Image 2 only when they agree with Image 1 and the surrounding scene context. The previous scene description is useful context but may contain mistaken identifications or important omissions. Compare its claims with the two images rather than repeating it automatically. A natural-looking detail in Image 2 does not by itself prove that the original scene contained that detail. If an earlier specific interpretation is not supported, broaden the description to the visible appearance; if another interpretation better explains Image 1, revise it. Keep important supported subjects and add any overlooked objects, regions, or relationships. Describe meaningful color, material, lighting, and characteristic regional texture without mistaking processing artifacts for real surface patterns. Avoid inventing hidden parts, tiny marks, individual hairs, or exact identities. For a small or ambiguous object, give its shape, color, and location rather than a speculative function. Quote brief text only when its characters are readable in Image 1; otherwise describe the lettering area without guessing missing words. Return one complete English paragraph of roughly 40 to 140 words describing the scene, from main subjects to secondary content. Include no image roles, comparisons, quality judgments, repair instructions, alternatives, heading, or preamble.

LQ restoration instruction (Subsequent rounds)
Inputs: [Image 1] Original LQ input; [Image 2] Most recent intermediate restoration result.
Text context: [UPDATED LQ CONTENT DESCRIPTION], followed by [PREVIOUS LQ RESTORATION INSTRUCTION].
Image 1 is the original degraded input and Image 2 is the current intermediate result. Review the whole intermediate result, including areas omitted from the earlier instruction. Use Image 1 to anchor subject identity, broad geometry, colors, and layout; treat clearer details in Image 2 as fallible when they disagree with that evidence. The updated scene description and previous instruction are useful hints, not fixed goals. Recheck both before deciding whether an earlier target remains valid. Look for incomplete recovery such as residual blur or lost regional detail, generated artifacts such as harsh edges or unnatural texture, and content or structural errors in identities, parts, attributes, or relationships. Select one to four important, supported local corrections and omit a global reconstruction clause. Each clause should specify the region, what to recover or correct, and the intended appearance. State a corrected identity or attribute explicitly when the visible evidence supports it. Merge related changes and leave adequate regions alone. Match texture scale, spacing, variation, and edge contrast to the material and context; a correction can reduce excessive sharpening or dense invented detail. Preserve supported colors, lighting, natural low-texture areas, and depth of field. Degradation blur alone does not establish intentional focus. Quote text only when its characters are readable in Image 1; otherwise preserve the text area's position and arrangement without guessing words. Before deciding nothing needs change, recheck prominent subjects for residual blur, smearing, or unmet supported goals. If no concrete correction remains, return exactly "Preserve the supported scene content and current regional appearance." Otherwise return one single-line English restoration instruction with semicolon-separated clauses, at most 180 words, without image roles, analysis, a heading, or a preamble.

Pairwise quality comparison
Inputs: [Image 1] Candidate restoration 1; [Image 2] Candidate restoration 2.
Compare image 1 and image 2. Choose the image with better overall perceptual quality and more natural, realistic details. Do not mistake artificial sharpness or generated artifacts for real detail. Choose tie if neither image is clearly better overall. Return exactly one lowercase string: image 1, image 2, or tie. Do not include any explanation or other text in the final answer.

Figure 12: Task prompt for generating $\boldsymbol{p}_c$ and $\boldsymbol{p}_r$, and quality comparison with LQ input only.