UNILENS: DEGRADATION-CENTRIC LEARNING AND EVALUATION FOR REAL-WORLD ALL-IN-ONE IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

All-in-one image restoration aims to handle multiple restoration tasks with a single model, but its generalization to real images is still difficult. A major obstacle is deciding which training examples to add, since simply scaling up dataset size does not guarantee coverage of real-world degradation patterns. In this paper, we propose **UniLens**, a unified pipeline that connects degradation-aware data construction, model adaptation, and task-conditioned evaluation for real-world all-in-one restoration. **First**, we formulate data construction as a degradation-coverage problem: how can we represent degradations in real-world images and use this representation to build training pairs? The key is to learn a degradation space that is less sensitive to scene content and reveals which real-world degradation patterns the training data miss. We then use this space to guide data synthesis that adds training pairs to cover these patterns. **Second**, building on this broader coverage, we augment paired supervision with instruction-dependent targets for single, joint, and selective removal, providing richer compositional supervision for all-in-one restoration under real-world composite degradations. An instruction-conditioned mixture of LoRA experts further supports these transformations, with a shared low-rank branch capturing common restoration knowledge and sparse experts providing request-dependent adaptation. **Finally**, we introduce a task-conditioned vision-language evaluator that assesses what degradation remains and how much the image has improved. Our method achieves state-of-the-art performance on real-world image restoration. Compared with FoundIR-v2, it improves task-conditioned restoration scores by 24.8% on public benchmarks and by 45.3% on real images with composite degradations.

1 INTRODUCTION

All-in-one image restoration aims to handle multiple restoration tasks with a single model (Li et al., 2022; Potlapalli et al., 2023; Conde et al., 2024). This is attractive in practice, where images can be degraded in many different ways. However, these models often struggle with real photographs. Even within one task, degradation patterns can vary widely. For example, noise varies in strength, motion blur in direction and shape, and haze in spatial distribution. Training data may cover only a small fraction of these variations, even when the dataset is large. This leaves a practical question: which real-world degradation patterns are missing, and what training examples would cover them?

Recent work has explored different sources of supervision for real-world image restoration (Li et al., 2025; Rajagopalan et al., 2025; Kong et al., 2026; Chen et al., 2026). Some methods collect real degraded–clean pairs, some synthesize degraded inputs from clean images, and others learn from unpaired clean and degraded images. Among these approaches, synthesis provides a practical way to create many training pairs with known clean targets. However, synthesized degradations may still miss patterns found in real images. To identify these missing patterns, we therefore need to compare real and synthesized degradations.

This comparison is difficult because image appearance reflects both degradation and scene content. Real degraded images also rarely have clean counterparts, making it difficult to isolate the degra-

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.

dation itself. We need a representation that captures degradation patterns directly from low-quality images while reducing the influence of scene content.

In this paper, we propose **UniLens**, a unified pipeline integrating data construction, model adaptation, and evaluation for real-world all-in-one restoration. We introduce **WildRestore-Bench**, a real-world benchmark built from application photographs and online images, to evaluate real-world restoration and validate no-reference assessment against human judgments.

UniLens first learns to represent real and synthesized images by their degradations rather than their scene content. This shared degradation space reveals which real degradation patterns the training data miss and guides the construction of additional pairs to cover them. To learn this representation, we train the **Cross-content Ordinal Degradation Encoder (CODE)**, which compares low-quality images by their degradations while reducing the influence of scene content. We use two complementary forms of supervision. First, we align images from the same degradation family across different scenes, such as a hazy street and a hazy forest, so that the encoder learns degradation characteristics shared across content. Second, we compare different degradation variants of the same scene, allowing the encoder to preserve finer differences within each family. Together, these two forms of supervision encourage CODE to capture degradation characteristics shared across scenes while preserving finer differences within each degradation family.

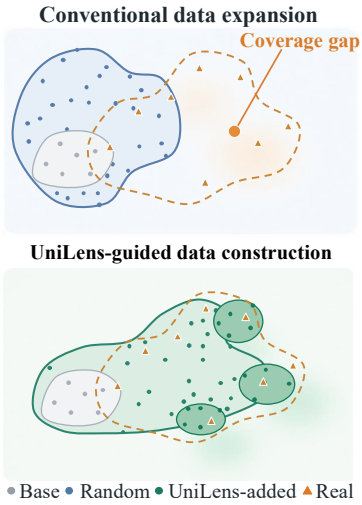

Figure 1: Data construction in UniLens. UniLens uses CODE to identify poorly covered degradation patterns and expands the training data to improve both broad and local coverage.

UniLens uses the coverage gaps identified by CODE to guide data construction, as illustrated in Figure 1. For under-covered degradation types, we construct offline training pairs using suitable degradation-specific methods and use online synthesis to diversify degradation strengths and types during training. For finer coverage gaps, UniLens uses CODE to retrieve related source pairs, generates degraded candidates from them, and selects those with similar degradation patterns. Together, these additions broaden the training distribution while improving support for degradation patterns that remain poorly covered.

Building on this broader coverage, we introduce instruction-dependent compositional supervision for single, joint, and selective removal. A real image may contain multiple degradations, while a restoration request may target only a subset of them. For example, a noisy and blurred image may require noise removal while retaining blur. By applying degradations in sequence and retaining intermediate images, we construct targets for single, joint, and selective removal from the same degraded input. To learn these behaviors, we adapt a pretrained image-editing model with shared low-rank updates for common restoration knowledge and instruction-conditioned sparse experts for request-specific adaptation. This enables a single model to perform request-dependent restoration under composite degradations.



	Output A	Output B
MUSIQ ↑	**55.70**	46.70
NIQE ↓	**3.74**	4.32
Human ↑	1.00	**4.71**
DRA (ours) ↑	2.50	**5.00**

Figure 2: A real-image flare-removal example where generic quality metrics disagree with human judgments. Generic metrics rank A higher, while DRA and human judgments rank B higher.

Finally, evaluating real-world restoration is challenging: images captured in the wild typically lack clean references, while generic metrics such as NIQE (Mittal et al., 2013) and MUSIQ (Ke et al., 2021) assess overall image quality rather than whether the requested degradation has been removed, and may therefore disagree with human judgments of restoration success. We introduce **Degradation-aware Restoration Assessment (DRA)**, a task-conditioned vision-language evaluator of residual target degradation and input-relative improvement. We compare metric preferences

in Figure 2. DRA agrees with human judgments in this example and shows higher development-set correlations than the compared quality metrics.

Experiments demonstrate cross-scene degradation matching and improved coverage of real degradations. Our model, trained with guided data expansion, improves task-conditioned restoration scores over FoundIR-v2 by 24.8% on public benchmarks and 45.3% on real images with composite degradations. With equal task weighting across 36 benchmarks and eight tasks, PSNR improves by 3.53 dB and DISTS decreases by 31.3%. Our contributions are summarized as follows:

- We propose **UniLens**, a unified pipeline for real-world all-in-one image restoration. We introduce **CODE** to compare low-quality images by their degradations rather than their scene content, revealing poorly covered real degradation patterns and guiding the construction of additional training pairs.
- We construct instruction-dependent targets for single, joint, and selective degradation removal, and adapt a pretrained image-editing model with shared low-rank updates and instruction-conditioned sparse experts for request-dependent restoration.
- We introduce **DRA**, a task-conditioned vision-language evaluator for residual target degradation and input-relative improvement without clean references, and **WildRestore-Bench**, a real-world benchmark with a human-rated development set for validating agreement with human judgments.

2 Related work

All-in-one real-world restoration. AirNet, PromptIR, and InstructIR condition a shared restoration model on degradation representations, learned prompts, and language, respectively (Li et al., 2022; Potlapalli et al., 2023; Conde et al., 2024). DA-CLIP separates content and degradation representations (Luo et al., 2024), while MoCE-IR uses conditional experts (Zamfir et al., 2025). FoundIR and FoundIR-v2 scale restoration through generalist–specialist modeling and optimized data mixtures (Li et al., 2025; Chen et al., 2026). GGT-100K adapts image-editing models with restoration pairs (Kong et al., 2026). UniLens pairs composite inputs with instruction-dependent targets and combines shared restoration knowledge with sparse request-dependent adaptation (Hu et al., 2022; Wu et al., 2024b).

Training data construction. BSRGAN and Real-ESRGAN compose synthetic operators, while GenDeg learns degradation synthesis (Zhang et al., 2021; Wang et al., 2021; Rajagopalan et al., 2025). FoundIR expands paired data, GGT-100K constructs generated targets for real inputs, and FoundIR-v2 optimizes mixture proportions (Li et al., 2025; Kong et al., 2026; Chen et al., 2026). These approaches expand supervision but do not by themselves establish support for individual real degradation patterns. Building on contrastive distortion learning (Madhusudana et al., 2022; Agnolucci et al., 2024), CODE compares training inputs with real observations. Its coverage diagnostics inform degradation-specific offline and online expansion, and its distances guide source retrieval and candidate selection for finer gaps.

Restoration assessment. Generic no-reference metrics assess naturalness or learned perceptual quality (Mittal et al., 2013; 2012; Ke et al., 2021; Yang et al., 2022; Wang et al., 2023). Q-Align, DepictQA, and DeQA-Score extend quality assessment with vision-language models (Wu et al., 2024a; You et al., 2024; 2025). DRA conditions its judgments on the requested task, assessing residual degradation before input-relative improvement. Additional background appears in Appendix G.

3 Method

UniLens addresses noise, blur, haze, rain streaks, raindrops, flare, low light, and low resolution. Table 14 defines their restoration goals and DRA criteria. The UniLens pipeline comprises three stages: degradation-aware data construction, instruction-dependent model adaptation, and task-conditioned assessment. First, CODE identifies degradation coverage gaps and guides pair construction in Figure 3. Second, instruction-dependent compositional targets and an instruction-conditioned mixture

of LoRA experts enable single, joint, and selective restoration. Finally, DRA assesses residual target degradation and input-relative improvement without clean references.

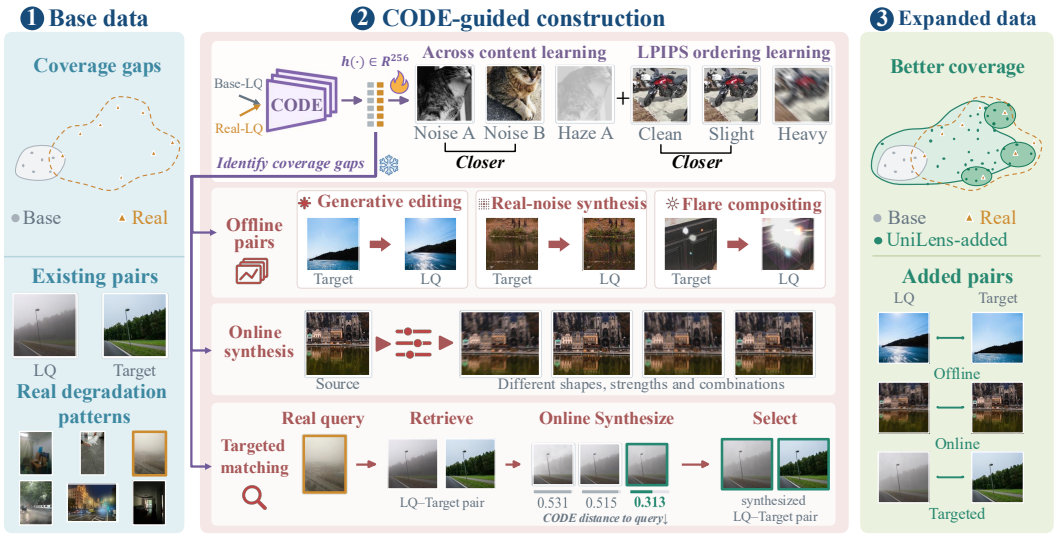

Figure 3: Data construction in UniLens. CODE identifies degradation coverage gaps between base data and real degradation exemplars, guiding offline pair construction, online synthesis, and targeted supplementation to expand training coverage. Coverage outlines are schematic.

3.1 Learning a degradation comparison space

CODE learns to compare degradation patterns across scenes using low-quality (LQ) images. Its unit embeddings define the distance:

$$d_E(\mathbf{x}, \mathbf{x}') = \|\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{x}')\|_2, \tag{1}$$

where $\mathbf{x}, \mathbf{x}'$ are low-quality RGB images and $\mathbf{h}(\cdot) \in \mathbb{R}^{256}$ denotes the CODE embedding to be learned. Smaller distances indicate more similar degradations.

The CODE block in Figure 3 uses two learning signals. *Across content* aligns images sharing a degradation family across scenes, with other families as negatives (Madhusudana et al., 2022; Agnolucci et al., 2024). *LPIPS ordering* holds content and family fixed: frozen LPIPS (Zhang et al., 2018) ranks variants as near or far relative to an anchor, and CODE learns their relative distances. The combined objective is:

$$\mathcal{L}_{\text{CODE}} = \mathcal{L}_{\text{type}} + \lambda_o \mathcal{L}_{\text{ord}}, \tag{2}$$

where $\mathcal{L}_{\text{type}}$ is the cross-content contrastive loss, $\mathcal{L}_{\text{ord}}$ is the within-family ranking loss, and $\lambda_o \geq 0$ balances them. The ordering encodes within-family perceptual differences. Trained independently of the restoration backbone, CODE is then frozen for comparisons without clean references. Full objectives appear in Appendix A.

We test content–degradation conflicts and real-image transfer in Table 1. CODE selects the degradation match in 98.45% of conflicts and achieves 80.44% retrieval mAP on five FoundIR-Test subsets, supporting cross-scene comparison.

3.2 Coverage-guided data construction

UniLens uses CODE to guide three construction routes for broad expansion and local matching, as outlined in Figure 3. We compare LQ inputs from the *Base mixture* with real references through distribution discrepancy and local-neighbor support. The task-wise diagnostics show sparse local support for noise and comparatively small distribution discrepancy for low light (Appendix B). These comparisons guide broad expansion of noise, blur, haze, rain, and flare; low light retains existing pairs during offline and online construction.

Table 1: Degradation similarity (%; higher is better). Conflict and real matching use two choices; FoundIR (Li et al., 2025) mAP measures retrieval.

Representation	Conflict	Real matching	FoundIR-test mAP
ResNet-50 (He et al., 2016)	32.09	**72.68**	**39.44**
Qwen VAE (Wu et al., 2025)	**44.38**	69.52	38.28
DINOv2 (Oquab et al., 2023)	0.03	60.25	16.27
CLIP (Radford et al., 2021)	0.79	62.73	21.36
CODE	**98.45**	**83.85**	**80.44**

To supply paired examples for these gaps, *offline construction* introduces blur, rain, and haze into clean targets through Seedream editing, transfers real-noise characteristics with PNG (Ko et al., 2026), and composites flare assets. PNG interpolates two conditions learned from real noisy–clean pairs, varying corruption while retaining the clean source as target. Real observations identify the degradation patterns to address, while source images provide the corresponding restoration targets.

Offline pairs capture fixed instances of these degradation patterns. *Online synthesis* samples operator parameters and combinations on clean images or paired inputs, retaining intermediate images for instruction-dependent targets (Section 3.3); super-resolution is also synthesized online. CODE comparisons in Figure 6 show the largest reduction in haze discrepancy after offline expansion and in blur discrepancy after online synthesis.

Broad expansion can still leave individual query neighborhoods poorly supported. *Targeted matching* therefore uses CODE within construction: for each real query in an augmented task, we retrieve paired sources with nearby LQ features, synthesize candidates from their degraded inputs or restoration targets, and select by candidate–query distance. Pair-consistency review and source-reuse limits constrain selection. These locally matched pairs complement broad offline and online additions; mixing the resulting UniLens-added pairs with the Base mixture forms the *UniLens data*. Appendix B.2 specifies the procedure.

3.3 Instruction-dependent supervision and model adaptation

Broader coverage brings diverse restoration demands: the same composite image may require joint or selective removal. We therefore pair each input with the output requested by its instruction, using a clean image for joint removal and an intermediate synthesis result for selective removal. With single-removal pairs, supervision is formulated as:

$$(\mathbf{x}, \mathbf{y}^*) = \begin{cases} (\mathrm{D}_a(\mathbf{y}), \mathbf{y}), & c: \text{ remove } a, \\ (\mathrm{D}_b(\mathrm{D}_a(\mathbf{y})), \mathbf{y}), & c: \text{ remove } a, b, \\ (\mathrm{D}_b(\mathrm{D}_a(\mathbf{y})), \mathrm{D}_a(\mathbf{y})), & c: \text{ remove } b, \text{ retain } a \end{cases}, \tag{3}$$

where $\mathbf{y} \in \mathbb{R}^{H \times W \times 3}$ is a clean RGB image of size $H \times W$, $\mathrm{D}_a, \mathrm{D}_b$ are distinct degradation operators, and $\mathbf{x}, \mathbf{y}^*$ are input and target. Here c is a structured text prompt (Appendix C.1).

Learning these targets requires shared restoration knowledge and request-specific updates. We therefore adapt Qwen-Image-Edit (Wu et al., 2025) with an instruction-conditioned mixture of LoRA experts (MoLE) (Hu et al., 2022; Wu et al., 2024b). A shared branch remains active for every request, complemented by sparse request-dependent experts. Expert selection depends on pooled image–instruction states, block, and denoising time. An adapted projection computes:

$$\mathbf{v}_{\ell,t} = \left[\mathbf{W}_{0,\ell} + \Delta \mathbf{W}_\ell^{\mathrm{sh}} + \sum_{e \in \mathcal{E}_{\ell,t}} \alpha_{\ell,t,e} \Delta \mathbf{W}_\ell^e \right] \mathbf{u}, \tag{4}$$

where $\mathbf{u} \in \mathbb{R}^{d_{\mathrm{in}}}$ and $\mathbf{v}_{\ell,t} \in \mathbb{R}^{d_{\mathrm{out}}}$ are input and output activations; the frozen $\mathbf{W}_{0,\ell}$ and its updates have shape $d_{\mathrm{out}} \times d_{\mathrm{in}}$. Here ℓ, t index block and denoising time, sh denotes the shared branch, $\mathcal{E}_{\ell,t}$ contains selected experts, and $\alpha_{\ell,t,e}$ gives their weights. Adapters and router are trained with flow matching and auxiliary intent/balancing losses, detailed in Appendix C.

3.4 TASK-CONDITIONED RESTORATION ASSESSMENT

Real images usually lack clean references, and overall quality does not establish whether a requested degradation was removed. Assessment must consider both the input and the task.

WildRestore-Bench. WildRestore-Bench combines application photographs with web-collected images visually screened for plausible degradation patterns. It covers seven single-task categories and their combinations, detailed in Appendix D.1. Each image is paired with a restoration request rather than a clean reference. Single requests one task; Composite requests removal of all listed degradations. A human-rated development set of 147 outputs from 49 inputs provides judgments of degradation removal, naturalness, and content fidelity for validating no-reference assessment. Collection and rating protocols appear in Appendix D.

Degradation-aware Restoration Assessment (DRA). Given a degraded input $\mathbf{x}$, restored output $\hat{\mathbf{y}}$, and requested task c, a frozen Qwen3-VL evaluator (Bai et al., 2025) first examines residual target degradation in the output under the task-specific criteria in Table 14. It then compares the input and output, using the first judgment and its explanation to assess improvement. These judgments distinguish remaining degradation from the benefit of processing. Composite requests consider all listed degradations. DRA summarizes the judgments as a task-conditioned restoration score without a clean reference. Prompting, calibration, score mapping, and the super-resolution variant appear in Appendix D; Section 4.3 tests agreement with human judgments.

4 EXPERIMENTS

4.1 RESTORATION ON PAIRED AND REAL IMAGES

We compare the restoration component of UniLens with restoration and image-editing systems in Table 2. The public suite contains 36 paired benchmarks and 9,662 images across seven single-task categories and composites. We average images within benchmarks, benchmarks within tasks, and then task means. PSNR, LPIPS (Zhang et al., 2018), and DISTS (Ding et al., 2022) assess fidelity; DRA evaluates requested restoration on both public and real images. WildRestore-Bench reports Single and the 69-request Composite image mean. Both system comparisons use the same model trained with targeted supplementation and twenty sampling steps. Component comparisons retain five steps; baselines use their native settings. Full protocols appear in Appendix E.

For GGT, we use the released Qwen-Image-Edit-2511 adapter. FoundIR-v2 uses twenty steps; Qwen-Image-Edit-2511 and GGT use forty. Processing geometry and measured inference costs are reported in Appendix E.2.

Table 2: System comparisons on 36 public benchmarks and WildRestore-Bench. Public results use eight-task means; Single and Composite are two DRA evaluation settings. UniLens uses the same model and twenty sampling steps in both comparisons. Red/blue indicate best/second-best.

	Public benchmarks				WildRestore-Bench DRA score↑	
Method	PSNR↑	LPIPS↓	DISTS↓	DRA↑	Single	Composite
FLUX.1-Kontext-dev (Black Forest Labs et al., 2025)	13.35	0.546	0.221	3.13	3.23	2.67
Step1X-Edit-v1.1 (Liu et al., 2025)	17.20	0.420	0.255	2.18	2.79	2.40
LongCat (Meituan LongCat Team et al., 2025)	17.30	0.346	0.192	3.30	3.18	2.88
FoundIR (Li et al., 2025)	**22.76**	0.230	0.145	2.79	2.40	2.09
FoundIR-v2 (Chen et al., 2026)	20.30	0.227	0.138	3.47	3.13	2.59
Qwen-Image-Edit-2511 (Wu et al., 2025)	19.12	0.282	0.172	3.62	3.37	2.98
GGT (Kong et al., 2026)	22.31	**0.202**	**0.121**	**4.01**	**3.83**	**3.17**
UniLens	**23.83**	**0.156**	**0.095**	**4.33**	**3.98**	**3.76**

The UniLens entry leads all reported aggregate metrics. Relative to FoundIR-v2, public PSNR increases by 3.53 dB, DISTS decreases by 31.3%, and DRA increases by 24.8%. On WildRestore-Bench, DRA improves by 0.85 for Single and 1.17 for Composite (45.3%). We evaluate the same

checkpoint in Table 21. On the remaining 9,491 public images, PSNR/LPIPS/DISTS/DRA are 24.0118/0.1530/0.0923/4.3188.

Visual comparisons. We compare two Composite cases and one Single case from WildRestore-Bench in Figure 4. In row (a), UniLens reduces headlight flare and improves road visibility; the inset shows fewer rain streaks around the bollard. In row (b), it reduces rain and haze around the bicycle and guardrail. The lower-left enlargement in row (c) reveals residual noise in competing outputs; UniLens produces cleaner chair, stool, and floor regions. These examples illustrate visible removal without clean references.



Figure 4: WildRestore-Bench comparisons: (a) low light, flare, and rain; (b) rain and haze; (c) noise. Insets enlarge rain streaks in (a) and the lower-left region in (c). Composite requests remove all listed degradations. Views and inset regions are matched across methods.

We compare paired restoration on Flare7K real and FoundIR-Test in Figure 5. UniLens preserves the ceiling lights while suppressing flare in (a), and reduces rain and haze in (b).

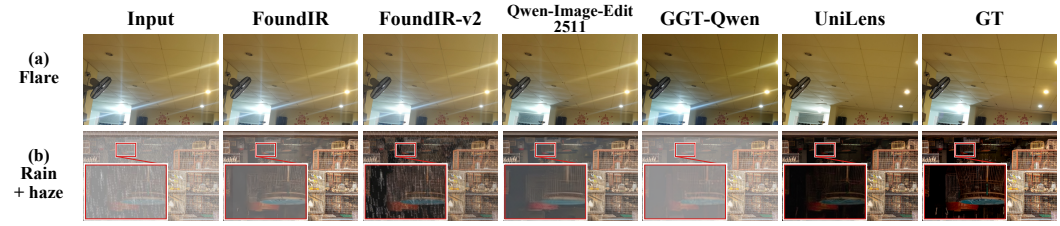

Figure 5: Paired comparisons: (a) Flare7K real (Dai et al., 2022); (b) FoundIR-Test (Li et al., 2025). Crops and insets are matched across methods; GT is shown on the right.

4.2 Data construction, supervision, and adaptation

Degradation coverage. We compare four coverage configurations at a fixed 24K-per-task budget in Table 3. MMD measures distribution discrepancy (Gretton et al., 2012), while neighbor distances and Density assess local support (Naeem et al., 2020). Targeted pairs replace 25% of the broad-expansion mixture, reducing single-/multi-scale MMD^2 by 10.25%/11.83% and 20-neighbor distance from 0.422886 to 0.406193. Against equal-budget random supplementation, targeted selection reduces multi-scale MMD^2 by 17.68%. The control and definitions appear in Appendix B.

Table 3: Distribution matching under a fixed 24K-per-task budget (six-task mean). Source proportions change while the total budget and low-light data remain fixed. Local statistics use unique inputs; global statistics retain sampling frequencies.

Statistic	Base mixture	+Offline	+Online	+Targeted
Mean pairwise ↓	0.991032	**0.952006**	0.994336	**0.966189**
Centroid ↓	0.469221	**0.445364**	0.451387	**0.413386**
Nearest distance ↓	0.362986	0.361501	**0.358947**	**0.349454**
20-NN distance ↓	0.426790	0.424956	**0.422886**	**0.406193**
MMD^2 (σ=0.30) ↓	0.121754	0.113579	**0.094964**	**0.085231**
Multi-scale MMD^2 ↓	0.106802	0.103354	**0.095185**	**0.083928**
$Density_{20}$ ↑	0.251687	0.245138	**0.293324**	**0.302813**

We compare how the construction routes change support for haze and blur in Figure 6. Offline construction gives the largest MMD^2 decrease for haze, whereas online synthesis gives the largest decrease for blur. Targeted additions further lower MMD^2 in both cases alongside the other sources. These complementary effects motivate task-specific expansion followed by query-matched supplementation; restoration outcomes are evaluated next.

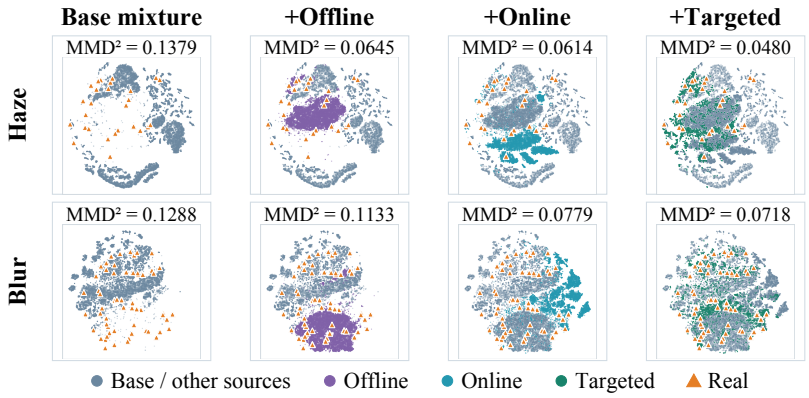

Figure 6: Coverage progression in shared t-SNE layouts. Colors highlight each added source; orange marks real observations. MMD^2 is measured in the original 256D CODE space ($\sigma = 0.30$).

Restoration after data expansion. We evaluate the construction progression with dense LoRA in Table 4(a). From the paired baseline to the targeted configuration, Single increases from 3.4587 to 3.8191 and Composite from 3.2174 to 3.5833. The targeted configuration improves on full online supervision in Single and all four FoundIR-Test metrics, with Composite at approximately 3.58.

Instruction-dependent targets. We compare supervision strategies at a common training budget in Table 4(b). Full compositional targets achieve the highest Single and Composite scores. Removing joint or selective targets reduces Composite to 2.8043 or 3.3116, respectively, and single-degradation online targets also yield lower scores. These comparisons support combining single, joint, and selective supervision when learning restoration under composite degradations.

Shared and sparse adaptation. We compare dense LoRA and MoLE with the same data and supervision in Table 4(c). MoLE raises Single from 3.8191 to 3.9338 and Composite from 3.5833 to 3.6993, and improves FoundIR-Test DRA from 4.4843 to 4.5252. The paired fidelity metrics complement these task-fulfillment scores. Together, the results support shared restoration knowledge with instruction-conditioned expert updates for single and composite requests.

4.3 Agreement with Human Judgments

We compare DRA with human judgments on 147 outputs in Table 5. On this development set, DRA exceeds five generic metrics, reaching PLCC/SRCC of 0.504/0.537 versus MUSIQ's 0.254/0.326.

Table 4: Component comparisons on WildRestore-Bench and FoundIR-Test (Li et al., 2025) (20-subset mean). Panels separate data progression, supervision, and adapter design. Best/second-best values are marked within each panel.

	WildRestore-Bench DRA↑		FoundIR-Test			
Configuration	Single	Composite	PSNR↑	LPIPS↓	DISTS↓	DRA↑
(a) Data and supervision progression: Dense LoRA						
Paired baseline	3.4587	3.2174	22.4401	0.2171	0.1106	4.2132
+ Offline pairs	3.6085	3.2283	22.1015	0.2199	0.1178	4.2749
+ Full online supervision	**3.7811**	**3.5942**	**22.8846**	**0.2110**	**0.1073**	**4.4770**
+ Targeted pairs	**3.8191**	**3.5833**	**23.1269**	**0.2069**	**0.1068**	**4.4843**
(b) Restoration targets: Dense LoRA						
Paired supervision only	**3.6085**	3.2283	22.1015	0.2199	0.1178	4.2749
Single-degradation online targets	3.4380	3.0870	22.4964	0.2250	0.1191	4.2188
Without joint-removal targets	3.5882	2.8043	22.5095	0.2315	0.1182	4.1288
Without selective-removal targets	3.5648	**3.3116**	**22.7458**	**0.2138**	**0.1109**	**4.2804**
Full compositional targets	**3.7811**	**3.5942**	**22.8846**	**0.2110**	**0.1073**	**4.4770**
(c) Adapter: training mixture with targeted pairs						
Dense LoRA (Hu et al., 2022)	**3.8191**	**3.5833**	**23.1269**	**0.2069**	**0.1068**	**4.4843**
Instruction-conditioned MoLE	**3.9338**	**3.6993**	**22.5056**	**0.2133**	**0.1070**	**4.5252**

Each row uses one checkpoint and five sampling steps. Panel (a) follows data and supervision progression; (b) shares a training budget; (c) fixes data and supervision. FoundIR-Test includes seven JPEG subsets outside the trained degradation families.

It agrees with human preferences on 77.9% of 140 non-tied within-input pairs, versus 70.0% for MUSIQ. The source-image bootstrap interval for SRCC is [0.388, 0.669]. The rating rubric covers degradation removal, naturalness, and content fidelity; calibration, panel selection, and sensitivity analyses are specified in Appendix D.

Table 5: Development-set correlations with human restoration-quality ratings (all 147 outputs; selected seven-rater consensus). Ratings include content fidelity. Higher is better; NIQE and BRISQUE are sign-reversed.

Score	PLCC	SRCC	KRCC
MUSIQ (Ke et al., 2021)	**0.254**	**0.326**	**0.233**
MANIQA (Yang et al., 2022)	0.036	0.175	0.132
NIQE (Mittal et al., 2013)	0.105	0.180	0.123
BRISQUE (Mittal et al., 2012)	0.179	0.145	0.099
CLIPIQA (Wang et al., 2023)	0.066	0.143	0.108
DRA	**0.504**	**0.537**	**0.403**

5 Conclusion

UniLens connects degradation-aware data construction, instruction-dependent restoration, and task-conditioned assessment. CODE guides offline, online, and targeted construction, improving distribution matching under a fixed budget. The restoration progression improves real-image scores and paired-benchmark performance. Compositional supervision and instruction-conditioned MoLE support single and composite requests. DRA evaluates residual degradation and input-relative improvement, with stronger human agreement than the compared generic metrics. Together, these results support degradation-aware learning and evaluation for real-world restoration.

AI USE STATEMENT

Generative AI assisted with literature discovery, language editing, LaTeX preparation, statistical analysis code, and organizing experimental results. Generated image pairs are used in data construction as described in Section 3.2.

REPRODUCIBILITY STATEMENT

The method and experiments specify the learning objectives, construction routes, and evaluation criteria. Appendices A–D follow the method components. Appendix E specifies evaluation protocols and extended quantitative results; Appendix F provides visual comparisons.

REFERENCES

Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *Proc. CVPR*, pp. 1692–1700, 2018.

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. ARNIQA: Learning distortion manifold for image quality assessment. In *Proc. WACV*, pp. 189–198, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv:2511.21631*, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv:2506.15742*, 2025.

Shuo Cao, Yihao Liu, Wenlong Zhang, Yu Qiao, and Chao Dong. GRIDS: Grouped multiple-degradation restoration with image degradation similarity. In *Proc. ECCV*, pp. 70–87, 2024.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proc. ECCV*, pp. 17–33, 2022.

Xiang Chen, Jinshan Pan, Jiangxin Dong, Jian Yang, and Jinhui Tang. FoundIR-v2: Optimizing pre-training data mixtures for image restoration foundation model. In *Proc. CVPR*, pp. 8471–8480, 2026.

Marcos V. Conde, Gregor Geigle, and Radu Timofte. InstructIR: High-quality image restoration following human instructions. In *Proc. ECCV*, pp. 1–21, 2024.

Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Flare7K: A phenomenological nighttime flare removal dataset. In *Proc. NeurIPS*, pp. 3926–3937, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, pp. 248–255, 2009.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2567–2581, 2022.

Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proc. CVPR*, pp. 1715–1723, 2017.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(25):723–773, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pp. 770–778, 2016.

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

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proc. CVPR*, pp. 5197–5206, 2015.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proc. ICCV*, pp. 5148–5157, 2021.

Jaekyun Ko, Dongjin Kim, Soomin Lee, Guanghui Wang, and Tae Hyun Kim. Diffusion-based sRGB real noise generation via prompt-driven noise representation learning. In *Proc. CVPR*, pp. 35956–35966, 2026.

Xiangtao Kong, Jixin Zhao, Lingchen Sun, Rongyuan Wu, and Lei Zhang. GGT-100K: Generative ground truth for generalizable real-world image restoration. *arXiv:2605.31039*, 2026.

Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.*, 28(1): 492–505, 2019.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-One image restoration for unknown corruption. In *Proc. CVPR*, 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 *Proc. ICCV*, pp. 12626–12636, 2025.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, et al. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Proc. ECCV*, pp. 430–448, 2024.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S. Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *arXiv:2507.20590*, 2025.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, et al. Step1X-Edit: A practical framework for general image editing. *arXiv:2504.17761*, 2025.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Controlling vision-language models for multi-task image restoration. In *Proc. ICLR*, 2024.

Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Trans. Image Process.*, 31:4149–4161, 2022.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. ICCV*, pp. 416–423, 2001.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, et al. LongCat-Image technical report. *arXiv:2512.07584*, 2025.

A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012.

A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proc. ICML*, pp. 7176–7185, 2020.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proc. CVPR*, pp. 257–265, 2017.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for All-in-One image restoration. In *Proc. NeurIPS*, pp. 71275–71293, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pp. 8748–8763, 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. In *Proc. CVPR*, pp. 28144–28154, 2025.

Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Proc. ECCV*, pp. 184–201, 2020.

Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proc. AAAI*, pp. 2555–2563, 2023.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proc. ICCV Workshops*, pp. 1905–1914, 2021.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep Retinex decomposition for low-light enhancement. In *Proc. BMVC*, pp. 155, 2018.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, et al. Qwen-Image technical report. *arXiv:2508.02324*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, et al. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proc. ICML*, pp. 54015–54029, 2024a.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *Proc. ICLR*, 2024b.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proc. CVPR Workshops*, pp. 1191–1200, 2022.

Wenhan Yang, Robby T. Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proc. CVPR*, pp. 1685–1694, 2017.

Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proc. CVPR*, pp. 3060–3069, 2020.

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

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proc. CVPR*, pp. 14483–14494, 2025.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, et al. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proc. CVPR*, pp. 25669–25680, 2024.

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 *Proc. CVPR*, pp. 12753–12763, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proc. CVPR*, pp. 5728–5739, 2022.

He Zhang and Vishal M. Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proc. CVPR*, pp. 695–704, 2018.

He Zhang, Vishwanath Sindagi, and Vishal M. Patel. Image de-raining using a conditional generative adversarial network. *IEEE Trans. Circuits Syst. Video Technol.*, 30(11):3943–3956, 2020.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proc. ICCV*, 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 *Proc. CVPR*, pp. 586–595, 2018.

Supplementary material. Appendices A–D detail the three stages of the UniLens pipeline: data construction, model adaptation, and assessment. CODE is the degradation comparison component within data construction. Appendices E and F describe evaluation protocols, additional quantitative results, inference cost, and visual comparisons.

A CODE: IMPLEMENTATION AND VALIDATION

A.1 TRAINING CONFIGURATION

CODE starts from ImageNet-1K-pretrained ResNet-50 (Deng et al., 2009; He et al., 2016), using 2,048-channel valid-region pooled features, a 256D linear readout, and a 256–256–128 contrastive head. BatchNorm statistics and affine parameters are frozen. SGD uses momentum 0.9, weight decay 10^{-4}, initial learning rate 0.0035, a cosine schedule with a 53,853-update horizon, and gradient clipping at 5. The encoder is trained for 20,000 updates and then held fixed for all comparisons. Training spans noise, blur, haze, rain, flare, low light, super-resolution, and raindrops. Evaluation excludes raindrops. Eight GPUs train with BF16 and seed 20260829; each GPU receives one positive pair per active family and a seven-image ordering group. The cosine schedule has a minimum learning rate of 10^{-6}.

The encoder is trained on 199,872 clean images and 718,934 pairs. Its training pool differs from that of the restoration model and excludes the controlled validation scenes and known real-image validation identities. Both validation sets support encoder development.

The contrastive and ordinal temperatures are both 0.1. The LPIPS separation threshold is 0.05, and the ordinal weight increases linearly from zero to 0.1 over the first 2,000 updates. LPIPS uses its frozen AlexNet v0.1 configuration on 512-pixel square inputs to rank candidates; it supplies no gradient or numeric regression target to CODE. CODE preserves aspect ratios within an approximately one-megapixel input budget and excludes padding during pooling.

A.2 LEARNING DEGRADATION SIMILARITY

Cross-content family alignment. Inspired by ARNIQA (Agnolucci et al., 2024), we align images with different scene content but a shared degradation family. Positive pairs can contain two synthesized images, one synthesized and one paired-data image, or two paired-data images for the raindrop family. They need not share the exact corruption operator or strength. Each per-device group contains one positive pair per active family. Thus, every non-self image other than the designated mate belongs to a different family and serves as a negative.

A training-only projector maps the pooled image embedding to a normalized contrastive vector. For a batch of positive pairs, the contrastive loss is formulated as:

$$\mathcal{L}_{\text{type}} = -\frac{1}{2B}\sum_{i=1}^{2B}\log\frac{\exp(\mathbf{z}_i^\top \mathbf{z}_{p(i)}/\tau)}{\sum_{\substack{j=1\\ j\neq i}}^{2B}\exp(\mathbf{z}_i^\top \mathbf{z}_j/\tau)}, \tag{5}$$

where B is the number of positive pairs, i and j index their $2B$ images, $p(i)$ identifies the positive mate of image i, $\mathbf{z}_i \in \mathbb{R}^{128}$ is its unit-normalized projected feature, and $\tau > 0$ is the temperature. The denominator includes the positive mate and all negative examples, masking only the anchor itself. Here B is seven or eight according to the active families. The loss is computed within each device group without gathering contrastive features across devices; ordinal candidates are separate. All image comparisons use $\mathbf{h}$, without the projector.

Within-family ordering. Family alignment alone does not constrain variation within a degradation family. For one scene and family, we compare an anchor with six variants using frozen LPIPS (Zhang et al., 2018). The closest two variants form a near set, the farthest two form a far set, and the middle two are omitted. A group is retained when the near and far perceptual-distance ranges have the required separation. The loss for these relative relations is formulated as:

$$\ell_{\text{ord}}(a) = \frac{1}{4}\sum_{n\in\mathcal{N}_a}\sum_{f\in\mathcal{F}_a}\log\left[1+\exp\left(\frac{s(a,f)-s(a,n)}{\tau_o}\right)\right], \tag{6}$$

where a indexes the anchor, $\mathcal{N}_a$ and $\mathcal{F}_a$ contain its two near and two far candidates, $\mathbf{h}_k = \mathbf{h}(\mathbf{x}_k) \in \mathbb{R}^{256}$ is the unit embedding of image k, $s(a,k) = \mathbf{h}_a^\top \mathbf{h}_k$ is cosine similarity, and $\tau_o > 0$ is the ordering temperature. Unit normalization gives $d_E(\mathbf{x}_a, \mathbf{x}_k)^2 = 2 - 2s(a,k)$, so the cosine ranking encourages the near–far ordering in the CODE block of Figure 3. Invalid groups contribute zero. The batch mean defines $\mathcal{L}_{\text{ord}}$. The losses are combined through Equation (2). Ordering candidates do not enter the contrastive negative set; LPIPS supplies relative supervision without regressing physical severity.

A.3 Comparison representations and evaluation protocol

DINOv2 (Oquab et al., 2023) uses ViT-B/14 CLS features (768D), CLIP (Radford et al., 2021) uses ViT-B/32 projected image features (512D), ImageNet uses pooled ResNet-50 features (2,048D), and Qwen VAE features use a 1,376D summary of its latents. All are normalized for cosine ranking. We use architecture-specific preprocessing and examine its effect below. Each scene–family cell retains 16 draws; 15 gallery entries share pixels with another entry.

All methods use their fixed feature definitions and L2 normalization. DINOv2 and CLIP use their recommended preprocessing for the main comparison. Native-resolution five-crop sensitivity yields controlled retrieval mAP of 32.59% and 44.54%, respectively; this affects the size of the gap but does not reverse the controlled content–degradation preference. Both use unchanged pretrained weights.

Matching and retrieval protocol. The controlled gallery comprises 128 scenes, seven families, and sixteen draws per scene–family (14,336 images); additional query sets contain 190 real images and 600 images from five FoundIR subsets. All representations receive only low-quality images. Cross-content matching evaluates all sixteen reference draws, pairing candidates with the same reference scene and draw index. Controlled retrieval excludes all 112 versions of the query scene; real and FoundIR queries exclude known identical identities and pixels. Exact score ties use the metric expectation over uniform tie orderings. Aggregation is scene-wise for controlled images, domain–family-wise for real images, and family-wise for FoundIR. Matching reports the fraction of correct two-choice decisions, with chance at 50%. Retrieval mAP averages precision at ranks occupied by same-family images, normalized by the number of relevant images; P@10 is the same-family fraction among the first ten results. The query populations and source image pool are shared, while matching uses candidate pairs and retrieval uses the full eligible ranking. Known duplicate identities and exact pixel matches are excluded.

A.4 Additional validation results

Content–degradation conflict. The conflict test in Table 1 compares a different-scene degradation match against a same-scene content match. Ties count as one half. CODE selects the degradation match in 98.45% of comparisons, supporting comparison across scenes.

Cross-content matching and retrieval. We evaluate transfer from controlled queries to real-image and FoundIR queries in Table 6. Matching compares two candidates from the same reference scene, only one of which shares the query's degradation family; chance is 50%. Retrieval ranks the full eligible gallery by feature distance. Mean average precision (mAP) measures the ranking of same-family examples, and precision at ten (P@10) measures their proportion among the first ten results. Identity exclusions and aggregation follow Section A.3.

CODE leads matching accuracy, mAP, and P@10 on all three query sets. On FoundIR, it reaches 89.61% matching, 80.44% mAP, and 85.67% P@10. DINOv2 and CLIP exceed chance in matching while rarely favoring degradation in the conflict test: their features retain degradation information, but scene similarity can dominate the comparison. These tests assess same-family matching with the complete CODE objective. Restoration performance is evaluated separately in Section 4.2.

Table 6: Cross-content degradation matching and retrieval (%, ↑). Match is a two-choice rate; mAP evaluates the complete ranked gallery. The tests share query populations and a source image pool; matching uses pairs and retrieval ranks the eligible gallery.

	Controlled		Real images		FoundIR		
Representation	Match	mAP	Match	mAP	Match	mAP	P@10
ResNet-50 (He et al., 2016)	**82.25**	**40.01**	**72.68**	**31.13**	**78.90**	**39.44**	**72.89**
Qwen VAE (Wu et al., 2025)	76.86	39.80	69.52	30.61	76.65	38.28	59.88
DINOv2 (Oquab et al., 2023)	60.41	16.36	60.25	16.89	58.76	16.27	22.02
CLIP (Radford et al., 2021)	67.48	21.12	62.73	20.73	62.76	21.36	46.69
CODE	**98.72**	**96.14**	**83.85**	**65.86**	**89.61**	**80.44**	**85.67**

B Coverage-guided data construction and distribution analysis

B.1 Reference observations and training sources

Real low-quality photographs provide an empirical reference for degradation patterns encountered in practical imaging. CODE compares their degradation features with training inputs, guiding the construction of pairs from other source images. The reference photographs are used as queries, without clean counterparts or supervised restoration targets. This role is distinct from the generated and existing pairs used to optimize the restoration model.

The controlled distribution study uses 250 image–task records from 249 photographs: noise 27, blur 82, haze 27, flare 41, rain 27, and low light 46. Figure 7 illustrates the scene variation. WildRestore-Bench and the evaluation use of real images are described in Appendix D.



Figure 7: Real degraded images used to guide data construction. Examples span six restoration demands and varied scene content.

The available paired pool contains 714,894 records from 33 sources, including public GGT Low-Light, Rain, and Haze pairs (Kong et al., 2026). Offline supplementation adds 117,159 source records (93,210 pairs with generated degradations and 23,949 flare pairs). The combined pool has 832,053 records and 828,631 unique degraded–target image pairs. We summarize the additional sources in Table 7. These counts describe available sources before training-specific filtering and sampling. The Base mixture samples this paired pool and is abbreviated as Base in the coverage plots.

Offline construction. For noise, PNG (Ko et al., 2026) transfers conditions learned from real noisy–clean pairs to new clean content. We interpolate two reference-noise conditions for each output. The 576-pixel and 1,024-pixel sources use the same generator; the latter blends four overlapping 576-pixel windows. Seedream image editing constructs blur, rain, and haze inputs from clean images under degradation-specific templates. Flare pairs use additional asset compositing. These

Table 7: Offline supplementation beyond the base archive, which already contains public GGT pairs (Kong et al., 2026). Counts precede deduplication and training resampling.

Task	Construction route	Records
Noise	PNG noise synthesis (Ko et al., 2026)	52,972
Blur	Image editing (Doubao/Seedream)	21,120
Rain	Synthetic rain generation	10,748
Haze	Generated haze appearances	8,370
Flare	Additional offline flare construction	23,949
Total	12 supplemental sources	117,159

routes generate degraded inputs with corresponding restoration targets; public GGT pairs instead supply generated targets for real degraded inputs and belong to the base paired pool.

B.2 CODE DIAGNOSTICS AND CONSTRUCTION PROCEDURE

We use frozen CODE embeddings to compare a task's real reference set $\mathcal{Q}$ with a sampled training multiset $\mathcal{T}$. For query $\mathbf{q} \in \mathcal{Q}$, local support is measured by:

$$L_k(\mathbf{q}, \mathcal{T}) = \frac{1}{k} \sum_{\mathbf{x} \in \mathrm{NN}_k(\mathbf{q}, \mathcal{T}^u)} d_E(\mathbf{q}, \mathbf{x}), \tag{7}$$

where $\mathcal{T}^u$ contains unique training inputs, NN_k returns the k nearest under d_E, and $1 \le k \le |\mathcal{T}^u|$. Larger L_k indicates weaker local support. Distribution discrepancy and local support summarize complementary aspects of coverage; the statistics are defined in Appendix B.4.

The task-wise results in Table 12 ground the construction choices. The Base mixture has the lowest $\mathrm{Density}_{20}$ for noise (0.0481), indicating sparse local support, while low light has the smallest single-scale MMD^2 (0.0700). We retain the low-light pool in broad offline/online expansion and in the controlled distribution study. The larger-scale targeted set also includes low light, as specified in Appendix B.6.

Core procedure. Given the Base mixture, real queries, frozen CODE, task-specific construction routes, and sampling budgets:

1. Encode LQ inputs from the Base mixture and real queries; summarize discrepancy by task and inspect each query's nearest-neighbor support.
2. Construct offline LQ–target pairs for the selected families using the generators described above.
3. Configure online sampling of operator parameters and compositions on clean or eligible paired inputs, retaining intermediate states for single, joint, and selective supervision.
4. For each query in the targeted task set, retrieve source pairs by LQ-feature distance and generate candidates from their LQ inputs or restoration targets.
5. Rank candidate LQs by CODE distance to the query and select under pair-consistency, identity, and source-reuse constraints, with balanced query allocations.
6. Mix the added pairs with the Base mixture and online sampling for restoration training; reassess the induced training distribution using the same CODE.

The task sets, budgets, and operator settings are specified below. Real queries determine matching conditions; their clean targets are not used to construct supervision.

B.3 CONSTRUCTION AND SAMPLING

We summarize the online families in Table 8. Low light uses existing paired supervision, and super-resolution uses clean-image online construction. Online augmentation can start from clean images or existing pairs; haze is excluded from the existing-pair branch. Within allowed families, stochastic sampling varies severity and composition. Removal and retention targets follow Section C.1.

Table 8: Online degradation families and relative sampling weights. An available family is sampled according to these weights, subject to input-specific exclusions.

Family	Weight	Construction and sampled variation
Noise	3	Fifteen branches spanning Gaussian, signal-dependent, impulsive, multiplicative, and compression-associated noise.
Blur	1	Eighteen branches varying kernel shape, direction, spatial support, and combinations of blur mechanisms.
Rain	2	Synthetic streak layers with variable length, density, width, direction, and spatial extent.
Haze	2	Depth-conditioned scattering with sampled atmospheric light and scattering coefficients.
Flare	2	Scattering and reflection assets with spatial transforms and repeated compositing.
Super-resolution	5	Real-ESRGAN-style blur, resampling, noise, and compression (Wang et al., 2021).

Targeted supplementation represents each augmented task through a balanced allocation across its real queries. CODE determines which source pairs and synthesized candidates support each query; the sampling budget preserves variety across the reference collection. The controlled study adds 6K pairs per augmented task, with at most five accepted pairs per source target and query. Low light reuses its existing paired pool. Explicit super-resolution and downsample–upsample blur are excluded from this distribution study; the restoration model's broader task coverage is evaluated separately.

B.4 Controlled comparison and quantitative diagnostics

The four coverage configurations are Base mixture (24K), +Offline (12K+12K), +Online (8K+8K+8K), and +Targeted (6K+6K+6K+6K), per task. All columns use data seed 20260914, the same 250 real reference queries, frozen CODE, preprocessing, and identity exclusions. We report unique-image counts in Table 9. Global statistics weight repeated samples by frequency, whereas local statistics use unique images. The four views use a common sample budget with different source proportions; unique-image counts record the number of distinct training inputs.

Table 9: Unique degraded images in the equal-budget comparison. Each configuration contains 24K samples per task, including repetitions.

Task	Base mixture	+Offline	+Online	+Targeted
Noise	21,318	23,630	23,840	23,908
Blur	19,620	23,538	23,991	23,995
Haze	21,438	20,326	22,637	23,463
Flare	24,000	24,000	24,000	24,000
Rain	23,980	19,975	22,060	22,899
Low light	18,953	18,953	18,953	18,953

All distribution statistics use the unit CODE embeddings from Equation (1) and the training/query sets defined in Section 3.2. Local proximity follows Equation (7); we average L_1 and L_{20} over queries to report nearest and 20-neighbor distance. Local metrics use unique images, while global metrics retain sampling-frequency weights.

To summarize global separation, mean pairwise and centroid distances are formulated as:

$$
\begin{aligned}
D_{\text{pair}} &= \frac{1}{|\mathcal{Q}||\mathcal{T}|} \sum_{\mathbf{q}\in\mathcal{Q},\mathbf{x}\in\mathcal{T}} d_E(\mathbf{q},\mathbf{x}), \\
D_{\text{center}} &= \left\| \frac{1}{|\mathcal{T}|} \sum_{\mathbf{x}\in\mathcal{T}} \mathbf{h}(\mathbf{x}) - \frac{1}{|\mathcal{Q}|} \sum_{\mathbf{q}\in\mathcal{Q}} \mathbf{h}(\mathbf{q}) \right\|_2 ,
\end{aligned}
\tag{8}
$$

where $|\mathcal{T}|$ counts training samples including repetitions, $|\mathcal{Q}|$ counts queries, and D_{pair} averages cross-distribution distances. The statistic D_{center} compares the two embedding means. Both decrease with closer global alignment.

We measure distribution discrepancy using the empirical squared maximum mean discrepancy (MMD) (Gretton et al., 2012), formulated as:

$$\widehat{\text{MMD}}^2_\sigma = \frac{1}{|\mathcal{T}|^2}\sum_{\mathbf{x},\mathbf{x}'\in\mathcal{T}} k_\sigma(\mathbf{h}(\mathbf{x}),\mathbf{h}(\mathbf{x}')) + \frac{1}{|\mathcal{Q}|^2}\sum_{\mathbf{q},\mathbf{q}'\in\mathcal{Q}} k_\sigma(\mathbf{h}(\mathbf{q}),\mathbf{h}(\mathbf{q}')) - \frac{2}{|\mathcal{T}||\mathcal{Q}|}\sum_{\mathbf{x}\in\mathcal{T},\mathbf{q}\in\mathcal{Q}} k_\sigma(\mathbf{h}(\mathbf{x}),\mathbf{h}(\mathbf{q})), \tag{9}$$

where $k_\sigma(\mathbf{u},\mathbf{v}) = \exp[-\|\mathbf{u}-\mathbf{v}\|_2^2/(2\sigma^2)]$ is a Gaussian kernel, $\mathbf{u},\mathbf{v}\in\mathbb{R}^{256}$ are embeddings, and $\sigma > 0$ is its bandwidth. This biased estimator includes diagonal terms and weights repeated training images by sampling frequency. We use $\sigma = 0.30$ for the primary measure. Multi-scale MMD averages squared MMD over nine fixed bandwidths: 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.60, 0.80, and 1.20. All samples are included, and MMD is not a selection objective.

We quantify adaptive neighborhood support using Density (Naeem et al., 2020), formulated as:

$$\text{Density}_{20}(\mathcal{Q},\mathcal{T}) = \frac{1}{20|\mathcal{Q}|}\sum_{\mathbf{q}\in\mathcal{Q}}\sum_{\mathbf{x}\in\mathcal{T}^u}\mathbb{I}[d_E(\mathbf{q},\mathbf{x}) < r_{20}(\mathbf{x},\mathcal{T}^u)]\,, \tag{10}$$

where $\mathbb{I}[\cdot]$ is the indicator function and $r_{20}(\mathbf{x},\mathcal{T}^u)$ is the distance from $\mathbf{x}$ to its twentieth other training neighbor. Radii are recomputed for each data composition. Density can exceed one and is not a coverage probability. We report six-task means for the fixed sampling seed; Section 4.2 reports restoration performance.

We compare the four mixtures in Table 3 and Figure 6. The following analyses examine whether changes in distribution matching persist across tasks and sampling choices.

B.5 Random control, source diversity, and sensitivity

We compare random and targeted additions at matched budgets in Table 10.

Table 10: Random versus targeted supplementation. Both retain the same 18K training samples and add 6K candidates, with matched generation and quality-check budgets. Candidate pools are independently generated. Values are six-task means.

Statistic	+Online	+Random	+Targeted
Mean pairwise ↓	**0.994336**	1.008561	**0.966189**
Centroid ↓	**0.451387**	0.482199	**0.413386**
Nearest distance ↓	**0.358947**	0.362046	**0.349454**
20-NN distance ↓	**0.422886**	0.427613	**0.406193**
MMD^2 (σ=0.30) ↓	**0.094964**	0.098624	**0.085231**
Multi-scale MMD^2 ↓	**0.095185**	0.101958	**0.083928**
Density_{20} ↑	0.293324	**0.303712**	**0.302813**

Targeted supplementation reduces single- and multi-scale MMD^2 relative to random supplementation by 13.58% and 17.68%, respectively, but its Density is slightly smaller (0.302813 versus 0.303712). Across the five augmented tasks, the mean number of distinct source target images among twenty neighbors falls from 19.16 to 13.64. Improved proximity therefore comes with more concentrated origin support. This comparison tests source retrieval and candidate selection together.

We examine two sampling sensitivities in Table 11: excluding the unchanged low-light task and assigning equal weight to unique images. Both retain lower single- and multi-scale MMD^2 after targeted supplementation. For unique-image weighting, the local-neighbor and Density statistics are unchanged because they already use unique inputs. At the narrowest individual bandwidth, 0.05, haze is an exception; the primary and multi-scale MMD measures improve across all five augmented tasks.

Table 11: Sensitivity of distribution matching to task averaging and image weighting. Panels use the same fixed pools and sampling seed as Table 3.

Statistic	Base mixture	+Offline	+Online	+Targeted
Five augmented tasks; low-light control excluded				
MMD^2	0.132111	0.122301	**0.099962**	**0.088284**
Multi-scale MMD^2	0.113309	0.109170	**0.099368**	**0.085859**
Six tasks; equal weight per unique degraded image				
MMD^2	0.125674	0.112121	**0.093595**	**0.083699**
Multi-scale MMD^2	0.110730	0.102542	**0.094383**	**0.082798**

We examine individual task responses in Table 12, reporting all seven statistics so that aggregate improvements can be assessed alongside mixed task-level effects.

Table 12: Distribution matching by task. Global metrics use sampling-frequency weights; local metrics use unique images. Colors rank data mixtures separately for each metric and task.

Statistic	Task	Base mixture	+Offline	+Online	+Targeted
Mean pairwise ↓	Noise	**1.229882**	1.254832	1.272940	**1.221521**
	Blur	**0.834816**	**0.792169**	0.892327	0.873092
	Haze	0.880018	0.689712	**0.686133**	**0.635476**
	Flare	**0.940583**	**0.967888**	1.024094	1.000440
	Rain	**0.991931**	**0.938474**	1.021559	0.997642
	Low light	**1.068960**	**1.068960**	**1.068960**	**1.068960**
Centroid ↓	Noise	**0.705267**	0.800713	0.791210	**0.703689**
	Blur	0.339381	**0.299604**	0.336551	**0.310789**
	Haze	0.453991	0.238609	**0.203496**	**0.149008**
	Flare	**0.518179**	0.579074	0.597455	**0.545521**
	Rain	0.347092	**0.302768**	0.328195	**0.319892**
	Low light	**0.451417**	**0.451417**	**0.451417**	**0.451417**
Nearest distance ↓	Noise	0.452628	0.461835	**0.448096**	**0.443572**
	Blur	0.384473	0.353621	**0.350872**	**0.342062**
	Haze	0.215880	0.207148	**0.204261**	**0.202426**
	Flare	**0.418825**	0.432366	0.429013	**0.416851**
	Rain	**0.344539**	0.352465	0.359868	**0.330240**
	Low light	**0.361570**	**0.361570**	**0.361570**	**0.361570**
20-NN distance ↓	Noise	**0.524758**	0.543681	0.532191	**0.510443**
	Blur	0.440803	0.405309	**0.404706**	**0.391770**
	Haze	0.245515	0.238200	**0.237594**	**0.229679**
	Flare	**0.501730**	0.523164	0.514991	**0.485985**
	Rain	0.423104	**0.414553**	0.423004	**0.394452**
	Low light	**0.424828**	**0.424828**	**0.424828**	**0.424828**
MMD^2 (σ=0.30) ↓	Noise	**0.147564**	0.167173	0.148148	**0.132590**
	Blur	0.128797	0.113337	**0.077936**	**0.071773**
	Haze	0.137915	0.064546	**0.061376**	**0.047961**
	Flare	0.159099	0.183530	**0.140915**	**0.120956**
	Rain	0.087181	0.082916	**0.071437**	**0.068139**
	Low light	**0.069970**	**0.069970**	**0.069970**	**0.069970**
Multi-scale MMD^2 ↓	Noise	**0.165821**	0.202822	0.189332	**0.159870**
	Blur	0.084497	0.070679	**0.059513**	**0.053757**
	Haze	0.111252	0.052217	**0.047308**	**0.037544**
	Flare	**0.130698**	0.153488	0.136765	**0.116838**
	Rain	0.074276	0.066646	**0.063925**	**0.061289**
	Low light	**0.074269**	**0.074269**	**0.074269**	**0.074269**

Statistic	Task	Base mixture	+Offline	+Online	+Targeted
Density$_{20}$ ↑	Noise	0.048148	0.072222	**0.168519**	**0.229630**
	Blur	0.154268	0.173780	**0.266463**	**0.287805**
	Haze	0.168519	0.209259	**0.275926**	**0.264815**
	Flare	**0.226829**	0.190244	**0.229268**	0.196341
	Rain	**0.535185**	0.448148	0.442593	**0.461111**
	Low light	**0.377174**	**0.377174**	**0.377174**	**0.377174**

B.6 Larger-scale data construction

The expanded restoration-training set contains 200,000 selected pairs: 33,334 each for noise and blur, and 33,333 each for haze, flare, rain, and low light. It spans 44,387 source targets and uses 1,101 real reference images. Pair review excludes confirmed correspondence errors. The reference queries guide matching; the resulting source-based pairs supply training supervision.

The expanded training set supplies the restoration configurations in Section 4.2; the separate 24K-per-task study compares distribution matching under a fixed budget. For the expanded set, nearest distances decrease across all six tasks, and Density varies by task. Evaluation protocols and reference-exclusion results appear in Appendix E.

B.7 Six-task visualization of the controlled mixtures

We compare all six tasks in Figures 8 and 9, using the fixed-budget mixtures from Section B.4. For each task, t-SNE is fitted jointly to unique inputs from all four mixtures and the real reference, with perplexity 30, 1,500 iterations, PCA initialization, and one random seed. Coordinates remain fixed across a row; all quantitative statistics use the original CODE embeddings. Colors identify construction sources using the same image exclusions as the quantitative analysis. Changing source proportions means that the columns are not nested sample sets.

The task-level statistics in Table 12 show different responses to broad expansion: offline supplementation reduces MMD^2 for haze but increases it for noise. Targeted supplementation reduces MMD^2 for all five augmented tasks relative to +Online. Low light starts with the lowest baseline MMD^2 (0.069970) under this reference and metric; its identical panels follow from reusing the same samples as a fixed control.

C Compositional supervision and MoLE

C.1 Compositional supervision

Conventional restoration pairs are augmented with the single, joint, and selective targets defined in Equation (3). These targets provide complementary transformations for learning restoration under degradation combinations.

Online synthesis from clean images samples single, joint, and selective removal with probabilities 0.3, 0.3, and 0.4. Existing-pair augmentation has probability 0.35, subject to task exclusions.

Structured text instructions. The conditioning variable c is the complete text prompt rendered in five fields: `TASK`, `SCOPE`, `REGIONS`, `ACTION`, and `PRESERVE`. `TASK` identifies the restoration task; `SCOPE` and `REGIONS` give task-based descriptions of the intended spatial extent, not segmentation masks. `ACTION` states the requested removal, while `PRESERVE` specifies retained degradations and content-preservation constraints. Training and UniLens inference use the same renderer and field semantics. Public and WildRestore-Bench evaluation fix the action text for each image–request pair and disable image-specific VLM planning. Thus, c remains a text instruction with an explicit structure.

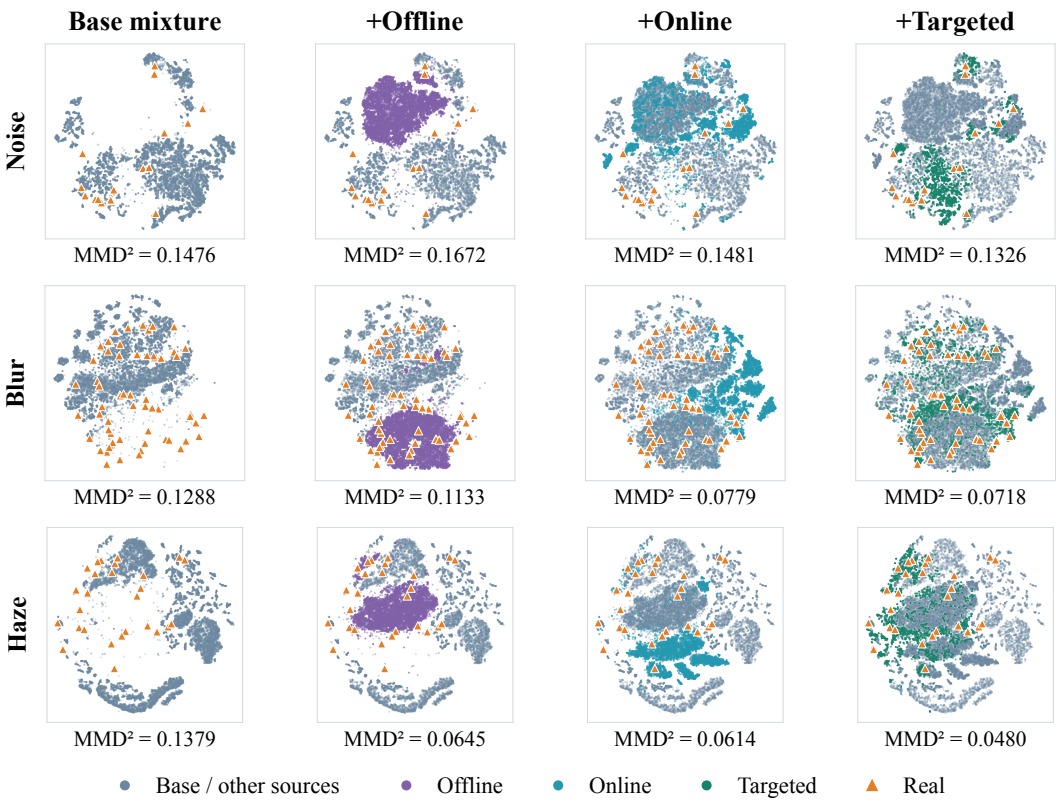

Figure 8: Four coverage configurations for noise, blur, and haze. Purple marks offline additions, cyan online synthesis, and green targeted additions when each source is introduced; other training samples are blue. Orange triangles mark the fixed real observations. The displayed MMD^2 uses $\sigma = 0.30$ in the original CODE space.

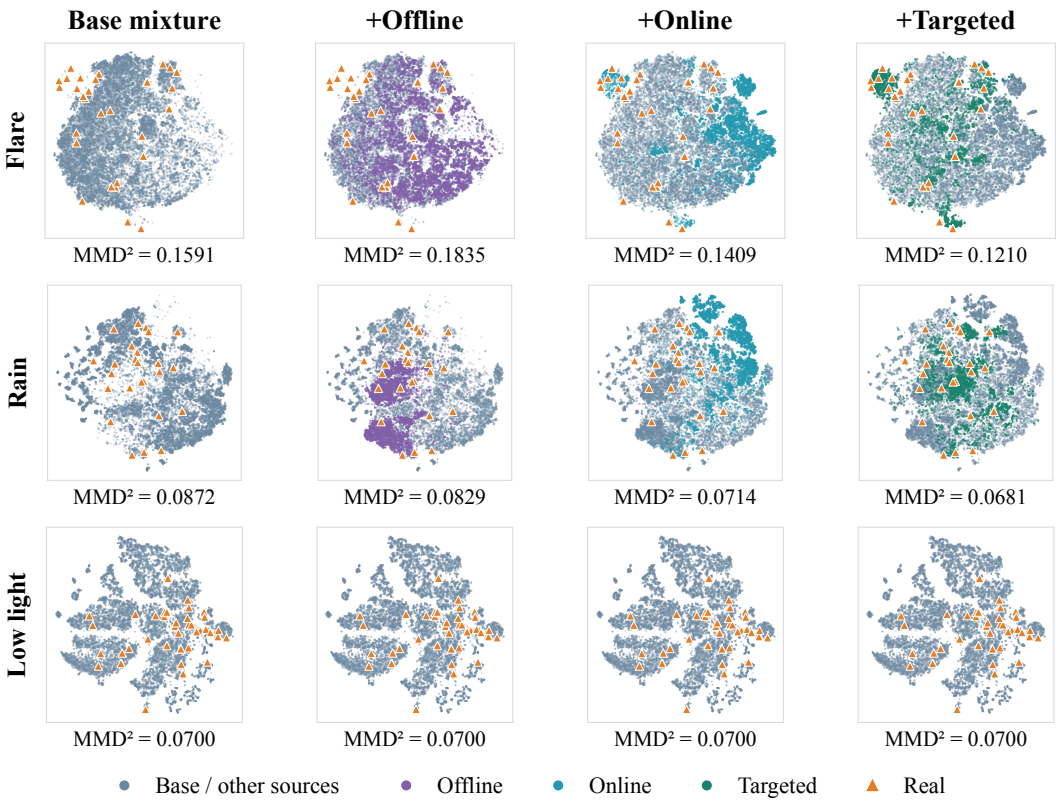

Figure 9: Four coverage configurations for flare, rain, and low light, using the same color convention. Low light is a fixed control: the same sampled training inputs are used in all four mixtures, so its points and statistics are identical. The fixed row provides a common control across configurations.

C.2 ADAPTER PLACEMENT AND PARAMETER BUDGET

Adapters are applied to twelve linear projections in each of the backbone's sixty Transformer blocks: query, key, value, and output projections in the image and condition attention streams; the output projections of both feed-forward branches; and the image and condition modulation projections. This gives 720 adapted linear modules. Each module owns its shared and expert matrices; the twelve modules in a block share that block's expert indices and weights. The decision is shared across tokens and recomputed at each denoising time.

The dense adapter has rank 256. MoLE uses a rank-128 shared branch and sixteen rank-8 experts, activating four experts per block. Each low-rank update factorizes as $\Delta\mathbf{W} = \mathbf{BA}$, with $\mathbf{A} \in \mathbb{R}^{r\times d_{\text{in}}}$, $\mathbf{B} \in \mathbb{R}^{d_{\text{out}}\times r}$, and branch rank r. The router projects the normalized pooled condition and time embedding to 64 dimensions, adds a learned 64-dimensional layer embedding, and produces sixteen expert scores through SiLU and a linear projection. Its parameters are shared across blocks. Both dense LoRA and MoLE store 1,887,436,800 adapter-matrix parameters; MoLE adds 412,770 routing and auxiliary-head parameters. Parameter counts describe stored adapter capacity; inference cost is measured in Appendix E.2. Zero-initialized output factors preserve the backbone's initial output.

C.3 ROUTING AND OPTIMIZATION

The router combines the pooled backbone condition with network position and denoising time. Its expert scores are formulated as:

$$\mathbf{g}_{\ell,t} = \mathbf{W}_g\,\mathrm{SiLU}[\mathrm{Proj}_c(\mathrm{LN}(\bar{\mathbf{c}}_\ell)) + \mathrm{Proj}_t(\mathrm{LN}(\mathbf{e}_t)) + \mathbf{e}_\ell] + \mathbf{b}_g, \tag{11}$$

where $\bar{\mathbf{c}}_\ell \in \mathbb{R}^{d_c}$ is the valid-position mean of image-and-instruction condition states, $\mathbf{e}_t \in \mathbb{R}^{d_t}$ is the time embedding, and $\mathbf{e}_\ell \in \mathbb{R}^{d_r}$ is a learned layer embedding. The affine maps $\mathrm{Proj}_c : \mathbb{R}^{d_c} \to \mathbb{R}^{d_r}$ and $\mathrm{Proj}_t : \mathbb{R}^{d_t} \to \mathbb{R}^{d_r}$ project the two inputs into a common space. Layer normalization is denoted by LN, and SiLU is the activation function. The matrix $\mathbf{W}_g \in \mathbb{R}^{E\times d_r}$ and bias $\mathbf{b}_g \in \mathbb{R}^E$ produce scores $\mathbf{g}_{\ell,t} \in \mathbb{R}^E$ for E experts. Intent conditioning allows the same degraded image to use different expert combinations for joint and selective restoration. Layer and time conditioning allow the combination to vary across network depth and the denoising trajectory. The router is evaluated at each block and denoising time; it does not prescribe a fixed degradation identity for any expert.

Sparse selection and mixture normalization are formulated as:

$$\mathcal{E}_{\ell,t} = \mathrm{TopK}(\mathbf{g}_{\ell,t}, K), \qquad \boldsymbol{\pi} = \mathrm{softmax}(\mathbf{g}_{\ell,t,\mathcal{E}_{\ell,t}}), \qquad \boldsymbol{\alpha} = \frac{\boldsymbol{\pi}}{\|\boldsymbol{\pi}\|_2}, \tag{12}$$

where $K = 4$ experts are selected from $E = 16$, TopK returns their indices, and $\boldsymbol{\pi}, \boldsymbol{\alpha} \in \mathbb{R}^K$ denote selected probabilities and normalized mixture weights. The component of $\boldsymbol{\alpha}$ associated with expert e supplies $\alpha_{\ell,t,e}$ in Equation (4). The shared branch bypasses this selection. CODE is used for degradation comparison and data construction; it is not an input to the evaluated model's router.

C.4 TRAINING OBJECTIVES

The training objective is formulated as:

$$\mathcal{L}_{\text{rest}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{i}}\mathcal{L}_{\text{intent}} + \lambda_{\text{b}}\mathcal{L}_{\text{bal}}, \tag{13}$$

where $\mathcal{L}_{\text{FM}}$ is the flow-matching loss for the requested target $\mathbf{y}^*$, $\mathcal{L}_{\text{intent}}$ predicts the annotated restoration goals, $\mathcal{L}_{\text{bal}}$ regularizes expert probabilities, and $\lambda_{\text{i}}, \lambda_{\text{b}}$ are nonnegative weights.

Restoration objective. For a training triple $(\mathbf{x}, c, \mathbf{y}^*)$, the frozen image encoder maps the requested target to a latent vector. The noised latent and flow-matching objective are formulated as:

$$\begin{aligned} \mathbf{z}_* &= \mathcal{V}(\mathbf{y}^*), \qquad \mathbf{z}_t = (1-\sigma_t)\mathbf{z}_* + \sigma_t\boldsymbol{\epsilon}, \\ \mathcal{L}_{\text{FM}} &= \mathbb{E}_{(\mathbf{x},c,\mathbf{y}^*),t,\boldsymbol{\epsilon}}\left[\frac{w(t)}{D_z}\left\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z}_t, t; \mathbf{x}, c) - (\boldsymbol{\epsilon} - \mathbf{z}_*)\right\|_2^2\right], \end{aligned} \tag{14}$$

where $\mathcal{V}$ is the pretrained latent encoder, $\mathbf{z}_*, \mathbf{z}_t, \boldsymbol{\epsilon} \in \mathbb{R}^{D_z}$ are vectorized target latents, noised latents, and standard Gaussian noise, and D_z is the latent dimension. The sampled denoising time t

determines the scheduler's noise level $\sigma_t \in [0, 1]$ and training weight $w(t)$. The velocity predictor $\mathbf{f}_{\boldsymbol{\theta}}$ conditions on the degraded input $\mathbf{x}$ and instruction c; $\boldsymbol{\theta}$ comprises the trainable adapter and router parameters while the pretrained backbone remains frozen. Thus, changing c and $\mathbf{y}^*$ changes the supervised transformation even for a fixed $\mathbf{x}$. The expectation averages training triples, scheduler time samples, and noise draws.

Intent annotations. Let $\mathcal{R}$ and $\mathcal{K}$ be the requested removal and retention sets. Each is encoded by seven binary entries for noise, super-resolution, flare, rain, haze, blur, and low light; rain streaks and raindrops share the rain entry. Three one-hot entries encode joint, selective, and single-task restoration, followed by the normalized removal count $\min(|\mathcal{R}|, 3)/3$. The concatenated intent target is $\mathbf{s} \in [0, 1]^{18}$. These labels come from sample construction and are used only for auxiliary training; no expert identity is annotated, and the router receives the backbone condition states at inference.

Auxiliary losses. For one training example at denoising time t, the full expert probabilities and layer average are formulated as:

$$\mathbf{p}_{\ell,t} = \mathrm{softmax}(\mathbf{g}_{\ell,t}), \qquad \bar{\mathbf{p}}_t = \frac{1}{L} \sum_{\ell=1}^{L} \mathbf{p}_{\ell,t}, \tag{15}$$

where $\mathbf{p}_{\ell,t}, \bar{\mathbf{p}}_t \in \mathbb{R}^E$ are probability vectors over the E experts, L is the number of blocks, and $\mathbf{g}_{\ell,t}$ is defined in Equation (11). A prediction head maps the layer average to intent logits. The auxiliary objectives are formulated as:

$$\mathcal{L}_{\mathrm{intent}} = \mathrm{BCEWithLogits}(\mathcal{H}(\bar{\mathbf{p}}_t), \mathbf{s}), \qquad \mathcal{L}_{\mathrm{bal}} = E \sum_{e=1}^{E} \bar{p}_{t,e}^{\,2} - 1, \tag{16}$$

where $\mathcal{H} : \mathbb{R}^E \to \mathbb{R}^{18}$ is a two-layer prediction head with a 64-dimensional SiLU hidden layer, $\mathbf{s}$ is the intent target, and $\bar{p}_{t,e}$ is component e of $\bar{\mathbf{p}}_t$. Binary cross-entropy is averaged over all target entries, including the fractional removal count. Both auxiliary coefficients are 0.01. The balancing term is computed per image across layers and averaged over training examples; it does not impose a uniform discrete selection frequency at every layer. Auxiliary losses use the full expert probabilities, whereas image restoration uses the selected and normalized weights in Equation (12). Gradients reach all expert scores through the full softmax. Shared and selected expert transformations are optimized by the restoration objective, and auxiliary gradients can also reach preceding adapters through the condition states. The intent prediction head is not used at inference.

C.5 Restoration optimization

For the targeted-adapter comparison, both models use eight GPUs, one image per GPU, BF16 model computation, AdamW with base learning rate 10^{-4}, weight decay 0.01, and seed 20260720. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. The first optimizer update uses one third of the base learning rate, followed by the base rate. For MoLE, weight decay applies only to experts selected by the distributed batch; inactive expert gradients and optimizer moments are zeroed. Shared parameters and the router use ordinary AdamW updates.

D Real-world evaluation data and DRA

D.1 WildRestore-Bench: real-image collection and task composition

WildRestore-Bench is a real-world restoration benchmark combining application photographs and web-collected images. Web images are selected by visual inspection for plausible degradation patterns and identifiable restoration needs. Web-image inclusion is based on visual plausibility; original capture histories are not verified for every image. It covers noise, blur, haze, flare, rain streaks, low light, low resolution, and their combinations without requiring clean references. The evaluated set contains 347 image–instruction requests: 278 Single requests across seven task categories and 69 Composite requests for joint removal. These are request counts; one photograph may support more than one task. We summarize the task composition in Table 13.

Table 13: Task composition of WildRestore-Bench. Images combine application photographs and web sources, without clean references. Counts are image–instruction pairs. Single names one restoration target; Composite requests removal of all listed degradations.

Single: seven task categories, 278 requests			
Degradation	Count	Degradation	Count
Noise	27	Rain streaks	27
Blur	82	Haze	27
Low light	47	Low resolution	27
Flare	41		
Composite: fourteen combinations, 69 requests			
Degradation combination	Count	Degradation combination	Count
Low light + noise	5	Noise + low resolution	2
Low light + blur	8	Blur + rain streaks	1
Low light + flare	7	Flare + rain streaks	2
Low light + rain streaks	10	Flare + haze	1
Low light + haze	11	Flare + low resolution	2
Low light + low resolution	1	Rain streaks + haze	11
Noise + blur	2	Low light + flare + rain streaks	6

Single draws 49 and 141 requests from two curated collections and 88 from mobile captures. Composite contains fourteen observed degradation combinations, and its reported score averages all 69 requests. restoration supervision uses source-based pairs. Public evaluation excluding shared-reference inputs is reported in Table 21.

D.2 Restoration tasks and assessment criteria

We summarize the model's eight single-degradation tasks and the visual evidence used by DRA in Table 14. WildRestore-Bench covers seven of these tasks, excluding raindrops. The public main comparison includes raindrops and Composite as a separate group, while super-resolution is evaluated on WildRestore-Bench. Thus, the two suites do not use identical task groupings.

Table 14: Restoration scope and task-specific DRA criteria. Observable task evidence supports both residual-degradation and input-relative judgments.

Task	Restoration goal	Evidence assessed by DRA
Noise	Reduce random luminance and color noise.	Residual luminance grain or colored speckles in flat and dark regions.
Blur	Recover edges and detail lost to motion or defocus.	Soft subject edges, motion smears, unreadable text, or lost texture. Intentional background bokeh is excluded.
Haze	Reduce the scattering veil and recover scene visibility.	Veiling, faded colors, low contrast, and obscured distant objects.
Rain streaks	Remove visible rain traces from the scene.	Streaks, splashes, and surface rain traces, excluding wet reflections alone.
Raindrops	Remove adherent lens or glass droplets.	Droplet occlusion and refraction on the lens or glass, distinct from falling streaks.
Flare	Reduce abnormal light streaks and ghosts.	Long light beams, star-shaped streaks, or colored ghosts. Compact natural halos are acceptable.
Low light	Restore underexposed regions while retaining natural illumination.	Scene-level underexposure. Nighttime darkness alone is not sufficient evidence.
Low resolution	Recover spatial detail through super-resolution.	Recovered structures in aligned input–output views; residual blur, ringing, and invented detail.

For each requested task, DRA assesses the residual evidence in the output and its reduction relative to the input. Joint requests consider all listed degradations. Task fulfillment is evaluated by DRA, with content fidelity additionally assessed by paired metrics and human judgments.

D.3 Task-conditioned scoring protocol

DRA asks two complementary questions: how much target degradation remains in the restored image, and how much the output improves over its degraded input for the requested task. An image can have little residual degradation yet show little improvement if its input was already mildly degraded; substantial improvement can also leave visible residual degradation. The two judgments characterize these different aspects of restoration without a clean reference.

For degradation-removal tasks, a frozen Qwen3-VL-8B-Instruct evaluator (Bai et al., 2025) first receives the output and task definition, returning a five-level residual-degradation grade and explanation. It then receives the input, output, and first assessment to judge negligible, partial, or global improvement. A task-dependent calibration contextualizes image changes and combines the judgments on a common five-point scale, with higher scores indicating stronger task fulfillment. Calibration is fixed after development; the evaluator is not fine-tuned on the human ratings.

Implementation. The output-only and input-relative prompts have generation budgets of 256 and 512 tokens at temperature 0.1, with up to three attempts for parsing or grade–explanation failures. Joint requests list all degradations in one rubric. A super-resolution-only request uses an NIQE-based output assessment (Mittal et al., 2013) and a 768-token relative comparison of aligned, contrast-normalized views, focusing on structural recovery, ringing, and invented detail. Composite requests containing super-resolution retain the joint vision-language rubric. Score aggregation for the reported comparisons is specified in Appendix E.2.

D.4 Human rating protocol

The human-rated development set contains seven restoration tasks, seven real inputs per task, and three anonymously named outputs per input. Fourteen ratings are available for each of the 147 outputs. The primary panel retains seven raters selected for higher agreement in the collected ratings.

Raters judged overall restoration quality, including degradation removal, natural appearance, and content fidelity. The rubric assigns: 5 for a clear qualitative improvement with little visible residual degradation and natural appearance; 4 for substantial correct restoration with minor residual or quality defects; 3 for substantial restoration with visible artifacts but preserved scene semantics; 2 for slight but acceptable improvement; 1 for negligible improvement with conspicuous residual problems; and 0 for severe structural or semantic distortion, poor fidelity, or worsening degradation. Thus, zero is a failure judgment rather than missing data.

For each output, consensus is zero when at least half of the panel assigns a failure grade of zero; otherwise it is the mean of the positive ratings, rounded to three decimals. This gives seven failure-consensus outputs for the selected panel and ten for all fourteen raters. We compare both panels in Table 15, retaining failures in the primary analysis and excluding them only as a sensitivity check.

Table 15: DRA agreement under different rater panels and failure handling. The seven-rater panel is the primary analysis; exclusion of failure labels is only a sensitivity check.

Panel	Failure labels	Outputs	PLCC	SRCC	KRCC
Selected seven	Included	147	0.504	0.537	0.403
Selected seven	Excluded	140	0.566	0.581	0.429
All fourteen	Included	147	0.417	0.474	0.352
All fourteen	Excluded	137	0.550	0.565	0.411

D.5 Human agreement and sensitivity analyses

Agreement statistics. We compute PLCC directly on evaluator scores and human consensus, without logistic fitting. SRCC uses average ranks for ties, and KRCC uses Kendall's tie-corrected

τ_b. NIQE and BRISQUE are sign-reversed so that higher scores always indicate better quality. All primary analyses retain zero-consensus failures; their exclusion is confined to the sensitivity comparison in Table 15.

Input-relative generic quality metrics. We test whether input-relative changes also help generic quality metrics in Table 16. These differences do not consistently improve agreement over output-only scores.

Table 16: Alternative no-reference comparisons. Output scores and signed output-minus-input differences use the same selected-panel consensus.

	Output score		Score difference	
Metric	PLCC	SRCC	PLCC	SRCC
MUSIQ (Ke et al., 2021)	0.254	0.326	0.247	0.231
MANIQA (Yang et al., 2022)	0.036	0.175	-0.013	0.065
NIQE (Mittal et al., 2013)	0.105	0.180	0.011	0.029
BRISQUE (Mittal et al., 2012)	0.179	0.145	0.039	0.044
CLIPIQA (Wang et al., 2023)	0.066	0.143	-0.002	-0.042

Sampling uncertainty and task-level agreement. We estimate conditional uncertainty using 2,000 bootstrap samples with seed 20260922. Within each task, seven source inputs are sampled with replacement, retaining all three outputs from each sampled input. This preserves within-input dependence. For DRA, the selected-panel 95% intervals are [0.339, 0.653] for PLCC and [0.388, 0.669] for SRCC. These intervals condition on the chosen panel and fixed evaluator scores. VLM scoring uses one sampled realization, so they do not estimate evaluator repeatability.

We compare task-specific agreement across methods in Table 17. DRA has the highest SRCC on six tasks; NIQE is slightly higher on super-resolution. Each task contributes 21 rated outputs to this comparison.

Table 17: SRCC by task for the selected panel, including all failures. Each task has 21 outputs from seven inputs. NIQE and BRISQUE are sign-reversed. The last two rows give DRA's additional correlations.

Score / statistic	Noise	Blur	Rain	Haze	Flare	Low light	SR
MUSIQ (Ke et al., 2021)	**0.348**	0.309	**0.210**	**0.187**	**0.166**	**0.614**	0.512
MANIQA (Yang et al., 2022)	0.304	0.582	0.089	-0.068	0.023	0.334	0.412
NIQE (Mittal et al., 2013)	-0.282	0.580	-0.027	0.130	0.110	0.359	**0.563**
BRISQUE (Mittal et al., 2012)	-0.258	0.507	0.066	-0.182	0.027	0.505	0.338
CLIPIQA (Wang et al., 2023)	-0.310	**0.691**	0.085	-0.055	-0.025	0.353	0.380
DRA	**0.476**	**0.703**	**0.606**	**0.842**	**0.564**	**0.783**	**0.543**
DRA PLCC	0.507	0.650	0.499	0.820	0.489	0.802	0.479
DRA KRCC	0.362	0.543	0.481	0.706	0.502	0.610	0.424

A complementary within-input comparison evaluates the three output pairs for each input, excludes human ties, and gives half credit to evaluator ties. Section 4.3 reports aggregate agreement.

E Restoration evaluation and extended results

E.1 Evaluation sets

Restoration evaluation. The public suite contains 43 paired subsets and 10,012 images. The main comparison covers 36 subsets (9,662 images) across noise, blur, rain streaks, raindrops, haze, flare, low light, and composites. Seven JPEG subsets (350 images) are evaluated separately. Scores are averaged within each benchmark, then equally across benchmarks within a task, and equally across eight tasks. RGB PSNR, LPIPS, and DISTS measure fidelity to clean ground truth; DRA assesses the requested removal without clean ground truth.

Real-image evaluation uses WildRestore-Bench without clean references. Single and Composite specify one-task and joint-removal requests, as detailed in Appendix D.1.

Benchmark sources. Denoising is evaluated on grayscale and color BSD68 (Martin et al., 2001), together with Urban100 (Huang et al., 2015). Deblurring uses the GoPro and RealBlur-J datasets (Nah et al., 2017; Rim et al., 2020). Indoor and outdoor haze removal use SOTS from RESIDE (Li et al., 2019). Rain removal covers Rain100H/L (Yang et al., 2017), Test100 (Zhang et al., 2020), Test1200 (Zhang & Patel, 2018), and Test2800 (Fu et al., 2017). We evaluate flare removal on Flare7K (Dai et al., 2022) and low-light enhancement on LOL and LOL-v2 (Wei et al., 2018; Yang et al., 2020). FoundIR supplies additional single and composite benchmarks (Li et al., 2025). The complete list appears in Appendix E.3.

E.2 Model configurations and inference

We specify the data, adaptation, and sampling steps for each comparison in Table 18. UniLens adapts the original Qwen-Image-Edit backbone; the Qwen and GGT baselines use Qwen-Image-Edit-2511. UniLens inference uses CFG 1 and seed 1. UniLens uses the structured text instructions defined in Appendix C.1, with the same format during training and evaluation. Each baseline retains its reported processing and sampling settings, so the system comparison includes these differences in computational cost.

Table 18: Restoration configurations used in the reported comparisons. Sampling steps refer to inference.

Evaluation	Data and adaptation	Steps
Public system comparison	Targeted pairs, MoLE	20
Real-image system comparison	Targeted pairs, MoLE	20
Sampling-step comparison	Targeted pairs, MoLE	5 / 20
Component comparisons on WildRestore-Bench and FoundIR-Test		
Data progression	Paired baseline, LoRA	5
Data / supervision	Offline pairs, LoRA	5
Data / supervision	Full online targets, LoRA	5
Supervision	Single online targets, LoRA	5
Supervision	Without joint targets, LoRA	5
Supervision	Without selective targets, LoRA	5
Data / adaptation	Targeted pairs, LoRA	5
Adaptation	Targeted pairs, MoLE	5
Supplementary reference	Broad expansion, MoLE	5

Score aggregation. DRA produces a per-request score on the five-point scale described in Appendix D. Single averages task means within three collections of 49, 141, and 88 requests, then combines collection means with weights 0.25, 0.50, and 0.25. Composite averages 69 joint-removal requests. Public scores average images within benchmarks, benchmarks within tasks, and then the eight task means.

Single-GPU inference latency. We compare complete inference calls under each method's evaluation settings in Table 19.

Table 19: Single-A800 restoration latency (seconds per image), averaged over 21 timed calls per input size. Model loading, disk I/O, and scoring are excluded. UniLens (main) uses the twenty-step evaluation configuration; the final two rows report five-step adapter timing configurations. Processing geometry and peak allocation are reported for a 2048^2 input.

		Input size				2048^2 input	
Model	Steps	512^2	1024^2	1536^2	2048^2	Geometry	GiB
UniLens (main)	20	27.26	27.24	58.00	58.04	1440^2	61.45
FoundIR (Li et al., 2025)	4	0.29	1.16	4.60	10.33	tiled	9.61
FoundIR-v2 (Chen et al., 2026)	20	16.85	19.03	30.19	45.73	tiled	29.46
Qwen-Image-Edit-2511 (Wu et al., 2025)	40	81.29	81.05	81.29	81.46	1024^2	57.94
LongCat (Meituan LongCat Team et al., 2025)	50	56.50	56.44	56.55	56.59	1024^2	29.74
FLUX.1-Kontext-dev (Black Forest Labs et al., 2025)	28	29.20	29.20	29.26	29.31	1024^2	33.83
Step1X-Edit (Liu et al., 2025)	28	59.35	59.32	59.38	59.40	1024^2	41.20
GGT-Qwen (Kong et al., 2026)	40	85.78	85.76	85.77	85.86	1024^2	57.96
MoLE (five-step reference)	5	2.68	7.20	15.17	15.21	1440^2	61.45
Dense LoRA (Hu et al., 2022) (reference)	5	2.29	6.54	14.21	14.25	1440^2	57.91

On one A800 80GB GPU, seven fixed images cover noise, blur, rain, haze, flare, low light, and blur with noise. Each size uses two warm-ups and three timed runs per image. CUDA-synchronized, in-memory input-to-output timing includes resizing, conditioning, sampling, decoding, and FoundIR-v2 caption generation; model loading, disk I/O, and DRA are excluded.

The main twenty-step UniLens model takes 27.24 s for 1024^2 inputs, versus 81.05 s for Qwen-Image-Edit-2511 and 85.76 s for GGT. FoundIR and FoundIR-v2 are faster at this size. The five-step rows report adapter timing configurations with the sampling and processing settings in Table 19.

UniLens processes 512^2/1024^2 inputs at 1024^2 and 1536^2/2048^2 inputs at 1440^2, then returns outputs to the requested size. Table 19 records processing geometry and peak memory. FoundIR and FoundIR-v2 retain the full input with 1024^2 image and latent tiles, respectively. FoundIR timing uses its four-step generalist, whereas quality evaluation uses its expert combination. Guidance is 1 for UniLens, 5 for FoundIR-v2, 4 for Qwen/GGT, 4.5 for LongCat, 2.5 for FLUX, and 6 for Step1X. FoundIR uses FP32, FoundIR-v2 FP16, and the others BF16. The timings reflect each method's inference configuration.

E.3 Additional quantitative comparisons

Degradation-level results. We compare task-level fidelity and DRA in Table 20, using the same eight-task aggregation as the main public comparison.

Sampling-step comparison. We evaluate the five-step alternative under the same resizing policy and checkpoint as the twenty-step main configuration in Table 21.

Adapter configurations with the broad-expansion mixture. We also compare five-step restoration with the broad-expansion mixture in Table 22. Dense LoRA and MoLE use 28,000 and 10,000 training updates, respectively, so these are additional system configurations.

Per-benchmark fidelity. We report every paired benchmark in Table 23, including the seven separately evaluated JPEG subsets.

Table 23: Per-benchmark PSNR↑ / LPIPS↓ for FoundIR (Li et al., 2025), FoundIR-v2 (Chen et al., 2026), GGT (Kong et al., 2026), UniLens, and LongCat-Image-Edit (Meituan LongCat Team et al., 2025). Colors rank displayed values within each benchmark.

Benchmark	FoundIR	FoundIR-v2	GGT	UniLens	LongCat
FoundIR: noise	**39.82** / **0.065**	31.66 / **0.071**	28.76 / 0.171	**34.13** / 0.072	19.52 / 0.282
BSD68 ($\sigma = 15$)	26.36 / 0.342	23.20 / 0.270	**28.33** / **0.098**	**26.79** / **0.105**	18.10 / 0.369
BSD68 ($\sigma = 25$)	22.10 / 0.550	22.42 / 0.324	**25.76** / **0.160**	**24.90** / **0.152**	17.43 / 0.526

Benchmark	FoundIR	FoundIR-v2	GGT	UniLens	LongCat
BSD68 ($\sigma = 50$)	16.10 / 0.878	19.80 / 0.467	**21.65** / **0.289**	**21.79** / **0.256**	14.63 / 0.775
CBSD68 ($\sigma = 15$)	**29.86** / 0.166	22.93 / 0.243	28.75 / **0.079**	**28.99** / **0.082**	18.92 / 0.210
CBSD68 ($\sigma = 25$)	24.48 / 0.385	22.58 / 0.310	**26.48** / **0.112**	**26.86** / **0.117**	18.50 / 0.282
CBSD68 ($\sigma = 50$)	17.56 / 0.836	20.93 / 0.453	**23.08** / **0.196**	**23.68** / **0.193**	17.05 / 0.469
Urban100 ($\sigma = 15$)	**29.32** / 0.137	22.76 / 0.151	**28.00** / **0.069**	26.53 / **0.082**	17.08 / 0.226
Urban100 ($\sigma = 25$)	24.34 / 0.300	22.34 / 0.200	**26.96** / **0.080**	**25.45** / **0.098**	16.74 / 0.272
Urban100 ($\sigma = 50$)	17.72 / 0.678	20.67 / 0.261	**24.68** / **0.122**	**23.44** / **0.134**	16.75 / 0.309
FoundIR: blur	**25.52** / **0.181**	**23.96** / **0.163**	21.22 / 0.220	19.93 / 0.194	18.23 / 0.538
GoPro	**25.87** / 0.270	23.73 / 0.203	**27.34** / **0.111**	24.60 / **0.140**	17.91 / 0.464
RealBlur-J	**26.50** / 0.216	21.37 / 0.206	**27.78** / **0.088**	24.22 / **0.139**	21.35 / 0.350
FoundIR: rain	**34.29** / **0.080**	24.62 / **0.202**	24.99 / 0.305	**25.77** / 0.211	16.82 / 0.451
Rain100H	12.58 / 0.602	18.87 / 0.283	**22.95** / **0.160**	**24.91** / **0.144**	18.52 / 0.273
Rain100L	**29.61** / 0.122	23.10 / 0.135	29.34 / **0.051**	**32.45** / **0.052**	23.31 / 0.103
Test100	**22.42** / 0.241	21.03 / 0.196	21.86 / **0.132**	**22.35** / **0.120**	18.01 / 0.212
Test1200	24.46 / 0.200	22.98 / 0.151	**25.90** / **0.122**	**26.22** / **0.121**	20.36 / 0.177
Test2800	**26.24** / 0.134	21.76 / 0.156	25.42 / **0.107**	**27.36** / **0.097**	18.98 / 0.171
FoundIR: raindrop	**29.72** / **0.102**	22.16 / 0.251	24.58 / 0.225	**26.00** / **0.150**	17.71 / 0.384
FoundIR: haze	19.07 / **0.225**	**19.43** / **0.221**	18.69 / 0.427	**21.78** / 0.288	12.95 / 0.556
SOTS indoor	12.50 / **0.211**	16.03 / 0.216	**17.68** / 0.220	**26.09** / **0.058**	11.51 / 0.310
SOTS outdoor	19.38 / 0.116	18.70 / 0.129	**24.53** / **0.063**	**25.50** / **0.064**	14.81 / 0.197
Flare7K real	22.56 / **0.114**	18.37 / 0.206	**23.27** / 0.130	**25.39** / **0.099**	21.66 / 0.134
Flare7K synthetic	22.72 / 0.093	19.35 / 0.167	**26.03** / **0.067**	**28.26** / **0.048**	21.55 / 0.110
FoundIR: low light	**18.24** / 0.239	**21.00** / **0.170**	14.70 / 0.369	17.62 / **0.228**	14.49 / 0.434
LOL-v1	12.83 / 0.276	16.17 / 0.235	**21.00** / **0.107**	**18.80** / **0.136**	14.15 / 0.269
LOL-v2 real	14.99 / 0.258	19.02 / 0.204	**19.31** / **0.116**	**20.62** / **0.147**	15.11 / 0.290
LOL-v2 synthetic	11.16 / 0.392	**13.57** / 0.315	12.32 / **0.282**	**18.23** / **0.161**	10.89 / 0.414
FoundIR: blur + noise	**23.49** / 0.247	**24.05** / **0.167**	20.43 / 0.227	20.89 / **0.188**	19.34 / 0.465
FoundIR: low light + haze	**23.00** / **0.240**	15.92 / **0.339**	15.18 / 0.495	**18.99** / 0.373	14.04 / 0.609
FoundIR: night rain	**33.97** / **0.215**	22.38 / 0.309	**23.61** / **0.245**	18.34 / 0.355	19.18 / 0.315
FoundIR: rain + haze	**17.73** / **0.286**	16.00 / 0.431	16.41 / 0.559	**19.82** / **0.371**	13.02 / 0.667
FoundIR: low light + blur	**23.87** / **0.228**	12.81 / 0.333	11.77 / 0.392	**19.85** / **0.239**	17.79 / 0.459
FoundIR: low light + noise	13.96 / 0.462	**21.03** / **0.253**	9.20 / 0.597	**15.65** / **0.393**	14.26 / 0.555
FoundIR: low light/blur/noise	**23.56** / **0.226**	12.83 / 0.283	10.73 / 0.375	**18.52** / **0.254**	17.00 / 0.462
FoundIR: blur + JPEG	**29.57** / **0.151**	**21.40** / 0.213	20.17 / 0.224	20.84 / **0.179**	16.70 / 0.463
FoundIR: blur + noise + JPEG	**27.92** / **0.160**	**22.27** / 0.198	20.04 / 0.212	20.46 / **0.163**	17.01 / 0.455
FoundIR: JPEG	**33.98** / **0.069**	24.66 / **0.116**	24.86 / 0.230	**28.68** / 0.149	16.09 / 0.367
FoundIR: noise + JPEG	**40.41** / **0.032**	27.77 / 0.075	25.90 / 0.111	**35.85** / **0.049**	16.22 / 0.322
FoundIR: low light + JPEG	**29.03** / **0.114**	16.47 / 0.230	18.63 / 0.294	**20.43** / **0.224**	13.63 / 0.432
FoundIR: low light/blur/JPEG	**22.56** / **0.210**	11.97 / 0.333	17.40 / 0.270	**17.99** / **0.245**	15.75 / 0.416
FoundIR: low light/noise/JPEG	**33.47** / **0.056**	16.82 / 0.137	23.63 / 0.147	**24.84** / **0.080**	15.86 / 0.293

Data progression. The dense LoRA progression adds offline pairs, online supervision, and targeted pairs. The paired baseline uses conventional restoration pairs without the GGT Low-Light, Rain, and Haze subsets. Online construction also supplies super-resolution supervision and composite targets. The targeted configuration adds 200,000 unique pairs, each sampled at twice the weight of an ordinary paired record. Table 18 lists the update budget for each configuration. The fixed-budget distribution study uses the equal-count mixtures specified in Appendix B.4.

Supervision comparison. All five dense variants are evaluated at 28,000 updates. Paired supervision uses the offline-expanded pool without online construction. Single-degradation online supervision excludes composite targets; the other variants remove joint targets, remove selective targets, or use the complete target set in Equation (3). The target-removal variants examine the contribution of each goal within the online supervision strategy.

F Additional visual comparisons

We compare all eight methods in Table 2 on WildRestore-Bench and paired academic benchmarks. Every example uses the same UniLens checkpoint and twenty-step inference configuration as the main results; baselines retain their reported settings. Methods follow the relative order of the main

Table 20: Fidelity and DRA by degradation. Each panel uses the same eight methods and task averages as Table 2. Colors rank methods within each task and metric.

Method	Noise	Blur	Rain	Raindrop	Haze	Flare	Low light	Composite
				PSNR↑				
FoundIR (Li et al., 2025)	24.7650	**25.9630**	24.9344	**29.7241**	16.9879	22.6414	14.3043	**22.7979**
FoundIR-v2 (Chen et al., 2026)	22.9289	23.0213	22.0607	22.1644	18.0505	18.8568	**17.4397**	17.8591
GGT (Kong et al., 2026)	**26.2462**	**25.4476**	**25.0761**	24.5809	**20.2980**	**24.6504**	16.8329	15.3327
Qwen-Image-Edit-2511 (Wu et al., 2025)	21.6914	22.1793	22.4799	18.8508	18.1220	24.5532	9.0404	16.0732
FLUX.1-Kontext-dev (Black Forest Labs et al., 2025)	12.2681	12.9906	13.9991	11.7046	11.5284	22.8176	10.0494	11.4421
Step1X-Edit-v1.1 (Liu et al., 2025)	20.2323	20.9054	19.6896	20.0062	11.0028	21.0341	9.4350	15.2693
LongCat-Image-Edit (Meituan LongCat Team et al., 2025)	17.4727	19.1606	19.3324	17.7100	13.0922	21.6060	13.6609	16.3741
UniLens	**26.2558**	22.9147	**26.5097**	**26.0023**	**24.4566**	**26.8268**	**18.8191**	**18.8650**
				LPIPS↓				
FoundIR	0.4337	0.2226	0.2297	**0.1021**	**0.1841**	0.1038	0.2911	**0.2720**
FoundIR-v2	0.2750	0.1910	0.1871	0.2507	0.1882	0.1865	0.2312	**0.3021**
GGT	**0.1377**	**0.1399**	**0.1461**	0.2248	0.2366	0.0986	**0.2183**	0.4129
Qwen-Image-Edit-2511	0.1888	0.2946	0.1759	0.2797	0.2532	**0.0810**	0.5515	0.4334
FLUX.1-Kontext-dev	0.5623	0.6000	0.5375	0.6646	0.5955	0.1070	0.5945	0.7102
Step1X-Edit-v1.1	0.4171	0.4449	0.3603	0.3877	0.5347	0.1308	0.5236	0.5591
LongCat-Image-Edit	0.3720	0.4507	0.2312	0.3843	0.3542	0.1219	0.3518	0.5047
UniLens	**0.1293**	**0.1576**	**0.1239**	**0.1503**	**0.1365**	**0.0734**	**0.1678**	0.3105
				DISTS↓				
FoundIR	0.2377	0.1346	0.1584	**0.0476**	0.1354	0.0854	0.2028	**0.1606**
FoundIR-v2	0.1811	0.1105	0.1218	0.1158	**0.1231**	0.1327	0.1487	0.1682
GGT	**0.1112**	**0.0632**	**0.1140**	0.0904	0.1394	0.0878	**0.1395**	0.2197
Qwen-Image-Edit-2511	0.1440	0.1572	0.1279	0.1120	0.1562	**0.0802**	0.3644	0.2363
FLUX.1-Kontext-dev	0.2361	0.2286	0.2230	0.2227	0.2126	0.0852	0.2877	0.2723
Step1X-Edit-v1.1	0.2472	0.2696	0.2091	0.1686	0.3513	0.0993	0.3527	0.3408
LongCat-Image-Edit	0.2219	0.2450	0.1352	0.1536	0.1994	0.0892	0.2227	0.2686
UniLens	**0.1070**	**0.0803**	**0.0947**	**0.0600**	**0.0804**	**0.0713**	**0.1182**	**0.1445**
				DRA↑				
FoundIR	2.9733	2.2642	2.8770	3.2950	2.6338	1.7988	2.9585	**3.5065**
FoundIR-v2	3.7789	3.1714	4.6082	3.6675	3.1216	2.4213	3.5950	3.3953
GGT	4.4334	**3.4646**	**4.8745**	**4.9725**	3.3136	4.0675	**3.8262**	3.1460
Qwen-Image-Edit-2511	**4.5107**	2.3134	4.7984	4.7350	3.0545	**4.4037**	1.7352	3.3969
FLUX.1-Kontext-dev	4.3246	2.7305	3.3988	3.4225	**3.4792**	2.1650	2.3181	3.2332
Step1X-Edit-v1.1	3.3297	1.6021	2.5150	2.1900	1.7079	2.0225	1.5725	2.4649
LongCat-Image-Edit	3.7677	1.6287	4.3020	3.6075	2.3704	**4.2287**	3.3131	3.1755
UniLens	**4.4371**	**3.6276**	**4.9805**	**4.8800**	**3.8294**	3.9737	**4.4588**	**4.4521**

Table 21: Five- and twenty-step inference with the same targeted-supplementation checkpoint, resizing policy, and scoring prompts. Public results use 36 benchmarks; real-image results use 278 Single and 69 Composite requests. The final row excludes 171 construction-reference inputs from the public evaluation (35 benchmarks, 9,491 images).

	Public benchmarks				WildRestore-Bench	
Setting	PSNR↑	LPIPS↓	DISTS↓	DRA↑	Single↑	Composite↑
Five steps	**24.5352**	**0.1497**	**0.0961**	**4.3389**	**3.9354**	**3.7717**
Twenty steps	**23.8312**	**0.1562**	**0.0945**	**4.3299**	**3.9813**	**3.7609**

Table 22: Five-step public results with broad expansion. Training updates are listed in Table 18.

Adapter	PSNR↑	LPIPS↓	DISTS↓	DRA↑
Dense LoRA (Hu et al., 2022)	**24.4955**	**0.1635**	**0.0973**	**4.0906**
MoLE	**24.8060**	**0.1453**	**0.0922**	**4.2871**

visual comparisons, with FLUX, Step1X, and LongCat added before UniLens. Images are read from left to right and top to bottom. Clean references accompany paired benchmarks only; matched detail views are identified in the captions.

F.1 REAL COMPOSITE RESTORATION

We compare low light and flare in Figure 10. Brighter foliage and reduced light streaks make the two restoration goals visible.

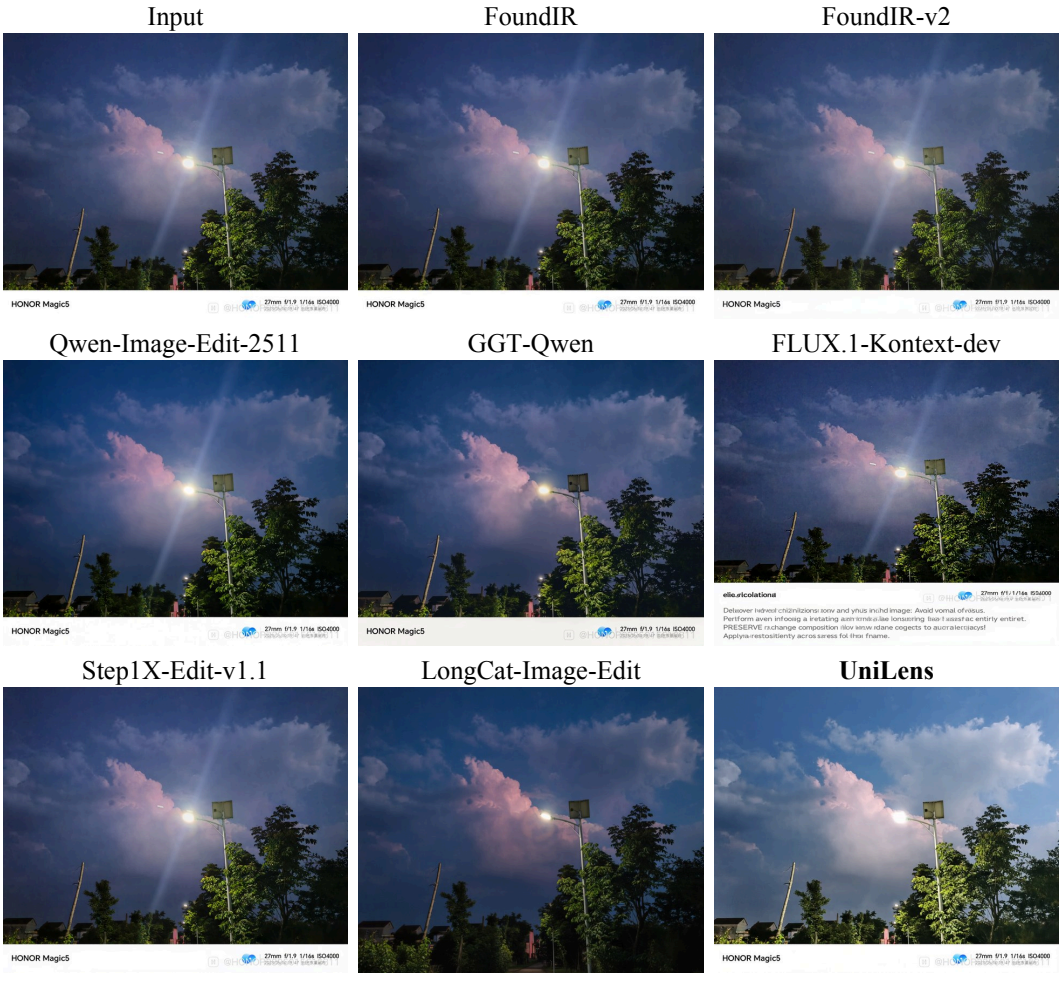

Figure 10: WildRestore-Bench Composite: low light and flare.

We compare low light and rain in Figure 11. The wall and shutter reveal differences in residual streaks and recovered exposure.

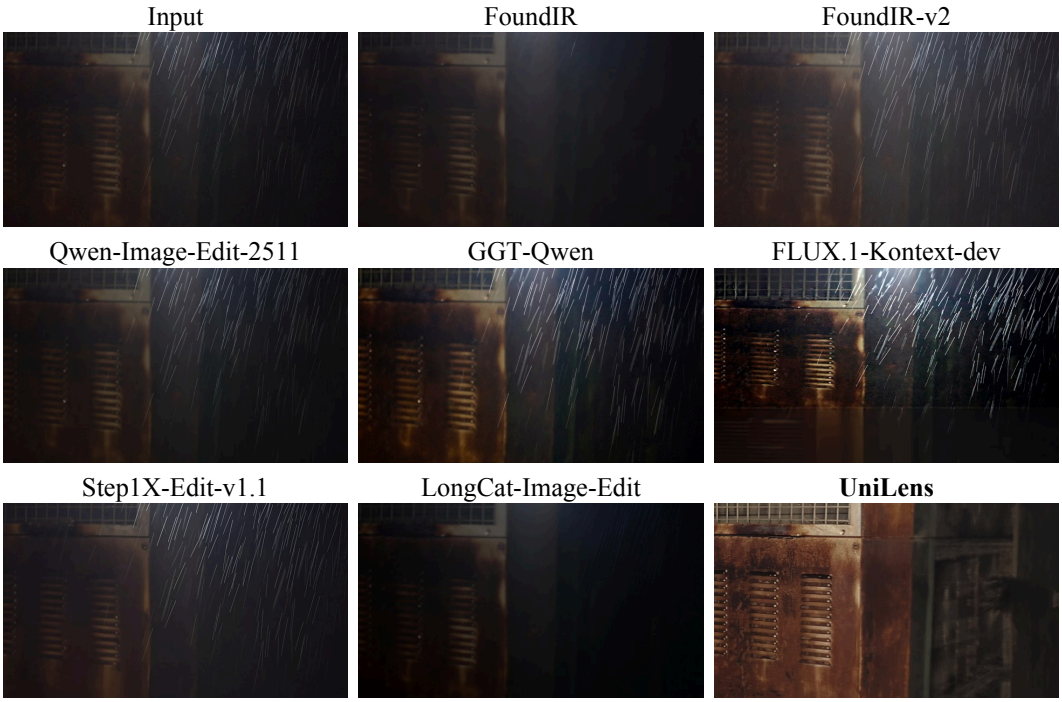

Figure 11: WildRestore-Bench Composite: low light and rain.

F.2 REAL SINGLE-TASK RESTORATION

We compare real motion blur in Figure 12. Matched close-ups of the facade expose residual blur and window-edge definition.

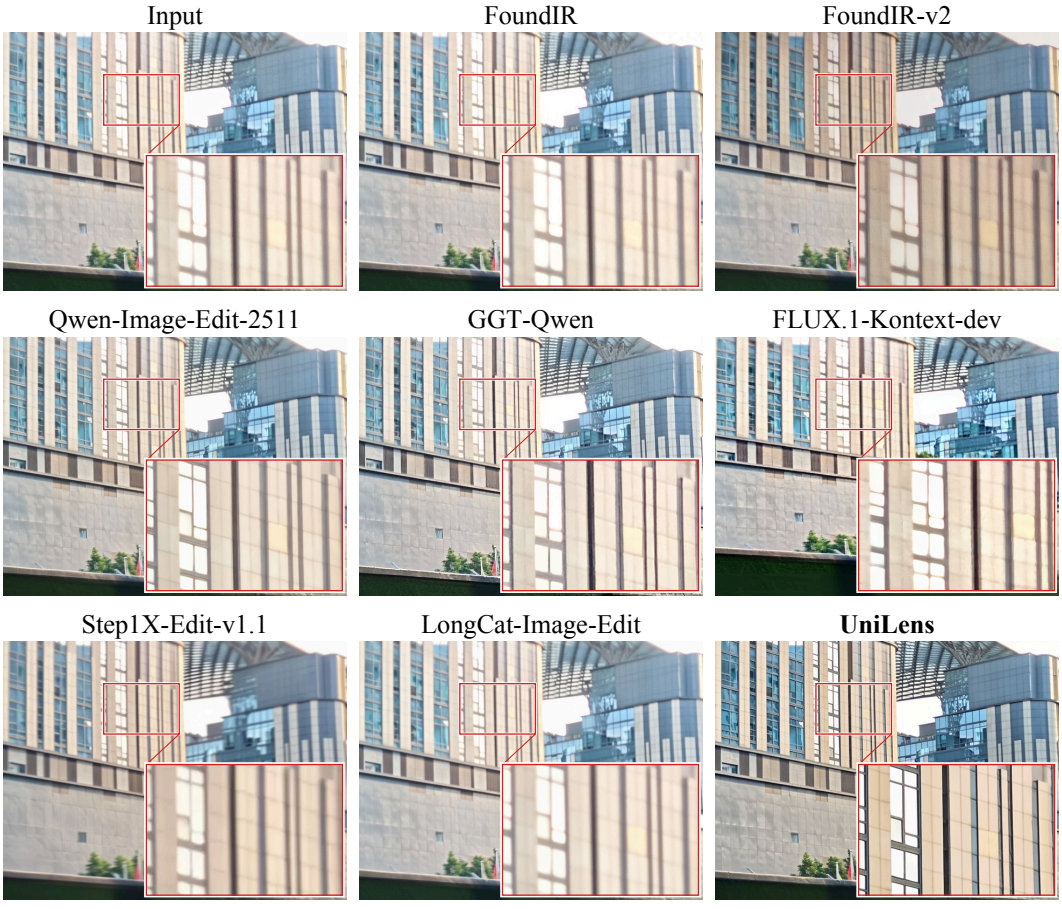

Figure 12: WildRestore-Bench Single: motion blur. Insets enlarge the same facade region across methods.

We compare real haze in Figure 13. Distant building outlines reveal differences in residual haze and local contrast.

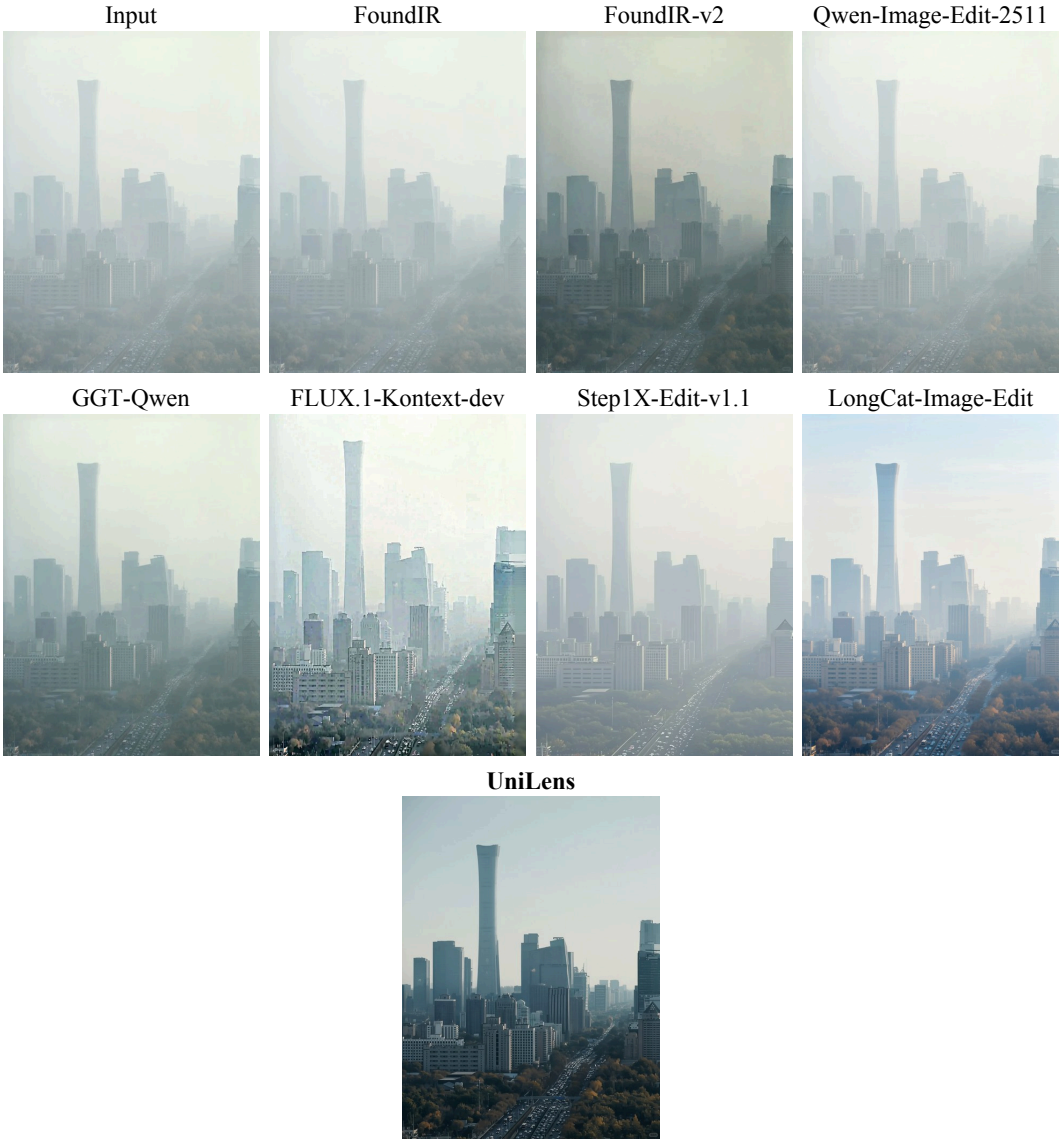

Figure 13: WildRestore-Bench Single: haze.

We compare real rain removal in Figure 14. Matched close-ups below the streetlight reveal residual rain streaks against the dark building.

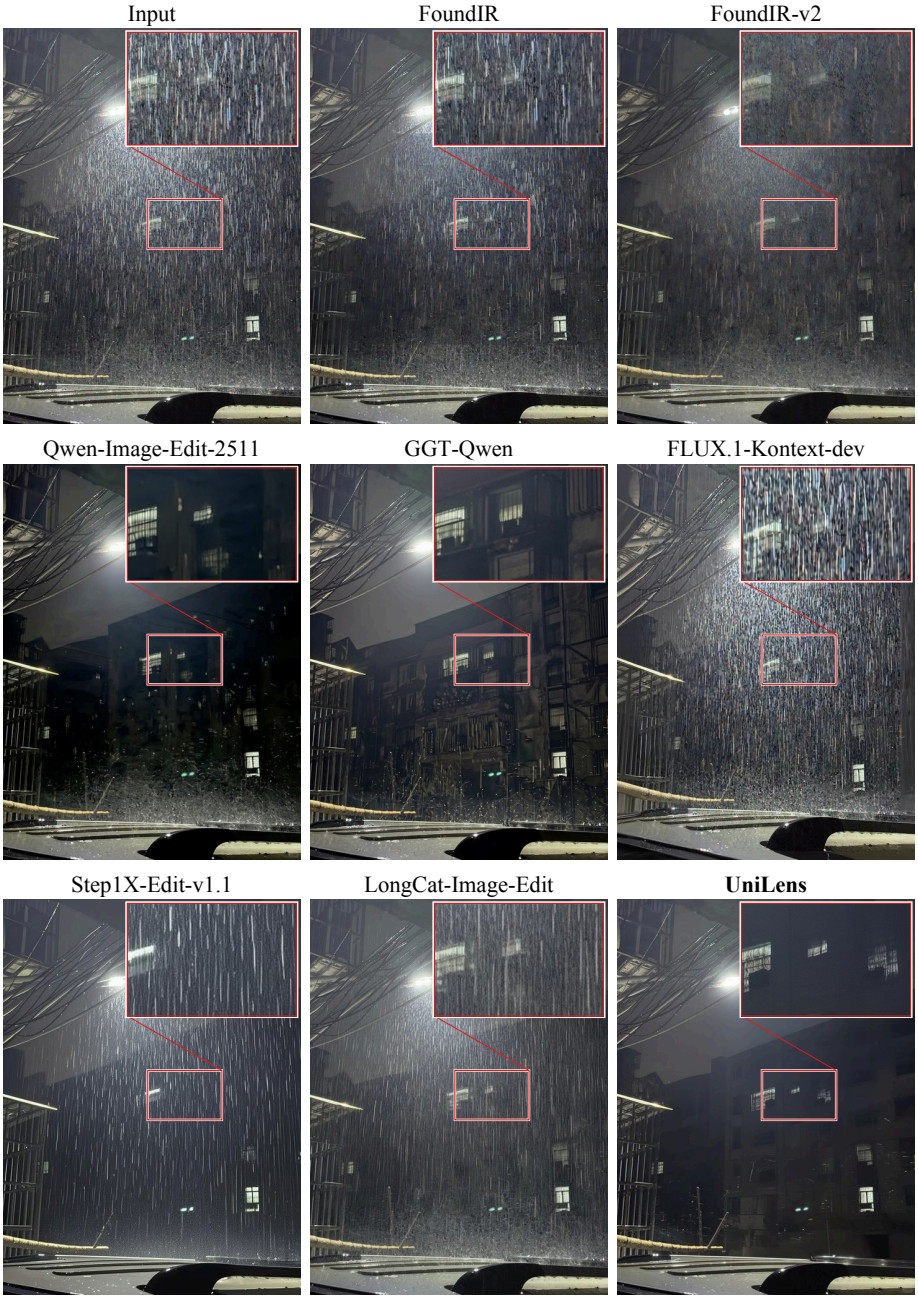

Figure 14: WildRestore-Bench Single: rain. Insets enlarge the same building region across methods.

We compare real flare in Figure 15. The lawn and tree canopy reveal differences in colored ghosts and light streaks.

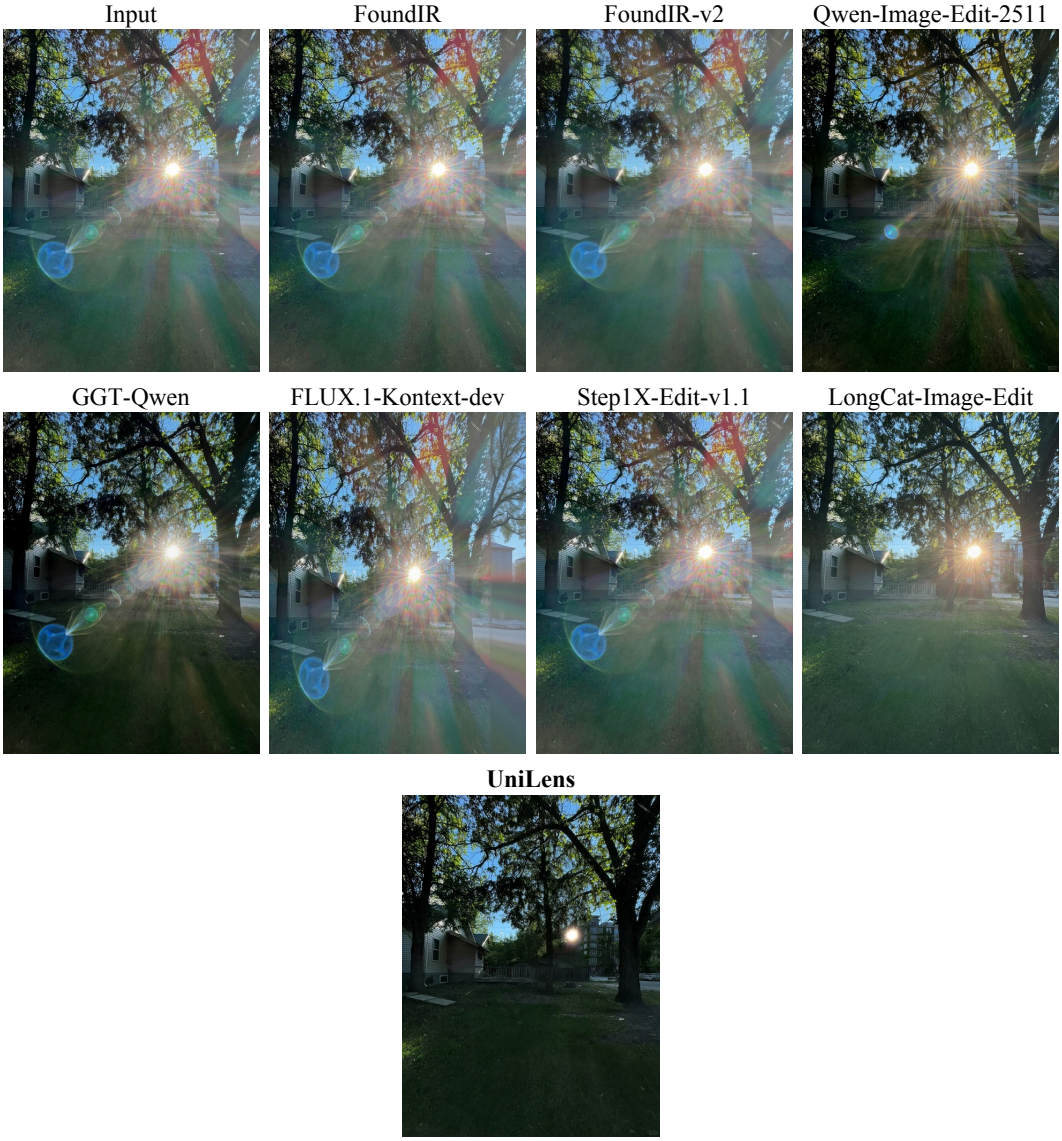

Figure 15: WildRestore-Bench Single: flare.

We compare real low light in Figure 16. The classroom exposes differences in recovered visibility, color, and contrast.



Figure 16: WildRestore-Bench Single: low light.

F.3 PAIRED ACADEMIC BENCHMARKS

We compare paired rain and haze in Figure 17. The roof and bridge reveal residual streaks, haze, and changes to scene geometry.

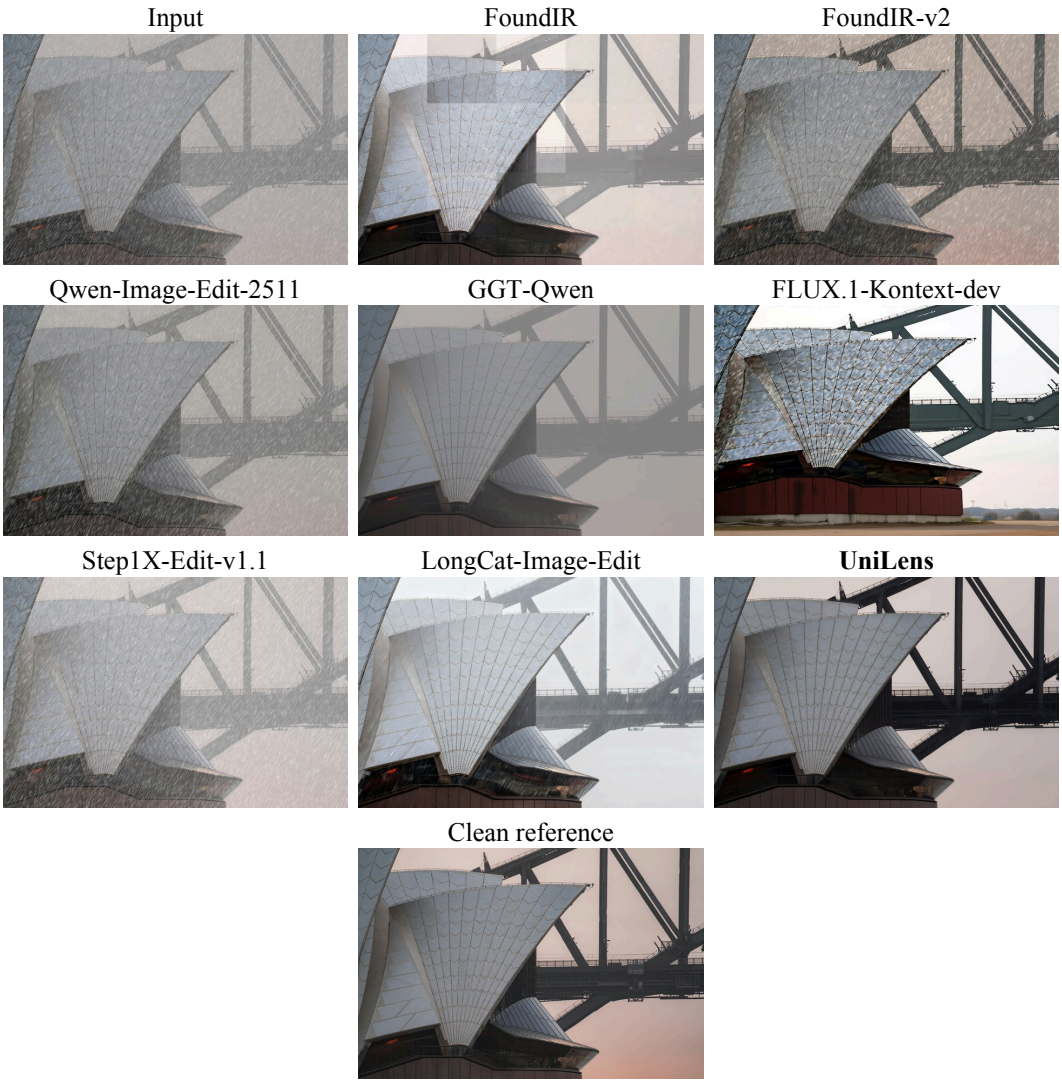

Figure 17: Rain and haze on FoundIR (Li et al., 2025).

We compare heavy rain removal on Rain100H in Figure 18. The bridge and river permit comparison of residual rain, scene geometry, and color against the clean reference.

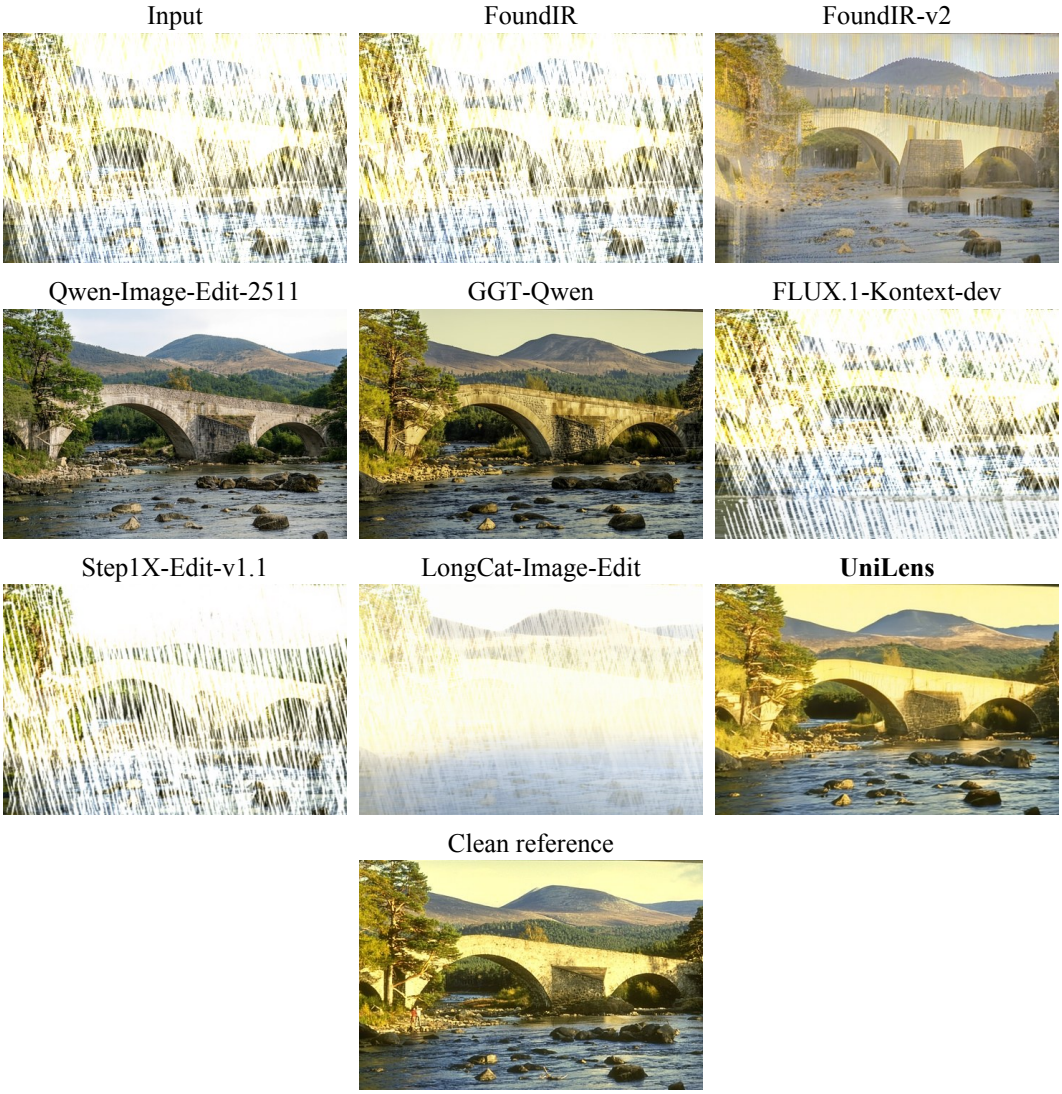

Figure 18: Rain removal on Rain100H (Yang et al., 2017).

We compare paired motion blur in Figure 19. The flowers expose differences in petal boundaries and preservation of the original arrangement.

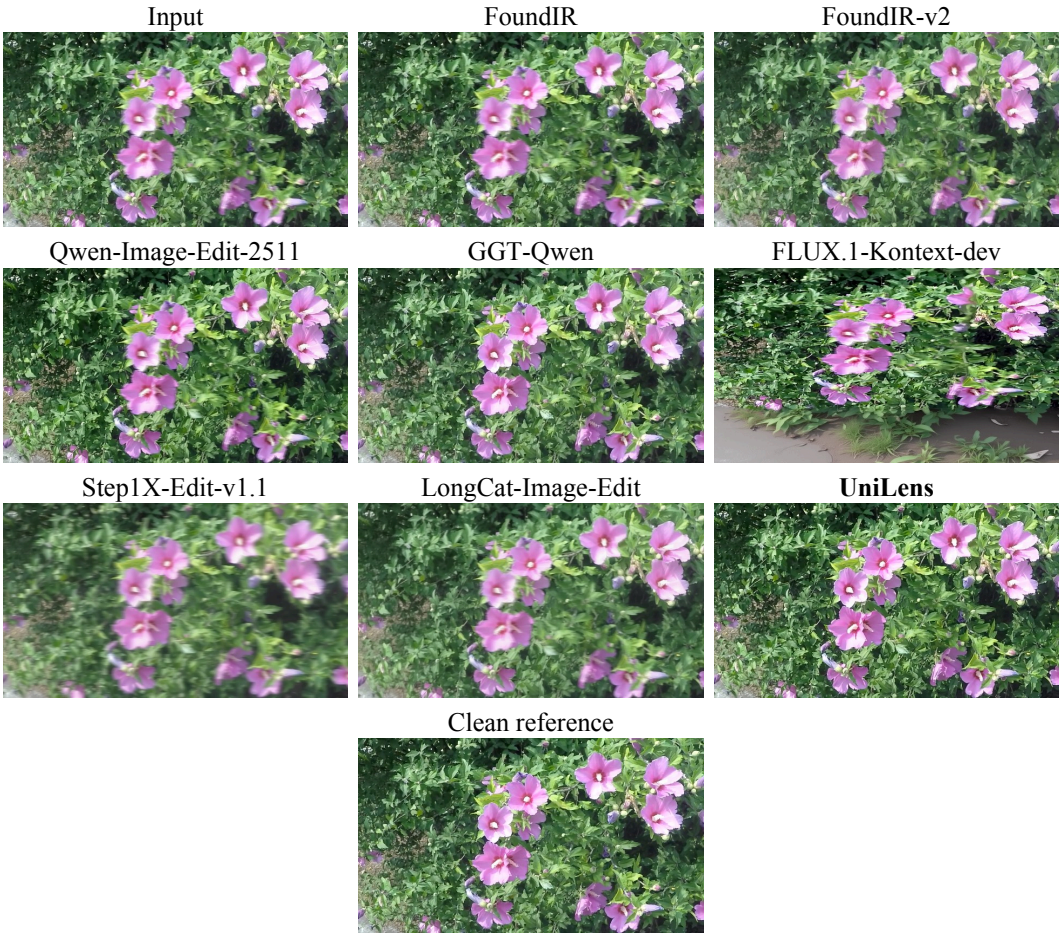

Figure 19: Motion blur on GoPro (Nah et al., 2017).

We compare noise removal on Urban100 in Figure 20. Matched close-ups expose residual grain and changes to the window-grid geometry relative to the clean reference.

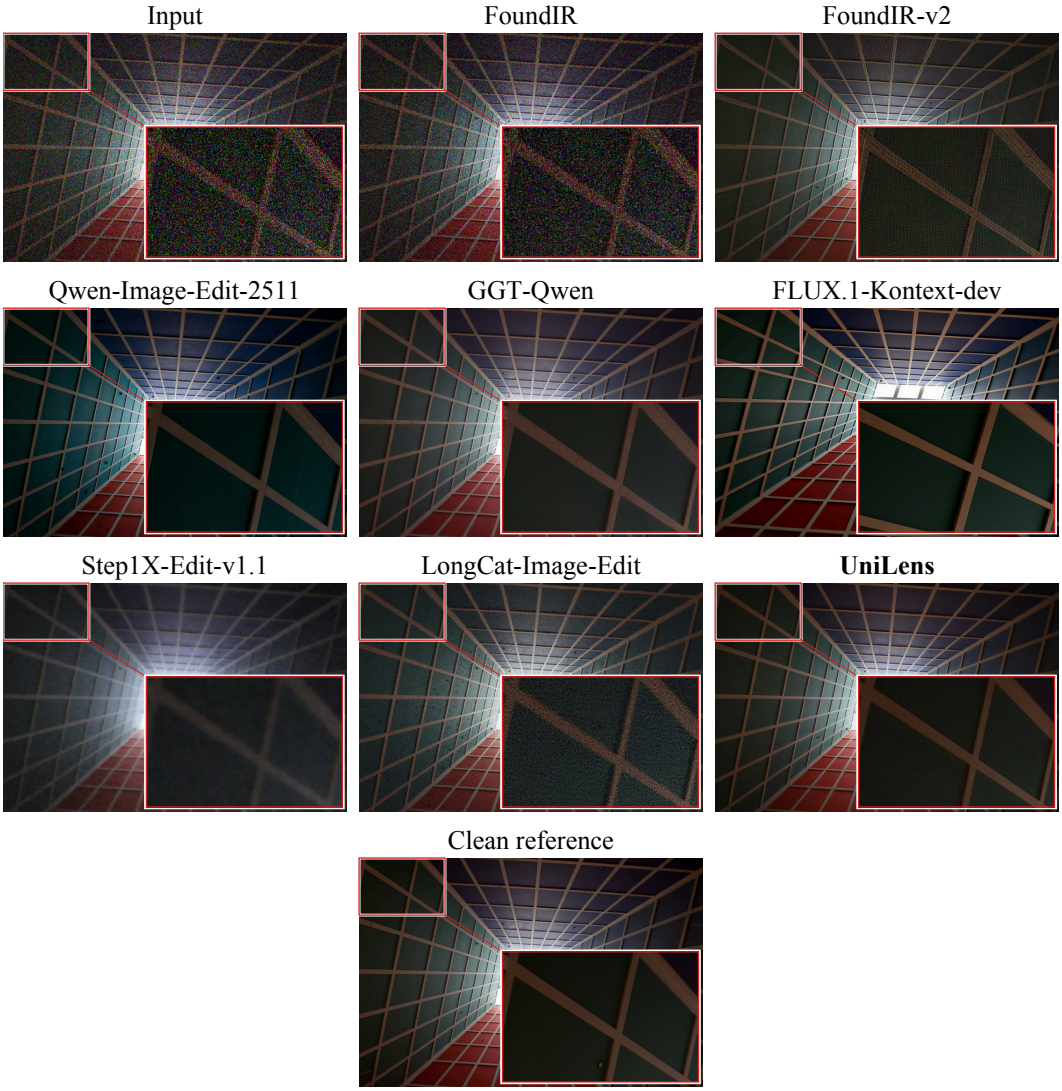

Figure 20: Noise on Urban100 (Huang et al., 2015), with $\sigma = 50$. Insets enlarge the same region across methods.

We compare paired haze in Figure 21. The path and foliage expose residual veiling and changes to the scene.



Figure 21: Haze on FoundIR (Li et al., 2025).

We compare paired rain in Figure 22. The railing and kitchen surfaces expose residual streaks and structural changes.



Figure 22: Rain on Test1200 (Zhang & Patel, 2018).

We compare paired flare in Figure 23. The ceiling lights distinguish suppression of streaks from removal of the light sources.

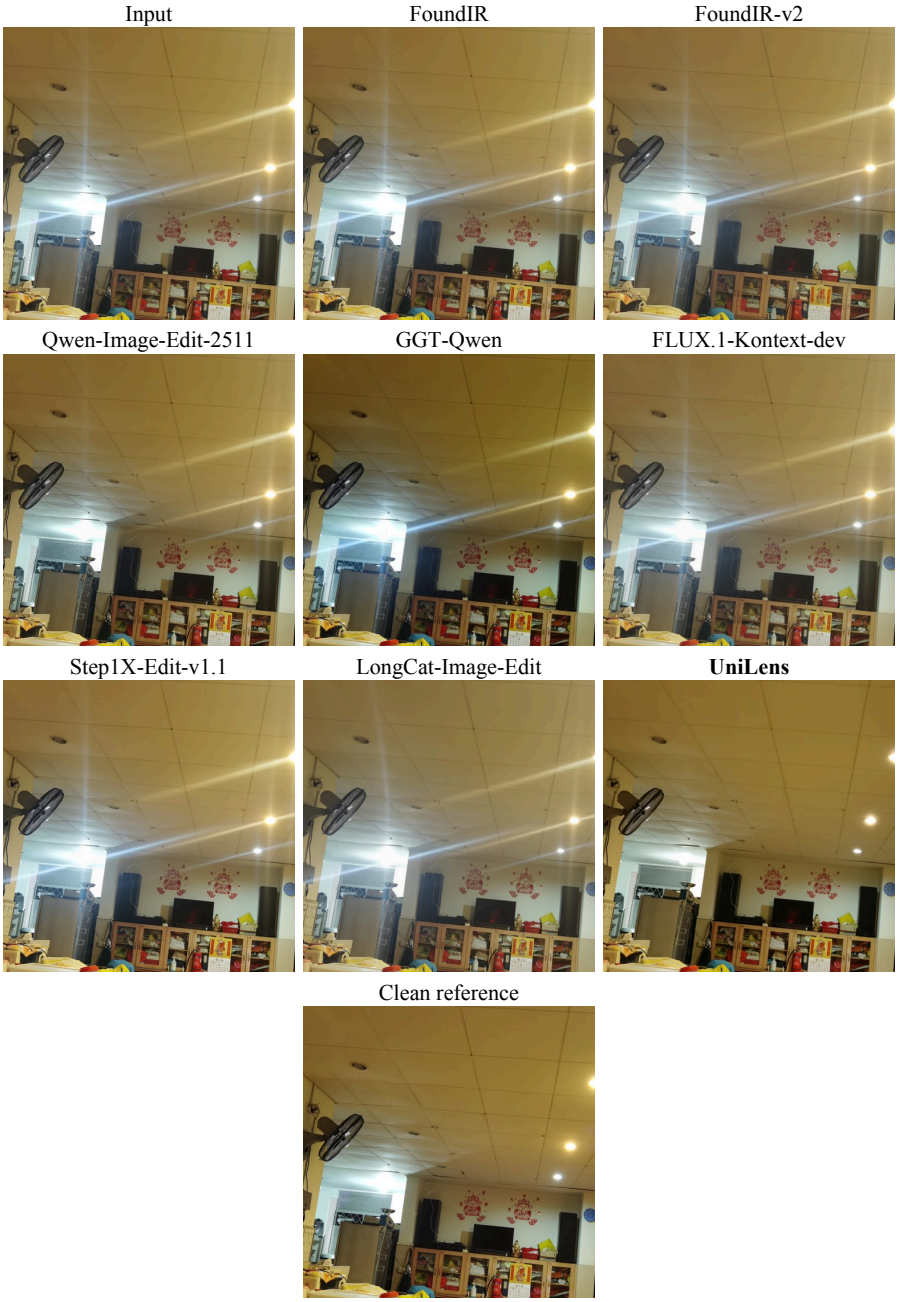

Figure 23: Flare on Flare7K real (Dai et al., 2022).

We compare paired low light in Figure 24. The landscape reveals differences in recovered shadow detail and exposure relative to the reference.



Figure 24: Low light on LOL-v2 synthetic (Yang et al., 2020).

We compare raindrop removal on FoundIR in Figure 25. The ground-region close-ups reveal residual droplet traces and changes to the underlying texture relative to the clean reference.

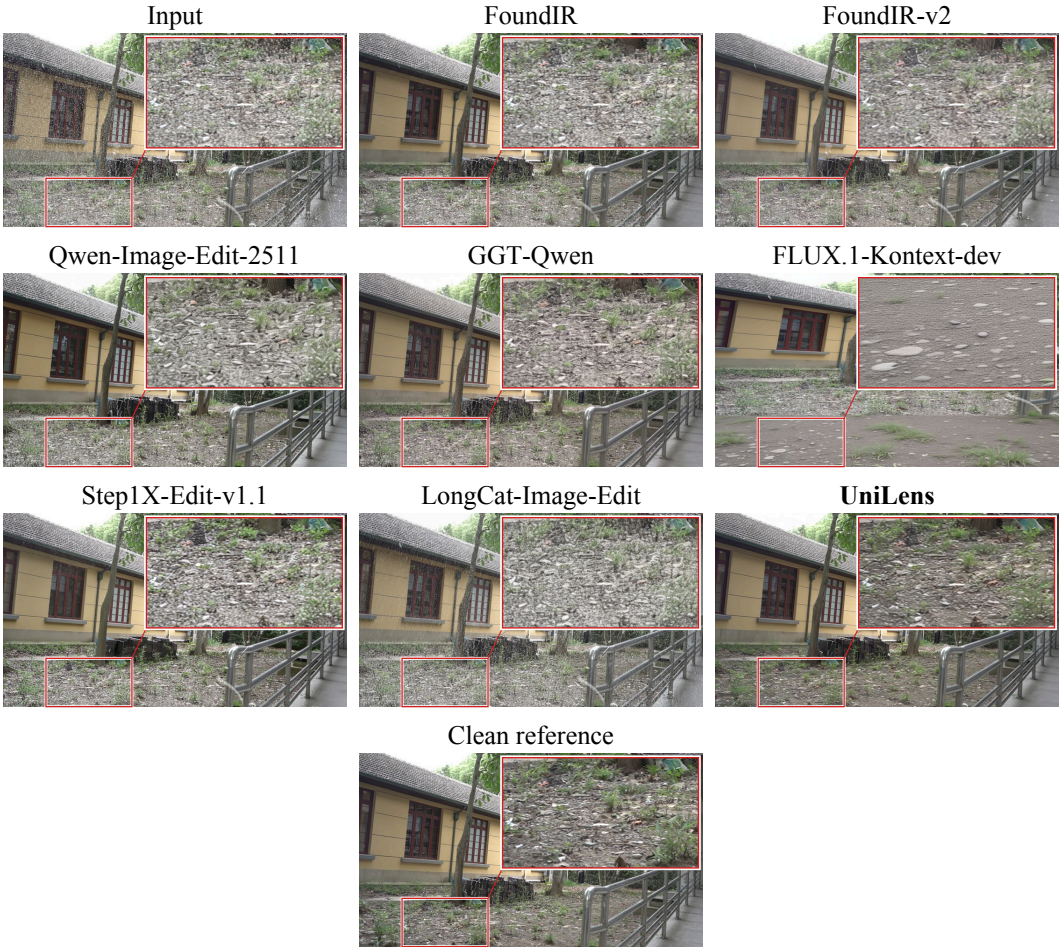

Figure 25: Raindrop removal on FoundIR (Li et al., 2025). Insets enlarge the same ground region across methods.

G Additional related work

Restoration priors and data. Convolutional networks and Transformers provide restoration backbones (Chen et al., 2022; Zamir et al., 2022). DiffBIR and SUPIR transfer diffusion priors to restoration, while HYPIR combines diffusion initialization with adversarial fine-tuning (Lin et al., 2024; Yu et al., 2024; Lin et al., 2025). Image-editing backbones include Step1X, FLUX Kontext, and Qwen-Image (Liu et al., 2025; Black Forest Labs et al., 2025; Wu et al., 2025). Captured pairs such as SIDD and RealBlur provide complementary real-image supervision (Abdelhamed et al., 2018; Rim et al., 2020); GRIDS uses degradation similarity to group restoration tasks (Cao et al., 2024). CODE instead compares training inputs with unpaired real observations to guide pair construction.

Quality and task fulfillment. Q-Align, DepictQA, and DeQA-Score extend image-quality assessment with vision-language models (Wu et al., 2024a; You et al., 2024; 2025). DRA focuses on the requested restoration: its output-only judgment assesses residual target degradation, and its comparative judgment evaluates improvement from the input. This task-conditioned view complements the fidelity and perceptual metrics reported for paired benchmarks.