CURE: CLEAN-GUIDED UNDERSTANDING RESTORATION FOR DEGRADED IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Image restoration aims to recover high-quality images from degraded observations, providing reliable inputs for downstream visual tasks. When these tasks are performed by large vision-language models (LVLMs), conventional restore-then-understand pipelines still optimize restoration primarily for pixel fidelity or perceptual quality, which does not directly guarantee improvements in understanding and may propagate restoration errors into model responses. To address this issue, we propose understanding restoration for degraded inputs and introduce CURE (Clean-Guided Understanding REstoration). CURE first lets the model generate responses to degraded images, then uses its predictive distributions under clean-image conditions to supervise each step of the generated responses. This supervision guides the model's generation decisions toward those under clean conditions, aiming to restore visual understanding impaired by image degradation. During inference, the adapted model directly processes degraded images without an additional image restoration module. On Qwen3.5-4B, CURE improves the average score on the real-world degradation benchmark R-Bench from 63.49 to 67.38, outperforming the evaluated two-stage restoration pipelines. It also achieves overall improvements in cross-dataset evaluations with synthetic degradations and evaluations on clean images. Furthermore, CURE can complement image restoration models in a two-stage pipeline to improve understanding of degraded images.

1 INTRODUCTION

Real-world images are often degraded by blur, noise, or compression (Hendrycks & Dietterich, 2019; Li et al., 2025a). Image restoration (Li et al., 2022; 2025b) aims to recover a high-quality image from a degraded observation, thereby providing reliable visual input for downstream tasks. However, the rapid development of large vision-language models (LVLMs) has enabled a single LVLM to perform visual understanding, reasoning, visual question answering, and image captioning, tasks that previously required separate models (Liu et al., 2024; OpenAI, 2024; Qwen Team, 2025). With their strong cross-task generalization, LVLMs are increasingly becoming important downstream consumers of restored images (Tang et al., 2026b). The ultimate value of restoration therefore depends on whether it helps LVLMs understand image content. This raises a central question: How should image restoration be optimized to better support visual understanding by LVLMs?

A common way to handle degraded images is to restore them before passing them to LVLMs for understanding (Tang et al., 2026b). This two-stage approach based on image restoration has two limitations. First, restoration models are typically optimized for pixel fidelity or perceptual quality, but these measures do not directly indicate whether a restored image helps LVLMs perform their tasks (Xing et al., 2025). Second, errors introduced during restoration can carry over to LVLMs (Hao et al., 2026). If the restoration model removes, distorts, or hallucinates task-relevant visual details, LVLMs may base their responses on incomplete or misleading evidence.

These limitations motivate us to reconsider what image restoration should recover when downstream tasks are performed by LVLMs in practice. Rather than requiring a clearer image as an intermediate result, we aim to bring LVLMs' responses to degraded images closer to their responses to the corresponding clean images. This goal can also be pursued alongside a reliable image restoration front end to support downstream understanding. Recent work on distillation has explored supervising models along their own generated responses using predictive distributions from stronger

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

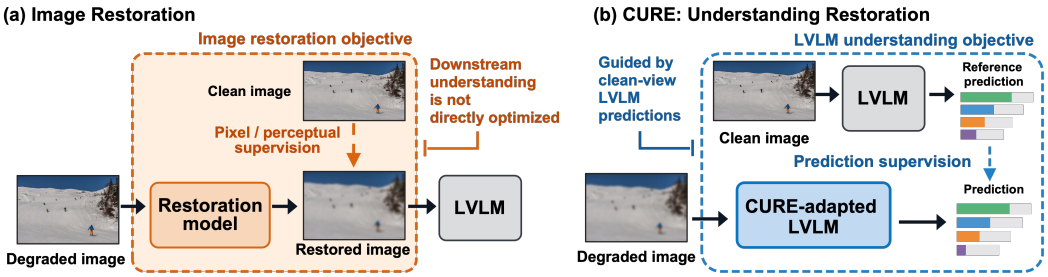


Figure 1: **Comparing image restoration with CURE.** (a) Pixel/perceptual supervision links clean and restored images. (b) CURE uses predictions from the base model on clean images as supervision during training, while adapted LVLMs directly process degraded images at inference.

models (Agarwal et al., 2024; Ko et al., 2024), as well as on-policy self-distillation under different informational conditions (Yuan et al., 2026).

Drawing on these distillation principles, we propose **CURE** (**C**lean-Guided **U**nderstanding **RE**storation), a method for improving the robustness of visual understanding in LVLMs by transferring clean-view predictive behavior to degraded inputs. During training, the model generates responses from degraded images and uses its predictive distribution on the paired clean image for the same task as a supervisory signal at each generation step. This supervision is applied to the full vision-language model, jointly encouraging visual representations and language-side task understanding and response-generation behavior to move toward their clean-view counterparts. During inference, the trained LVLM processes degraded images directly, without explicitly generating an intermediate restored image or invoking an additional restoration module.

On the real-world degradation benchmark R-Bench (Li et al., 2025a), CURE raises Qwen3.5-4B's average score from 63.49 to 67.38, outperforming the evaluated two-stage restoration pipelines. CURE also improves average performance on cross-dataset synthetic degradation and clean-image understanding benchmarks. Our contributions are as follows:

- We introduce CURE, which uses predictions from paired clean images to supervise responses to degraded images. This shifts the restoration target from pixel reconstruction to LVLMs' understanding and behavior.
- CURE improves aggregate scores across real-world degradation, cross-dataset synthetic degradation, and clean-image understanding evaluations. At inference, the adapted LVLMs follow a single-model path without an additional image restoration module.
- Through representation analysis and module ablation on Qwen3.5-4B, we find that CURE increases clean–degraded representation similarity within the language decoder. Language adaptation accounts for most of the performance gain, while visual adaptation provides complementary benefits.

2 Related Work

Image restoration (Li et al., 2022; 2025b) has long aimed to recover a clean image from a degraded one. All-in-one restoration models handle multiple types of degradation with a single model (Li et al., 2022). Other methods use natural language instructions to guide restoration for different degradations (Xu et al., 2024; Yang et al., 2026). Restored images are often evaluated using PSNR and SSIM, which measure their similarity to clean references. Yet the tradeoff between perceptual quality and distortion (Blau & Michaeli, 2018) shows that high scores on these metrics do not guarantee that a restored image looks natural. This matters when restored images are used by LVLMs. Benchmarks show large drops in LVLM accuracy on degraded images (Li et al., 2025a; Saxena et al., 2026), and recent work finds that higher image quality scores do not always lead to better LVLM understanding (Xing et al., 2025).

To improve LVLM performance on degraded images, a common approach is to restore the image before passing it to the LVLM. This approach has two limitations. Its focus on pixel similarity may not reflect the visual information needed for the task. Restoration errors can also affect the LVLM's

answer, and false details added during restoration may be more misleading than the original degradation (Tang et al., 2026b). Some studies aim to make LVLMs more robust to degraded inputs rather than produce better restored images. FARE (Schlarmann et al., 2024) trains the vision encoder to produce similar features for clean images and images with adversarial perturbations. These methods improve visual features, but they do not directly guide how the decoder uses them to generate an answer. Robust-LLaVA (Malik et al., 2025) incorporates vision encoders pretrained with adversarial examples on large datasets. It then trains the visual projector and language model to use these robust features. Some methods focus on task responses. Standard supervised training on degraded images updates the model using target answers, while Robust-R1 (Tang et al., 2026a) uses reinforcement learning to encourage explicit structured reasoning about degradation properties. Both approaches leave an important middle stage unaddressed. As the model generates a response, it makes a sequence of conditional decisions that gradually turns visual evidence into an answer. Degradation can change any of these decisions, yet neither feature alignment nor supervision on the final output answer directly corrects these changes at each generation step.

3 METHOD

This section describes how CURE helps LVLMs restore their understanding of degraded images for a given task. We first compare next-token distributions conditioned on clean and degraded images under the same task and response prefix, and examine how the resulting prediction shifts vary with degradation type and severity. We then describe how training pairs are constructed and how predictions from clean images supervise response generation from degraded images. Figure 2 illustrates the CURE framework.

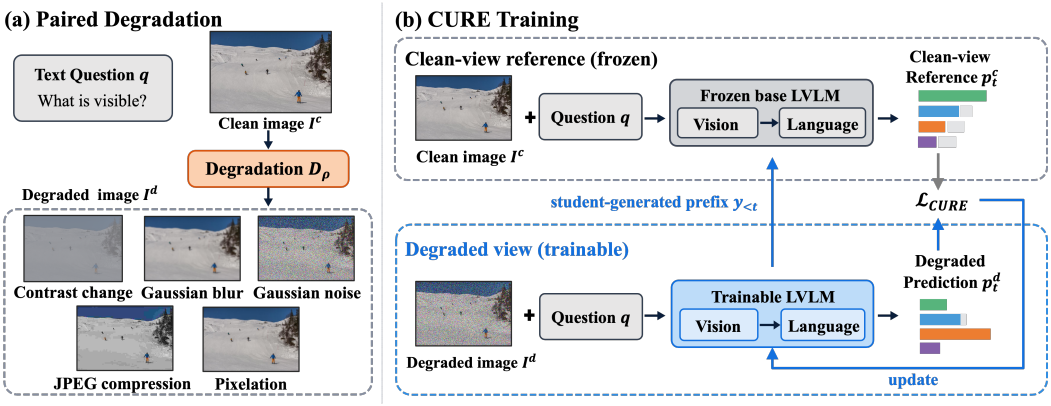


Figure 2: **The CURE framework.** (a) Synthetic degradation maps a clean image I^c to a degraded image I^d for the same question q. (b) The LVLM generates an answer prefix $y_{<t}$ from the degraded view. The frozen clean-view branch then provides a next-token distribution for the same question and prefix. CURE updates the model by aligning its prediction under degradation with this clean-view reference.

3.1 TASK-CONDITIONED UNDERSTANDING RESTORATION

For LVLMs, image understanding restoration concerns their ability to complete a task when given a degraded image. Given a clean image I^c, its corresponding degraded view I^d, and a task instruction q, CURE aims to use the model's prediction for the same task on I^c as a reference to improve its prediction on I^d.

Therefore, image understanding restoration differs from image restoration because its goal must be defined with respect to a specific task. The clean and degraded views depict the same scene, but degradation may affect the visual evidence available to the model, while the task instruction determines which evidence is relevant to the response. We therefore need to consider how degradation changes the model's predictions for the same task and how they manifest during response generation.

Given an image I and a text task instruction q, LVLMs generate responses autoregressively. The conditional probability of a response $y = (y_1, \ldots, y_T)$ factorizes as

$$p_\Theta(y \mid I, q) = \prod_{t=1}^{T} p_\Theta(y_t \mid I, q, y_{<t}), \tag{1}$$

where Θ denotes the model parameters, and $y_{<t}$ is the response prefix generated before step t. At each generation step, the model produces a next-token distribution conditioned on the image, the text instruction, and the response prefix. To examine how degradation changes this distribution, we hold the model, instruction, and prefix fixed and vary only the image. The resulting distributions are

$$p_t^c = p_\Theta(\cdot \mid I^c, q, y_{<t}), \qquad p_t^d = p_\Theta(\cdot \mid I^d, q, y_{<t}). \tag{2}$$

Because p_t^c and p_t^d share the same model, task instruction, and response prefix, their difference reflects how degradation changes the model's prediction at step t. Task-conditioned understanding restoration aims to reduce this difference, bringing predictions under degraded inputs closer to those under clean inputs.

3.2 Degradation-Induced Prediction Shifts

To understand how degradation affects task predictions, we compare the next-token distributions under paired clean and degraded views. Specifically, we quantify their difference at each generation step using Jensen–Shannon divergence (JSD)

$$\delta_t = \mathrm{JSD}(p_t^c, p_t^d). \tag{3}$$

Here, δ_t measures the prediction shift for the same task instruction and response prefix. It does not directly indicate whether the answer is correct.

Then, we evaluate changes across degradation types and severity levels using δ_t. Specifically, we randomly sample examples from the GQA dataset (Hudson & Manning, 2019) and create synthetic degradations based on ImageNet-C (Hendrycks & Dietterich, 2019). We apply five operators, Contrast, GaussianBlur, GaussianNoise, JpegCompression, and Pixelate, to create paired clean and degraded images. Each operator is applied at three severity levels, labeled s1, s2, and s3. This gives 15 settings with 300 examples per setting. Using the Qwen3.5-4B base model (Qwen Team, 2026), we generate responses from degraded views. At each generation step, we compute p_t^c and p_t^d under the same task instruction and generated prefix, then record the resulting δ_t.

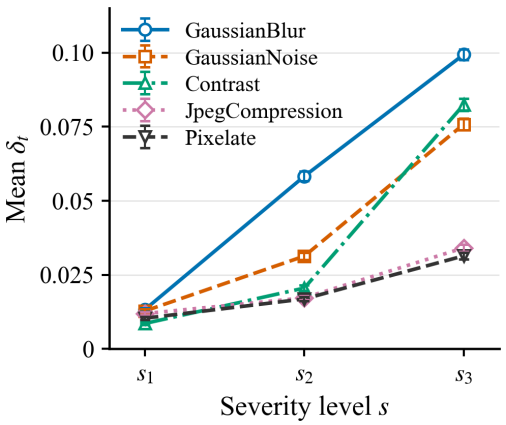


Figure 3: Mean prediction shift δ_t between clean and degraded views across degradation operators and severity levels, measured with Qwen3.5-4B.

Figure 3 shows that mean δ_t increases with severity for all five operators, although the magnitude of the shift varies across operators. Contrast remains well below GaussianBlur at the middle level but shows the largest increase from the middle to the highest level. GaussianNoise also rises substantially, while JpegCompression and Pixelate produce smaller shifts at the two higher levels. The differences among operators at the same assigned severity suggest that severity alone does not capture how strongly degradation changes task predictions.

3.3 Clean-Guided Training with Paired Images

CURE uses clean-view next-token distributions to guide response generation from degraded images under the same task instruction and response prefix. We first construct paired clean and degraded training examples, then use the clean-view reference to train on degraded inputs. Specifically, we

construct each training pair by applying a degradation operator o at severity level s to a clean image I^c. This produces a degraded view $I^d = \mathsf{D}_\rho(I^c)$, where $\rho = (o, s)$ specifies the degradation setting. We associate both views with the same task instruction q to form a training tuple (I^c, I^d, q).

To build the paired training set, we randomly sample 32,000 GQA (Hudson & Manning, 2019) examples with clean images and construct one candidate pair from each example. Using Qwen3.5-4B, we compute the mean δ_t for each pair and filter out pairs with excessively high or low mean prediction shifts. From the remaining pairs, we sample equal numbers for each of the five degradation operators, yielding a paired training set $\mathcal{D} = \{(I_i^c, I_i^d, q_i)\}_{i=1}^N$, where $N = 6{,}400$.

Each selected pair provides a degraded input and a clean-view reference for the same task. Unlike the diagnostic comparison in Section 3.2, which uses the same fixed model for both views, training uses a trainable degraded-view branch with parameters θ and a frozen clean-view branch with parameters $\bar{\theta}$. Both branches are initialized from the same base-model checkpoint. Given I^d and q, the trainable branch generates a response y. At each response position, the frozen branch provides a next-token distribution conditioned on I^c, q, and the same student-generated prefix $y_{<t}$. We minimize the JSD between the two next-token distributions, averaging over valid response positions and training examples. The resulting objective is

$$\mathcal{L}_{\mathrm{CURE}}(\theta) = \mathbb{E}_{\substack{(I^c, I^d, q)\sim\mathcal{D} \\ y\sim p_\theta(\cdot|I^d,q)}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \mathrm{JSD}\big(\mathrm{sg}[p_{\bar{\theta}}(\cdot \mid I^c, q, y_{<t})], p_\theta(\cdot \mid I^d, q, y_{<t})\big) \right], \tag{4}$$

where sg denotes stop-gradient and $|y|$ is the number of valid response positions included in the loss. The response generated from the degraded view supplies the prefix $y_{<t}$ used by both branches. This allows the model to bring its degraded-view predictions closer to the clean-view reference at prefixes it actually generates. During training, we update both the vision and language modules. At inference, the trained LVLM generates a response directly from the degraded image and task instruction using a single model.

4 Experiments

In this section, we evaluate whether CURE can effectively restore models' ability to understand degraded inputs. Section 4.1 describes the experimental setup. Section 4.2 evaluates CURE under real-world degradations, synthetic degradations across benchmarks, and clean inputs. It also compares CURE with two-stage restore-then-understand pipelines. Section 4.3 examines how the visual and language components contribute to understanding restoration, using module freezing, weight recombination, and layerwise representation analysis.

4.1 Experimental Setup

Training Configuration We use Qwen3.5-4B (Qwen Team, 2026) as the base model for the main experiments with thinking mode disabled. The training data comprise 6,400 clean–degraded image pairs constructed from GQA (Hudson & Manning, 2019). For each training example, we generate a single response rollout. We update all model parameters and train for 100 steps using a global batch size of 64 and a learning rate of 2×10^{-6}.

Benchmarks and Metrics We evaluate understanding restoration from three perspectives:

- **Real-world degradations.** R-Bench (Li et al., 2025a) evaluates multiple-choice questions (MCQ), visual question answering (VQA), and image captioning (CAP) under real-world degradation. Each task is tested at low, medium, and high degradation levels. Qwen3.6-35B-A3B serves as the judge model for scoring responses. We report Overall as the mean accuracy across the resulting nine settings.
- **Synthetic degradations across benchmarks.** Following the evaluation setting of Robust-R1 (Tang et al., 2026a), we evaluate MMMB (Sun et al., 2025), MMStar (Chen et al., 2024), and RealWorldQA (xAI, 2024) on clean images and at synthetic degradation settings labeled 25%, 50%, and 100%. RVE is the mean of the 12 benchmark-condition scores.
- **Clean-image understanding.** We evaluate POPE (Li et al., 2023), HallusionBench (Guan et al., 2024), VstarBench (Wu & Xie, 2024), and HRBench-4K/8K (Wang et al., 2025) to

assess performance on clean inputs, including hallucination, fine-grained perception, and high-resolution understanding. GVE is the mean score across these five benchmarks.

Comparison Methods We compare CURE with three groups of methods. (1) **General-purpose LVLMs** include Gemma3-4B (Gemma Team, 2025), Qwen3-VL-4B/8B (Qwen Team, 2025), and Qwen3.5-4B/9B. These models provide degradation-robustness baselines across different base-model scales. (2) **Robustness-enhanced models** include Robust CLIP (Schlarmann et al., 2024) and Robust LLaVA (Malik et al., 2025), which use robust vision encoders, and Robust-R1 (Tang et al., 2026a), which employs structured degradation-aware reasoning fine-tuning. (3) **Restoration-based two-stage methods** include two-stage pipelines that first restore images with BaryIR (Tang et al., 2025) or RestoreVAR (Rajagopalan et al., 2025) and then pass them to either the original Qwen3.5-4B or CURE for understanding. We also evaluate Robust-U1 (Tang et al., 2026b), a unified model for visual understanding and generation.

4.2 Overall Effectiveness and Generalization

Understanding Restoration under Real-World Degradations. Table 1 reports the R-Bench results. CURE improves the overall score of Qwen3.5-4B from 63.49 to 67.38, a gain of 3.89 points. It outperforms the base model in all nine task–severity settings, ranking first in four and second in two. Its overall score also slightly exceeds that of Qwen3.5-9B (67.22), showing that the gains from adapting the 4B model extend to understanding under real-world degradations.

We next compare CURE with two-stage restore-then-understand pipelines. With RestoreVAR or BaryIR as the restoration front end, the original Qwen3.5-4B scores 62.19 and 62.72, respectively, below its score of 63.49 when processing degraded images directly. Using CURE as the understanding model raises the corresponding scores to 65.60 and 66.47, but both remain below CURE alone (67.38). Although CURE+BaryIR improves individual settings such as MCQ-low and VQA-high, neither restoration front end improves the overall R-Bench score of its corresponding understanding model. Both front ends nevertheless perform better with CURE than with the original base model, indicating that CURE's understanding-side adaptation also benefits these two-stage pipelines.

Table 1: R-Bench accuracy (%) on MCQ, VQA, and captioning (CAP) at low, medium, and high real-degradation levels; Overall is the mean over the nine task–intensity pairs. Robust-U1 jointly performs restoration and understanding; RestoreVAR and BaryIR are restoration front ends evaluated with the original and CURE-fine-tuned Qwen3.5-4B models. Bold and underlined values indicate the best and second-best results in each column.

Category	Method	MCQ			VQA			CAP			Overall
		low	mid	high	low	mid	high	low	mid	high	
General MLLM	Gemma3-4B	58.82	50.31	48.78	52.64	47.58	48.92	52.84	49.29	44.41	50.40
	Qwen3-VL-4B	68.82	67.70	57.93	57.18	52.85	48.56	66.30	66.26	57.20	60.31
	Qwen3-VL-8B	71.57	70.60	59.55	61.06	58.93	56.09	64.81	64.43	55.75	62.53
	Qwen3.5-4B	70.59	69.98	60.57	59.47	58.32	59.34	67.18	67.32	58.64	63.49
	Qwen3.5-9B	**76.86**	**78.88**	**68.90**	64.38	<u>61.66</u>	<u>61.38</u>	**68.62**	65.81	58.50	<u>67.22</u>
Robust Fine-tuned	Robust CLIP	44.71	44.10	45.12	47.85	44.30	44.55	37.41	35.62	32.56	41.80
	Robust LLaVA	37.65	46.58	38.41	51.66	47.21	51.14	34.07	31.72	31.72	41.13
	Robust-R1	57.06	50.93	50.00	52.52	49.94	46.41	53.89	49.23	43.79	50.42
Restoration based	Robust-U1	73.53	70.81	62.20	44.11	39.94	41.08	49.32	49.45	50.59	53.45
	RestoreVAR	69.02	68.12	63.21	59.94	58.95	56.29	66.19	63.25	54.76	62.19
	BaryIR	70.59	69.77	63.21	58.67	57.37	55.99	67.35	65.07	56.49	62.72
Ours	CURE	74.12	<u>77.85</u>	67.68	**65.40**	**63.45**	60.68	<u>68.56</u>	**68.27**	**60.42**	**67.38**
	CURE+RestoreVAR	75.29	74.53	<u>68.50</u>	63.89	61.45	56.71	65.84	66.65	57.56	65.60
	CURE+BaryIR	<u>75.88</u>	73.08	66.26	<u>64.74</u>	60.77	**61.68**	67.63	<u>68.06</u>	<u>60.13</u>	66.47

Generalization to Synthetic Degradations. Table 2 shows that CURE raises the RVE average of Qwen3.5-4B from 73.89 to 75.25. CURE reduces the performance drop from clean images to the 100% degradation setting on all three benchmarks. CURE also outperforms the base model at 100% degradation on all three benchmarks. These results indicate that the gains extend to the three evaluated benchmarks under synthetic degradation.

The effect of a two-stage pipeline depends on its restoration front end. With RestoreVAR, the base model and CURE achieve RVE averages of 71.36 and 72.74, respectively, below their scores without restoration (73.89 and 75.25). With BaryIR, the corresponding averages rise to 76.20 and 76.86; the latter is the highest in the table and exceeds CURE alone by 1.61 points. At 100% degradation on MMMB, MMStar, and RealWorldQA, CURE+BaryIR scores 86.11, 68.47, and 75.56, compared with 82.17, 58.93, and 73.20 for CURE alone. Thus, BaryIR can complement CURE in these synthetic degradation settings, while the same overall benefit is not observed on real-world R-Bench.

Table 2: Accuracy (%) on MMMB, MMStar, and RealWorldQA for clean images and images with 25%, 50%, or 100% synthetic degradation; Avg. is the mean over these 12 benchmark–condition pairs. RestoreVAR and BaryIR are evaluated with the original and CURE-fine-tuned Qwen3.5-4B models. Bold and underlined values mark the best and second-best results in each column.

Category	Method	MMMB				MMStar				RealWorldQA				Avg.
		clean	Intensity			clean	Intensity			clean	Intensity			
			25%	50%	100%		25%	50%	100%		25%	50%	100%	
General MLLM	Gemma3-4B	70.96	71.06	70.45	68.38	44.07	44.20	43.73	42.93	55.29	53.73	54.64	52.03	55.96
	Qwen3-VL-4B	84.34	84.04	82.98	80.40	62.53	62.20	59.40	56.27	70.72	70.85	69.41	66.67	70.82
	Qwen3-VL-8B	85.25	84.49	83.99	81.21	64.93	64.20	62.27	57.33	72.42	71.90	70.59	69.28	72.32
	Qwen3.5-4B	86.06	85.05	83.94	81.52	66.33	65.20	63.00	56.20	77.25	76.08	75.42	70.59	73.89
	Qwen3.5-9B	**86.16**	85.51	84.34	81.77	65.47	64.73	62.73	57.93	77.12	**76.86**	75.29	72.94	74.24
Robust Fine-tuned	Robust CLIP	67.47	66.67	65.61	61.67	33.00	32.00	31.47	30.60	48.76	46.14	46.41	43.14	47.74
	Robust LLaVA	62.68	62.93	61.82	59.09	28.93	28.53	28.33	28.20	48.37	49.15	48.50	47.71	46.19
	Robust-R1	79.75	78.64	77.53	74.65	53.60	53.20	52.73	49.00	67.19	66.27	64.84	61.70	64.93
Restoration based	RestoreVAR	84.85	84.65	84.09	80.66	65.13	62.33	60.80	55.47	71.11	71.76	68.24	67.19	71.36
	BaryIR	85.35	85.66	85.45	85.61	68.20	65.60	66.13	66.13	75.16	76.73	**77.65**	**76.73**	76.20
Ours	CURE	**86.16**	85.66	85.25	82.17	67.13	**67.87**	65.87	58.93	**77.52**	76.47	76.73	73.20	75.25
	CURE+RestoreVAR	86.01	84.70	83.89	81.92	67.13	65.87	63.13	57.13	72.16	72.68	71.76	66.54	72.74
	CURE+BaryIR	85.86	**85.86**	**86.31**	**86.11**	**69.80**	67.20	**68.60**	**68.47**	76.21	75.69	76.60	75.56	**76.86**

Performance on Clean Inputs. Table 3 reports performance on clean-image benchmarks. CURE raises the GVE average of Qwen3.5-4B from 77.91 to 81.26 and achieves the highest score on POPE, VstarBench, HRBench-4K, and HRBench-8K. Together with the degraded-input results, this shows that CURE's gains are accompanied by higher aggregate performance on clean-image understanding benchmarks.

Table 3: Accuracy (%) on five clean-image benchmarks. Avg. is the mean across the five benchmarks; CURE is fine-tuned from Qwen3.5-4B. Bold and underlined values mark the best and second-best results in each column.

Category	Method	POPE	HallusionBench	VstarBench	HRBench4K	HRBench8K	Avg.
General MLLM	Qwen2.5-VL-3B	85.81	63.30	74.35	66.50	62.00	70.39
	Gemma3-4B	85.21	58.15	36.65	45.75	41.38	53.43
	Qwen3-VL-4B	89.18	69.72	79.58	76.50	73.63	77.72
	Qwen3-VL-8B	87.93	73.40	83.77	76.63	70.87	78.52
	Qwen3.5-4B	87.28	71.82	82.72	75.75	72.00	77.91
	Qwen3.5-9B	88.84	**73.61**	83.77	80.13	76.25	80.52
Robust Fine-tuned	Robust CLIP	78.64	41.54	40.31	49.00	46.00	51.10
	Robust LLaVA	82.17	47.11	38.22	35.25	33.75	47.30
	Robust-R1	87.34	66.46	71.20	65.50	61.88	70.48
Ours	CURE	**89.27**	71.40	**86.39**	**81.87**	**77.38**	**81.26**

Qualitative Examples. Figure 4 presents two VQA examples with paired clean and degraded images. In Case A, the base model gives a valid answer for the clean view but, when shown the blurred image, only states that the cat is watching TV without identifying the program's content. CURE identifies the space-related content in both views. In Case B, the base model correctly identifies the umbrella as open in the clean view but incorrectly calls it closed in the darker degraded view; CURE

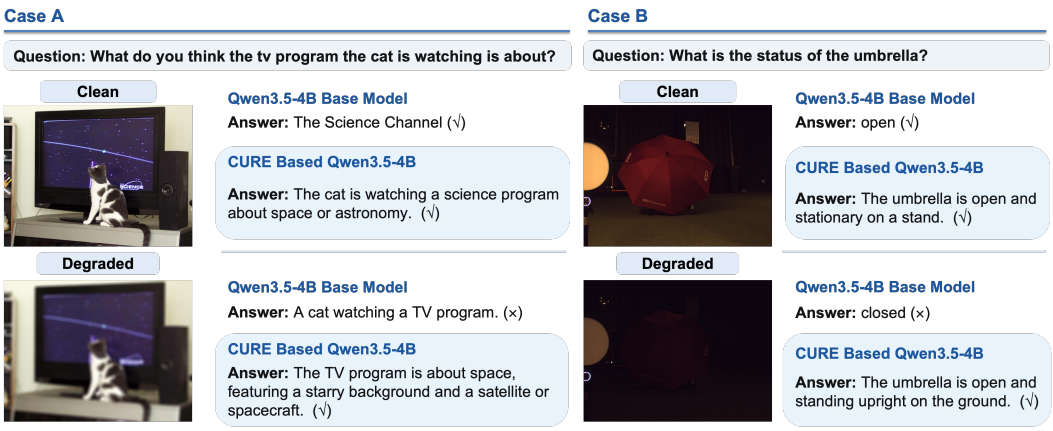


Figure 4: Two VQA examples with clean and degraded images. In Case A, blur makes the television content harder to discern; in Case B, degradation darkens the scene, making the umbrella's state harder to determine. Each case shows responses from the Qwen3.5-4B base model and CURE on both views. Check marks and crosses indicate correct and incorrect responses, respectively.

identifies it as open in both. These examples illustrate cases in which CURE answers the questions correctly despite image degradation.

4.3 Visual and Language Adaptation in Understanding Restoration

We examine the roles of the visual and language components in understanding restoration through controlled parameter updates and representation analysis. Table 4 compares module-freezing and post-training weight-recombination experiments using the same training data, budget, and evaluation protocol. Figure 5 provides a complementary view of clean–degraded representation changes at visual and text token positions within the language decoder.

Training Scope and Weight Recombination. In Table 4, LLM denotes the language model, and Vision Projector denotes the multimodal projection layer. LLM only updates the language component; LLM + Vision Projector additionally updates the projector; and Vision only updates the visual encoder and projector. On R-Bench, LLM only improves the base score from 63.49 to 66.95, retaining approximately 89% of the gain achieved by full fine-tuning (63.49 to 67.38). Adding the projector yields 66.92, whereas Vision only reaches 64.02. On RVE, LLM only and Vision only score 74.69 and 74.71, both close to the full fine-tuning score of 75.25. On GVE, they score 81.94 and 79.16, respectively, compared with 81.26 for full fine-tuning. Thus, language-side adaptation carries most of the R-Bench gain, while the effects of training scope differ across RVE and GVE.

Post-training weight recombination provides a further test. Retaining the language component from the fully fine-tuned model while restoring the base visual component yields 66.69 on R-Bench. The reverse combination scores 63.59, close to the base model's 63.49. Combining the language component from the independently trained LLM-only model with the visual component from the Vision-only model yields 67.15 on R-Bench, 74.88 on RVE, and 82.58 on GVE; the GVE score is the highest in the table. These results show that the main R-Bench gain persists when the fine-tuned language weights are retained. They also show that independently adapted visual and language components can be combined, with additional benefits on some metrics.

Layerwise Representation Analysis. Figure 5 plots the change relative to the base model in cosine similarity between paired clean and degraded representations. We measure visual and text token positions within the Qwen3.5-4B language decoder on real-world degraded examples from R-Bench. Full fine-tuning produces larger increases at both position types in the middle and later layers. At text token positions in later layers, LLM only shows a larger increase than Vision only. At visual token positions, their curves are relatively close through much of the middle and later layers, despite their different R-Bench scores. The representation analysis therefore shows that both train-

Table 4: Module-freezing controls (upper block) and weight-surgery combinations (lower block) for Qwen3.5-4B, reporting scores (%) on R-Bench (real degradation), RVE (synthetic degradation), and GVE (clean images) under a shared protocol. Bold and underlined values mark the best and second-best results among the ablation arms.

Trainable or combined modules	R-Bench	RVE	GVE
Qwen3.5-4B (base)	63.49	73.89	77.91
Trainable scope			
Full FT	**67.38**	**75.25**	81.26
LLM only	66.95	74.69	81.94
LLM + Projector	66.92	74.83	81.61
Vision only	64.02	74.71	79.16
Weight surgery			
Full-FT LLM + base Vision	66.69	74.60	<u>82.03</u>
Base LLM + Full-FT Vision	63.59	74.30	79.64
LLM-only LLM + Vision-only Vision	<u>67.15</u>	<u>74.88</u>	**82.58**

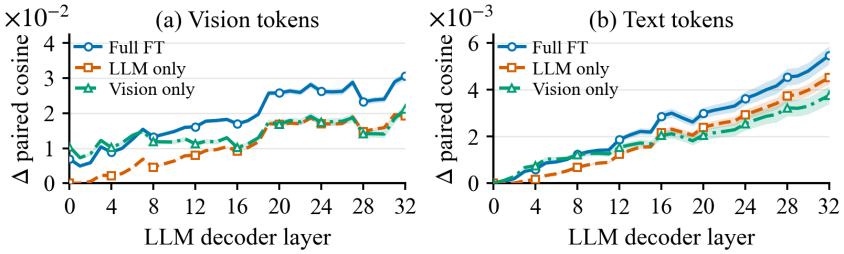


Figure 5: Layerwise change in the cosine similarity between clean- and degraded-view representations at corresponding input-token positions on R-Bench, relative to the Qwen3.5-4B base model. Panels (a) and (b) show visual and text token positions within the LLM decoder, respectively; shaded bands denote 95% confidence intervals.

ing scopes alter the relationship between clean and degraded inputs inside the decoder; the similarity curves should be interpreted alongside the performance and weight-recombination results.

Overall, language-side adaptation carries most of the gain in understanding under real-world degradations, while visual-side updates provide complementary benefits on some metrics. CURE also increases the similarity of clean and degraded representations within the language decoder, consistent with improved understanding of degraded inputs.

5 Conclusion

The value of image restoration depends not only on whether it can reconstruct a clear image, but also on whether the result helps downstream models understand the image correctly. However, when the downstream model is an LVLM, the common practice of restoring an image before passing it to the model for understanding may be misaligned with the model's needs. For these models, we shift the restoration target from image appearance to the ability to understand degraded images and propose CURE. Using the model's predictions on paired clean images as a reference, CURE guides the recovery of this ability, allowing the model to operate at inference time without generating an intermediate restored image. On a benchmark of real image degradations, CURE outperforms the evaluated pipelines that restore images before passing them to a model for understanding. In evaluations with synthetic degradations, it can also be combined with an image restoration front end to further improve understanding. These results show that image restoration for LVLMs can act directly on the model's understanding process while remaining compatible with methods that restore images at the pixel level. They also point toward more precise recovery of specific downstream task capabilities from degraded images.

AI USE STATEMENT

We used generative AI tools to aid and polish the English writing, assist with code generation, and support the retrieval and discovery of related work. We used Qwen3.6-35B-A3B as an LLM judge to score model responses in benchmark evaluations. AI-generated code was reviewed and tested for correctness by two authors. The authors reviewed the AI-assisted content and take responsibility for the final manuscript, code, and reported results.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6228–6237. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00652. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Gemma Team. Gemma 3 technical report. *CoRR*, abs/2503.19786, 2025. doi: 10.48550/ARXIV.2503.19786. URL https://doi.org/10.48550/arXiv.2503.19786.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14375–14385. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01363. URL https://doi.org/10.1109/CVPR52733.2024.01363.

Xiangzhao Hao, Zefeng Zhang, Zhenyu Zhang, Linhao Yu, Yao Chen, Yiqian Zhang, Haiyun Guo, Shuohuan Wang, and Yu Sun. CLEAR: unlocking generative potential for degraded image understanding in unified multimodal models. *CoRR*, abs/2604.04780, 2026. doi: 10.48550/ARXIV.2604.04780. URL https://doi.org/10.48550/arXiv.2604.04780.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-In-One image restoration for unknown corruption. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 17431–17441. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01693. URL https://doi.org/10.1109/CVPR52688.2022.01693.

Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiongkuo Min, Xiaohong Liu, Weisi Lin, Xiao-Ping Zhang, and Guangtao Zhai. R-bench: Are your large multimodal model robust to real-world corruptions? *IEEE J. Sel. Top. Signal Process.*, 19 (7):1349–1361, 2025a. doi: 10.1109/JSTSP.2025.3558652. URL https://doi.org/10.1109/JSTSP.2025.3558652.

Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement: A comprehensive survey. *Int. J. Comput. Vis.*, 133(11):8078–8108, 2025b. doi: 10.1007/S11263-025-02570-9. URL https://doi.org/10.1007/s11263-025-02570-9.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL https://doi.org/10.18653/v1/2023.emnlp-main.20.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Hashmat Shadab Malik, Fahad Shamshad, Muzammal Naseer, Karthik Nandakumar, Fahad Khan, and Salman Khan. Robust-llava: On the effectiveness of large-scale robust image encoders for multi-modal large language models. *CoRR*, abs/2502.01576, 2025. doi: 10.48550/ARXIV.2502.01576. URL https://doi.org/10.48550/arXiv.2502.01576.

OpenAI. GPT-4o system card. *CoRR*, abs/2410.21276, 2024. doi: 10.48550/ARXIV.2410.21276. URL https://doi.org/10.48550/arXiv.2410.21276.

Qwen Team. Qwen3-VL technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Sudarshan Rajagopalan, Kartik Narayan, and Vishal M. Patel. RestoreVAR: Visual autoregressive generation for all-in-one image restoration. *CoRR*, abs/2505.18047, 2025. doi: 10.48550/ARXIV.2505.18047. URL https://doi.org/10.48550/arXiv.2505.18047.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. Vlm-robustbench: A comprehensive benchmark for robustness of vision-language models. *CoRR*, abs/2603.06148, 2026. doi: 10.48550/ARXIV.2603.06148. URL https://doi.org/10.48550/arXiv.2603.06148.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine*

Learning, ICML 2024, Vienna, Austria, July 21-27, 2024, volume 235 of *Proceedings of Machine Learning Research*, pp. 43685–43704. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/schlarmann24a.html.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Parrot: Multilingual visual instruction tuning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/sun25ad.html.

Jiaqi Tang, Jianmin Chen, Wei Wei, Xiaogang Xu, Runtao Liu, Xiangyu Wu, Qipeng Xie, Jiafei Wu, Lei Zhang, and Qifeng Chen. Robust-r1: Degradation-aware reasoning for robust visual understanding. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 9421–9429. AAAI Press, 2026a. doi: 10.1609/AAAI.V40I11.37902. URL https://doi.org/10.1609/aaai.v40i11.37902.

Jiaqi Tang, Jianmin Chen, Youyang Zhai, Wei Wei, Runtao Liu, Mengjie Zhao, Xiangyu Wu, Qingfa Xiao, and Qifeng Chen. Robust-U1: Can MLLMs self-recover corrupted visual content for robust understanding? In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=I6W6cxVVts.

Xiaole Tang, Xiaoyi He, Xiang Gu, and Jian Sun. BaryIR: Learning multi-source unified representation in continuous barycenter space for generalizable all-in-one image restoration. *CoRR*, abs/2505.21637, 2025. doi: 10.48550/ARXIV.2505.21637. URL https://doi.org/10.48550/arXiv.2505.21637.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 7907–7915. AAAI Press, 2025. doi: 10.1609/AAAI.V39I8.32852. URL https://doi.org/10.1609/aaai.v39i8.32852.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13084–13094. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01243. URL https://doi.org/10.1109/CVPR52733.2024.01243.

xAI. RealWorldQA. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Shuo Xing, Lanqing Guo, Hongyuan Hua, Seoyoung Lee, Peiran Li, Yufei Wang, Zhangyang Wang, and Zhengzhong Tu. Demystifying the visual quality paradox in multimodal large language models. *CoRR*, abs/2506.15645, 2025. doi: 10.48550/ARXIV.2506.15645. URL https://doi.org/10.48550/arXiv.2506.15645.

Jiaqi Xu, Mengyang Wu, Xiaohu You, Chi-Wing Fu, Qi Dou, and Pheng-Ann Heng. Towards real-world adverse weather image restoration: Enhancing clearness and semantics with vision-language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVIII*, volume 15076 of *Lecture Notes in Computer Science*, pp. 147–164. Springer, 2024. doi: 10.1007/978-3-031-72649-1_9. URL https://doi.org/10.1007/978-3-031-72649-1_9.

Cuixin Yang, Rongkang Dong, and Kin-Man Lam. Vision-language model guided image restoration. *Image Vis. Comput.*, 173:106091, 2026. doi: 10.1016/J.IMAVIS.2026.106091. URL https://doi.org/10.1016/j.imavis.2026.106091.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *CoRR*, abs/2605.18740, 2026. doi: 10.48550/ARXIV.2605.18740. URL https://doi.org/10.48550/arXiv.2605.18740.