USIR: UNIFIED SINGLE-STEP IMAGE RESTORATION VIA REWARD-GUIDED PRIVILEGED SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image restoration is challenging due to diverse and coupled degradations. Existing all-in-one methods often rely on explicit degradation understanding or task-specific guidance, and their performance declines under complex degradations, while recent agentic systems further introduce iterative assessment, planning, and tool execution, incurring substantial computational overhead and potential error accumulation across stages. To this end, we propose **USIR**, a Unified Single-step Image Restoration framework that directly restores degraded inputs in a single forward pass. At its core is **Reward-Guided Privileged Self-Distillation** (RPSD), a post-training framework that combines self-distillation under asymmetric information with reward-guided policy improvement. Specifically, reward-guided self-distillation progressively improves a degradation-aware privileged policy through reward-guided positive and negative targets, while privileged self-distillation transfers its preferred restoration behavior into a blind generator conditioned on the low-quality input. In this way, degradation-aware restoration knowledge and reward preference are progressively internalized into the single-step deployment model. Extensive experiments demonstrate that USIR outperforms both all-in-one and agentic restoration methods in perceptual quality and fidelity, while achieving significantly higher inference efficiency than agentic methods. Code and model weights are available at `https://anonymous.4open.science/r/anonymous-744F`.

1 INTRODUCTION

Image restoration aims to recover a high-quality image from a degraded observation affected by noise, blur, compression artifacts, adverse weather, low illumination, or other degradations (Jiang et al., 2025b). Despite the strong performance of task-specific restoration models, they typically assume that the degradation type is known and handled by a dedicated model (Li et al., 2022). This assumption is difficult to satisfy in real-world scenarios, where degradations can be unknown, mixed, spatially varying, and entangled with image content (Ai et al., 2024). All-in-one image restoration therefore seeks a unified model that can restore images across diverse degradation types and levels without manually selecting a task-specific restorer (Conde et al., 2024).

A key challenge in all-in-one restoration is how to obtain and use degradation-aware guidance. Recent methods address this challenge by learning degradation representations or prompts that adapt the restoration network to the input corruption (Yao et al., 2023; Luo et al., 2023; Jiang et al., 2023). Other methods introduce language instructions, vision-language priors, or automatic quality assessment modules to infer the restoration intent before or during generation (Liu et al., 2026b; Jiang et al., 2025a; Luo et al., 2025). More recently, agentic restoration systems use multimodal large language models to assess degradations, plan task sequences, select restoration tools, and refine outputs through multi-loop execution (Jia et al., 2026; Zhu et al., 2025; Chen et al., 2024). These studies show that explicit degradation-aware guidance is valuable for complex degradations.

However, producing such guidance at test time also complicates the restoration pipeline. A degradation estimator, prompt generator, or MLLM-based planner must be maintained together with the restoration generator. Such multi-module inference increases memory usage, computation, and la-

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.

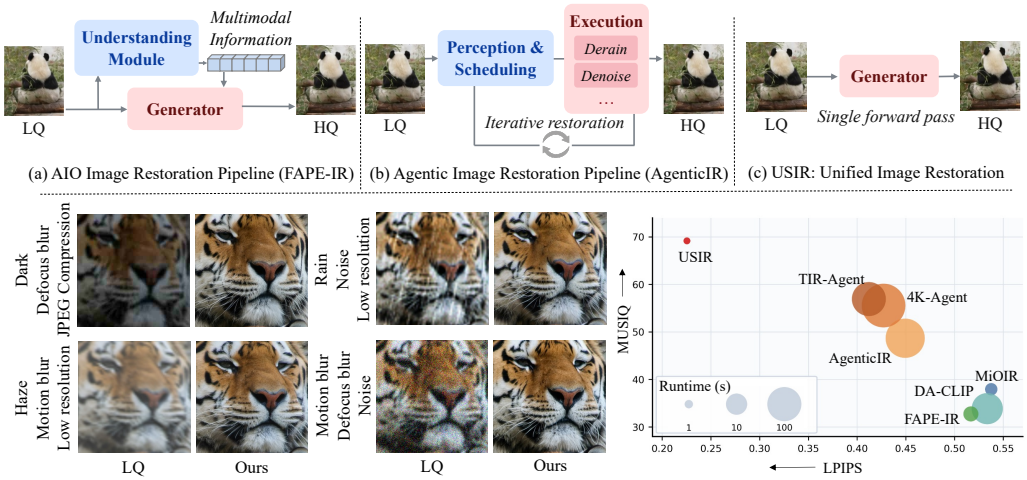

Figure 1: **Motivation and performance of USIR. Top:** USIR provides a unified single-step alternative to conventional all-in-one and agentic restoration pipelines. **Bottom-left:** USIR delivers consistent, high-quality restoration across diverse and complex degradations. **Bottom-right:** USIR bridges restoration quality, fidelity and inference efficiency, achieving strong perceptual performance without the substantial inference overhead of agentic pipelines.

tency, particularly when restoration is performed through iterative assessment–planning–execution loops (Lu et al., 2026; Liu et al., 2025; 2026c). Moreover, agentic methods may accumulate errors across stages, as early-stage mistakes can propagate through subsequent restoration steps. (Li et al., 2026a; Zhang et al., 2026). These limitations motivate a training-rich but inference-simple paradigm: use degradation-aware supervision during training and deploy a single generator conditioned on the low-quality (LQ) input during inference.

To this end, we propose **USIR**, a unified single-step image restoration framework that internalizes degradation-aware capabilities during training and deploys a single LQ-conditioned generator for direct restoration. As shown in Figure 1, USIR requires neither a degradation understanding module nor iterative tool execution at inference time. To achieve this, we first train the single-step generator to restore images with known degradations through noise-hybrid adversarial post-training and then use it as the privileged policy in the subsequent stage. Building upon this initialization, we develop **Reward-Guided Privileged Self-Distillation (RPSD)**, a post-training framework tailored for image restoration that couples reward-guided self-distillation with privileged self-distillation. Specifically, reward-guided self-distillation progressively aligns this privileged policy with restoration preference through positive and negative reward-guided targets, while privileged self-distillation transfers both its degradation-aware restoration capabilities and learned preference into an LQ-only blind generator under fixed deployment conditions. RPSD reuses the same candidate group and reward evaluations across the updates of these two distillation processes, avoiding redundant rollout and scoring. In this way, USIR internalizes degradation-aware capability and restoration preference into the deployment model. At inference time, all privileged conditions and reward-related components are not required, leaving only the LQ-conditioned generator for single-step image restoration.

Our contributions can be summarized as follows:

- We introduce USIR, a single-step image restoration framework that internalizes degradation-aware capability during training and deploys one LQ-conditioned generator without explicit test-time understanding or iterative tool execution.
- We propose RPSD, a post-training framework that couples reward-guided self-distillation for preference alignment with privileged self-distillation for transferring degradation-aware restoration capabilities and learned preference into a blind generator. The two processes share grouped privileged samples and reward evaluations, reducing redundant computation.
- Extensive experiments demonstrate that USIR outperforms both all-in-one and agentic restoration methods in perceptual quality and fidelity, while achieving substantially faster inference than agentic methods.

2 Related Work

2.1 Image Restoration

All-in-one image restoration aims to recover high-quality images across diverse degradation types and levels. Existing methods mainly improve the adaptability of the restoration network through degradation-aware representations, prompt learning, or task-shared architectures (Valanarasu et al., 2022; Potlapalli et al., 2023; Yao et al., 2023; Tian et al., 2025). More recent approaches further introduce language or vision-language guidance, automatic degradation assessment, or agentic planning to infer restoration intent and handle compound degradations (Conde et al., 2024; Jiang et al., 2023; Tang et al., 2026; Cheng et al., 2026; Wu et al., 2025). These methods demonstrate the importance of degradation-aware guidance, but they often require explicit prompts, assessment modules, or multi-stage reasoning at test time. Agentic methods further decompose restoration into iterative perception, planning, tool execution, and reflection (Jia et al., 2026; Zhu et al., 2025; Chen et al., 2024; Zuo et al., 2025). Although capable of handling complex degradations, these methods require multiple calls to multimodal language models and restoration tools, resulting in substantial computational overhead and inference latency. Moreover, errors in degradation assessment or task scheduling may propagate through subsequent restoration stages, potentially degrading the final quality. In contrast, USIR uses privileged degradation-aware guidance only during training and deploys a single LQ-only generator for direct, single-step inference.

2.2 RL and Distillation for Diffusion Models

Reinforcement learning and on-policy optimization have recently been introduced to align diffusion and flow-based generative models with reward functions. Flow-GRPO performs group-relative policy optimization for flow matching models, while DiffusionNFT improves diffusion policies through reward-guided positive and negative objectives (Liu et al., 2026a; Zheng et al., 2025). In parallel, recent on-policy distillation methods train the student on samples generated by the evolving policy rather than relying solely on fixed offline targets (Li et al., 2026b; Jiang et al., 2026; Zhou et al., 2026). Both paradigms commonly rely on group sampling, where multiple on-policy candidates are generated for each input to estimate relative preference or construct distillation targets. Such sampling can be computationally expensive due to repeated model rollouts and reward evaluations. RPSD instead couples reward-guided self-distillation with privileged self-distillation under asymmetric conditions, avoiding redundant sampling and reward evaluation while progressively internalizing privileged restoration knowledge into the single-step deployment model.

3 Method

3.1 Overview

USIR aims to directly model the blind restoration mapping from a complex degraded observation to its restored output:

$$I_{\mathrm{HQ}} = v_{\mathrm{blind}}(I_{\mathrm{LQ}}), \tag{1}$$

without explicit degradation understanding or iterative tool execution at inference time. Achieving this mapping is challenging because the desired restoration behavior depends strongly on the degradation type, severity, and composition.

We address this challenge through a two-stage training framework. As illustrated in Figure 2, in stage 1, we initialize a task-aware privileged generator using noise-hybrid adversarial post-training. With access to known degradation information during training, this stage equips a single-step generator with task-addressable restoration capabilities across diverse degradations. In stage 2, we remove the dependence on such privileged information through **Reward-Guided Privileged Self-Distillation (RPSD)**. RPSD couples reward-guided self-distillation, which progressively aligns the privileged policy with restoration preference, with privileged self-distillation, which transfers its degradation-aware restoration capabilities and learned preference into an LQ-only blind generator.

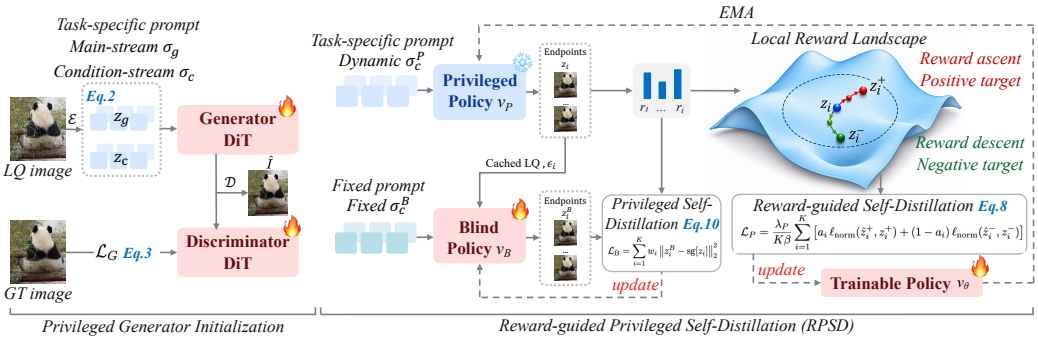

Figure 2: Overview of USIR. We first initialize a task-aware privileged generator through noise-hybrid adversarial post-training. RPSD then couples reward-guided self-distillation with privileged self-distillation to progressively transfer degradation-aware capabilities and learned preference into the blind policy. Only the blind policy is retained for inference.

3.2 Privileged Generator Initialization

We first initialize a privileged restoration policy that leverages known degradation information to guide single-step restoration. The initialization combines task-specific prompts with a noise-hybrid dual-stream design, where the main stream preserves the pretrained generative prior and the condition stream encodes degradation severity. Given the LQ latent $z_{\mathrm{LQ}} = \mathrm{Enc}(I_{\mathrm{LQ}})$ and shared Gaussian noise ϵ, we construct

$$z_g = (1-\sigma_g)z_{\mathrm{LQ}} + \sigma_g\epsilon, \qquad z_c = (1-\sigma_c)z_{\mathrm{LQ}} + \sigma_c\epsilon. \tag{2}$$

The main-stream noise level σ_g is fixed to retain the pretrained generative prior, whereas the condition-stream noise level σ_c is adaptively determined from the LQ–GT distance to reflect restoration difficulty, providing an explicit cue for the required restoration strength (Fang et al., 2026). A task-specific prompt p_{task} further specifies the known degradation type. The two latent streams are concatenated and processed by a single Generator DiT conditioned on p_{task}.

Adversarial post-training (Wang et al., 2026) is adopted for privileged generator initialization. The generator is optimized with a combination of reconstruction, perceptual, and adversarial objectives:

$$\mathcal{L}_G = \lambda_{\mathrm{MSE}}\mathcal{L}_{\mathrm{MSE}}(\hat{I}, I_{\mathrm{GT}}) + \lambda_{\mathrm{LPIPS}}\mathcal{L}_{\mathrm{LPIPS}}(\hat{I}, I_{\mathrm{GT}}) + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{BCE}}(D(\hat{z}), 1), \tag{3}$$

where $\hat{I}$ denotes the restored image, $\hat{z}$ is its latent representation, and D denotes the discriminator. The resulting generator acquires task-addressable single-step restoration capability under privileged degradation conditions and serves as the initialization for RPSD.

3.3 RPSD

Typical reward-based policy optimization and on-policy self-distillation methods for diffusion models rely on group sampling, where multiple on-policy candidates are generated for each input to derive reward-guided optimization signals or construct distillation targets. This sampling procedure is computationally expensive, as it requires repeated model rollouts and reward evaluation (Zhou et al., 2026; Jiang et al., 2026). We observe that the same candidate group can serve both purposes: reward-guided self-distillation uses these samples to estimate relative preference and construct reward-guided targets, while privileged self-distillation reuses the resulting privileged outputs as supervision for the blind generator.

As illustrated in Figure 2, RPSD starts from the degradation-aware privileged policy initialized in Section 3.2 and couples two complementary self-distillation processes. Reward-guided self-distillation progressively aligns the privileged policy with designed restoration preference, while privileged self-distillation transfers its degradation-aware restoration capabilities and learned preference into an LQ-only blind generator. Alternating these two processes continuously improves the privileged policy while progressively transferring its restoration capability to the blind policy, thereby reducing the latter's reliance on training-time privileged conditions.

Reward-Guided Self-Distillation. The privileged policy operates under degradation-aware conditions $c_P = (I_{\mathrm{LQ}}, p_{\mathrm{task}}, \sigma_c^P)$, where p_{task} specifies the known degradation and the difficulty-aware condition level σ_c^P is calibrated from the paired $(I_{\mathrm{LQ}}, I_{\mathrm{GT}})$.

For each LQ input, the privileged policy v_P samples K restoration candidates using independent noises $\{\epsilon_i\}_{i=1}^K$. With $z_{g,i} = (1-\sigma_g)z_{\mathrm{LQ}} + \sigma_g \epsilon_i$, the corresponding privileged endpoint is

$$z_i = z_{g,i} - \sigma_g v_P(z_{g,i}, c_P, \sigma_g) \tag{4}$$

where v_P denote the privileged policy. After decoding, each candidate is evaluated by a fixed reward model as $r_i = R_\phi(\mathrm{Dec}(z_i))$. We standardize the rewards within each candidate group and map the resulting relative advantage to an optimality coefficient $a_i \in [0,1]$:

$$\alpha_i = \frac{r_i - \mu_r}{\sigma_r + \varepsilon}, \qquad a_i = \mathrm{clip}_{[0,1]}\left(\frac{1}{2} + \frac{\mathrm{clip}(\alpha_i, -\kappa, \kappa)}{2\kappa}\right). \tag{5}$$

Here, μ_r and σ_r denote the within-group reward mean and standard deviation, respectively, and κ controls the clipping range of the relative advantage.

To convert image-level preference into explicit self-distillation supervision, we construct positive and negative targets from each behavior endpoint through normalized reward-gradient ascent and descent within a relative trust region $\mathcal{B}_i$ (Zhou et al., 2026):

$$z_{i,j+1}^{\pm} \leftarrow \Pi_{\mathcal{B}_i}\left(z_{i,j}^{\pm} \pm h \frac{\nabla_z \widetilde{R}_\phi(z_{i,j}^{\pm})}{\|\nabla_z \widetilde{R}_\phi(z_{i,j}^{\pm})\|_2 + \varepsilon}\right), \qquad z_{i,0}^{+} = z_{i,0}^{-} = z_i, \tag{6}$$

where $\widetilde{R}_\phi(z) = R_\phi(\mathrm{Dec}(z))$. The resulting targets are detached, separating reward-guided target construction from policy realization.

After that, let v_θ denote the trainable policy, at the same query state $s_i = (z_{g,i}, c_P, \sigma_g)$, we construct

$$v_i^{+} = (1-\beta)v_P(s_i) + \beta v_\theta(s_i), \qquad v_i^{-} = (1+\beta)v_P(s_i) - \beta v_\theta(s_i), \tag{7}$$

with corresponding endpoints $\hat{z}_i^{\pm} = z_{g,i} - \sigma_g v_i^{\pm}$. The privileged policy is then optimized with advantage-weighted normalized endpoint regression:

$$\begin{aligned} \mathcal{L}_P &= \frac{\lambda_P}{K\beta}\sum_{i=1}^{K}\left[a_i\,\ell_{\mathrm{norm}}(\hat{z}_i^{+}, z_i^{+}) + (1-a_i)\,\ell_{\mathrm{norm}}(\hat{z}_i^{-}, z_i^{-})\right], \\ \ell_{\mathrm{norm}}(\hat{z}, z) &= \frac{\|\hat{z} - \mathrm{sg}[z]\|_2^2}{\mathrm{sg}[\max(\|\hat{z} - z\|_1, \delta)]}. \end{aligned} \tag{8}$$

Here, $\delta > 0$ is a small constant for numerical stability, and λ_P controls the overall loss scale. Both the target and the normalization denominator are detached, such that gradients are propagated only through the predicted endpoint. In this way, reward-guided self-distillation progressively aligns the degradation-aware privileged policy with the designed restoration preference.

Privileged Self-Distillation. Although reward-guided self-distillation improves the privileged policy, its restoration behavior still relies on degradation information unavailable at deployment. We therefore introduce a blind policy under the fixed condition $c_B = (I_{\mathrm{LQ}}, p_{\mathrm{fix}}, \sigma_c^B)$, where both the generic prompt and condition-stream noise level are fixed across all inputs.

To transfer both restoration capability and learned preference, we reuse the same privileged candidate group and convert its rewards into normalized preference weights. The blind policy is evaluated with the same LQ inputs and noise samples, producing paired endpoints

$$z_i^B = z_{g,i} - \sigma_g v_B(z_{g,i}, c_B, \sigma_g) \tag{9}$$

and the blind policy is optimized by a reward-weighted objective:

$$w_i = \frac{\exp(r_i/\tau)}{\sum_{j=1}^{K}\exp(r_j/\tau)}, \qquad \mathcal{L}_{\mathrm{B}} = \sum_{i=1}^{K} w_i \left\|z_i^B - \mathrm{sg}[z_i]\right\|_2^2, \tag{10}$$

where τ controls the concentration on preferred privileged endpoints. Consequently, high-reward restoration behaviors receive stronger self-distillation weights, allowing the blind generator to absorb not only the degradation-aware capabilities of the privileged policy but also the restoration preference acquired through reward-guided self-distillation.

3.4 Reward Modeling

To guide restoration preference, we design a composite reward that jointly considers restoration fidelity, perceptual quality, and visual aesthetics:

$$R = \lambda_{\text{fid}} R_{\text{fid}} + \lambda_{\text{qual}} R_{\text{qual}} + \lambda_{\text{aes}} R_{\text{aes}}, \tag{11}$$

where R_{fid} measures consistency with the ground truth using complementary perceptual, pixel-level, and structural criteria, R_{qual} evaluates the perceptual quality of the restored image, and R_{aes} captures high-level aesthetic preference. Together, these complementary signals define a restoration-oriented preference that encourages perceptual improvement while preserving content fidelity.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation. For training, we construct a paired mixed-degradation dataset comprising 13,110 LQ–GT pairs across eight degradation categories, with one to five degradation types per image and three severity tiers: easy, medium, and hard. Stage 1 combines this dataset with a 48,380-pair single-degradation subset built with the FAPE-IR training corpus (Liu et al., 2026b) and additional high-quality synthetic pairs through adversarial post-training, while Stage 2 applies RPSD exclusively to the mixed-degradation training split. Appendix A details our training set. For testing, we conduct evaluations on MiO100 (Zhu et al., 2025) and Real-1000 (Li et al., 2025). Following prior work (Zhu et al., 2025; Zuo et al., 2025; Jia et al., 2026), we report three full-reference (FR) metrics PSNR, SSIM, and LPIPS, and three no-reference (NR) metrics MANIQA, CLIPIQA, and MUSIQ to assess restoration performance.

Implementation Details. We build USIR on FLUX.2-Klein-Base-9B and train on 8 types of degradation (Zhu et al., 2025) and their combinations. Stage 1 performs noise-hybrid adversarial post-training for $30{,}000$ iterations on four NVIDIA A100 80GB GPUs with a global batch size of 4. Stage 2 applies RPSD on 1 NVIDIA A100 with the groupsize K=8. The privileged condition noise σ_c^P is dynamically determined from LPIPS distance of the LQ–GT pair, while the blind generator uses a fixed condition noise $\sigma_c^B = 0$. Other details are provided in Appendix B.

4.2 Comparison with SOTA Methods

Baselines. We compare USIR with representative agentic and all-in-one image restoration methods using the same GPU hardware and evaluation protocol as TIR-Agent (Jia et al., 2026). The agentic baselines include AgenticIR (Zhu et al., 2025), MAIR (Jiang et al., 2025c), 4K-Agent (Zuo et al., 2025), and TIR-Agent (Jia et al., 2026). The all-in-one baselines include AirNet (Li et al., 2022), PromptIR (Potlapalli et al., 2023), MiOIR (Kong et al., 2024), DA-CLIP (Luo et al., 2023), InstructIR (Conde et al., 2024), AutoDIR (Jiang et al., 2023), and FAPE-IR (Liu et al., 2026b). Specifically, we retrain DA-CLIP and FAPE-IR on the same training set as USIR for a controlled comparison, following their official open-source implementations and released model configurations. The resulting variants are denoted as DA-CLIP* and FAPE-IR*.

Quantitative Comparison. Table 1 presents the comparison on the three groups of MiO100 (Zhu et al., 2025). Here, we report two main observations. 1) *Restoration Quality*: USIR achieves strong performance on both full-reference fidelity and no-reference perceptual quality metrics. Compared with the strongest competing results, USIR substantially reduces LPIPS from 0.2874/0.2725/0.4120 to 0.1456/0.1458/0.2253 on Groups A/B/C. It also achieves the best CLIP-IQA scores of 0.6778, 0.6818, and 0.6715, improving over the corresponding second-best results by 0.1234, 0.1313, and 0.1220. 2) *Inference Efficiency*: Unlike agentic methods that require iterative assessment and tool execution, USIR performs restoration with a single generator call and takes 0.77–0.78 seconds per image, placing its runtime comparable to conventional all-in-one methods and yielding approximately 79×, 92×, and 127× speedups over TIR-Agent, the fastest agentic baseline with reported runtimes, on Groups A, B, and C, respectively. Moreover, the single-generator inference pipeline of USIR keeps the runtime essentially constant across Groups A–C, rather than increasing with degradation complexity as in iterative agentic pipelines. The results on Real-1000 in Table 2 further demonstrate the robustness of USIR under increasingly complex degradations.

Table 1: Quantitative comparison on Groups A, B, and C of the MiO100 dataset. The best and second-best performances of each metric are highlighted in **bold** and underlined.

Dataset	Type	Method	PSNR↑	SSIM↑	LPIPS↓	MANIQA↑	CLIPIQA↑	MUSIQ↑	Time(s)↓
Group A	Agent	AgenticIR (Zhu et al., 2025)	21.04	0.6818	0.3148	0.3071	0.4474	56.88	184.57
		MAIR (Jiang et al., 2025c)	21.02	0.6715	0.2963	0.3330	0.4751	59.19	–
		4KAgent (Zuo et al., 2025)	21.48	0.6720	0.3019	0.3748	0.5544	63.19	333.14
		TIR-Agent (Jia et al., 2026)	22.07	0.6935	0.2874	0.3907	0.5454	63.69	61.58
	All-in-One	AirNet (Li et al., 2022)	19.13	0.6019	0.4283	0.2581	0.3930	42.46	1.75
		PromptIR (Potlapalli et al., 2023)	20.06	0.6088	0.4127	0.2633	0.4013	42.62	1.26
		MiOIR (Kong et al., 2024)	20.84	0.6558	0.3715	0.2451	0.3933	47.82	**0.69**
		DA-CLIP (Luo et al., 2023)	19.58	0.6032	0.4266	0.2418	0.4139	42.51	44.87
		DA-CLIP* (Luo et al., 2023)	19.78	0.6284	0.3071	0.3067	0.4510	52.66	44.87
		InstructIR (Conde et al., 2024)	18.03	0.5751	0.4429	0.2660	0.3528	45.77	1.45
		AutoDIR (Jiang et al., 2023)	19.64	0.6286	0.3967	0.2500	0.3767	47.01	128.64
		FAPE-IR (Liu et al., 2026b)	20.16	0.6345	0.3927	0.2469	0.4085	44.33	2.91
		FAPE-IR* (Liu et al., 2026b)	20.34	0.6465	0.3573	0.3255	0.4817	46.99	2.91
		USIR (Ours)	**22.69**	**0.7279**	**0.1456**	**0.4446**	**0.6778**	**70.26**	0.78
Group B	Agent	AgenticIR (Zhu et al., 2025)	20.55	0.7009	0.3072	0.3204	0.4648	57.57	201.25
		MAIR (Jiang et al., 2025c)	20.92	0.7004	0.2788	0.3544	0.5084	60.98	–
		4KAgent (Zuo et al., 2025)	20.95	0.6727	0.3017	0.3734	0.5505	62.69	409.49
		TIR-Agent (Jia et al., 2026)	**22.80**	0.7340	0.2725	0.3997	0.5488	65.24	71.09
	All-in-One	AirNet (Li et al., 2022)	19.31	0.6567	0.3670	0.2882	0.4274	47.88	1.83
		PromptIR (Potlapalli et al., 2023)	20.47	0.6704	0.3370	0.2893	0.4289	48.10	1.30
		MiOIR (Kong et al., 2024)	20.56	0.6905	0.3243	0.2638	0.4330	51.87	1.09
		DA-CLIP (Luo et al., 2023)	18.56	0.5946	0.4405	0.2435	0.4154	43.70	47.41
		DA-CLIP* (Luo et al., 2023)	19.73	0.6574	0.2901	0.3002	0.4394	53.48	47.41
		InstructIR (Conde et al., 2024)	18.34	0.6235	0.4072	0.3022	0.3790	50.94	1.50
		AutoDIR (Jiang et al., 2023)	19.90	0.6643	0.3542	0.2534	0.3986	49.64	174.48
		FAPE-IR (Liu et al., 2026b)	19.01	0.6179	0.3928	0.2470	0.4211	45.10	2.95
		FAPE-IR* (Liu et al., 2026b)	19.86	0.6379	0.3660	0.3254	0.4940	47.79	2.95
		USIR (Ours)	22.74	**0.7407**	**0.1458**	**0.4467**	**0.6818**	**70.36**	**0.77**
Group C	Agent	AgenticIR (Zhu et al., 2025)	18.82	0.5474	0.4493	0.2698	0.3948	48.68	249.70
		MAIR (Jiang et al., 2025c)	19.42	0.5544	0.4142	0.2798	0.4239	51.36	–
		4KAgent (Zuo et al., 2025)	19.77	0.5629	0.4271	0.3545	0.5233	55.56	555.12
		TIR-Agent (Jia et al., 2026)	19.53	0.5643	0.4120	0.3739	0.5495	56.93	98.92
	All-in-One	AirNet (Li et al., 2022)	17.95	0.5145	0.5782	0.1854	0.3113	30.12	2.07
		PromptIR (Potlapalli et al., 2023)	18.51	0.5166	0.5756	0.1906	0.3104	29.71	1.36
		MiOIR (Kong et al., 2024)	15.63	0.4896	0.5376	0.1717	0.2891	37.95	2.06
		DA-CLIP (Luo et al., 2023)	18.53	0.5320	0.5335	0.1916	0.3476	33.87	59.64
		DA-CLIP* (Luo et al., 2023)	17.99	0.4935	0.4966	0.2101	0.3300	36.54	59.64
		InstructIR (Conde et al., 2024)	17.09	0.5135	0.5582	0.1732	0.2537	33.69	1.90
		AutoDIR (Jiang et al., 2023)	18.61	0.5443	0.5019	0.2045	0.2939	37.86	240.34
		FAPE-IR (Liu et al., 2026b)	19.47	0.5627	0.5167	0.2043	0.3337	32.76	3.34
		FAPE-IR* (Liu et al., 2026b)	19.76	0.5767	0.4265	0.2825	0.3999	39.51	3.34
		USIR (Ours)	**20.80**	**0.6116**	**0.2253**	**0.4297**	**0.6715**	**69.17**	**0.78**

While USIR already achieves strong performance on single-degradation inputs, its advantage becomes more pronounced as the number of degradations increases, achieving the best results on all six metrics under two degradations and on five out of six metrics under three degradations.

Qualitative Comparison. Fig. 3 further demonstrates the clear qualitative advantage of USIR under complex mixed degradations. Conventional all-in-one methods often leave visible residual artifacts or over-smooth fine structures, while agentic methods may introduce inconsistent restoration effects through sequential assessment and tool execution. In contrast, USIR consistently removes diverse degradations, recovers sharper structures and textures, and better preserves the underlying image content. These results demonstrate that USIR achieves stronger restoration quality with a single LQ-conditioned generator.

Table 2: Quantitative comparisons on Real-1000 (Li et al., 2025). The three blocks from top to bottom correspond to one, two, and three degradations The best and second-best results are highlighted in **bold** and underline.

Method	PSNR	SSIM	LPIPS	MANIQA	CLIPIQA	MUSIQ
DA-CLIP	19.62	0.6930	0.3049	0.2870	0.3509	55.75
InstructIR	21.72	0.7796	0.2374	0.2922	0.3572	55.89
AutoDIR	20.72	0.7157	0.3704	0.2395	0.3040	43.56
MiOIR	22.66	0.7798	0.1717	0.4005	0.4648	61.72
AgenticIR	18.68	0.6625	0.2937	0.2977	0.3599	55.34
FAPE-IR	**24.34**	0.8084	0.0990	0.3867	0.4369	68.98
Ours	23.95	**0.8218**	**0.0950**	**0.4452**	**0.5690**	**72.20**
DA-CLIP	17.85	0.6680	0.3786	0.2871	0.3673	52.47
InstructIR	18.09	0.6488	0.3970	0.2610	0.3504	50.36
AutoDIR	18.79	0.5931	0.5182	0.2102	0.2545	40.38
MiOIR	17.48	0.5461	0.3202	0.3225	0.4048	51.44
AgenticIR	15.37	0.4850	0.5010	0.2400	0.2950	42.10
FAPE-IR	19.79	0.6900	0.1789	0.3419	0.4207	63.83
Ours	**22.42**	**0.7685**	**0.1213**	**0.4169**	**0.5571**	**70.25**
DA-CLIP	15.04	0.7082	0.2940	0.2303	0.2377	43.75
InstructIR	16.36	**0.7250**	0.2799	0.2157	0.2318	42.52
AutoDIR	18.06	0.7122	0.3245	0.2145	0.2378	42.64
MiOIR	13.27	0.5404	0.2432	0.2912	0.2792	54.70
AgenticIR	14.80	0.5230	0.3500	0.2400	0.2600	45.00
FAPE-IR	16.04	0.6440	0.1790	0.3330	0.3231	64.28
Ours	**20.34**	0.7046	**0.0938**	**0.4191**	**0.4840**	**73.66**

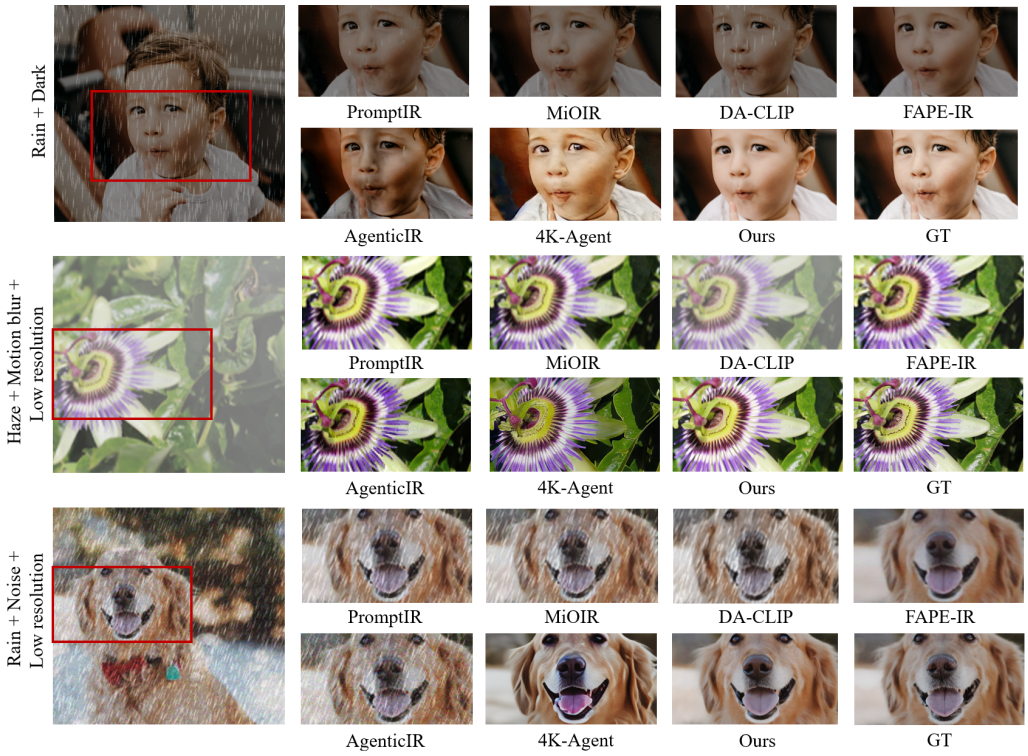

Figure 3: Qualitative comparison on MiO100 under complex mixed degradations.

Table 3: Comparison of joint and sequential training strategies for RPSD on MiO100 Group B. GPU-hours measure the Stage 2 training cost.

Training Strategy	PSNR↑	SSIM↑	LPIPS↓	MANIQA↑	CLIPIQA↑	MUSIQ↑	GPU-h↓
Privileged SD → Reward-Guided SD	22.66	0.7401	**0.1433**	0.4394	0.6766	**70.51**	21.16
Reward-Guided SD → Privileged SD	21.48	0.7216	0.1472	**0.4470**	0.6708	70.36	21.45
RPSD	**22.74**	**0.7407**	0.1458	0.4467	**0.6818**	70.36	**14.85**

4.3 RPSD Analysis

We further analyze RPSD from two complementary perspectives: the training schedule of the two self-distillation processes and the optimization strategy for the privileged policy.

Joint versus Sequential Self-Distillation. Reward-guided self-distillation and privileged self-distillation both rely on grouped samples from the privileged policy. RPSD jointly performs the two updates so that the same candidate group and reward evaluations can be reused. We compare this joint design against two sequential baselines: privileged self-distillation followed by reward-guided self-distillation, and reward-guided self-distillation followed by privileged self-distillation. For a fair comparison, all strategies use the same number of reward-guided and privileged self-distillation updates. As shown in Table 3, the joint strategy achieves comparable restoration quality while requiring substantially fewer GPU-hours, demonstrating that reusing candidate samples and reward evaluations effectively reduces post-training overhead.

Privileged Policy Optimization. We further compare different strategies for improving the privileged policy while leaving the blind policy alone. Specifically, we compare our reward-guided self-distillation with direct reward optimization (Xu et al., 2023), group-based policy optimization (Liu et al., 2026a), and negative-aware fine-tuning (Zheng et al., 2025). As shown in Table 4, reward-guided self-distillation provides more effective restoration-oriented policy improvement by converting reward signals into explicit positive and negative supervision, rather than directly opti-

Table 4: Comparison of privileged policy optimization strategies on MiO100 Group B. Results of this table are tested with the privileged policy under privileged condition.

Policy Optimization	PSNR↑	SSIM↑	LPIPS↓	MANIQA↑	CLIPIQA↑	MUSIQ↑
ReFL (Xu et al., 2023)	**23.86**	0.7486	0.1383	0.4384	0.6712	69.93
Flow-GRPO (Liu et al., 2026a)	23.48	0.7382	0.1384	0.4175	0.6578	68.99
DiffusionNFT (Zheng et al., 2025)	23.44	0.7399	0.1376	0.4215	0.6615	69.31
Reward-Guided Self-Distillation	23.86	**0.7500**	**0.1371**	**0.4502**	**0.7170**	**70.43**

Table 5: Ablation studies on MiO100 Group B. We evaluate (a) the degradation-aware conditions used for privileged initialization and (b) the components of RPSD. Part (a) is evaluated under training-aligned conditions, whereas Part (b) is evaluated under the blind deployment setting.

Ablation	Configuration	PSNR↑	SSIM↑	LPIPS↓	MANIQA↑	CLIPIQA↑	MUSIQ↑
(a) Initialization	Fixed prompt + Fixed σ_c	19.95	0.6868	0.1806	0.3973	0.5987	65.47
	Task prompt + Fixed σ_c	20.31	0.6990	0.1725	0.3902	0.6128	66.13
	Task prompt + Dynamic σ_c	**23.36**	**0.7363**	**0.1390**	**0.4189**	**0.6595**	**69.14**
(b) RPSD components	Baseline	20.18	0.6867	0.1841	0.3775	0.6036	66.10
	+ *Privileged Self-Distillation*	21.91	0.7257	0.1574	0.4287	0.6715	69.57
	+ *Reward-Guided Self-Distillation*	20.56	0.6970	0.1720	0.3977	0.6106	67.30
	+ *RPSD w/o Reward Weighting*	22.53	0.7371	0.1486	0.4307	0.6666	69.64
	RPSD (Ours)	**22.74**	**0.7407**	**0.1458**	**0.4467**	**0.6818**	**70.36**

mizing the policy from scalar rewards or from group relative advantages. Further details and analysis of this experiment are provided in Appendix C.1 and C.2.

4.4 Ablation Studies

Initialization. We train three variants with progressively richer conditioning: fixed prompt with fixed condition noise, task-specific prompt with fixed condition noise, and task-specific prompt with dynamic condition noise. For each variant, inference follows the same conditioning setting used during training. As shown in Table 5, introducing task-specific prompts improves restoration capability, while difficulty-aware condition noise further provides stronger degradation guidance. The full configuration therefore achieves the strongest privileged initialization.

Components of RPSD. We progressively introduce the components of RPSD to the blind baseline. As shown in Table 5, privileged self-distillation substantially improves the blind policy by transferring degradation-aware restoration capabilities from the privileged policy, increasing PSNR from 20.18 to 21.91 dB and CLIPIQA from 0.6036 to 0.6715. Reward-guided self-distillation independently improves the privileged policy and also yields consistent gains over the baseline. Combining the two processes further improves restoration performance, reaching 22.53 dB PSNR and 0.1486 LPIPS even without reward weighting. Finally, incorporating reward-based weighting into privileged self-distillation yields the complete RPSD, which achieves the best performance across all six metrics. These results demonstrate the complementary roles of privileged self-distillation and reward-guided self-distillation, as well as the benefit of emphasizing preferred privileged behaviors during distillation.

5 Conclusion

We presented USIR, a unified single-step image restoration framework that internalizes degradation-aware restoration capability during training and deploys a single LQ-conditioned generator for direct inference. At its core, RPSD couples reward-guided self-distillation for privileged policy improvement with privileged self-distillation for transferring restoration capabilities and learned preference into the blind generator, while reusing privileged rollouts and reward evaluations across both processes to reduce post-training overhead. Extensive experiments demonstrate that USIR achieves strong restoration fidelity and perceptual quality under diverse and complex degradations, while substantially improving inference efficiency over agentic restoration pipelines.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript by correcting grammatical errors, refining wording, and improving overall language fluency. We have not used generative AI tools for any of the research subtasks for which disclosure is required, including research methodology or experimental design, hypothesis formulation, mathematical claims or proofs, and result interpretation; and the rest of the required disclosure tasks are not applicable to this work. All AI-assisted edits were manually reviewed and verified by the authors to ensure that they preserved the intended meaning and did not introduce unsupported claims or factual errors. All figures were created by the authors. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on general-purpose image restoration and does not collect personally identifiable information. All datasets used in our experiments are publicly available and are used in accordance with their respective licenses and terms of use. While image restoration techniques may potentially be applied to altered or misleading visual content, our method is intended for recovering degraded visual information rather than modifying image semantics. We follow standard research practices and report experimental results without manual alteration.

REPRODUCIBILITY STATEMENT

We provide descriptions of the proposed method and training objectives in the main paper. The appendix includes implementation details, training hyperparameters, dataset and degradation settings. We further provide an anonymous supplementary repository containing inference code and model weights. The complete training code, configuration files, and evaluation scripts will be released upon acceptance.

REFERENCES

Abdullah Abuolaim and Michael S. Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pp. 111–126, 2020.

Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness, generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25432–25444, 2024. doi: 10.1109/CVPR52733.2024.02403.

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. *arXiv preprint arXiv:2407.18035*, 2024.

Zihan Cheng, Liangtai Zhou, Dian Chen, Ni Tang, Xiaotong Luo, Yuan Xie, and Yanyun Qu. Unildiff: Unlocking the power of diffusion priors for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37465–37475, 2026.

Marcos V. Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, 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*, 2026. URL https://arxiv.org/abs/2511.17138.

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

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015.

Guoli Jia, Yisheng Zhang, Haote Hu, Shanxu Zhao, Kaikai Zhao, Long Sun, Xinwei Long, Kai Tian, Che Jiang, Zhaoxiang Liu, Kai Wang, Shiguo Lian, Kaiyan Zhang, and Bowen Zhou. TIR-Agent: Training an explorative and efficient agent for image restoration. *arXiv preprint arXiv:2603.27742*, 2026. URL https://arxiv.org/abs/2603.27742.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, et al. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026.

Jiangang Jiang, Zhe Chen, Yuxin Su, Pancheng Zhang, and Yihui Hu. All-in-one image restoration via diffusion models with degradation perception and semantic enhancement. *Neurocomputing*, pp. 131517, 2025a.

Junjun Jiang, Zengyuan Zuo, Gang Wu, Kui Jiang, and Xianming Liu. A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11892–11911, 2025b. doi: 10.1109/TPAMI.2025.3598132.

Xu Jiang, Gehui Li, Bin Chen, and Jian Zhang. Multi-agent image restoration. *arXiv preprint arXiv:2503.09403*, 2025c. URL https://arxiv.org/abs/2503.09403.

Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. *arXiv preprint arXiv:2310.10123*, 2023.

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.

Xiangtao Kong, Chao Dong, and Lei Zhang. Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy. *arXiv preprint arXiv:2401.03379*, 2024. URL https://arxiv.org/abs/2401.03379.

Bingchen Li, Xin Li, Yiting Lu, and Zhibo Chen. Hybrid agents for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22636–22647, 2026a.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17452–17462, 2022. doi: 10.1109/CVPR52688.2022.01693.

Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. FoundIR: Unleashing million-scale training data to advance foundation models for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12626–12636, 2025.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026b.

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.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

Jingren Liu, Shuning Xu, Qirui Yang, Yun Wang, Xiangyu Chen, and Zhong Ji. FAPE-IR: Frequency-aware planning and execution framework for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15561–15573, 2026b.

Junyu Liu, Tianyu Li, Lanyue Liang, Gang Fu, Guoqing Wang, Quan Rui, Xiongxin Tang, Shuyuan Zhu, and Yang Yang. Rir-agent: An interactive framework for effective and adaptive restoration of remote sensing imagery. *Expert Systems with Applications*, 324:132495, 2026c. doi: 10.1016/j.eswa.2026.132495.

Lu Liu, Chunlei Cai, Shaocheng Shen, Jianfeng Liang, Weimin Ouyang, Tianxiao Ye, Jian Mao, Huiyu Duan, Jiangchao Yao, Xiaoyun Zhang, Qiang Hu, and Guangtao Zhai. Moa-vr: A mixture-of-agents system toward all-in-one video restoration. *IEEE Journal of Selected Topics in Signal Processing*, 19:1822–1837, 2025. doi: 10.1109/JSTSP.2025.3627128.

Jianglin Lu, Yuanwei Wu, Ziyi Zhao, Hongcheng Wang, Felix Jimenez, Abrar Majeedi, and Yun Fu. Restore-r1: Efficient image restoration agents via reinforcement learning with multimodal llm perceptual feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8629–8639, 2026.

Wenyang Luo, Haina Qin, Zewen Chen, Libin Wang, Dandan Zheng, Yuming Li, Yufan Liu, Bing Li, and Weiming Hu. Visual-instructed degradation diffusion for all-in-one image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12764–12777. IEEE, 2025.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjolund, and Thomas B. Schon. Controlling vision-language models for multi-task image restoration. *arXiv preprint arXiv:2310.01018*, 2023.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration. In *Advances in Neural Information Processing Systems*, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Ni Tang, Xiaotong Luo, Zihan Cheng, Liangtai Zhou, Dongxiao Zhang, and Yanyun Qu. Diffusion once and done: Degradation-aware lora for all-in-one image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9448–9456, 2026.

Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28165–28175, 2025. doi: 10.1109/CVPR52734.2025.02623.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. Seedvr2: One-step video restoration via diffusion adversarial post-training. In *ICLR*, 2026.

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.

Zhijian Wu, Wenhui Liu, Jingchao Wang, Jun Li, and Dingjiang Huang. Freprompter: Frequency self-prompt for all-in-one image restoration. *Pattern Recognition*, 161:111223, 2025. doi: 10.1016/j.patcog.2024.111223.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. URL https://arxiv.org/abs/2304.05977.

Mingde Yao, Ruikang Xu, Yuanshen Guan, Jie Huang, and Zhiwei Xiong. Neural degradation representation learning for all-in-one image restoration. *arXiv preprint arXiv:2310.12848*, 2023.

Xuanyu Zhang, Weiqi Li, Qunliang Xing, Jingfen Xie, Bin Chen, Junlin Li, Li Zhang, Jian Zhang, and Shijie Zhao. Vq-jarvis: Retrieval-augmented video restoration agent with sharp vision and fast thought. *arXiv preprint arXiv:2603.22998*, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

Wei Zhou, Xiongwei Zhu, Lingdong Kong, Bo Chen, Lei Zhang, Yongyuan Liang, Xiaoxia Hou, Ye Tian, Xian Sun, Yingshuo Wang, Linfeng Li, Shengqiong Wu, Leigang Qu, Feng Li, Wei Liu, Julian McAuley, and Tat-Seng Chua. On-policy self-distillation in diffusion models. *arXiv preprint arXiv:2608.24646*, 2026. URL https://arxiv.org/abs/2608.24646.

Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for complex image restoration problems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3RLxccFPHz.

Yushen Zuo, Qi Zheng, Mingyang Wu, Xinrui Jiang, Renjie Li, Jian Wang, Yide Zhang, Gengchen Mai, Lihong V. Wang, James Zou, Xiaoyu Wang, Ming-Hsuan Yang, and Zhengzhong Tu. 4KAgent: Agentic any image to 4k super-resolution. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5476.

A Training-Set Construction

Dataset overview. Our training set consists of two complementary components: a *single-degradation subset* and a *mixed-degradation subset*. Both cover eight restoration tasks, including haze, rain, motion blur, defocus blur, low light, low resolution, noise, and JPEG compression. The single-degradation subset combines real paired data with controlled synthetic degradations to provide task-specific restoration supervision, whereas the mixed-degradation subset composes multiple degradations with different severity levels to approximate more complex real-world degradation processes. The two subsets contain $48{,}380$ and $13{,}110$ paired examples, respectively.

Single-degradation subset. For haze, low light, motion blur, noise, and rain, we primarily use valid LQ–GT pairs from the FAPE-IR (Liu et al., 2026b) training set. For motion blur, we additionally discard samples whose GT mean intensity is below 0.15 to reduce the influence of extremely dark content. For JPEG compression, we use 200 LQ–GT pairs from Urban100 (Huang et al., 2015) and supplement them with $1{,}000$ synthetic pairs generated using the standard degradation pipeline (Zhu et al., 2025). For defocus blur, we use valid source–target pairs from the official DPDD (Abuolaim & Brown, 2020) mapping.

Low-resolution restoration is constructed entirely with synthetic degradations. We sample $12{,}000$ GT images from LSDIR (Li et al., 2023) and $3{,}000$ from FFHQ (Karras et al., 2019), and generate the corresponding LQ images using the two-stage degradation pipeline (Wang et al., 2021; Fang et al., 2026), which includes blur, resampling, noise injection, and JPEG compression. Among the resulting $15{,}000$ pairs, $11{,}250$ use the $4\times$ degradation setting and $3{,}750$ use the $2\times$ setting.

We allocate a larger proportion of training samples to low-resolution restoration, as super-resolution is a fundamental low-level vision task and its degradation process naturally encompasses multiple corruption types, including blur, resampling, noise, and JPEG compression. In contrast, we include fewer standalone JPEG-compression samples because real paired data for this task are relatively limited, while JPEG degradation is extensively covered by the synthetic super-resolution pipeline.

After checking image readability, spatial-size consistency, and task assignment, candidate samples are ranked by PSNR in ascending order, such that stronger but valid degradations are preferentially retained. The final composition of the single-degradation subset is summarized in Table 6.

Mixed-degradation subset. To expose the model to diverse compound degradations, we construct a mixed-degradation subset containing $13{,}110$ paired examples from $3{,}450$ unique GT images. Each sample contains between one and five degradation types and is assigned an overall severity of easy,

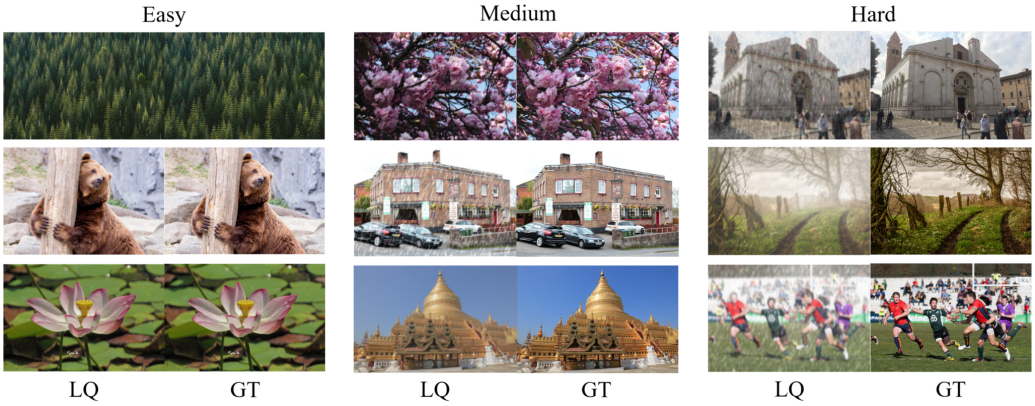

Figure 4: Representative examples from the mixed-degradation subset. The dataset covers diverse combinations of degradation types and severity levels.

Table 6: Composition of the single-degradation subset. "Real/reference" denotes paired examples collected from existing datasets, while "Synthetic" denotes samples generated using controlled degradation pipelines.

Degradation	**Real/reference**	**Synthetic**	**Total**
JPEG compression	200	1,000	1,200
Defocus blur	352	2,000	2,352
Haze	3,000	2,000	5,000
Low light	1,589	1,589	3,178
Low resolution	0	15,000	15,000
Motion blur	5,000	3,450	8,450
Noise	3,000	3,000	6,000
Rain	3,600	3,600	7,200
Total	**16,741**	**31,639**	**48,380**

medium, or hard. Representative examples are shown in Fig. 4, illustrating the diversity of compound degradation patterns and severity levels covered by the dataset. All splits are determined at the GT-image level, preventing different degraded versions of the same content from appearing across training and validation.

We parameterize each degradation using four levels of increasing strength. The overall sample severity jointly determines the supported number of degradations k and the distribution from which individual degradation levels are sampled. The complete composition rules are summarized in Table 7. For samples with multiple degradations, individual levels are sampled according to severity-specific categorical distributions, subject to an additional constraint on their average level. This design allows the overall difficulty to increase through both the number and the strength of constituent degradations, while avoiding excessively strong compositions in easier samples.

The four levels are instantiated using task-specific degradation parameters, as summarized in Table 8. For noise degradation, we consider both Gaussian and Poisson noise. Increasing the level consistently corresponds to stronger corruption, although the underlying physical parameter varies across degradation types. For example, JPEG quality and low-resolution scale decrease with severity, whereas noise magnitude, blur strength, haze density, and rain strength increase.

To balance the occurrence of individual degradation types and their co-occurrences, each candidate degradation combination $\mathcal{C}$ is selected according to

$$S(\mathcal{C}) = \sum_{t \in \mathcal{C}} [n_{\text{local}}(t) + 0.15 n_{\text{global}}(t)] + 0.10 \sum_{\{u,v\} \subset \mathcal{C}} n_{\text{pair}}(u, v), \tag{12}$$

where n_{local}, n_{global}, and n_{pair} measure the frequency of each degradation within the current split/severity group, its global frequency, and the co-occurrence frequency of each degradation pair, respectively. Candidate combinations with lower scores are preferentially selected, promoting bal-

Table 7: Composition rules for the mixed-degradation subset. The "Level sampling" column applies to samples with $k > 1$. Repeat-JPEG denotes the probability of applying a second JPEG compression when JPEG degradation is included.

Severity	Supported k	Sampling Level	Composition constraint	Repeat-JPEG
Easy	$\{1, 2\}$	$P(L_1, L_2) = (0.75, 0.25)$	$\overline{L} \leq 1.5$	0.10
Medium	$\{1, 2, 3, 4\}$	$P(L_1, L_2, L_3, L_4) = (0.10, 0.40, 0.40, 0.10)$	$1.75 \leq \overline{L} \leq 3.0$	0.30
Hard	$\{1, 2, 3, 4, 5\}$	$P(L_2, L_3, L_4) = (0.15, 0.50, 0.35)$	$\overline{L} \geq 2.8$, at least one L_4	0.50

Table 8: Task-specific parameterization of the four degradation levels used in the mixed-degradation subset.

Degradation	Level 1	Level 2	Level 3	Level 4
JPEG compression	$q \sim \mathcal{N}(50, 2)$	$q \sim \mathcal{N}(30, 2)$	$q \sim \mathcal{N}(20, 2)$	$q \sim \mathcal{N}(10, 2)$
Gaussian noise	$\sigma \in [2, 6]$	$\sigma \in [10, 20]$	$\sigma \in [25, 35]$	$\sigma \in [40, 50]$
Poisson noise	$s \in [0.4, 0.6]$	$s \in [0.8, 1.2]$	$s \in [1.7, 2.3]$	$s \in [2.7, 3.3]$
Low resolution	$r \in [0.50, 0.67]$	$r \in [0.40, 0.50]$	$r \in [0.33, 0.40]$	$r \in [0.20, 0.30]$
Defocus blur	radius $= 2$, alias $= 0.1$	radius $= 3$, alias $= 0.1$	radius $= 4$, alias $= 0.5$	radius $= 6$, alias $= 0.5$
Motion blur	radius $= 5$, $\sigma = 2$	radius $= 10$, $\sigma = 3$	radius $= 15$, $\sigma = 5$	radius $= 15$, $\sigma = 8$
Haze	$A \approx 0.6$, $\beta \approx 0.4$	$A \approx 0.7$, $\beta \approx 0.6$	$A \approx 0.8$, $\beta \approx 1.2$	$A \approx 1.0$, $\beta \approx 1.8$
Low light	shift ≈ 15, $\gamma \approx 0.8$, max ≈ 230	shift ≈ 30, $\gamma \approx 0.7$, max ≈ 190	shift ≈ 50, $\gamma \approx 0.6$, max ≈ 150	shift ≈ 70, $\gamma \approx 0.5$, max ≈ 120
Rain	strength ≈ 5	strength ≈ 60	strength ≈ 75	strength ≈ 85

anced coverage of both individual degradations and their co-occurrences. The same GT image is not assigned the same degradation combination more than once.

The selected degradations are applied according to a physically motivated image-formation order:

$$\text{haze/rain} \rightarrow \text{motion/defocus blur} \rightarrow \text{low light} \rightarrow \text{low resolution} \rightarrow \text{noise} \rightarrow \text{JPEG}.$$

Operations within the same stage are randomly permuted. When JPEG compression is present, a second JPEG operation is applied with probabilities 0.10, 0.30, and 0.50 for easy, medium, and hard samples, respectively. All GT images are resized to a maximum long edge of 1024 pixels, while the generated LQ images retain the same spatial resolution as their corresponding GTs.

Before degradation synthesis, each source GT is compared against the AgenticIR test set using a standardized 64×64 RGB SHA256 digest and perceptual dHash. An exact digest match or a dHash distance no larger than 2 is regarded as potential data leakage and excluded. After synthesis, we perform automatic quality control to detect abnormal degradation outcomes. A sample is flagged if its PSNR is below 6 dB, if the relative increase of black or white regions exceeds 0.4 compared with the GT, or if its normalized intensity standard deviation is below 0.02. For each flagged sample, the degradation types, severity levels, and operation order are kept fixed, while only the physical degradation parameters are resampled for up to 50 attempts. A final audit further verifies file validity, LQ/GT spatial consistency, content-level split integrity, degradation coverage, and task balance.

Prompt Construction Strategy. For each degradation category, we define a fixed task-specific prompt that explicitly specifies the observed corruption and the desired restoration operation. The prompts are designed with two complementary objectives: identifying the degradation to be removed and constraining the model to preserve image content unrelated to that degradation. In particular, each instruction describes the corresponding restoration intent, such as removing noise, haze, rain streaks, or blur, while explicitly discouraging unnecessary changes to content, color, composition, or already well-preserved regions. This design provides consistent degradation-aware guidance to the privileged generator without introducing sample-dependent language variation. The task-specific prompts used in our experiments are summarized in Table 9.

B Implementation Details

Training Configuration. We use FLUX.2-klein-base-9B as the backbone. Only the Transformer is fine-tuned with LoRA using rank 64 and scaling factor 64. The complete Stage 1 optimization and system configuration is summarized in Table 10. For RPSD, each outer iteration samples 16 conditioning images and generates 8 privileged candidates per image. The privileged and non-privileged branches are optimized separately with AdamW using a learning rate of 5×10^{-5} and no scheduler. For reward-guided self-distillation, we use a latent-space trust-region radius of $\rho = 0.1$

Table 9: Task-specific prompts used for degradation-aware restoration. Each prompt specifies the target degradation and constrains the restoration to preserve unrelated image content.

Degradation	**Task-Specific Prompt**
Compression	This is an image that has been compressed, resulting in visible artifacts such as blurring, pixelation, or loss of detail. Please restore the image by removing the compression artifacts and recovering lost details while keeping all other elements unchanged.
Defocus blur	This is an image with partial blurring and lack of clarity due to camera defocus. Transform this image into an all-in-focus photograph. Please carefully observe this picture to enhance the sharpness of the blurred areas while keeping all other elements consistent.
Haze	This is a hazy image with reduced clarity and contrast. Please completely remove the atmospheric haze while meticulously preserving all other elements without alteration.
Low light	This is a low-light image. Please restore it to normal illumination while keeping all other elements unchanged.
Low resolution	Please upscale this low-resolution image to the original resolution. Perform super-resolution processing to significantly enhance clarity, remove pixelation, and sharpen textures, while strictly preserving the original content, composition, and colors.
Motion blur	This is an image with blurring and lack of clarity due to motion. Transform this image into a sharp, static photograph. Please carefully observe this picture to eliminate motion streaks and recover details in blurred areas while keeping originally sharp elements consistent.
Noise	This is a noisy image. Please remove the noise while keeping all other elements unchanged.
Rain	This is a rainy image. Please remove the rain streaks while keeping all other elements, including the original color tone, lighting, and atmosphere, unchanged.

Table 10: Stage 1 adversarial post-training configuration.

Hyperparameter	**Value**
Generator learning rate	Constant, 5×10^{-5}
Discriminator learning rate	Constant, 5×10^{-5}
Batch size	Global batch 4 (1 per GPU on 4 GPUs)
Generator optimizer	RMSprop ($\alpha = 0.9$, $\mathrm{momentum} = 0$)
Discriminator optimizer	RMSprop ($\alpha = 0.9$, $\mathrm{momentum} = 0$)
MSE loss weight	1.0
LPIPS loss weight	1.0
Adversarial loss weight	0 for the first 200 iterations, then 0.1
LoRA rank	64
LoRA alpha	64
Training iterations	30,000
Training resources	4× NVIDIA A100 80GB PCIe
Gradient checkpointing	Enabled for both the generator and discriminator

and 2 reward-gradient ascent steps with step size $\eta = 1.0$. Specifically, the actual target update step is computed as $h_{\mathrm{step}} = \eta\rho\|z_i\|_2/M_{\mathrm{rgs}}$, where M_{rgs} denotes the number of reward-gradient steps. The non-privileged branch performs reward-weighted distillation over the 8 privileged endpoints with temperature $\tau_{distill} = 1.0$. The δ in Eq. 8 is set to 10^{-5}.

Training algorithm. Algorithm 1 summarizes Stage 2 RPSD post-training. The privileged, blind, and behavior policies are initialized from the same Stage 1 generator. Within each outer window, the behavior EMA samples a complete privileged candidate group, whose endpoints are scored and perturbed through bounded reward ascent/descent. The resulting targets are used to update the online privileged policy via Eq. 8. The same candidate group and its normalized reward weights are then reused to update the blind policy through reward-weighted matching in Eq. 10. After the privileged

Algorithm 1 RPSD post-training for USIR

Require: Stage 1 generator v_{init}, fixed reward model R_ϕ, training set $\mathcal{D} = \{(I_{\text{LQ}}, I_{\text{GT}}, p_{\text{task}})\}$
Require: Group size K, reward temperature τ, EMA decay η, and RPSD hyperparameters (β, κ, h)
Initialize: $v_P \leftarrow v_{\text{init}}$, $v_B \leftarrow v_{\text{init}}$, and $v_\theta \leftarrow v_{\text{init}}$
for each outer training window **do**
  Sample $(I_{\text{LQ}}, I_{\text{GT}}, p_{\text{task}}) \sim \mathcal{D}$
  Construct $c_P = (I_{\text{LQ}}, p_{\text{task}}, \sigma_c^P)$ and $c_B = (I_{\text{LQ}}, p_{\text{fix}}, \sigma_c^B)$
  Encode $z_{\text{LQ}} = \text{Enc}(I_{\text{LQ}})$ and sample $\{\epsilon_i\}_{i=1}^K$
  for $i = 1, \dots, K$ **do** ▷ Shared privileged rollout
    $z_{g,i} \leftarrow (1 - \sigma_g) z_{\text{LQ}} + \sigma_g \epsilon_i$
    $z_i \leftarrow z_{g,i} - \sigma_g v_P(z_{g,i}, c_P, \sigma_g)$
    $r_i \leftarrow R_\phi(\text{Dec}(z_i))$
  end for
  Compute optimality coefficients $\{a_i\}$ using Eq. 5
  Compute reward weights $\{w_i\}$ using Eq. 10
  Cache $\{z_{g,i}, z_i, r_i, w_i\}_{i=1}^K$ for both updates
  for $i = 1, \dots, K$ **do**
    Initialize $z_i^+ = z_i^- = z_i$ and construct detached targets $z_i^\pm$ using Eq. 6
    Evaluate behavior and online privileged predictions at $(z_{g,i}, c_P, \sigma_g)$ and construct $v_i^\pm$ using Eq. 7
    $z_i^\pm \leftarrow z_{g,i} - \sigma_g v_i^\pm$
  end for
  Update v_θ by minimizing $\mathcal{L}_P$ in Eq. 8
  EMA update v_P ▷ Privileged policy update
  for $i = 1, \dots, K$ **do**
    $z_i^B \leftarrow z_{g,i} - \sigma_g v_B(z_{g,i}, c_B, \sigma_g)$
  end for
  Update v_B by minimizing $\mathcal{L}_B$ in Eq. 10 with cached $\{z_i, w_i\}$ ▷ Blind policy update
end for
Output: blind restoration policy v_B

update, the behavior EMA is refreshed for the next outer window. After post-training, only the blind policy is retained for deployment.

C Additional Experiments and Experimental Details

C.1 Privileged Policy Optimization Baselines

We compare our reward-guided self-distillation with three reward-optimization baselines under the same LQ-centered single-step restoration setting. All methods are initialized from the same 30k-step FLUX.2-Klein-9B LoRA checkpoint and optimize only the privileged policy. We use rank-64 LoRA, bfloat16 precision, a learning rate of 5×10^{-5}, 1024×1024 inputs, and a single restoration step, with the same frozen reward model across all methods. Each optimization step aggregates 16 groups with $K = 8$ candidates per group. The blind branch is disabled throughout these comparisons.

ReFL. We adapt ReFL to the single-step restoration setting with an additional latent-space GT regularization term. Given the predicted endpoint $\hat{\mathbf{z}}$, GT latent $\mathbf{z}_{\text{gt}}$, and differentiable reward $R(\hat{\mathbf{z}})$, we optimize

$$\mathcal{L}_{\text{ReFL}} = \lambda_r \max\left(0,\, m - \frac{R(\hat{\mathbf{z}}) - \mu_r}{s_r}\right) + \lambda_{\text{gt}} \left\|\hat{\mathbf{z}} - \text{sg}(\mathbf{z}_{\text{gt}})\right\|_2^2 . \tag{13}$$

We use raw reward values with $\mu_r = 0$ and $s_r = 1$, margin $m = 0.65$, $\lambda_r = 1.0$, and $\lambda_{\text{gt}} = 0.1$. The GT latent is encoded once per group and detached from gradient computation. This baseline directly optimizes the differentiable reward and does not use reward-guided targets, trust-region projection, within-group reward normalization, or an additional reference LoRA.

One-step Flow-GRPO. We adapt Flow-GRPO to our single-step setting using a single Flow-SDE transition from the LQ-centered latent state to the restoration endpoint. For each group, rewards are standardized over the $K = 8$ candidates:

$$A_i = \mathrm{clip}\left(\frac{r_i - \bar{r}}{\mathrm{std}(r) + \epsilon}, -5, 5\right). \tag{14}$$

The policy is then optimized with the clipped likelihood-ratio objective

$$\mathcal{L}_{\mathrm{GRPO}} = \frac{1}{K}\sum_i \max\left[-A_i\rho_i,\ -A_i\,\mathrm{clip}\left(\rho_i, 1-10^{-4}, 1+10^{-4}\right)\right], \tag{15}$$

where ρ_i denotes the likelihood ratio between the current policy and the rollout policy for the sampled transition. We use Flow-SDE dynamics with noise level $\eta = 0.2$. This adaptation preserves the group-relative policy optimization mechanism of Flow-GRPO while matching the single-step restoration budget of USIR.

One-step DiffusionNFT. We adapt DiffusionNFT to deterministic single-step ODE sampling. For each rollout endpoint $\mathbf{z}_i$, we construct a second LQ-centered noisy state and evaluate both the EMA behavior velocity and the current-policy velocity. These predictions define positive and negative implicit-policy endpoints:

$$\mathbf{z}_i^+ = \tilde{\mathbf{x}}_i - \sigma\left[(1-\beta)\mathbf{v}_i^{\mathrm{old}} + \beta\mathbf{v}_i^{\theta}\right], \qquad \mathbf{z}_i^- = \tilde{\mathbf{x}}_i - \sigma\left[(1+\beta)\mathbf{v}_i^{\mathrm{old}} - \beta\mathbf{v}_i^{\theta}\right]. \tag{16}$$

The group-standardized and clipped reward is mapped to an optimality weight $w_i \in [0, 1]$, which balances the positive and negative endpoint-matching objectives. We use $\beta = 1.0$, an NFT loss scale of 2.0, and clip the standardized reward to $[-5, 5]$. Since this formulation operates directly on implicit endpoints, it avoids likelihood estimation and remains compatible with deterministic ODE sampling.

C.2 Additional Analysis of Table 4.

As shown in Table 4, reward-guided self-distillation achieves more effective privileged-policy optimization than the compared reward-optimization baselines. Compared with ReFL (Xu et al., 2023), which directly backpropagates a scalar reward through the generated output, our approach converts the reward signal into explicit, bounded positive and negative targets and then fits these targets as detached supervision. This separates reward-guided target construction from policy realization, providing a more structured optimization signal for single-step restoration. Flow-GRPO (Liu et al., 2026a) and DiffusionNFT (Zheng et al., 2025), in contrast, rely more strongly on relative reward differences among candidates within each rollout group. In our single-step restoration setting, the generation space around a given LQ input is relatively constrained, resulting in limited variation among sampled candidates and weaker within-group reward separation. This makes the relative advantages in Flow-GRPO and the reward-based positive–negative separation in DiffusionNFT less discriminative (Fei et al., 2026). Consequently, these group-relative objectives provide weaker optimization signals, whereas reward-guided self-distillation explicitly constructs improvement directions for each sampled endpoint and remains effective even when the original candidate group exhibits limited diversity.

C.3 Reward Design

We instantiate the composite reward in Eq. 11 using five complementary metrics:

$$R = 0.25R_{\mathrm{LPIPS}} + 0.20R_{\mathrm{PSNR}} + 0.15R_{\mathrm{SSIM}} + 0.25R_{\mathrm{CLIPIQA+}} + 0.15R_{\mathrm{Aes}} \tag{17}$$

Here, $R_{\mathrm{CLIPIQA+}}$ measures no-reference perceptual quality, R_{Aes} captures high-level aesthetic preference, while R_{LPIPS}, R_{PSNR}, and R_{SSIM} jointly measure perceptual, pixel-level, and structural fidelity to the GT, respectively. Accordingly, the fidelity component in Eq. 11 is instantiated by the LPIPS, PSNR, and SSIM terms, while the quality and aesthetic components correspond to CLIP-IQA+ and the aesthetic score from Schuhmann et al. (2022). This combination assigns substantial weight to fidelity while retaining explicit perceptual-quality and aesthetic signals, providing a balanced restoration-oriented reward for policy optimization.

Perceptual Fidelity. We use LPIPS to measure perceptual consistency between the restored image I_r and the ground-truth image I_g. Since LPIPS is a perceptual distance, we convert it into a higher-is-better reward using an exponential mapping:

$$R_{\text{LPIPS}} = \exp\left(-\frac{\text{LPIPS}(I_r, I_g)}{\tau_{lpips}}\right), \qquad \tau_{lpips} = 0.25. \tag{18}$$

This transformation provides a bounded reward and places greater emphasis on fine perceptual differences when the restored image is already close to the ground truth.

No-Reference Perceptual Quality. We use a frozen CLIP-IQA+ model to assess the perceptual quality of the restored image independently of the ground truth. Its output is normalized as

$$R_{\text{CLIP-IQA+}} = \text{clip}\left(\frac{\text{CLIP-IQA+}(I_r) - a_{\min}}{a_{\max} - a_{\min}}, 0, 1\right), \tag{19}$$

where $a_{\min} = 0$ and $a_{\max} = 1$ in our implementation. This component provides a no-reference quality signal that penalizes visually undesirable effects such as residual noise, blur, compression artifacts, and unnatural textures.

Pixel Fidelity. We additionally use PSNR to provide an explicit pixel-level reconstruction signal. Given the RGB-space mean squared error

$$\text{MSE} = \frac{1}{3HW}\left\|I_r - I_g\right\|_2^2, \tag{20}$$

the corresponding PSNR is computed as

$$\text{PSNR} = 10\log_{10}\frac{1}{\max(\text{MSE}, \epsilon)}. \tag{21}$$

We linearly normalize PSNR from $[10, 50]$ dB to $[0, 1]$:

$$R_{\text{PSNR}} = \text{clip}\left(\frac{\text{PSNR} - 10}{50 - 10}, 0, 1\right). \tag{22}$$

This term explicitly encourages pixel-level consistency in color, luminance, and image content.

Structural Fidelity. We further employ SSIM to measure structural consistency between the restored image and the ground truth:

$$R_{\text{SSIM}} = \text{SSIM}(I_r, I_g). \tag{23}$$

Since SSIM naturally lies in a bounded range with larger values indicating higher structural similarity, we directly use it as a reward component. It complements LPIPS and PSNR by emphasizing local structural preservation.

Aesthetic Prior. To complement perceptual image quality with higher-level aesthetic preference, we employ the frozen LAION Improved Aesthetic Predictor. The restored image is first encoded by a frozen OpenAI CLIP ViT-L/14 image encoder, and the resulting feature is ℓ_2-normalized:

$$\mathbf{f}_r = \frac{\text{CLIP}_{\text{ViT-L/14}}(I_r)}{\left\|\text{CLIP}_{\text{ViT-L/14}}(I_r)\right\|_2}. \tag{24}$$

A frozen aesthetic MLP then maps the normalized feature to a raw aesthetic score:

$$q_{\text{AES}} = h_{\text{AES}}(\mathbf{f}_r). \tag{25}$$

The output is normalized using the nominal AVA score range:

$$R_{\text{AES}} = \text{clip}\left(\frac{q_{\text{AES}} - 1}{10 - 1}, 0, 1\right). \tag{26}$$

This component encourages visually pleasing color, composition, and overall appearance beyond low-level restoration fidelity.

Table 11: Reward-choice ablation for full RPSD on MiO100 Group B. All variants use the same training and evaluation protocol. The RPSD reference results are repeated from Table 5.

Reward / configuration	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**MANIQA↑**	**CLIPIQA↑**	**MUSIQ↑**
LPIPS	22.69	0.7400	0.1374	0.4050	0.6508	69.12
CLIPIQA	18.31	0.5659	0.2697	0.6898	0.8464	75.22
DISTS + CLIPIQA+	22.71	0.7236	0.1466	0.4862	0.7306	71.60
Ours	22.74	0.7407	0.1458	0.4467	0.6818	70.36

C.4 Reward Choice.

We study the influence of the reward signal used in RPSD by replacing our default composite reward with several alternatives, including LPIPS, CLIP-IQA, and DISTS+CLIPIQA+, while keeping the remaining training configuration unchanged. Each reward is consistently used for both reward-guided target construction and preference weighting in privileged self-distillation.

As shown in Table 11, RPSD remains effective under different reward formulations, demonstrating that its optimization mechanism is not tied to a specific reward design. Single-objective rewards exhibit clear preferences: full-reference rewards primarily improve restoration fidelity, whereas no-reference rewards place greater emphasis on perceptual quality. Combining complementary signals allows the reward to capture multiple aspects of restoration preference simultaneously, leading to a better balance between fidelity and perceptual quality. Across these different choices, RPSD consistently achieves strong restoration performance, highlighting its ability to accommodate diverse reward signals and optimization objectives.

D User Study

To complement the quantitative evaluation, we conduct a user study to assess human preference over restoration results. The study involves 12 participants and uses 100 samples from MiO100 Group B, covering four degradation categories with 25 images per category. For each sample, we present the degraded input and the ground-truth image together with five anonymously shuffled restoration results produced by PromptIR, MiOIR, FAPE-IR, 4K-Agent, and USIR. The five results are randomly assigned as options A–E to avoid method-specific bias. Participants are asked to select the restoration result they prefer most in a blind setting. The evaluation interface is illustrated in Fig. 6. In total, we collect 1200 valid responses, with an average decision time of 17.15 seconds per question.

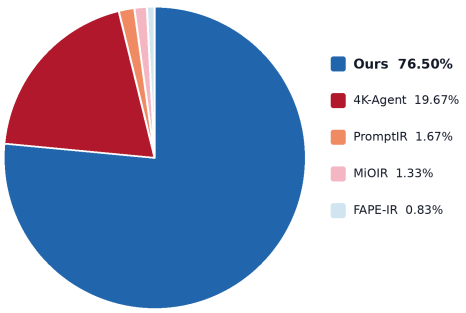

Figure 5: Overall human preference distribution in the user study. USIR receives the dominant preference share, substantially surpassing all compared baselines.

The overall preference statistics are shown in Fig. 5. USIR receives the highest preference share of 76.50%, substantially outperforming all competing methods. Among the baselines, 4K-Agent ranks second with 19.67%, while PromptIR, MiOIR, and FAPE-IR obtain only 1.67%, 1.33%, and 0.83%, respectively. These results demonstrate that the restoration outputs of USIR are consistently favored by human observers, validating its advantage in visual quality beyond standard full-reference and no-reference metrics.

E Additional Qualitative Results

We provide additional qualitative results to further evaluate USIR under challenging mixed degradations and real-world inputs. As shown in Fig. 7, USIR consistently removes severe compound

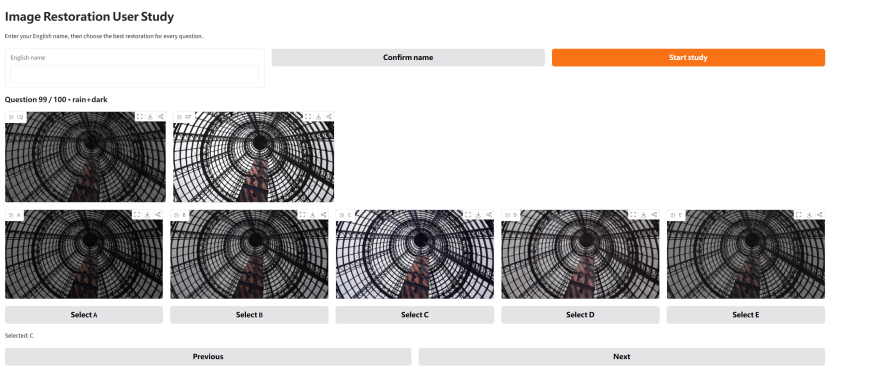

Figure 6: User-study interface. For each question, participants are shown the degraded input, the ground truth, and five anonymously shuffled restoration results from different methods, and are asked to select the preferred result.

degradations on MiO100 Group C while preserving image structures and fine details. Compared with existing all-in-one and agentic restoration methods, USIR produces cleaner and more natural restorations with fewer residual artifacts.

We further evaluate USIR on the Real-1000 benchmark to examine its generalization to real-world degraded images. As illustrated in Fig. 8, USIR effectively handles diverse real-world degradation patterns and produces visually clean and coherent restorations. These results further demonstrate the robustness of the learned single-step restoration capability beyond synthetic mixed-degradation settings.

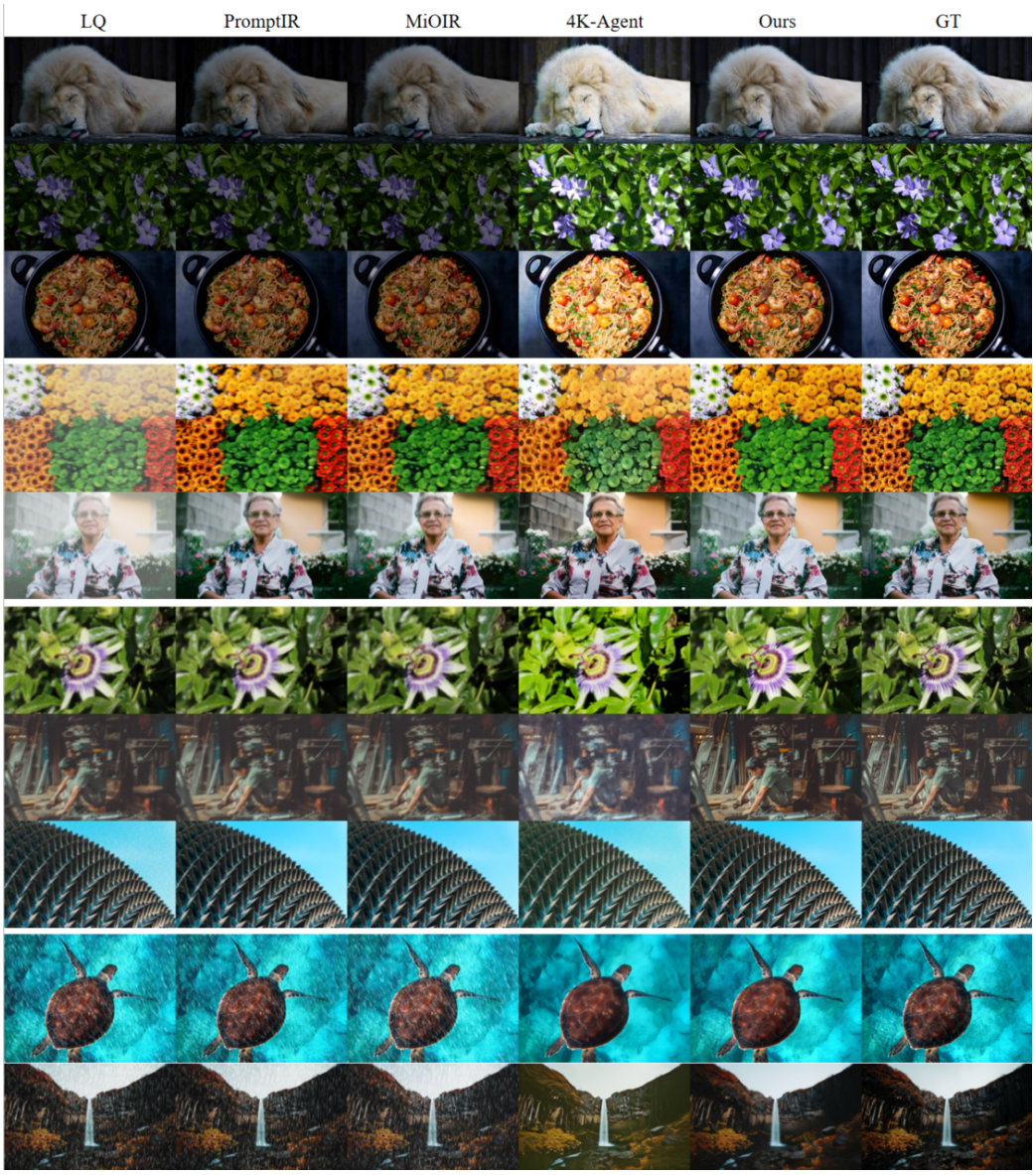

Figure 7: Additional qualitative comparisons on MiO100 Group C under challenging compound degradations. USIR more effectively removes mixed degradations while preserving image structures and fine details, producing cleaner and more natural restoration results.

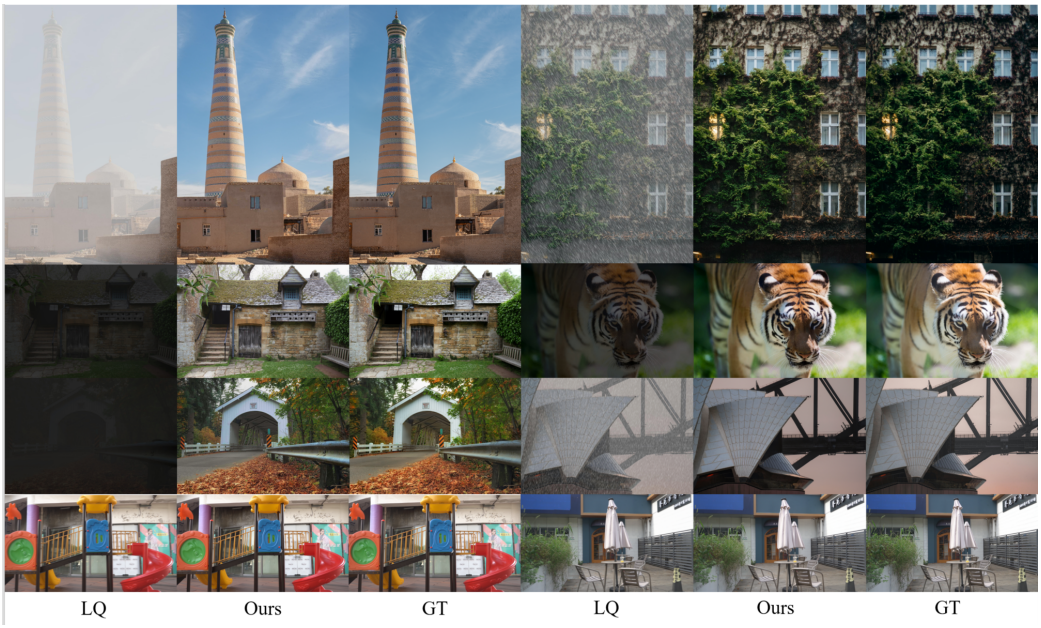

Figure 8: Qualitative results of USIR on Real-1000. USIR effectively restores images affected by diverse real-world degradations, producing clean and visually coherent results while preserving the underlying image content.