# REALRESTORER: TOWARDS GENERALIZABLE REAL-WORLD IMAGE RESTORATION WITH LARGE-SCALE IMAGE EDITING MODELS

**Anonymous authors**
Paper under double-blind review




Figure 1: RealRestorer effectively restores diverse real-world image degradations, including deblurring, moiré pattern removal, compression restoration, reflection removal, hazing removal, rain removal, deflare, and low-light enhancement.

## ABSTRACT

Image restoration under real-world degradations is critical for downstream tasks such as autonomous driving and object detection. However, existing restoration models are often limited by the scale and distribution of their training data, resulting in poor generalization to real-world scenarios. Recently, large-scale image editing models have shown strong generalization ability in restoration tasks, especially for closed-source models like Nano Banana Pro, which can restore images while preserving consistency. Nevertheless, achieving such performance with those large universal models requires substantial data and computational costs. To address this issue, we construct a large-scale dataset covering nine common real-world degradation types and train a state-of-the-art open-source model to narrow the gap with closed-source alternatives. Furthermore, we introduce RealIR-Bench, which contains 464 real-world degraded images and tailored evaluation metrics focusing on degradation removal and consistency preservation. Extensive experiments demonstrate our model ranks first among open-source methods, achieving state-of-the-art performance.

## 1 INTRODUCTION

Image restoration Liang et al. (2021); Gunturk & Li (2018); Lehtinen et al. (2018); Zamir et al. (2021); Li et al. (2023) aims to recover high-quality images from degraded observations and serves as a fundamental building block for downstream applications such as autonomous driving Hu et al.



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.

(2023); Caesar et al. (2020), remote sensing Wang et al. (2024), detection Hu et al. (2018); Joseph et al. (2021), and 3D reconstruction Yao et al. (2018). However, real-world images often suffer from diverse and co-existing degradations Liang et al. (2022); Huang et al. (2020); Flusser et al. (2015); Damera-Venkata et al. (2000); Zhang et al. (2021); Dufaux et al. (2000); Jiang et al. (2021a); Liu et al. (2021); Guo & Hu (2023); He et al. (2021); Hermann (2012); Al-Salem et al. (2015); Tong & White (1996); Benz (2017), including blur, rain, noise, low-light, moiré patterns, haze, compression artifacts, reflection, and flare. This complexity goes beyond the single degradation and single model paradigm.

To address this, recent all-in-one restoration methods Li et al. (2022); Zamfir et al. (2025); Potlapalli et al. (2023); Huang et al. (2024) attempt to handle multiple degradations within a unified framework. Nevertheless, they often rely on a limited set of synthetic degradation distributions, while collecting large-scale real degraded-clean pairs remains expensive and difficult. As a result, these models can generalize poorly to real-world scenarios. In parallel, large image editing models trained on massive editing datasets have recently demonstrated strong restoration capabilities Zuo et al. (2025), such as Nano Banana Pro Team et al. (2023) and GPT-Image-1.5 OpenAI (2025). However, these models are typically trained with closed-source data and compute, which makes them hard to reproduce and limits their utility for the research community. Despite this, leveraging the strong priors learned by image editing models provides a promising path to overcome the key limitation of traditional restoration approaches.

However, conventional restoration datasets typically cover a narrow degradation distribution, which is often not representative of real-world conditions. As shown in Figure 2, small-scale models trained only on fixed synthetic datasets often struggle to generalize to real-world degradations. In contrast, large-scale image editing models exhibit strong zero-shot enhancement capability and produce more natural results. This observation highlights the need for a more realistic dataset generation pipeline that can produce diverse and generalizable training data. Moreover, evaluation protocols that rely primarily on reference-based metrics further exacerbate this issue, since such metrics may fail to reflect perceptual quality, robustness across diverse degradations, and detail consistency in real scenes.

To bridge these gaps, we design a comprehensive degradation synthesis pipeline to generate high-quality training data, aiming to narrow the gap between synthetic and real-world degradations. Based on this dataset, we fine-tune an open-source image editing model **RealRestorer** across nine restoration tasks, and further introduce a new benchmark **RealIR-Bench** to evaluate restoration performance under real-world degradations.

In summary, our contributions are threefold:

- We introduce **RealIR-1M**, a high-quality image restoration training dataset that covers common degradation types and incorporates strictly filtered real-world degradation pairs with low hallucination artifacts.
- We develop a new benchmark, **RealIR-Bench**, grounded in real-world cases, to evaluate both degradation restoration and consistency preservation. By addressing the lack of reliable evaluation protocols for real-world restoration, it enables more authentic and comprehensive assessment of restoration models.
- We develop **RealRestorer**, an open-source real-world image restoration model that sets a new state of the art and achieves performance highly comparable to closed-source systems. We will release the model to facilitate future research in real-world restoration.

## 2 Related Work

### 2.1 Single-Degradation Restoration

Single-degradation restoration methods typically focus on removing one specific type of degradation under constrained and well-defined scenarios. With the rapid development of deep learning, numerous works Nah et al. (2017); Li et al. (2018); Cai et al. (2023); Zhao et al. (2025); Huang et al. (2024) have achieved impressive performance on individual tasks such as deblurring, haze removal, low-light enhancement, deflare, and reflection removal. These approaches often rely on carefully designed architectures and degradation-specific priors, enabling strong performance.




Figure 2: For small-scale models, EnlightenGAN Jiang et al. (2021b) (114.35M) and Retinexformer Cai et al. (2023) (1.61M), trained on fixed datasets such as LOL Yang et al. (2021), often suffer from over-enhancement artifacts when generalized to real-world low-light images. In contrast, GPT-Image-1.5 OpenAI (2025) and our model produce more natural and visually plausible results, as indicated by the red-box regions in the figure.

However, most single-degradation models are built upon task-specific assumptions, where the degradation type is predefined and relatively homogeneous, which makes models trained for a single degradation tend to generalize poorly and may even introduce secondary artifacts when encountering unseen or compound degradations.

Moreover, many existing methods are trained and evaluated primarily on synthetic datasets with simplified degradation models, which may not faithfully represent the complexity of real-world data distributions. This gap between synthetic training data and real-world testing scenarios further limits their robustness and practical applicability. Consequently, while single-degradation methods achieve strong performance on benchmark datasets, their effectiveness in real-world applications remains constrained.

### 2.2 All-in-One Image Restoration

All-in-one approaches Li et al. (2025); Lin et al. (2024); Potlapalli et al. (2023); Luo et al. (2023); Li et al. (2022); Cui et al. (2025); Zamfir et al. (2025); Guo et al. (2024) aim to handle multiple degradations within a unified network by balancing shared representations and task-specific components. Nevertheless, many of these methods still rely heavily on synthetic datasets with limited and overly simplified degradation patterns. Such a narrow training distribution often results in weak robustness and poor generalization to real-world degradations, where corruption characteristics are diverse, complex, and domain-dependent.

Meanwhile, large diffusion or flow-matching image editing models Lipman et al. (2022); Esser et al. (2024); Peebles & Xie (2023); Rombach et al. (2022) have recently demonstrated strong semantic priors for image enhancement and restoration. Trained on massive image–text pairs, these image editing models Labs et al. (2025); Liu et al. (2025); Wu et al. (2025); Team et al. (2025) can leverage semantic conditioning and often generalize better to real-world data than small specialized restoration networks. Therefore, transferring and exploiting the priors of large image editing models provides a promising direction for building restoration systems with stronger real-world generalization.

Motivated by this observation, we develop a high-quality and realistic degradation synthesis pipeline covering nine major degradations and use it to fine-tune open-source image editing models for robust real-world restoration while maintaining strong content consistency. Furthermore, to evaluate real-world restoration performance in the absence of clean references, we curate a benchmark of 464 real images spanning nine single-degradation categories, and propose new evaluation metrics that measure both degradation removal ability and consistency with the input content. Based on the proposed dataset and metrics, our fine-tuned model achieves state-of-the-art performance among open-source methods and is competitive with closed-source systems, while qualitative results further demonstrate strong generalization to real-world degradations.

## 3 RealRestorer

### 3.1 RealIR-1M

Existing image restoration datasets Li et al. (2025); Guo et al. (2024) often rely on a single degradation model to synthesize degraded images and use a fixed composition strategy to explicitly disentangle degradation features for representation learning. These modeling approaches are effective for specific degradation settings. However, in real-world scenarios, degradations are far more complex and diverse. Simple synthetic degradation models are usually insufficient to approximate real degradation

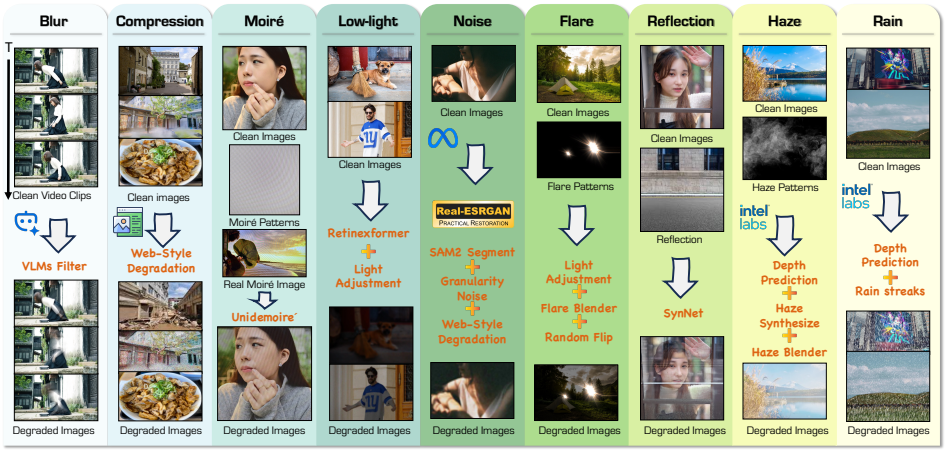


Figure 3: Overview of our large-scale Synthetic Degradation Data pipeline. We construct nine representative degradation types, including blur, compression artifacts, moiré patterns, low-light, noise, flare, reflection, haze, and rain. Compared with previous synthetic-only pipelines, our upgraded framework incorporates granular noise modeling, segment-aware perturbations, and web-style degradation processes, significantly narrowing the gap between synthetic and real-world distributions. This comprehensive pipeline enables more robust and generalizable restoration learning.

distributions, and they are often not robust enough for large-scale training that aims at strong generalization.

To address this limitation, we adopt two main strategies to construct high-quality paired data for nine image restoration tasks. The Real-World Degradation Pipeline plays a key role in improving generalization to real-world degradations, while the Synthetic Degradation Pipeline serves as a complementary component that covers common degradation patterns and helps prevent overfitting.

**Synthetic Degradation Data**: We synthesize degradations from clean images, which offers high scalability but cannot fully capture the diversity of real-world degradations. Still, synthetic data provides useful supervision for transferring image editing priors and learning foundational restoration knowledge. We use SAM-2 Ravi et al. (2024) and MiDaS Ranftl et al. (2022) to filter unsuitable samples and provide structural cues, such as semantic masks and depth maps, for degradation synthesis.

To ensure data quality, we further employ Vision-Language Models (VLMs) and quality assessment models Minderer et al. (2023) to remove low-quality or watermarked images. After constructing image pairs, we verify the degradation alignment between degraded inputs and restored references to ensure reliable and learnable paired supervision. The synthetic data generation pipeline is illustrated in Figure 3.

**Real-World Degradation Data**: Collect real degraded images and generate corresponding clean images by removing degradations using high performance restoration models. Compared with synthetic pairing, this approach is more likely to preserve the true degradation statistics of real-world data, enabling restoration models trained on such pairs to generalize better to real scenarios. To bridge the gap between synthetic and real-world degradations, we collect real degraded images from the web and pair them with high-quality references.

During web data collection, we first use CLIP Radford et al. (2021) to retrieve images with degradation-related semantic cues, followed by watermark detection and quality assessment filters to remove unsuitable samples. After generating clean references with high-performance image generation models, we further conduct an edge-shift analysis to verify pairwise alignment.

Specifically, given an RGB degraded image $I$, we first convert it to luminance:

$$Y = 0.299I_R + 0.587I_G + 0.114I_B, \tag{1}$$

where $I_R$, $I_G$, and $I_B$ denote the red, green, and blue channels, respectively. We then compute the Sobel edge response:

$$E(I) = |S_x(Y)| + |S_y(Y)|, \tag{2}$$

where $S_x$ and $S_y$ denote the horizontal and vertical Sobel filters, respectively. Given a degraded image $A$ and its clean reference $B$, the edge-difference map is defined as:

$$I_{\text{edge-diff}}(A, B) = |E(A) - E(B)|. \tag{3}$$

This metric mainly reflects local structural discrepancies while suppressing illumination and color variations. In our curated real-world pairs, fewer than $20\%$ of samples exhibit edge shifts larger than $5$ pixels, indicating strong visual alignment. A subset of pairs is further manually reviewed to ensure consistent degradation type and severity. These high-quality real-world degradation pairs help the model better adapt to realistic data distributions Podell et al. (2023); Team (2025b); Team et al. (2025).

Representative visual examples of our training set are presented in Figure 4.




Figure 4: Examples from our training dataset containing both synthetic and real-world degradation pairs. The upper rows with gray labels show synthesized degradations generated by our pipeline, while the bottom rows highlighted with orange labels correspond to real-world degraded images paired with clean references.

### 3.2 Method and Training Strategy

We fine-tune the base model Step1X-Edit Liu et al. (2025) built on a large Diffusion in Transformer (DiT) backbone Peebles & Xie (2023), which is effective for generation. It is equipped with QwenVL Bai et al. (2025) as a text encoder that injects high-level semantic extraction into the DiT denoising pathway. The reference image and output image are both encoded into latent space by Flux-VAE Labs (2024). Starting from the original image editing model, we fine-tune on nine restoration tasks in two stages: a Transfer-training stage for large-scale restoration transfer and a Supervised Fine-tuning stage for constraining the manifold of the final model distribution.

**Transfer Training Stage:** In the first stage, we use synthetic paired data to transfer high-level knowledge and priors from image editing to image restoration. Since we initialize from a pretrained backbone, we eschew progressive resolution schedules Liu et al. (2025) for training. Instead, we adopt a high-resolution setting of 1024×1024 throughout the entire training process. The learning rate is kept constant at $1e^{-5}$, and the global batch size is set to 16. After several steps of transfer training, **RealRestorer** begins to exhibit signs of knowledge transfer from high-level image editing tasks to image restoration tasks, which is insufficient in the base model.

Although **RealRestorer** gradually acquires the basic capability to handle simple degradation patterns across all nine tasks, its ability to distinguish and model diverse real-world degradation patterns remains limited. This observation motivates us to introduce a second training stage aimed at improving generalization and restoration quality under real-world degradation scenarios.

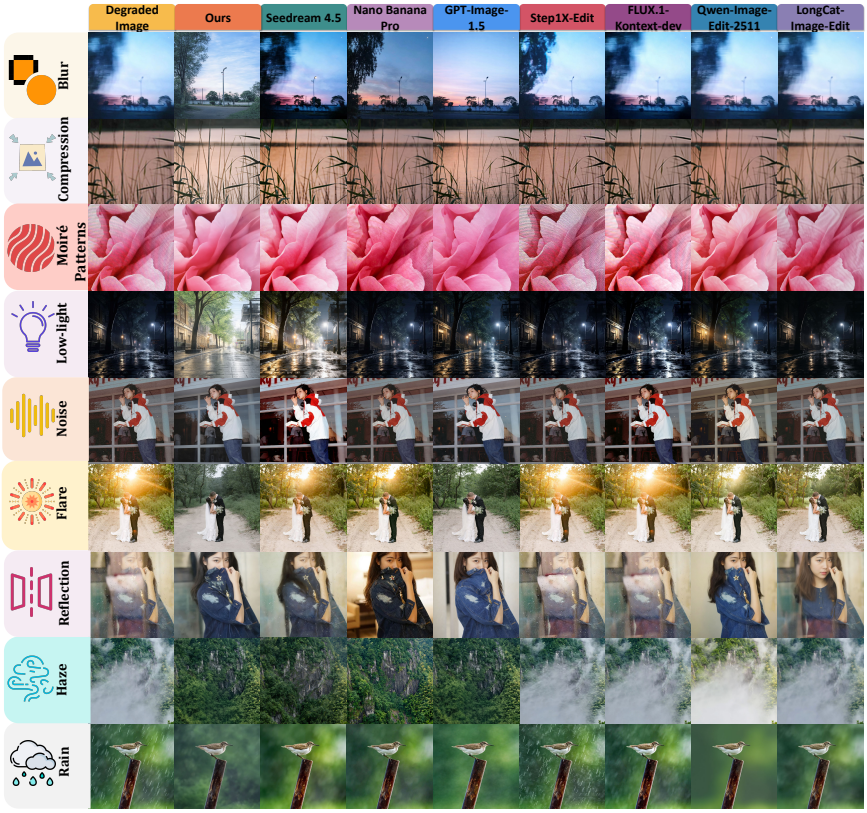


Figure 5: Comparison with state-of-the-art image editing models across nine real-world degradations, including blur, compression artifacts, moiré patterns, low-light, noise, flare, reflection, haze, and rain. We compare our method with large-scale image editing models, such as Seedream 4.5, Nano Banana Pro, GPT-Image-1.5, Step1X-Edit, FLUX.1-Kontext-dev, Qwen-Image-Edit-2511, and LongCat-Image-Edit.

**Supervised Fine-tuning Stage:** For the second training stage, we incorporate real-world degradation data to further enhance restoration fidelity and improve generalization under real-world degradation scenarios Wu et al. (2025); Team (2025b); Team et al. (2025). Compared with the first stage, this stage emphasizes adaptation to complex and authentic degradation patterns. We adopt a cosine annealing learning rate schedule, where the learning rate is gradually decayed to zero, using the same initial learning rate as in the first stage.

Importantly, instead of completely replacing synthetic data, we adopt a **Progressively-Mixed** training strategy, which retains a small proportion of synthetic paired samples during the second stage. **RealRestorer** is first exposed to diverse synthetic degradations to build broad generalization, and then gradually adapted to real-world degradations while maintaining exposure to synthetic distributions. Throughout the two-stage training process, we select the intermediate checkpoint with the best generalization capability to maintain a balanced performance across multiple tasks and ensure strong overall performance of the final model. All our experiments are conducted on 8 NVIDIA H800 GPUs. More implementation details can be found in Appendix B.

## 4 Benchmark and Evaluation

### 4.1 RealIR-Bench

Traditional image restoration benchmarks primarily focus on single-degradation tasks with synthetic corruptions or limited degradation patterns, which makes them insufficient for evaluating model performance in real-world applications Guan et al. (2025); Guo et al. (2024); Rajagopalan et al. (2024); Li et al. (2025). Such benchmarks often fail to capture the complexity, diversity, and unpredictability of degradations encountered in practical scenarios.

To properly evaluate restoration performance under real-world degradations, we construct a new benchmark composed entirely of internet-sourced, naturally degraded images. The proposed benchmark spans nine common restoration tasks and covers a wide range of degradation types frequently observed in real-world photography, including blur, rain, noise, low-light, moiré patterns, haze, compression artifacts, reflection, and flare, which collectively represent the most common forms of real-world image degradation.

In total, the benchmark contains 464 non-reference degraded images for testing. To ensure a fair and consistent evaluation protocol, we adopt a fixed enhancement instruction for all samples. This design minimizes the influence of instruction variation and allows the evaluation to focus more directly on a model's restoration capability and its ability to preserve image consistency. More details about the benchmark construction and data statistics are provided in Appendix C.

Table 1: Quantitative comparison on the Rain Removal, Deblurring, Low-light Enhancement, Haze Removal, and Reflection Removal tasks. We compared state-of-the-art (SOTA) image editing models. For each task, we reported LPS (↓), RS (↑), and FS (↑). The best result is marked in **bold**, and underline indicates the second-best result. The best and second-best open-source results are highlighted with yellow and blue backgrounds, respectively.

| Method | Open-source | Rain Removal | | | Deblurring | | | Low-light Enhancement | | | Haze Removal | | | Reflection Removal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LPS↓ | RS↑ | FS↑ | LPS↓ | RS↑ | FS↑ | LPS↓ | RS↑ | FS↑ | LPS↓ | RS↑ | FS↑ | LPS↓ | RS↑ | FS↑ |
| *Image Editing Methods* | | | | | | | | | | | | | | | | |
| Nano Banana Pro Team et al. (2023) | No | 0.429 | 2.063 | **0.236** | 0.326 | 1.068 | 0.144 | 0.467 | 0.720 | 0.077 | 0.492 | **1.920** | **0.195** | 0.358 | 1.368 | 0.176 |
| GPT–Image–1.5 OpenAI (2025) | No | 0.535 | **2.120** | 0.197 | 0.532 | 1.667 | 0.156 | 0.523 | 1.048 | 0.100 | 0.558 | 1.840 | 0.163 | 0.468 | **2.320** | **0.247** |
| Seedream 4.5 Seedream et al. (2025b) | No | 0.438 | 1.500 | 0.169 | 0.254 | 0.255 | 0.038 | 0.423 | 0.600 | 0.069 | 0.418 | 1.140 | 0.133 | 0.291 | 1.156 | 0.164 |
| LongCat–Image–Edit Team et al. (2025) | Yes | 0.381 | 1.302 | 0.161 | 0.200 | 0.000 | 0.000 | 0.158 | 0.120 | 0.020 | 0.236 | 0.000 | 0.000 | 0.254 | 1.060 | 0.158 |
| Qwen–Image–Edit–2511 Wu et al. (2025) | Yes | 0.435 | 1.736 | 0.196 | 0.170 | 0.240 | 0.040 | 0.122 | 0.080 | 0.014 | 0.337 | 0.060 | 0.008 | 0.333 | 1.820 | 0.243 |
| FLUX.1-Kontext-dev Labs et al. (2025) | Yes | **0.244** | 0.673 | 0.102 | **0.090** | 0.104 | 0.019 | **0.108** | 0.160 | 0.029 | **0.058** | 0.020 | 0.004 | **0.048** | 0.127 | 0.024 |
| Step1X–Edit Liu et al. (2025) | Yes | 0.282 | 0.019 | 0.003 | 0.321 | 0.906 | 0.123 | 0.306 | 0.340 | 0.047 | 0.194 | 0.190 | 0.031 | 0.247 | 0.080 | 0.012 |
| **RealRestorer (ours)** | Yes | 0.371 | 1.076 | 0.135 | 0.582 | **1.900** | **0.159** | 0.597 | **1.360** | **0.110** | 0.339 | 0.680 | 0.090 | 0.290 | 1.620 | 0.230 |

## 4.2 Experimental Results on RealIR-Bench

Based on RealIR-Bench, we evaluate a diverse set of large image editing models' ability in image restoration towards the real world, covering state-of-the-art closed-source systems such as GPT-Image-1.5 OpenAI (2025), Nano Banana Pro Team et al. (2023), Seedream 4.5 Seedream et al. (2025b), as well as strong open models including Qwen-Image-Edit-2511 Wu et al. (2025), FLUX.1-Kontext-dev Labs et al. (2025), LongCat-Image-Edit Team et al. (2025) and Step1X-Edit Liu et al. (2025). We provide nine major degradation tasks for evaluation: deblurring, rain removal, denoise, low-light enhancement, moiré patterns removal, haze removal, compression restoration, reflection removal, and deflare, with task-specific English instructions for each model to remove the corresponding degradation.

Table 2: Quantitative comparison on the Deflare, Demoiré, Denoise, and Compression-restoration tasks. The average results of all 9 tasks are reported in the last column. We compared state-of-the-art (SOTA) image editing models. For each task, we reported LPIPS (↓), RS (↑), and FS (↑). The best result is marked in **bold**, and underline indicates the second-best result. The best and second-best open-source results are highlighted with yellow and blue backgrounds, respectively.

| Method | Open-source | Deflare | | | Moiré Patterns Removal | | | Denoise | | | Compression Restoration | | | Avg Total (9) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LPS↓ | RS↑ | FS↑ | LPS↓ | RS↑ | FS↑ | LPS↓ | RS↑ | FS↑ | LPS↓ | RS↑ | FS↑ | LPS↓ | RS↑ | FS↑ |
| *Image Editing Methods* | | | | | | | | | | | | | | | | |
| Nano Banana Pro Team et al. (2023) | No | 0.214 | 1.222 | 0.192 | 0.562 | 1.560 | 0.137 | 0.386 | 0.712 | 0.087 | 0.483 | 1.122 | **0.116** | 0.413 | 1.306 | **0.153** |
| GPT–Image–1.5 OpenAI (2025) | No | 0.336 | 1.415 | 0.188 | 0.646 | 1.633 | 0.116 | 0.496 | **0.993** | **0.100** | 0.633 | **1.167** | 0.086 | 0.525 | **1.578** | 0.150 |
| Seedream 4.5 Seedream et al. (2025a) | No | 0.225 | 1.104 | 0.171 | 0.548 | 1.600 | **0.145** | 0.387 | 0.770 | 0.094 | 0.529 | 1.136 | 0.107 | 0.390 | 1.029 | 0.125 |
| LongCat–Image–Edit Team et al. (2025) | Yes | 0.241 | **1.717** | **0.261** | 0.420 | 1.200 | 0.139 | **0.350** | 0.471 | 0.061 | **0.188** | 0.083 | 0.014 | 0.270 | 0.661 | 0.097 |
| Qwen–Image–Edit–2511 Wu et al. (2025) | Yes | 0.222 | 1.660 | 0.258 | 0.595 | **1.660** | 0.135 | 0.429 | 0.824 | 0.094 | 0.242 | 0.300 | 0.046 | 0.320 | 0.931 | 0.127 |
| FLUX.1-Kontext-dev Labs et al. (2025) | Yes | **0.064** | 0.264 | 0.049 | **0.348** | 0.540 | 0.070 | 0.429 | 0.628 | 0.072 | 0.429 | 0.628 | 0.072 | **0.202** | 0.349 | 0.056 |
| Step1X–Edit Liu et al. (2025) | Yes | 0.173 | 0.000 | 0.000 | 0.654 | 0.410 | 0.028 | 0.409 | 0.098 | 0.012 | 0.344 | 0.383 | 0.050 | 0.325 | 0.270 | 0.036 |
| **RealRestorer (ours)** | Yes | 0.239 | 1.623 | 0.247 | 0.563 | 1.620 | 0.142 | 0.478 | 0.863 | 0.090 | 0.547 | 1.067 | 0.097 | 0.445 | 1.312 | 0.146 |

Unlike full-reference metrics such as PSNR and SSIM Wang et al. (2004), which require paired clean reference images for evaluation, RealIR-Bench is built entirely from non-reference images collected from diverse real-world scenarios. In these cases, obtaining perfectly aligned clean targets is infeasible, making conventional full-reference evaluation protocols unsuitable. Therefore, instead of relying on pixel-wise fidelity measures, we adopt a non-reference evaluation framework to assess

Table 3: Quantitative comparison on the FoundIR dataset across various real-world degradations. We report PSNR (↑) and SSIM (↑). The best results are highlighted in **bold**, and the second-best results are underlined.

| Method | Blur | | Rain | | Raindrops | | Noise | | Low-light | | Haze | | Compression | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| Nano Banana Pro Team et al. (2023) | **20.16** | **0.68** | 21.55 | 0.65 | 23.29 | 0.66 | 24.78 | 0.89 | 15.73 | 0.73 | 17.63 | 0.50 | 19.48 | 0.58 | 20.37 | 0.67 |
| GPT–Image–1.5 OpenAI (2025) | 13.25 | 0.41 | 12.52 | 0.34 | 13.89 | 0.31 | 14.15 | 0.68 | 11.00 | 0.55 | 12.39 | 0.28 | 13.06 | 0.39 | 12.89 | 0.42 |
| Seedream 4.5 Seedream et al. (2025b) | 17.94 | 0.61 | 17.61 | 0.57 | 18.01 | 0.58 | 15.57 | 0.78 | 13.64 | 0.68 | 15.65 | 0.44 | 14.29 | 0.54 | 16.10 | 0.60 |
| LongCat–Image–Edit Team et al. (2025) | 18.53 | 0.65 | 17.09 | 0.52 | 18.22 | 0.48 | 19.77 | 0.81 | 15.24 | 0.70 | 10.52 | 0.36 | 15.64 | 0.47 | 16.43 | 0.57 |
| Qwen–Image–Edit–2511 Wu et al. (2025) | 15.60 | 0.55 | 14.68 | 0.46 | 15.19 | 0.42 | 19.73 | 0.81 | 7.13 | 0.17 | 7.20 | 0.34 | 18.29 | 0.61 | 13.98 | 0.48 |
| FLUX.1-Kontext-dev Labs et al. (2025) | 10.73 | 0.41 | 10.48 | 0.35 | 11.25 | 0.31 | 15.10 | 0.69 | 10.34 | 0.53 | 10.74 | 0.29 | 10.74 | 0.34 | 11.34 | 0.42 |
| Step1X–Edit Liu et al. (2025) | 16.38 | 0.63 | 19.91 | 0.63 | 19.42 | 0.58 | 27.18 | **0.91** | 15.72 | 0.70 | 14.20 | 0.45 | **22.81** | **0.75** | 19.37 | 0.66 |
| **RealRestorer (ours)** | 18.99 | 0.59 | **23.72** | **0.71** | **23.64** | **0.71** | **28.15** | 0.90 | **17.59** | **0.77** | **17.66** | **0.56** | 20.40 | 0.66 | **21.45** | **0.70** |

how well image editing models handle real-world degradations, with particular emphasis on both degradation removal capability and consistency preservation.

To evaluate both restoration effectiveness and content fidelity, we report two complementary metrics: Restoration Score (RS) and LPIPS-based Perceptual Similarity (LPS) Zhang et al. (2018).

**Restoration Score.** Inspired by non-reference evaluation methods such as VIEScore Ku et al. (2024) and Q-Bench Wu et al. (2024), which demonstrate that native VLMs can serve as image quality assessors without task-specific training, we employ Qwen3-VL-8B-Instruct Team (2025a) as a zero-shot VLM evaluator to assess degradation severity. Specifically, the VLM assigns a score from 0 to 5 to both the degraded input and the clean output, where 0 indicates severe degradation and 5 indicates no visible degradation. The Restoration Score is then defined as the improvement after restoration:

$$RS = S_{\text{clean}} - S_{\text{degraded}}, \tag{4}$$

where $S_{\text{degraded}}$ and $S_{\text{clean}}$ denote the VLM scores of the degraded and clean images, respectively. A higher RS indicates stronger perceived restoration improvement.

**Perceptual Similarity.** To evaluate content fidelity, we use LPIPS Zhang et al. (2018) to measure the perceptual similarity between the degraded input $I_{\text{degraded}}$ and the clean output $I_{\text{clean}}$. Since perceptual similarity metrics are sensitive to structural shifts between aligned image pairs Ghildyal & Liu (2022), LPIPS serves as an indicator of whether the restoration process preserves the original scene structure and semantic content. As lower LPIPS values indicate higher similarity, we convert it into a perceptual similarity score as $LPS = 1 - \text{LPIPS}(I_{\text{degraded}}, I_{\text{clean}})$.

**Final Score.** To jointly assess restoration effectiveness and content fidelity, we combine RS and LPS into a unified Final Score after normalizing them to the same scale, defined as $FS = 0.2 \cdot LPS \cdot RS$. This score penalizes methods that either fail to improve degradation quality or introduce structural and semantic inconsistencies, thus providing a balanced measure of restoration performance. The Final Score jointly reflects restoration improvement and content fidelity, where poor performance in either aspect leads to a lower overall score.

Table 1 and Table 2 demonstrate the strong restoration capability of RealRestorer. It consistently outperforms existing open-source image editing models and achieves performance comparable to leading closed-source systems. Across all nine tasks, RealRestorer achieves the best performance on deblurring and low-light enhancement and ranks second on moiré pattern removal. Figure 5 presents qualitative comparisons, further demonstrating that RealRestorer produces visually cleaner and more consistent restoration results compared to other state-of-the-art image editing methods.

### 4.3 Extra Benchmark Evaluation and Zero-shot Generalization

To further evaluate restoration performance on a traditional all-in-one benchmark, we additionally evaluate the same set of image editing models on the FoundIR test set Li et al. (2025). FoundIR contains 20 real-world degradation settings with paired clean references, including 7 isolated degradations (blur, rain, noise, low-light, raindrops, haze, and compression artifacts) and 13 coupled degradation combinations. We report results on the 7 isolated degradation subsets, which also overlap with RealIR-Bench, resulting in a total of 750 paired image pairs with an average resolution of 2514 × 1516. For the editing prompt, we use the same prompt set as RealIR-Bench.

Table 3 shows that RealRestorer achieves strong restoration performance on these 7 tasks, obtaining the best PSNR and SSIM on 5 out of 7 degradations. More qualitative results, evaluations on additional public benchmarks, and detailed comparisons are provided in Appendix D, together with further visualizations and analysis.

### 4.4 Ablation and User Studies

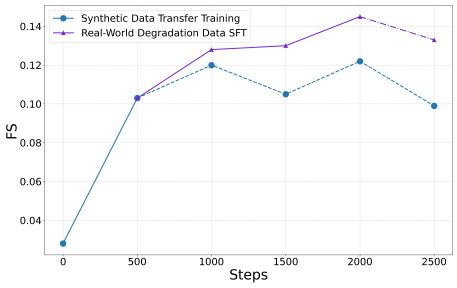


Figure 6: Model performance with varying training steps and training data on RealIR-Bench. The blue line shows transfer training on synthetic degradation data, where the model gradually acquires basic restoration capability. The blue dashed line indicates performance degradation after prolonged training due to the limited diversity of synthetic data. The purple line represents supervised fine-tuning with real-world degradation data, which rapidly improves performance and generalization. The purple dashed segment indicates the onset of overfitting after around 2.5K steps.




Figure 7: Qualitative comparison of different training strategies. Models trained only with synthetic degradation data show limited ability to restore complex real-world degradations. In contrast, models trained solely on real-world degradation data tend to overfit, which may harm structural consistency. which means our two-stage training strategy effectively balances restoration capability and content consistency.

We conduct an ablation study on training data and training stages to validate the proposed two-stage strategy. In the first stage, the model is trained on Synthetic Degradation Data with about 1M samples. As shown in Figure 6, it learns basic restoration ability and reaches a peak FS of 0.122, but still shows limited generalization on rare real-world cases.

We further study the effect of adding Real-World Degradation Data with about 100K samples in the second stage. As shown in Figure 7, models trained only on synthetic data struggle with complex real-world degradations, while models trained solely on real-world data tend to overfit and may reduce structural consistency. In contrast, our two-stage training strategy better balances restoration capability and content fidelity, produce better visual results.

Furthermore, we conduct an ablation study on the Progressively-Mixed training strategy. Without this component, the final FS score decreases by 0.004 points under the same training configuration, confirming its effectiveness also in a qualitative perspective.

We conduct a user study to evaluate both the reliability of the proposed RealIR-Bench metrics and the perceptual performance of our model from a human perspective. Specifically, we recruit 32 participants to rank 3,200 groups of generated images produced by five high-performing models according to two criteria: restoration quality and content consistency. Specifically, Nano Banana Pro achieves the highest first-ranking rate of 32.02%, followed by GPT-Image-1.5 with 23.83%, while our method attains 21.54%. This trend is consistent with the average overall scores reported in Table 2. Moreover, we perform a statistical analysis of the proposed metrics and observe a moderate alignment with human judgments across all evaluation measures ($p < 0.01$). Further details on the study design, ranking protocol, and statistical analysis are provided in the Apeendix F.

## 5 Conclusion

In this paper, we introduce **RealRestorer**, a robust open-source image editing model for complex real-world image restoration. To reduce the synthetic-to-real domain gap, we propose a comprehensive data generation pipeline with high-quality real degraded data to construct **RealIR-1M** and a two-stage progressively mixed training strategy that combines synthetic and real-to-clean pairs. We further present **RealIR-Bench**, a non-reference benchmark with authentic degraded images and a VLM-based evaluation framework for real-world restoration. The extensive experiments on many evaluation sets demonstrate that our **RealRestorer** achieves the open-source state-of-the-art performance.

## AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments and to assist with translation. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or assist in the writing of proofs, and the remaining tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools to create or edit software code and to edit the research paper to improve readability. We have reviewed all AI-assisted work. All AI-assisted content, including experimental suggestions, translations, code, and manuscript edits, was manually reviewed by the authors to ensure correctness and consistency with the intended methodology and technical meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES


SM Al-Salem, G Abraham, OA Al-Qabandi, and AM Dashti. Investigating the effect of accelerated weathering on the mechanical and physical properties of high content plastic solid waste (psw) blends with virgin linear low density polyethylene (lldpe). *Polymer Testing*, 46:116–121, 2015.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Arnold O Benz. Flare observations. *Living reviews in solar physics*, 14(1):2, 2017.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12504–12513, 2023.

Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 183–192, 2020.

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *13th International Conference on Learning Representations, ICLR 2025*, pp. 57335–57356. International Conference on Learning Representations, ICLR, 2025.

Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, and Xiaojuan Qi. Video demoireing with relation-based temporal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Yihang Luo, and Chen Change Loy. Flare7k++: Mixing synthetic and real datasets for nighttime flare removal and beyond. 2023.

Niranjan Damera-Venkata, Thomas D Kite, Wilson S Geisler, Brian L Evans, and Alan C Bovik. Image quality assessment based on a degradation model. *IEEE transactions on image processing*, 9(4):636–650, 2000.

Alain Dufaux, Laurent Besacier, Michael Ansorge, and Fausto Pellandini. Automatic sound detection and recognition for noisy environment. In *2000 10th European Signal Processing Conference*, pp. 1–4. IEEE, 2000.

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.

Jan Flusser, Sajad Farokhi, Cyril Höschl, Tomáš Suk, Barbara Zitova, and Matteo Pedone. Recognition of images degraded by gaussian blur. *IEEE transactions on Image Processing*, 25(2):790–806, 2015.

Abhijay Ghildyal and Feng Liu. Shift-tolerant perceptual similarity metric. In *European Conference on Computer Vision*, pp. 91–107. Springer, 2022.

Qiyuan Guan, Qianfeng Yang, Xiang Chen, Tianyu Song, Guiyue Jin, and Jiyu Jin. Weatherbench: A real-world benchmark dataset for all-in-one adverse weather image restoration. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12607–12613, 2025.

Bahadir Gunturk and Xin Li. *Image restoration*. CRC Press, 2018.

Xiaojie Guo and Qiming Hu. Low-light image enhancement via breaking down the darkness. *International Journal of Computer Vision*, 131(1):48–66, 2023.

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.

Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation*, 90:103712, 2023.

Feng He, Yongjian Zhou, Zefang Ye, Sang-Hyeok Cho, Jihoon Jeong, Xianghai Meng, and Yaguo Wang. Moiré patterns in 2d materials: a review. *ACS nano*, 15(4):5944–5958, 2021.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

Klaus Hermann. Periodic overlayers and moiré patterns: theoretical studies of geometric properties. *Journal of Physics: Condensed Matter*, 24(31):314210, 2012.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3588–3597, 2018.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17853–17862, 2023.

Yi Huang, Jiancheng Huang, Jianzhuang Liu, Mingfu Yan, Yu Dong, Jiaxi Lv, Chaoqi Chen, and Shifeng Chen. Wavedm: Wavelet-based diffusion models for image restoration. *IEEE Transactions on Multimedia*, 26: 7058–7073, 2024.

Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. Real-world person re-identification via degradation invariance learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14084–14094, 2020.

Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Zheng Wang, Xiao Wang, Junjun Jiang, and Chia-Wen Lin. Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining. *IEEE Transactions on Image Processing*, 30:7404–7418, 2021a.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021b.

KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5830–5840, 2021.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

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

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.

Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE transactions on image processing*, 28(1):492–505, 2018.

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.

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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1775–1787, 2023.

Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, pp. 574–591. Springer, 2022.

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

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, pp. 430–448. Springer, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jiaying Liu, Dejia Xu, Wenhan Yang, Minhao Fan, and Haofeng Huang. Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, 129(4):1153–1184, 2021.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018*, 3(8), 2023.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection, 2023.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, July 2017.

OpenAI. Introducing 4o image generation, 2025. URL `https://openai.com/index/introducing-4o-image-generation/`.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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. *arXiv preprint arXiv:2307.01952*, 2023.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sudarshan Rajagopalan, Nithin Gopalakrishnan Nair, Jay N. Paranjape, and Vishal M. Patel. Gendeg: Diffusion-based degradation synthesis for generalizable all-in-one image restoration, 2024. URL `https://arxiv.org/abs/2411.17687`.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Team Seedream, :, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guang Shi, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Wenxu Wu, Yonghui Wu, Xin Xia, Xuefeng Xiao, Shuang Xu, Xin Yan, Ceyuan Yang, Jianchao Yang, Zhonghua Zhai, Chenlin Zhang, Heng Zhang, Qi Zhang, Xinyu Zhang, Yuwei Zhang, Shijia Zhao, Wenliang Zhao, and Wenjia Zhu. Seedream 4.0: Toward next-generation multimodal image generation, 2025a. URL https://arxiv.org/abs/2509.20427.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025b.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Qwen Team. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025b.

Li Tong and JR White. Photo-oxidation of thermoplastics in bending and in uniaxial compression. *Polymer degradation and stability*, 53(3):381–396, 1996.

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

Yan Wang, Ling Yang, Xinzhan Liu, and Pengfei Yan. An improved semantic segmentation algorithm for high-resolution remote sensing images based on deeplabv3+. *Scientific reports*, 14(1):9716, 2024.

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

Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, 2024.

Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30: 2072–2086, 2021.

Zemin Yang, Yujing Sun, Xidong Peng, Siu Ming Yiu, and Yuexin Ma. Unidemoiré: Towards universal image demoiréing with data generation and synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9354–9362, 2025.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12753–12763, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.

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

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

Hao Zhao, Mingjia Li, Qiming Hu, and Xiaojie Guo. Reversible decoupling network for single image reflection removal. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26430–26439, 2025.

Jialong Zuo, Haoyou Deng, Hanyu Zhou, Jiaxin Zhu, Yicheng Zhang, Yiwei Zhang, Yongxin Yan, Kaixing Huang, Weisen Chen, Yongtai Deng, et al. Is nano banana pro a low-level vision all-rounder? a comprehensive evaluation on 14 tasks and 40 datasets. *arXiv preprint arXiv:2512.15110*, 2025.

# Appendix

## A Data Construction Details

### A.1 Synthetic Degradation Data

For the Synthetic Degradation Data, we collect the clean image from the internet and synthesize the nine major degradation patterns: blur, rain, noise, low-light, moiré patterns, haze, compression artifacts, reflection, and flare. We will release the pipeline.

**Blur:** The motion blur dataset is primarily synthesized using temporal averaging over video clips to simulate realistic motion trajectories. Both the target and source images are filtered to ensure consistent blur patterns. In addition, web-style degradation, including common blur operations, such as Gaussian blur and standard motion blur, is incorporated to better approximate real-world motion blur characteristics.

**Compression Artifacts:** We simulate compression artifacts using JPEG compression and image resizing to approximate common web compression effects. In addition to standard JPEG degradation, we also incorporate web-style compression processes to better reflect the wide range of compression artifacts found in online images.

**Moiré Patterns:** Following UniDemoiré Yang et al. (2025), we generate 3,000 moiré patterns at multiple scales and randomly fuse one to three patterns into clean images. This strategy substantially improves the diversity and generalization capability of the model for moiré pattern removal.

**Low-Light:** We simulate low-light conditions by applying brightness attenuation and gamma correction to reduce pixel intensity. Moreover, we train a separate model Cai et al. (2023) using paired datasets such as LOL Yang et al. (2021) and LSRW Hai et al. (2023), reversing the low-exposure and high-exposure image pairs. This trained model is then applied to clean images to better mimic realistic low-light distributions.

**Noise:** We adopt web-style degradation as the primary noise synthesis pipeline. Compared with the degradation strategy used in Real-ESRGAN Wang et al., we further introduce granular noise for web images. Additionally, we incorporate segment-aware noise, which significantly improves performance on real-world denoising tasks.

**Flare:** We collect more than 3,000 glare patterns and adapt them to clean images for realistic blending. In addition, random horizontal and vertical flipping is applied to further enhance the diversity of the generated data pairs.

**Reflection:** For reflection degradation synthesis, we collect two sources of clean images. The first source mainly consists of portrait images, which are treated as transmission layers. The second source contains diverse scenes with human faces, which are used as reflection layers. To increase the diversity of the paired data, we randomly swap a few portions of the image pairs, using human portraits as reflection layers instead of transmission layers. The overall synthesis pipeline follows SynNet Wen et al. (2019).

**Haze:** We synthesize hazy images based on the classic atmospheric scattering model by estimating depth from clean images and generating fog accordingly He et al. (2010). To better simulate real haze, we collect nearly 200 haze patterns and randomly blend them with the synthesized haze, making the results closer to real-world haze distributions.

**Rain:** To synthesize realistic rain degradation, we not only add rain streaks but also incorporate splashes and simulate physical effects such as perspective distortion and droplet sputtering. Furthermore, we collect 200 real rain patterns and randomly blend them into clean images to enhance diversity and realism. Besides, we also adopt the rain category from the FoundIR dataset Li et al. (2025), which contains about 70K paired samples.

### A.2 Real-World Degradation Data

For the real-world degradation data, we collect clean images from high-quality open-source image websites, including Pexels and Pinterest, covering six types of degradation: blur, rain, low light, haze, reflection, and flare. These degradation types often exhibit a substantial gap between real-world degradations and synthesized patterns.

Table 4: Semantic prompts used for CLIP-based degradation filtering.

| Degradation Type | CLIP Text Prompt |
|---|---|
| Flare | a photo with lens flare, bright streaks of light |
| Haze | a hazy photo, foggy atmosphere, low contrast |
| Rain | a rainy photo with rain streaks or raindrops |
| Low-light | a dark photo, underexposed, low illumination |
| Blur | a blurry photo with motion blur or out-of-focus regions |
| Reflection | a photo with glass or mirror-like reflection artifacts |

To construct a high-quality real-world degradation dataset, we first employ the CLIP model Radford et al. (2021) to filter images based on degradation-related semantic cues, as shown in Table 4. Second, we apply a watermark detection filter Minderer et al. (2023) to remove watermarked images, thereby retaining samples suitable for obtaining paired clean data. After selecting appropriate editing models to generate a large amount of raw paired data, additional filtering is required to remove failure cases. Specifically, we use a skeleton-shift-based method Cheng et al. (2020) to remove pixel pairs with alignment errors. Finally, after strictly filtering the raw dataset, we further performed human curation on the remaining subset to construct the final dataset. Three trained human experts participated in the annotation and verification process.

### A.3 Training Dataset statistics

Table 5 summarizes the statistics of the two components of our training data across different degradation types. Visual examples of the data are presented in Figure 4.

Table 5: Statistics of our dataset across different degradation types. The table reports the number of training image pairs from Synthetic Degradation Data and Real-World Degradation Data. The total column indicates the combined number of samples for each degradation category.

| Degradation | Synthetic | Real | Total |
|---|---|---|---|
| Rain | 84,968 | 43,415 | 128,383 |
| Blur | 1,014,229 | 13,458 | 1,027,687 |
| Low-light | 5,000 | 7,005 | 12,005 |
| Hazy | 103,971 | 8,147 | 112,118 |
| Reflection | 68,227 | 7,604 | 75,831 |
| Flare | 59,520 | 7,956 | 67,476 |
| Moire | 99,085 | 0 | 99,085 |
| Noise | 64,492 | 0 | 64,492 |
| Compression | 68,000 | 0 | 68,000 |
| **Total** | **1,567,492** | **87,585** | **1,655,077** |

## B Implementation Details

During training, we treat the DiT blocks as trainable components, while freezing both the VAE and text encoders. In the Transfer Training Stage, we train the model using the Synthetic Degradation

Table 6: Training hyperparameters used in the two training stages.

| **Hyperparameters** | **Transfer Training Stage** | **Supervised Fine-Tuning Stage** |
|---|---|---|
| Learning rate | $1 \times 10^{-5}$ | $1 \times 10^{-5} \rightarrow 0$ |
| LR scheduler | Constant | Cosine |
| Weight decay | 0.0 | 0.01 |
| Gradient norm clip | 1.0 | 1.0 |
| Optimizer | AdamW ($\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1e^{-8}$) | |
| Warm-up steps | 100 | 100 |
| Frozen layers | None | First $1/4$ SingleStreamBlocks |
| Training steps | 500 | 1500 |
| Training samples | 1.5M | 80K |
| Resolution | $1024 \times 1024$ | $1024 \times 1024$ |
| Synthetic : Real data ratio | 1 : 0 | 2 : 8 |

Table 7: Category statistics of RealIR-Bench. The benchmark contains 456 real-world non-reference degraded images across nine degradation categories.

| Category | Blur | Compression | Flare | Haze | Low-light | Moiré | Noise | Rain | Reflection |
|---|---|---|---|---|---|---|---|---|---|
| # Images | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 56 |

Dataset, which covers nine degradation types. To balance the learning across tasks, we adopt an average sampling strategy over all nine degradation categories. In this stage, the bucket resolution is fixed at 1024 × 1024, and the global batch size is set to 16.

After approximately 500 training steps, the model begins to transfer knowledge from high-level editing capabilities to low-level restoration tasks. However, it still struggles to handle more complex degradations, often producing artifacts in the restored results. To address this limitation, we introduce a Supervised Fine-Tuning Stage. In this stage, we adopt a Progressively-Mixed training strategy, combining Real-World Degradation Data with a small portion of Synthetic Degradation Data. This strategy helps constrain the model toward the data manifold of real-world restoration tasks while retaining the robustness learned from synthetic degradations.

Additionally, we freeze the first one-fourth of the SingleStreamBlocks in the DiT architecture to stabilize training. The global batch size is increased to 32, and a cosine annealing learning rate schedule is applied, where the learning rate gradually decays to zero while maintaining the same initial learning rate as in the first stage. This stage lasts for 1.5K training steps, allowing the model to converge to a balanced and generalizable checkpoint. All experiments are conducted on NVIDIA H800 GPUs, and the entire training process takes approximately one day on 8 H800 GPUs. More detailed training hyperparameters are provided in Table 6.

## C REALIR-BENCH AND METRICS DETAILS

The distribution of **RealIR-Bench** is summarized in Table 7, and representative examples from the benchmark are shown in Figure 10.

Specifically, we evaluate models using two complementary metrics: Restoration Score (RS), which reflects the perceptual restoration quality, and LPIPS Zhang et al. (2018), which measures perceptual similarity to assess consistency after restoration.

### C.1 RESTORATION SCORE

The Restoration Score (RS) is designed to evaluate the ability of a model to remove degradations without explicitly considering content consistency. Inspired by VIEScore, we employ Qwen3-VL-8B-Instruct as a vision-language evaluator to assess the degradation severity of both degraded images and restored images. The Restoration Score (RS) is then defined as the improvement in the degradation level after restoration. The detailed system instruction for Qwen3-VL-8B-Instruct is shown in the Figure 11.

Table 8: Quantitative comparison on extra public benchmarks for real-world image restoration tasks. We report PSNR (↑) and SSIM (↑). The evaluation is conducted on multiple datasets including **Flare-R** from Flare7K++ for deflare, **UHDM** for moiré pattern removalremoval, and the **SIR$^2$+** reflection removal benchmark with three subsets: PostcardDataset, SolidObjectDataset, and WildScene. Flare-R contains real captured flare images, while UHDM provides ultra-high-definition paired moiré images. The SIR$^2$+ subsets represent different reflection scenarios with diverse scene contents. The best results are highlighted in **bold**, and the second-best results are underlined.

| Method | Flare | | Moiré | | Refl (Postcard) | | Refl (Solid) | | Refl (Wildscene) | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| Nano Banana Pro | **25.28** | **0.889** | **18.86** | 0.726 | 20.11 | 0.821 | **24.53** | **0.880** | 22.27 | 0.835 | **22.21** | **0.830** |
| GPT–Image–1.5 | 16.86 | 0.526 | 11.84 | 0.549 | 13.51 | 0.465 | 13.97 | 0.451 | 14.59 | 0.527 | 14.15 | 0.504 |
| Seedream 4.5 | 23.88 | 0.831 | 15.30 | 0.643 | 21.38 | 0.807 | 22.16 | 0.809 | 21.45 | 0.783 | 20.83 | 0.775 |
| LongCat–Image–Edit | 22.57 | 0.809 | 15.79 | 0.670 | 20.28 | 0.706 | 17.12 | 0.632 | 21.28 | 0.805 | 19.41 | 0.724 |
| Qwen–Image–Edit–2511 | 22.45 | 0.869 | 16.07 | 0.673 | 21.07 | 0.839 | 20.17 | 0.864 | **23.77** | **0.896** | 20.71 | 0.828 |
| FLUX.1-Kontext-dev | 23.27 | 0.859 | 10.36 | 0.516 | 13.43 | 0.428 | 11.60 | 0.374 | 14.62 | 0.520 | 14.66 | 0.539 |
| Step1X–Edit | 18.74 | 0.772 | 14.31 | 0.583 | 19.29 | 0.841 | 18.56 | 0.783 | 21.12 | 0.864 | 18.40 | 0.769 |
| **Ours** | 22.05 | 0.745 | 17.62 | **0.743** | **22.67** | **0.904** | 20.35 | 0.797 | 21.72 | 0.826 | 20.88 | 0.803 |

## D More Qualitative Results and Benchmark Evaluation

We present additional visualization results in the following pages to further demonstrate the strong restoration capability of our model compared with other image editing models on **RealIR-Bench** across nine degradation types.

Furthermore, to evaluate **RealRestorer** on public benchmarks for deflare, reflection removal, and demoiré, we conduct additional experiments on several widely used datasets. For deflare evaluation, we use the Flare-R subset from the Flare7K++ dataset Dai et al. (2023), which contains 100 paired real-world flare images. The Flare7K++ dataset combines synthetic and real flare data and provides a comprehensive benchmark for nighttime flare removal tasks.

For moiré pattern removalevaluation, we adopt the UHDM test set Dai et al. (2022), which contains 500 paired real moiré images captured at ultra-high resolution. For reflection removal, we evaluate on the SIR²+ benchmark, which includes three subsets: SolidObjectDataset, PostcardDataset, and WildScene, containing 50, 50, and 101 paired images respectively. These datasets include real-world scenes with complex reflective patterns and are widely used for evaluating single-image reflection removal methods. The quantitative comparison results are presented in Table 8. The results show that **RealRestorer** achieves the second-best performance in PSNR and the third-best performance in SSIM on average across the five evaluation datasets.

## E Ablation Study Details

Besides comparing the proposed two-stage training strategy composed of the synthetic degradation transfer training stage and the real-world degradation SFT stage, we further analyze an alternative setting where the model is trained using only real-world degradation data. For a fair comparison, we train the model for the same number of iterations as in the synthetic transfer training stage. At the peak point, the model trained only on real-world data tends to overfit the degradation patterns, which harms the structural consistency of the restored images. This often results in artifacts such as object deformation, body shifting, and unrealistic enhancement. These observations further confirm the importance of the proposed two-stage training strategy.

Specifically, at 2.5K training steps, the model trained with only synthetic degradation data still shows limited ability to handle complex real-world degradations, while the model trained solely on real-world degradation data can partially restore degradations but often fails to preserve content consistency. The model produces overly enhanced results, such as removing natural light sources, as shown in Figure 6. In contrast, our two-stage training strategy effectively balances restoration capability and structural consistency, leading to more stable and generalizable performance.

## F User Study Details

Our user study evaluates the results of the five best-performing models on **RealIR-Bench**. All the 32 participants receive a brief tutorial beforehand to ensure they understand the task and the evaluation criteria. The interface used in the user study is illustrated in Figure 16. To further analyze the reliability of the proposed metric, we compute the Kendall's $\tau_b$, Spearman Rank Correlation Coefficient (SRCC), and Pearson Linear Correlation Coefficient (PLCC) between the metric score (FS) and human judgments. These correlation measures are widely used to evaluate the consistency between automatic metrics and human evaluation. The results demonstrate that the proposed metric achieves moderate statistical alignment with human judgments ($p < 0.01$) across all evaluation settings, as shown in Table 9.

Table 9: Consistency evaluation between RS metric and subjective human perception.

| **Metric** | **Correlation Coefficient** | $p$**-value** |
|---|---|---|
| Kendall's $\tau_b$ | 0.2493 | $2.96 \times 10^{-124}$ |
| SRCC | 0.3010 | $4.62 \times 10^{-128}$ |
| PLCC | 0.2919 | $3.21 \times 10^{-120}$ |

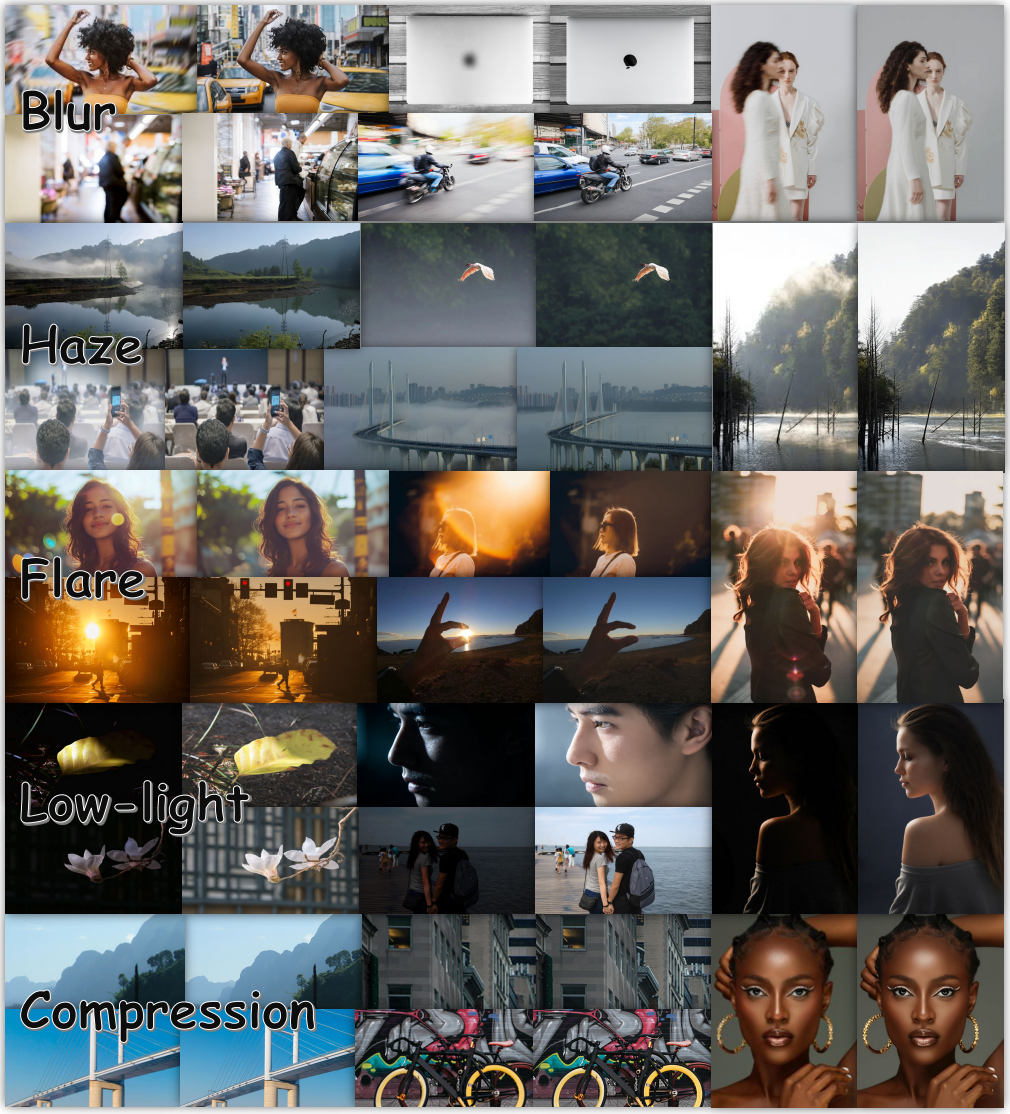


Figure 8: Additional qualitative results of RealRestorer under real-world degradations. Please zoom in for better visualization of details.




Figure 9: Additional qualitative results of RealRestorer under real-world degradations. Please zoom in for better visualization of details.

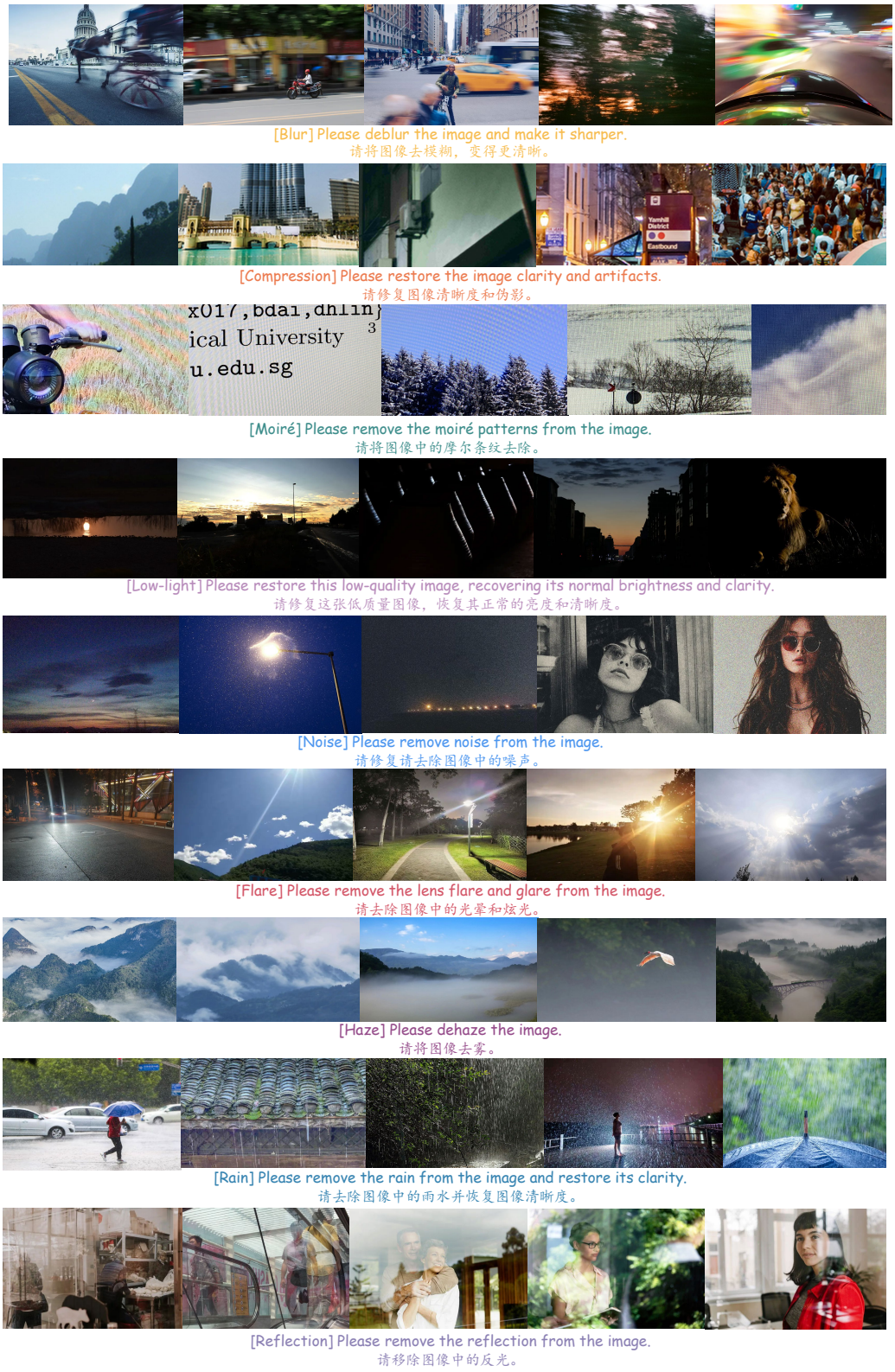


Figure 10: Examples from our RealIR-Bench. Each degradation category is evaluated using a fixed bilingual prompt.

## Instruction for Degradation Evaluation

- **Degradation Evaluation:** You will evaluate whether an image exhibits the degradation type {task} and determine its severity level. The assessment should focus exclusively on the specified degradation type, while ignoring unrelated image-quality issues, semantic content, and aesthetic preference. Your goal is to provide a consistent and objective judgment of how strongly the target degradation affects the image.

- **Evaluation Procedure:**

  1. Divide the image into several local regions (e.g., a $3 \times 3$ grid) to ensure a systematic inspection.
  2. Examine each region carefully and identify whether the degradation type {task} is present.
  3. Estimate both the severity of the degradation and the proportion of the image area that is affected.
  4. Summarize the regional observations into a single overall score that best reflects the image-level degradation.
  5. If the degradation is spatially uneven, place greater emphasis on the regions that are more strongly affected.
  6. For borderline cases, choose the nearest score level based on the dominant visual impression.
  7. Do not output any intermediate reasoning, explanation, or analysis.

- **Scoring Scale (1–5):** Use the following scale to measure the severity of the target degradation. Higher scores indicate cleaner images with less visible degradation, while lower scores indicate stronger and more widespread corruption.

  - 5 = No {task}; the image is essentially clean with no noticeable degradation.
  - 4 = Mild degradation; only a small portion of the image is affected ($\leq 20\%$ area), and the overall visual quality remains largely intact.
  - 3 = Moderate degradation; the degradation is clearly visible and affects a noticeable part of the image (20–50% area).
  - 2 = Severe degradation; the corruption is strong and influences a large portion of the image (50–80% area).
  - 1 = Extreme degradation; the degradation dominates most of the image ($> 80\%$ area) and seriously harms visual quality.

- **Output Format:** Return only in the format "Degradation Score: <1–5>".

Figure 11: System instruction used for degradation evaluation.

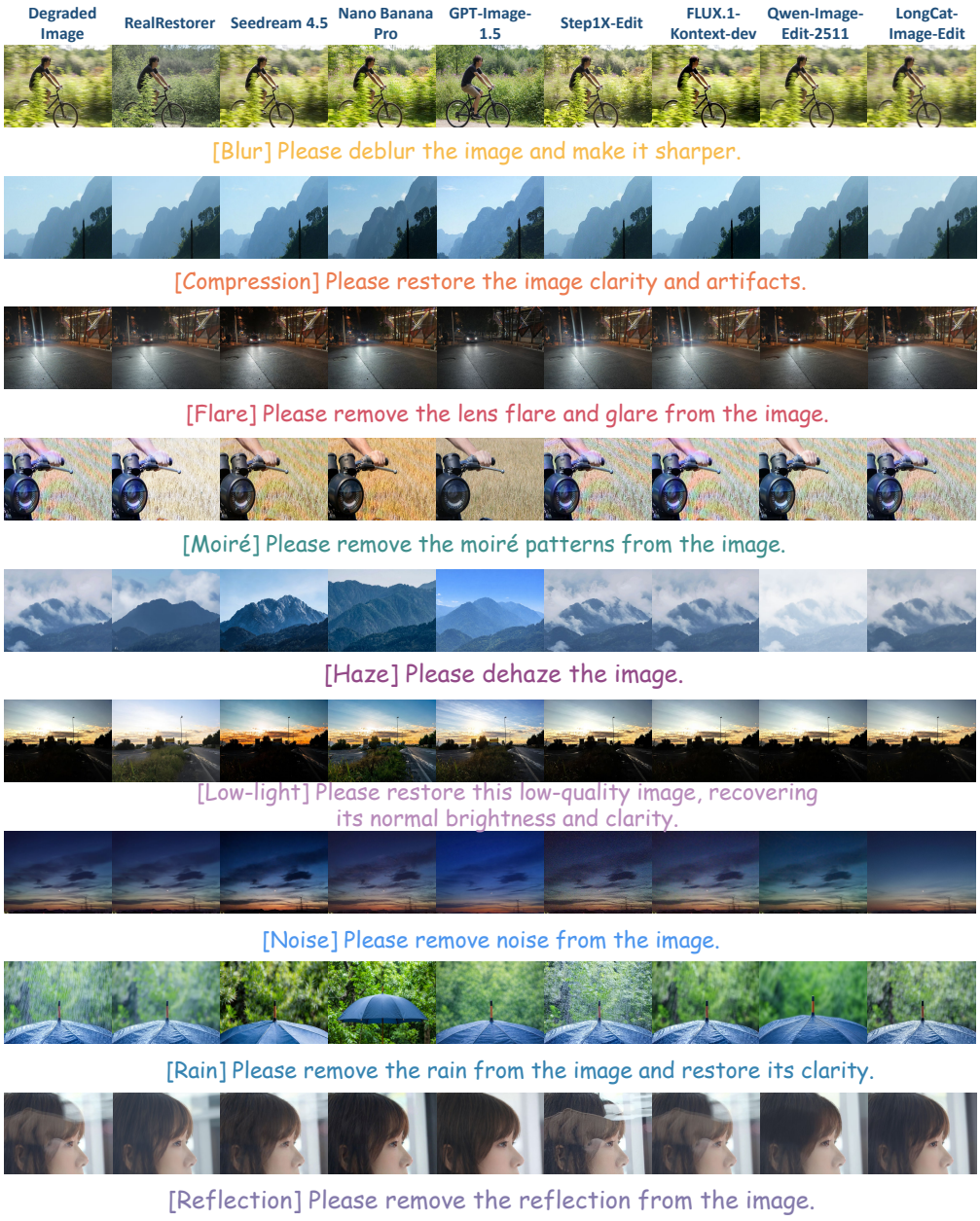


Figure 12: More qualitative comparison results on RealIR-Bench. Zoom in to see more details.

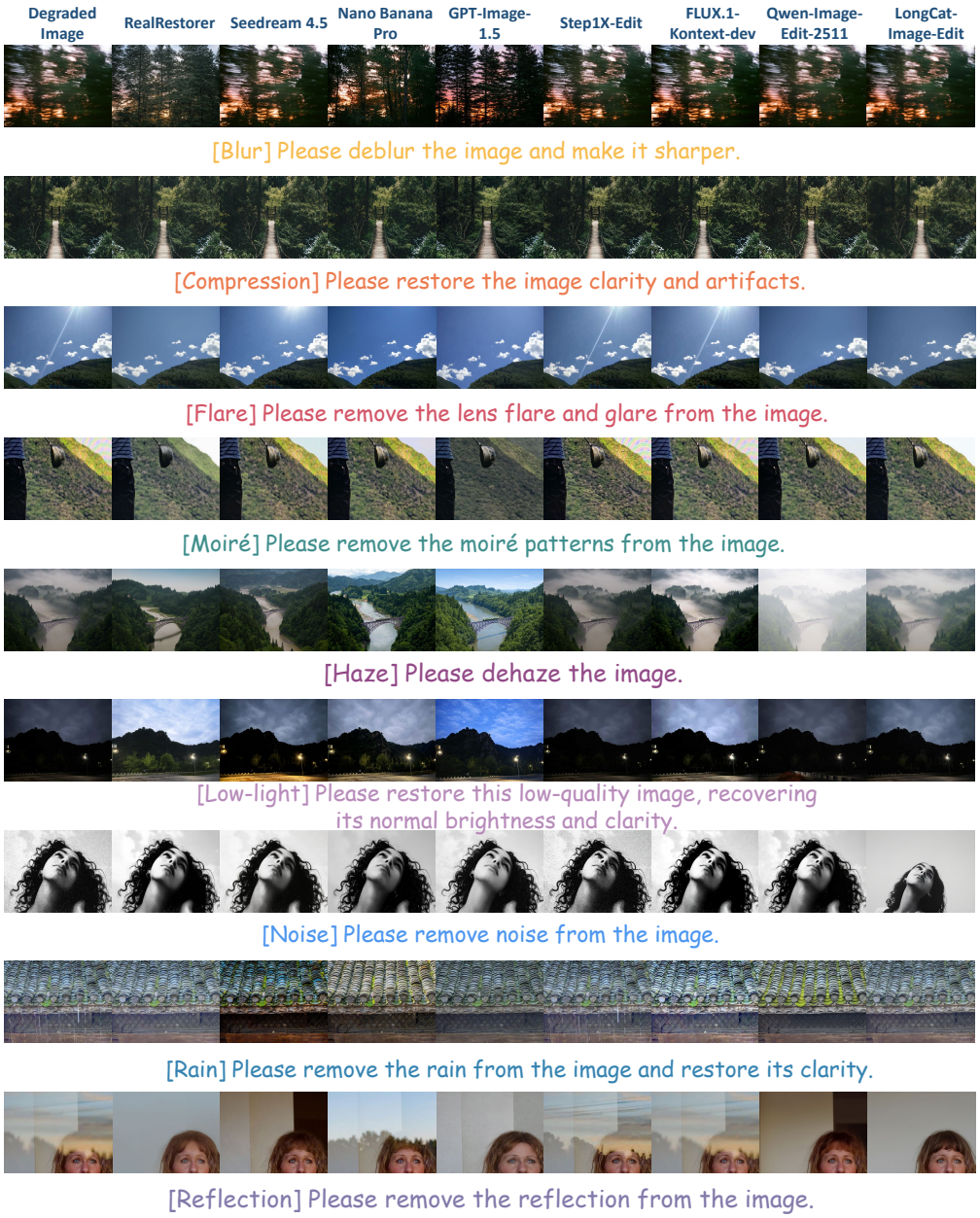


Figure 13: More qualitative comparison results on RealIR-Bench. Zoom in to see more details.

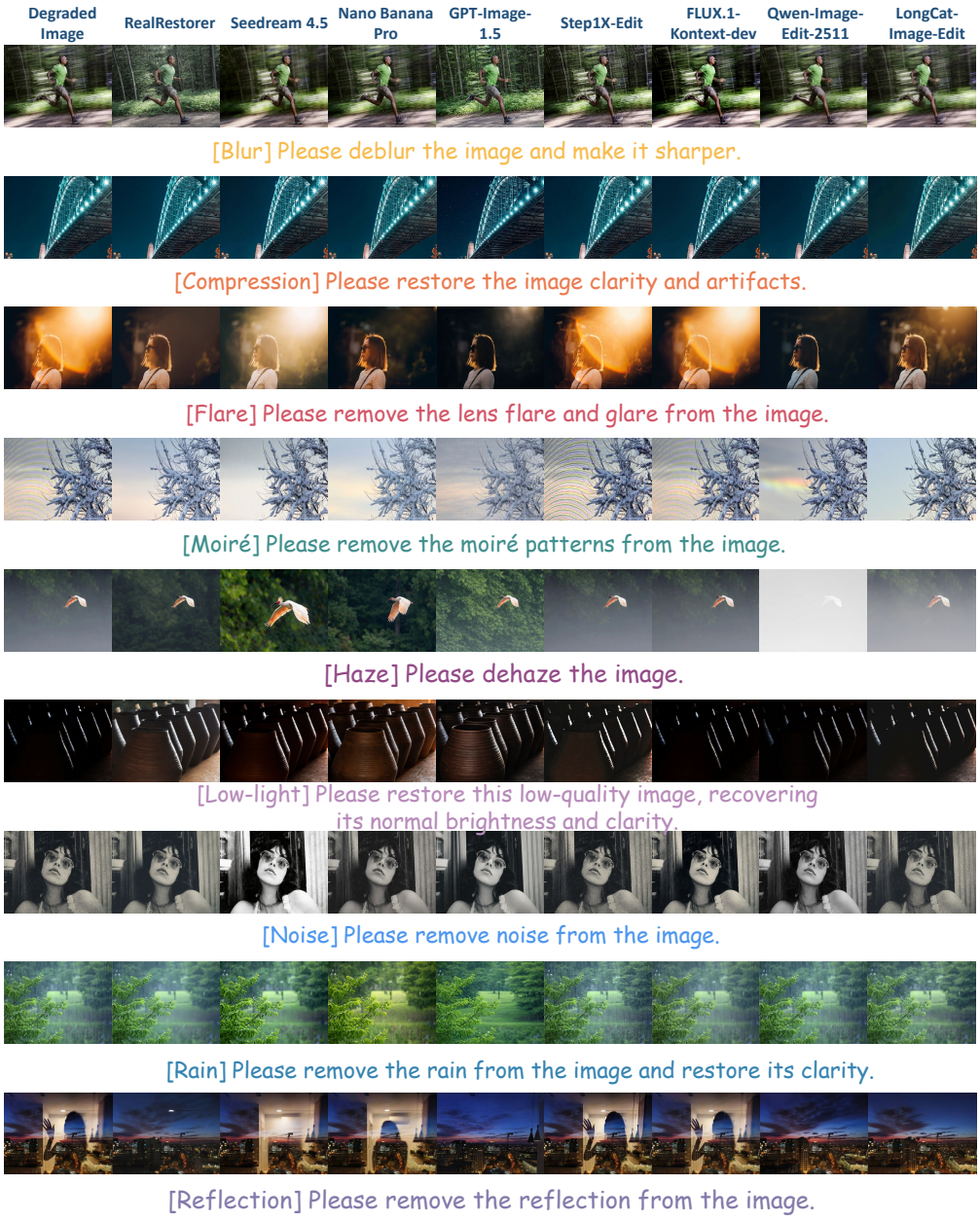


Figure 14: More qualitative comparison results on RealIR-Bench. Zoom in to see more details.

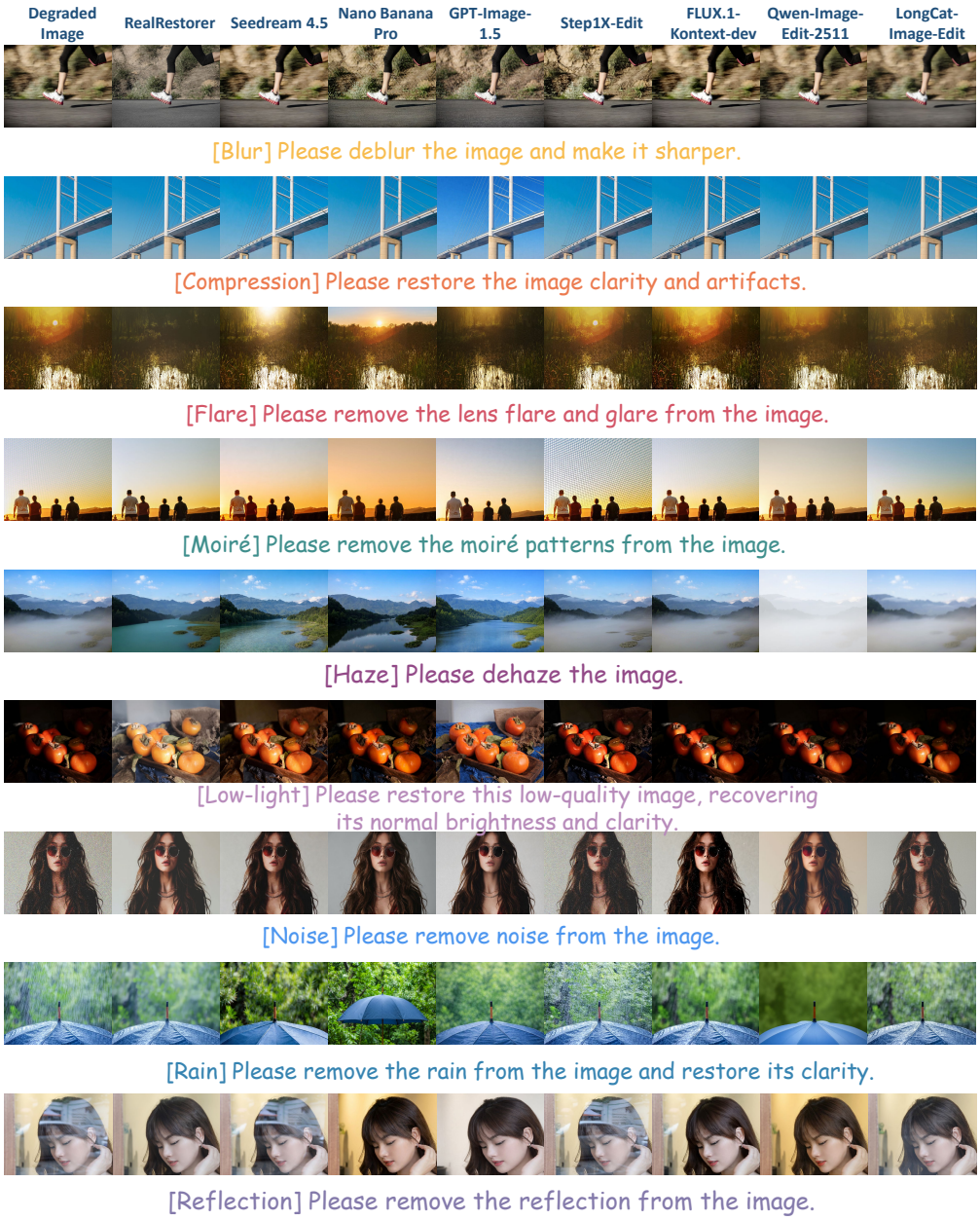


Figure 15: More qualitative comparison results on RealIR-Bench. Zoom in to see more details.

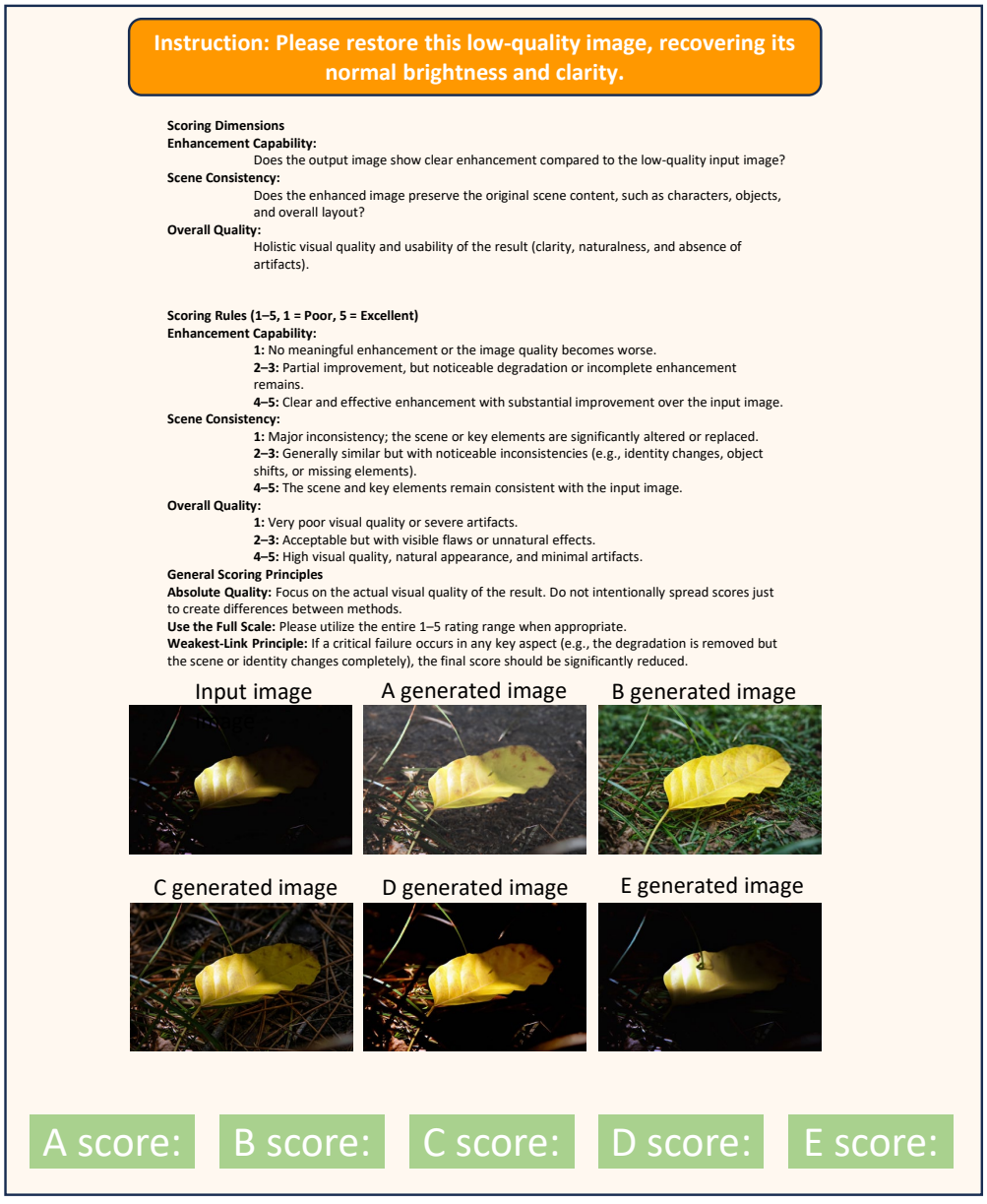


Figure 16: User study interface used to evaluate the restoration results. Participants are presented with one degraded input image and five restored results generated by different models and are asked to rate them based on restoration quality and consistency.