All-Weather VLM: Enhancing Vision Language Models' Robustness Under Adverse Imaging Conditions

Tianfu Wang[1*] **Mingyang Xie**[1*] **Haoming Cai**[1] **Tianyi Xiong**[1] **Xiyao Wang**[1]
Dongdong Fu[2] **Guan-Ming Su**[2] **Paola Cascante-Bonilla**[1,3†] **Christopher Metzler**[1†]

[1]University of Maryland [2]Dolby Laboratories Inc. [3]Stony Brook University

Abstract

Visual Language Models (VLMs) have shown strong multimodal inference capabilities, yet their robustness under real-world degradations remains underexplored. We study VLM performance under common adverse imaging conditions, including rain, fog, snow, haze, motion blur, low light, turbulence, and adherent raindrops, without prior knowledge of degradation type. Many of these conditions are critical for autonomous driving safety and related applications. We show that naive supervised fine-tuning fails to generalize across degradation types and that existing VLMs suffer substantial performance drops. A natural solution is to pre-process degraded images with restoration models before VLM inference; however, we show that these methods often hallucinate details that mislead the VLM and further hurt performance. Surprisingly, we found that skipping restoration and directly applying VLM to degraded images sometimes yields better results. Building on this finding, we propose a multi-stage training framework that integrates degradation awareness into the VLM reasoning process. A degradation-aware encoder predicts corruption type and severity to condition generation through a unified reasoning trace before the final response. Direct Preference Optimization (DPO) further improves factual grounding by treating responses from clean images as a more reliable reference for the scene's content, while penalizing degradation-induced errors. In visual question answering (VQA) tasks, our approach yields $\sim 10\%$ gains on real-world driving scenes and $\sim 4\%$ gains on real-world camera-degradation benchmarks for generic scenes, while showing that robustness can be maintained without sacrificing general capability.

1 Introduction

Vision-Language Models (VLMs) Hurst et al. (2024); Bai et al. (2025); Chen et al. (2024c); Li et al. (2024); Liu et al. (2023) have emerged as strong general-purpose reasoners, demonstrating strong zero-shot performance across image captioning, visual question answering, and multimodal reasoning tasks. However, these capabilities are mostly evaluated on clean, well-lit, artifact-free images, leaving a gap for real-world deployment. This gap is especially concerning for autonomous driving, a safety-critical downstream task. Under common adverse conditions, such as fog, heavy rain, low light, or atmospheric turbulence, structured visual corruption obscures semantic content and undermines the perceptual grounding VLMs rely on. Despite the frequency of these scenarios in practice, VLM robustness under such settings remains under-studied relative to the rapid progress on clean-image benchmarks. The consequences of this gap are concrete: a model that correctly identifies a pedestrian in clear conditions may misidentify them as a static object in thick fog, or hallucinate lane markings from a motion-blurred signal.

A natural solution is to preprocess degraded images with an all-in-one restoration model Luo et al. (2024); Li et al. (2025b); Guo et al. (2024); Jiang et al. (2024) before inference. We evaluate

[*]Equal contribution.
[†]Equal advising.

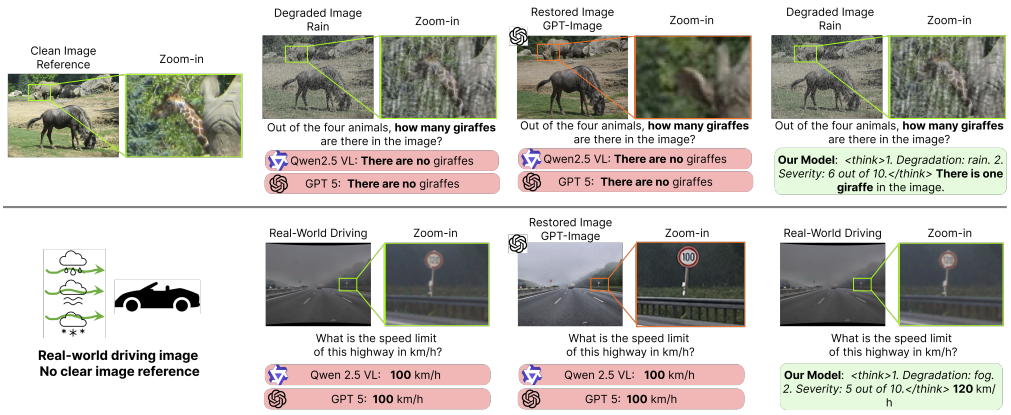


Figure 1: **Our model enhances robust visual reasoning under severe image degradation.** Standard VLMs are not robust to heavy weather degradation: failing to identify a giraffe in a rainy image or reading a speed limit through thick haze. Applying an image restoration model does not solve the problem—instead, restoration hallucinates away critical visual evidence, effectively removing the giraffe or misrendering a 120 km/h sign as 100 km/h. Our model instead first reasons about the degradation (type and severity) and then answers the visual question, enabling more reliable reasoning even under severe degradation.

this strategy and find it consistently underperforms. While restoration models produce visually plausible outputs, they often hallucinate textures and structures from corrupted signals. These artifacts induce greater semantic drift in the downstream VLM than the original degradations. This finding reframes the problem: rather than *attempting to recover* a clean image before reasoning, the model must *learn to reason* robustly through degradation.

In this paper, we propose a robust training framework that enables VLMs to accurately perceive and describe their surroundings even under severe visual interference from adverse weather conditions (Fig. 1). Naïvely fine-tuning on synthetically degraded images is not enough: it delivers only modest robustness improvements and can even degrade the base model's performance on clean images due to overspecialization. Instead, we integrate degradation awareness directly into the model's reasoning process through a multi-stage training strategy. Our approach introduces a degradation-aware encoder and a Unified Chain-of-Thought (CoT) Wei et al. (2022) mechanism, where the model is trained to identify the type and severity of visual corruptions within a dedicated reasoning process before generating a final response. To address hallucination failures that supervised fine-tuning alone cannot resolve, we apply Direct Preference Optimization (DPO) Rafailov et al. (2023), constructing preference pairs from clean and degraded views of the same scene and using clean-image responses as grounded references. Together, our multi-stage training pipeline yields a unified degradation-robust reasoning system.

For visual question answering (VQA) tasks, our proposed model yields $\sim$ 10% gains on real-world driving scenes Bijelic et al. (2020) and $\sim$ 4% gains on real-world camera-degradation benchmarks Li et al. (2025a) for generic images — while largely preserving performance on clean images, showing that the model maintains scene-level understanding without sacrificing the integrity of the original visual data or relying on external restoration.

We summarize the contributions of our paper as follows:

- We evaluate VLMs' VQA performance under 8 common image degradations (e.g., rain, snow, motion blur, and low light) and observe a significant performance drop.
- We show that a naive restoration-first pipeline (restore, then VLM) is ineffective and can even hurt downstream visual understanding due to hallucinated artifacts.
- We propose a unified multi-stage training pipeline with DPO to improve degradation-aware reasoning and robust visual understanding.

- We evaluate on both real-world autonomous driving scenes (Tab. 1 and Fig. 3) and real-world camera-degradation benchmarks for generic images (Tab. 2), where our model achieves 10% and 4% performance gains over the base model respectively.

2 Related Work

2.1 Vision Language Models

Vision language models (VLMs) extend large language models to perceive and reason over multiple modalities (e.g., text and images). Early works focus on representation-centric pretraining, where models learn general-purpose, aligned image-text embeddings at scale (Schuhmann et al., 2022; Li et al., 2022; 2023a; Zhai et al., 2023). Building on these foundational representations, the rise of LLMs shifts the focus from learning transferable embeddings to enabling general-purpose multimodal reasoning and instruction following. This motivates LLM-centric designs that couple a pretrained vision encoder with an LLM via lightweight, token-efficient connectors that preserve fine-grained visual detail without overwhelming context length (Alayrac et al., 2022; Liu et al., 2023; Bai et al., 2025; Xiong et al., 2025b). A complementary line of work focuses on post-training to elicit stronger reasoning and better alignment: chain-of-thought-style supervision (Kojima et al., 2022; Wei et al., 2022; Shao et al., 2024; Qin et al., 2025) encourages multi-step inference, while preference- and RL-based methods (e.g., DPO/GRPO-style training) (Guo et al., 2025; Rafailov et al., 2023; Xiong et al., 2025a; 2026) further shape outputs toward human preferences and more reliable long-horizon reasoning on complex tasks. In this paper, we improve adverse-weather robustness by using a degradation-aware encoder with unified CoT and DPO to align degraded and clean representations.

2.2 Adding robustness to Vision Language Models

Despite their impressive generative capabilities, vision-language models still exhibit fragility across the full perception-to-reasoning pipeline. Previous studies have highlighted the prevalence of hallucinations stemming from text-image misalignment, leading to methods Sun et al. (2024) that ground generations more strictly in visual evidence. Models also show contextual fragility, where irrelevant or misleading surrounding context distracts the model and degrades predictions; Qi et al. (2024) mitigates this by training with explicit distractors. A separate line of research studies robustness and safety under adversarially perturbed image inputs (Xu et al., 2024; Malik et al., 2025). In contrast, our work targets intrinsic visual robustness: we incorporate degradation awareness into internal llm reasoning while aligning clean and degraded image representations.

2.3 Vision Language Models under Adverse Imaging Conditions

Real-world deployments expose VLMs to diverse corruptions like motion blur and adverse weather, revealing a robustness gap even when clean-image performance is strong. While recent benchmarks evaluate this vulnerability under simulated conditions (Hendrycks & Dietterich, 2019; Qiu et al., 2025; Liao et al., 2025; Xie et al., 2025), efforts to directly address it remain limited. Recent works tackle robustness against camera-induced sensor degradations Li et al. (2025a); Tang et al. (2026), however native visual understading of more complex and severe adverse weather environments directly from raw inputs remains largely unexplored. Other works incorporate VLMs in downstream restoration pipelines Cai et al. (2025); Shao et al. (2025); Yang et al. (2024); Luo et al. (2024); Zhang et al. (2025); Chen et al. (2024a), but they use VLMs for auxiliary prompt generation or degradation evaluation. In contrast, our approach trains the VLM to explicitly perceive and reason through severe degradation, enabling robust multimodal understanding without relying on external restoration or preprocessing.

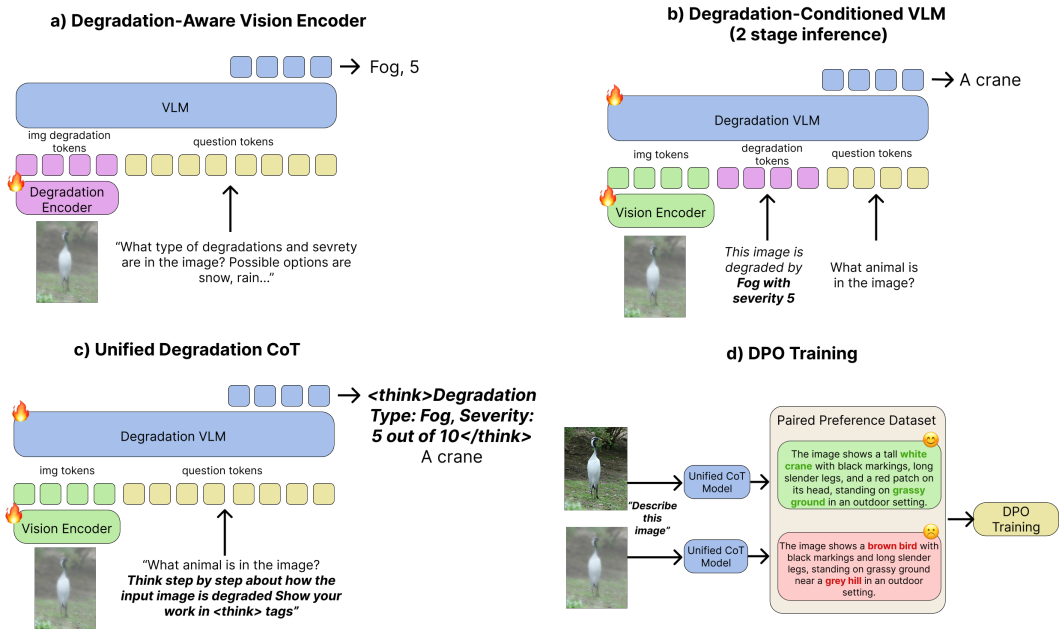


Figure 2: **Multistep training to a unified degradation-robust VLM.** We transform a pre-trained VLM into a single-pass degradation-robust reasoning pipeline through sequential fine-tuning. (a) Degradation encoder: we fine-tune the vision encoder to predict degradation type and severity from the degraded image. (b) Degradation-conditioned VLM: fine-tune the language model to perform captioning/QA conditioned on the explicit degradation tokens, yielding robustness but requiring two-step inference (predict degradation, then answer). (c) Unified degradation CoT: we unify the pipeline into one forward pass by training the model to first verbalize degradation type/severity within and then produce the final answer, using the generated reasoning trace as implicit conditioning.

3 Method

3.1 Problem Setting

We study robust visual understanding under adverse imaging conditions. Let x^c denote a clean image and $x^d = \mathcal{T}(x^c; t, s)$ its degraded counterpart, where t and s denote the degradation type and severity. At inference time, the model is given only x^d and a query q, without access to the clean image or degradation type.

Our goal is to train a vision-language model f_θ that produces a response y that remains semantically faithful to the underlying clean scene despite severe visual corruption: $y = f_\theta(x^d, q)$. A robust model should meet two goals: it should make grounded, informative predictions even when the image is partially corrupted, and it should still perform strongly on clean, standard inputs.

Restoration-first Approaches does not work. A natural baseline is to apply restoration before visual reasoning, but we have found that this restoration-first strategy does not reliably help due to haullucination from restoration itself. As we show in Fig. 1, even using strong closed-soucre image models, restoration can hallucinate away or distort critical visual evidence, misleading the downstream VLM further away from the correct semantic interpretation rather than closer to it. We later test state-of-the-art all-in-one restoration pipelines explicitly in the ablation study in Tab. 3, and it does not outperform direcly feeding the degraded image to the base model. This motivates a training pipeline in which degradation awareness and reasoning is built directly into the model internally, rather than handled by a separate restoration front-end.

Our approach converts a pretrained vision-language model into a degradation-aware reasoner. Our method has three components: (1) a degradation simulation library for generating large-scale paired clean/degraded training data, (2) supervised fine-tuning for injecting explicit degradation awareness into the model, and (3) Direct Preference Optimization (DPO) for reducing degradation-induced hallucinations.

3.2 Dataset curation - Degradation Simulation Library

The success of our training pipeline depends on a massive paired dataset of clean and degraded images, but diverse, accurately captioned degradation pairs at scale are scarce. To address this, we built a Python simulation library that parametrically synthesizes a range of degradations with light to severe intensities. Our library is designed to model real-world physical and optical phenomena in multiple categories to promote robust generalization, including adverse weather (fog, haze, procedural snow, motion-blurred rain streaks, adherent fisheye raindrops), low light, trajectory-based motion blur, and atmospheric optical turbulence. All degradation operators are parametric and support a range of intensities from light to severe. For dataset construction, we scale severity from 1 to 10, creating a smooth visual corruption curriculum and providing explicit degradation-type and severity labels required for subsequent training stages.

3.3 Degradation-Aware Visual Reasoning Pipeline

Our training recipe begins by establishing explicit degradation awareness through a sequential, two-stage evaluation and captioning framework. The primary objective of the first stage is to train a degradation-aware vision encoder, while the second stage uses degradation information from the first stage for robust general image understanding. Specifically, in the first stage, we directly fine tune the VLM's vision encoder so that it predicts the correct degradation type and severity of the input image. Following this perceptual alignment, the second stage introduces a conditional generation paradigm. In this phase, the language model is fine tuned to ingest the explicit degradation tokens produced by the first stage alongside the standard visual embeddings and textual query tokens. By grounding the generative process in the image's specific degradation state, the model gains a contextual prior that conditions downstream semantic tasks. This allows the system to dynamically adjust its interpretation of ambiguous visual boundaries and calibrate its confidence based on the identified interference.

While this sequential pipeline successfully establishes foundational degradation awareness, it inherently introduces a computationally inefficient, disjointed inference process. To resolve this limitation, we transition into a Unified Chain-of-Thought (CoT) pipeline, effectively collapsing the two-step inference into a single, cohesive forward pass. In this step, the VLM is instructed to generate a structured internal reasoning trail within explicit `<think>` tags immediately prior to answering the user's primary query. During this think phase, the model explicitly verbalizes the degradation type and a numerical severity rating. For clean inputs, the degradation taxonomy includes an explicit *no degradation* category, so the model is trained to generate this label with minimal severity in its Unified CoT trace rather than forcing an adverse degradation type. This explicitly verbalized Unified CoT serves as a vital cognitive anchor for the language model. Rather than passing degradation parameters as isolated control signals from a separate stage, the autoregressive generation of the reasoning trail effectively conditions the subsequent downstream text generation on the recognized degradation profile. By forcing the language model to sequentially articulate its understanding of the visual noise before attempting to extract semantic meaning, the network learns to dynamically adjust its semantic confidence and recalibrate its understanding of the scene, allowing the VLM to achieve robust understanding directly from the raw, degraded input.

3.4 Facilitating Robustness via Direct Preference Optimization

While Supervised Fine-Tuning effectively teaches the model to adopt the unified CoT format, it does not explicitly penalize the model for hallucinating non-existent objects obscured

GPT-4o Hurst et al. (2024)	LLaVA-OneVision-1.5-8B Li et al. (2024)	InternVL-2.5-8B Chen et al. (2024c)	Qwen2.5-VL Bai et al. (2025)	Ours
61.9	58.7	55.6	63.5	73.0

Table 1: **Performance on the curated real-world autonomous driving benchmark from Bijelic et al. (2020) (accuracy, %).** We construct a question-answering benchmark from *real-world* adverse-weather autonomous-driving video data containing snow, fog, low illumination, and mixed conditions with multiple co-occurring degradations. Our method outperforms all compared baselines.

	Cam-MCQ	Env-MCQ	Wild-MCQ	Cam-VQA	Env-VQA	Wild-VQA	Cam-CAP	Env-CAP	Wild-CAP	Overall
Qwen2.5-VL-7B	73.3	75.6	74.7	48.9	57.8	54.3	30.0	24.0	27.6	52.2
Ours	**76.7**	**83.3**	**80.0**	**56.7**	**61.1**	**59.3**	**30.3**	**27.3**	**29.1**	**56.1**

Table 2: **Robustness on real-world in-the-wild images in R-Bench Li et al. (2025a) wild group.** We evaluate performance under *real-world* camera degradations (e.g., lens blur, camera shake, occlusions) and environmental interferences (e.g., motion blur, low-light/dark illumination) across all task categories—multiple-choice (MCQ), visual question answering (VQA), and captioning (CAP)—reported on Cam/Env/Wild splits. Our method consistently outperforms the Qwen2.5-VL-7B baseline and with $\sim 4\%$ overall performance gains.

by severe degradation. To actively steer the model toward robust and reliable output, we introduce an additional reinforcement learning stage (Fig. 2d) utilizing Direct Preference Optimization (DPO) Rafailov et al. (2023). DPO provides a mechanism to directly reward factual accuracy while penalizing generative hallucinations commonly triggered by visual noise. To construct the preference dataset, we generate paired responses for both clean and degraded versions of the same image using an identical, open-ended prompt (e.g., "Describe what is in the image"). Rather than relying on narrow, task-specific questions that might mask errors, this generic prompt allows us to fully probe the model's comprehensive scene understanding and explicitly expose any degradation-induced hallucinations. The preferred response (y_w) is generated from the clean image, providing a hallucination-free baseline. The rejected response (y_l) is generated from the degraded image, capturing the base model's tendency to hallucinate. Crucially, during training, the visual input x provided to the model is always the degraded image. We optimize the language model using the standard DPO objective,

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right]$$

, where π_θ is the active training policy, π_{ref} is the reference unified CoT policy, and β is a hyperparameter controlling the deviation from the reference. Because each preference pair is constructed by applying the same generic, open-ended prompt to clean and degraded versions of the same image, the preferred response provides a holistic, comprehensive description of the scene, while the rejected response captures degradation-induced errors (semantic drift, or hallucination). Optimizing this objective therefore increases the likelihood of behavior that remains consistent with the clean image while suppressing failure modes introduced by visual noise. In this way, DPO directly reinforces factual accuracy and mitigates degradation-triggered hallucinations beyond what SFT alone can achieve.

4 Experiments

4.1 Experiment setup

We use Qwen-2.5-VL-7B Bai et al. (2025) as the base model and study whether a general-purpose VLM can be made robust to adverse imaging conditions without sacrificing its clean-image capability. We use LoRA fine-tuning on both the vision encoder and the language model, with rank 256 and alpha 128. For the degradation encoder, degradation-conditioned VLM, and unified degradation CoT, we perform synthetic fine tuning on the

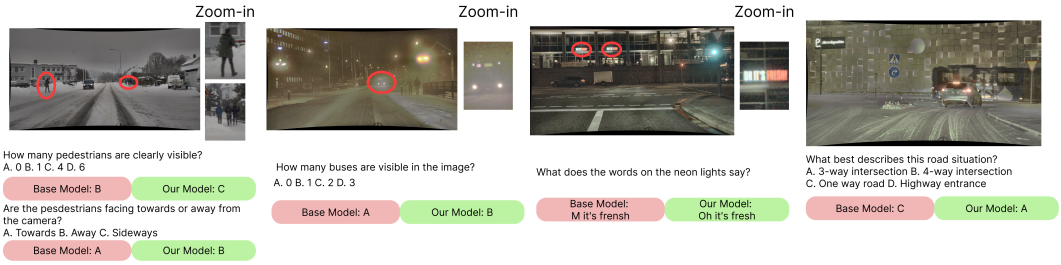


Figure 3: **Qualitative examples on real-world autonomous driving data from Bijelic et al. (2020).** We compare model performance on complex *real-world* autonomous driving scenarios captured under adverse weather conditions. Many of the examples are especially challenging as they exhibit multiple mixed degradations (snow + low light, haze + motion blur). For each manually annotated test case, we contrast the predicted outputs from the baseline Qwen-2.5-7B (red) against our proposed model (green). The comparisons demonstrate our model's robustness when processing severe environmental visual degradations.

	fog	haze	low light	motion blur	rain	raindrop	snow	turbulence
Qwen-2.5-VL-7B	61.7	58.7	59.7	57.4	55.8	55.7	51.2	55.4
+ Degradation Label	60.8	58.4	57.8	56.5	56.1	52.3	50.6	56.5
Restoration first								
+ DAClip Luo et al. (2024)	58.7	58.8	60.8	57.1	52.8	49.4	52.8	56.1
+ OneRestore Guo et al. (2024)	N/A	58.1	60.9	57.3	59.0	54.3	50.2	N/A
+ AutoDIR Jiang et al. (2024)	N/A	59.3	61.3	58.5	57.5	54.1	51.6	N/A
+ FoundIR Li et al. (2025b)	N/A	59.2	60.9	56.7	59.6	56.0	50.8	N/A
Ours CoT + DPO	**66.4**	**62.9**	**64.3**	**60.3**	**60.8**	**59.9**	**53.9**	**59.6**

Table 3: **Comparison between inference with degradation labels, restoration first methods, and our method on RealWorldQA XAI (2024) (accuracy, %).** For restoration first baselines, we only evaluate degradations that the restoration models are trained on. None of the restoration first methods outperforms the baseline across all degradation types. In contrast our method vastly outperforms these preporcessing baselines.

LoRA layers for 20,000 iterations with a batch size of 16 and learning rate of 1e-6. For DPO training, we first generate 20,000 caption pairs with images/question pairs from LLaVA-NeXT Liu et al. (2024b) Dataset and randomly sampled degradations. We then perform DPO training with a batch size of 4, a learning rate of 1e-6, and $\beta = 0.3$. All of our experiments are conducted on a 4xA6000 compute node.

4.2 Restoration first does not work

We first show in Tab. 3 that preprocessing without fine-tuning the base model is ineffective. Simply appending degradation labels does not reliably improve performance, as the model is not trained to meaningfully exploit such signals. We also evaluate several state-of-the-art all-in-one restoration pipelines as a preprocessing step to the degraded image, but none delivers consistent and significant gains over the base model across all degradation types.

4.3 Real-world VQA Results

Adverse weather autonomous driving scenes. We test our robustness on adverse weather on Bijelic et al. (2020), a real-world video driving dataset collected under adverse weather and low-visibility conditions. Since the dataset does not provide target QA annotations, we manually construct questions to probe scene understanding under degradation. Specifically we manually inspect video sequences captured under strong degradation conditons (i.e. low light, snow, fog, rain), select one image per scene, and manually label question sets consisting of multiple choice questions and short visual questions. Overall, our model outperforms the

	clean	fog	haze	low light	motion blur	rain	raindrop	snow	turbulence	Avg Deg
gpt4o Hurst et al. (2024)	68.0	59.6	59.5	62.5	59.5	55.2	55.9	52.9	49.8	56.9
InternVL2-8B Chen et al. (2024d)	65.3	59.6	55.0	61.5	54.3	53.2	52.9	50.7	51.7	54.9
InternVL2.5-8B Chen et al. (2024c)	69.8	63.6	59.2	63.5	58.3	57.2	55.8	53.3	55.7	58.3
LLaVA-1.5-7B Liu et al. (2024a)	56.1	52.1	53.1	57.8	51.9	45.5	48.1	50.6	52.9	51.5
LLaVa-OneVision-1.5-8B Li et al. (2024)	68.0	63.1	56.7	62.9	57.3	56.1	55.7	53.3	55.2	57.5
QWEN2-VL-7B Wang et al. (2024)	66.8	54.8	58.8	59.3	59.0	57.6	55.3	52.8	55.8	56.7
QWEN2-5-VL-7B Bai et al. (2025)	69.0	61.7	58.7	59.7	57.4	55.8	55.7	51.2	55.4	57.0
Ours CoT + DPO	**70.7**	**66.4**	**62.9**	**64.3**	**60.3**	**60.8**	**59.9**	**53.9**	**59.6**	**61.0**

Table 4: **RealWorldQA XAI (2024) performance under degradations (accuracy, %).** We report accuracy on clean images and across multiple degradation types (fog, haze, low light, motion blur, rain, raindrop, snow, turbulence). *Avg Deg* is the mean accuracy over all degraded conditions (excluding clean). Best and second-best results in each column are highlighted with **bold** and underline, respectively (ties share the same marking).

Benchmark	Method	clean	fog	haze	low light	motion blur	rain	raindrop	snow	turbulence	Avg Deg
Pope Li et al. (2023b)	Qwen2.5-VL-7B	87.4	79.9	80.3	81.1	76.1	76.0	79.0	73.7	80.5	78.3
	Ours	88.5	81.9	82.5	83.0	77.9	78.2	81.0	75.8	82.0	80.3
	Improvement	1.1	2.0	2.2	1.9	1.8	2.2	2.0	2.1	1.5	2.0
MMStar Chen et al. (2024b)	Qwen2.5-VL-7B	62.7	52.3	53.2	52.4	40.7	52.1	50.2	47.3	53.1	50.2
	Ours	62.5	53.7	55.8	54.6	44.0	54.7	52.1	49.5	55.9	52.5
	Improvement	-0.2	1.4	2.6	2.2	3.3	2.6	1.9	2.2	2.8	2.4
MMHal-Bench Sun et al. (2024)	Qwen2.5-VL-7B	3.63	2.31	3.15	2.91	3.05	2.10	2.80	2.45	2.59	2.67
	Ours	3.58	2.51	3.21	3.04	3.25	2.27	2.97	2.63	2.72	2.81
	Improvement	-0.05	0.20	0.06	0.14	0.20	0.17	0.17	0.18	0.13	0.14

Table 5: **Robustness evaluation on three highly challenging benchmarks POPE Li et al. (2023b), MMStar Chen et al. (2024b), and MMHal-Bench Sun et al. (2024)**. These benchmarks are difficult even on clean images. Our method consistently improves performance over the base model Qwen2.5-VL-7B, across degradation types.

Qwen base model by $\sim$ 10% in this evaluation set (Tab. 1). As we show in Fig. 3, compared with the baseline, our model produces more reliable answers in these real-world degraded scenes, including more accurate object counting and better understanding of road structure. This setting shows that our robustness gains transfer to practical conditions where reliable perception matters most.

Camera-induced degradations. We further evaulate our method on R-Bench Li et al. (2025a), which contains real in-the-wild corrupted images. This setting includes both environmental degradations, such as motion blur and low light, and camera degradations, such as lens blur, occlusion, and camera shake. Our method achieves $\sim$ 4% improvements compared to the baseline overall (Tab. 2).

4.4 Evaluation on Images with Synthetic Degrdations

In Tab. 4 we evaluate current vision-language models on RealWorldQA XAI (2024) under multiple synthetically added degradations types and find that robustness to degraded visual inputs remains a clear weakness of existing models. We evaluate our data using images with degradations higher than moderate severity (level 4 and above), since mild corruptions often do not sufficiently stress current VLMs and therefore understate the robustness gap. While strong baselines degrade substantially under adverse conditions, our final model achieves the best overall performance on degraded images while also improving clean-image accuracy. In Tab. 5 we evaluate the model on POPE Li et al. (2023b), MMStar Chen et al. (2024b), and MMHal-Bench Sun et al. (2024) under synthetic degradation. These benchmarks are already challenging even in their clean form, as they are designed to stress hallucination robustness, grounding, and fine-grained reasoning. Under added degradation, the difficulty increases further, yet our method consistently improves robustness without materially hurting clean-image performance.

We provide further qualitative examples in Fig. 4 on MMHal-Bench Sun et al. (2024). Compared with the baseline, our model remains better grounded and is less likely to drift toward

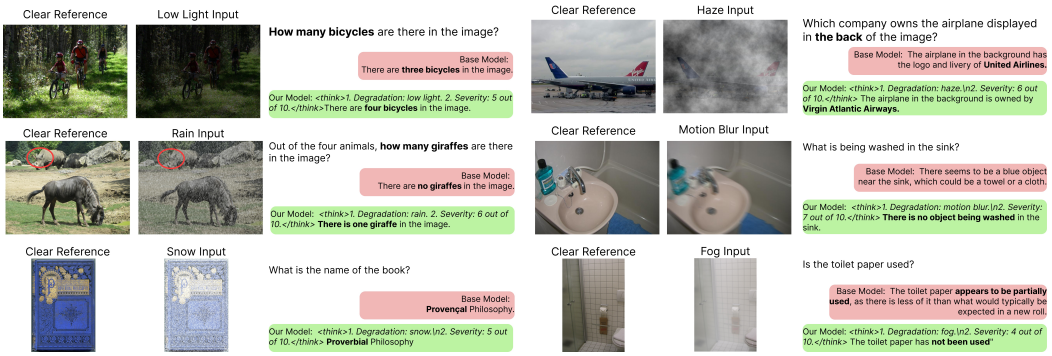


Figure 4: **Qualitative examples on MMHalBench Sun et al. (2024) with synthetically added degradations.** The models only receive degraded images as inputs. The clean images here are shown only as a reference and are not provided to the models. The colored text boxes report responses from the base model (red) and our model (green), both evaluated on the same degraded input, and show that our model is more robust against image degradations.

	clean	fog	haze	low light	motion blur	rain	raindrop	snow	turbulence	Avg Deg
Qwen-2.5-VL-7B	69.0	61.7	58.7	59.7	57.4	55.8	55.7	51.2	55.4	57.0
Simple SFT	68.6	63.0	59.5	61.2	58.6	57.0	56.3	51.7	56.3	58.0
Ours 2 Stage Eval	70.1	64.9	62.2	62.6	60.1	59.1	58.4	53.3	58.3	60.0
Improvement	1.1	3.2	3.5	2.9	2.7	3.3	2.7	2.1	2.9	3.1
Ours Unified CoT (Skip Stage 2)	69.3	65.2	60.9	62.3	58.5	59.3	58.0	51.8	57.5	59.2
Improvement	0.3	3.5	2.2	2.6	1.1	3.5	2.3	0.6	2.1	2.2
Ours CoT + DPO (Skip Stage 2)	70.0	65.0	61.8	62.7	57.9	60.1	58.3	53.5	58.3	59.7
Improvement	1.0	3.3	3.1	3.0	0.5	4.3	2.6	2.3	2.9	2.7
Ours Unified CoT	69.5	**66.5**	61.8	63.4	59.1	**60.9**	58.8	52.3	57.8	60.1
Improvement	0.5	4.8	3.1	3.7	1.7	5.1	3.1	1.1	2.4	3.2
Ours CoT + DPO	**70.7**	66.4	**62.9**	**64.3**	**60.3**	60.8	**59.9**	**53.9**	**59.6**	**61.0**
Improvement	1.7	4.7	4.2	4.6	2.9	5.0	4.2	2.7	4.2	4.0

Table 6: **Ablation study on RealWorldQA XAI (2024) performance across training stages (accuracy, %).** We compare the base Qwen-2.5-VL-7B model, Simple SFT (directly training on degraded image/text pairs without degradation reasoning), and our training variants with and without Stage 2. *Improvement* is measured relative to the base Qwen-2.5-VL-7B model. *Avg Deg* is the mean accuracy over all degraded conditions. Best and second-best results in each column are highlighted with **bold** and underline, respectively; ties receive the same highlighting.

plausible but incorrect interpretations. Additionally, we improve robustness across diverse tasks, including scene understanding, spatial reasoning, fine-grained recognition, and OCR.

4.5 Ablation on Training Strategy

We ablate the key components of our method in Tab. 6. Explicitly incorporating degradation reasoning into training already improves over simple SFT. The unified CoT formulation yields additional gains by merging degradation identification and semantic interpretation into a single reasoning process, enabling the model to solve both jointly rather than in isolation. Additionally, skipping Stage 2 reduces model robustness even after preference optimization, suggesting that the degradation-conditioned Stage 2 remains useful. Finally, RL/DPO achieves the best overall performance, despite being trained only on open-ended prompts without task-specific QA data. This highlights that preference optimization is crucial for anchoring the model to robust image semantics.

4.6 Discussion

Across all experiments, three findings emerge consistently. First, our method improves robustness across a broad range of real-world and synthetic degradations, not just a narrow subset. Second, these gains come with little to no penalty on clean images, and in several

cases even improve clean performance. Third, our method translates well to real-world data and degradations, showing that the benefits observed in controlled settings carry over beyond synthetic benchmarks. Together, the results indicate that robustness under adverse imaging conditions is best achieved not by external restoration, but by teaching the VLM to explicitly recognize degradation and reason through it in a grounded way.

5 Conclusion

We study robust visual reasoning for VLMs under adverse imaging conditions, where existing models often fail. We show that naive restoration-first inference is unreliable and can worsen downstream understanding by introducing hallucination-induced semantic drift. We therefore propose a unified multi-stage framework that combines degradation-aware reasoning with DPO. Across real-world autonomous-driving settings and generic camera-degradation benchmarks, our model consistently improves degraded-image performance while largely preserving clean-image capability.

6 Acknowledgement

We sincerely thank Yuancheng Xu for his valuable insights and suggestions that improved this project. T.W., and M.X. are funded in part by Dolby Laboratories - UMD Seed Grant Program. C.M. is funded in part by the UMD AIM Seed Grant Program, and NSF CAREER grant no. 2339616.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

Jie Cai, Kangning Yang, Jiaming Ding, Lan Fu, Ling Ouyang, Jiang Li, Jinglin Shen, and Zibo Meng. Degradation-aware image enhancement via vision-language classification. *arXiv preprint arXiv:2506.05450*, 2025.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024d.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiongkuo Min, Xiaohong Liu, Weisi Lin, et al. R-bench: Are your large multimodal model robust to real-world corruptions? *IEEE Journal of Selected Topics in Signal Processing*, 2025a.

Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Foundir: Unleashing million-scale training data to advance foundation models for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12626–12636, 2025b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Dacheng Liao, Mengshi Qi, Peng Shu, Zhining Zhang, Yuxin Lin, Liang Liu, and Huadong Ma. Robodrivevlm: A novel benchmark and baseline towards robust vision-language models for autonomous driving. *arXiv preprint arXiv:2512.01300*, 2025.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

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

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. *arXiv preprint arXiv:2502.01576*, 2025.

Zhiyuan Mao, Nicholas Chimitt, and Stanley H Chan. Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14759–14768, 2021.

Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jin Di, Yu Cheng, Qifan Wang, and Lifu Huang. Rora-vlm: Robust retrieval-augmented vision language models. *arXiv preprint arXiv:2410.08876*, 2024.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Xinkuan Qiu, Meina Kan, Yongbin Zhou, and Shiguang Shan. Benchmarking multimodal large language models against image corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9014–9023, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS Systems Datasets and Benchmarks Track*, 2022.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Qianyi Shao, Yuanfan Zhang, Renxiang Xiao, and Liang Hu. Vlm-augmented degradation modeling for image restoration under adverse weather conditions. In *2025 30th International Conference on Automation and Computing (ICAC)*, pp. 1–6. IEEE, 2025.

Vera Soboleva and Oleg Shipitko. Raindrops on windshield: Dataset and lightweight gradient-based detection algorithm, 2021.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

XAI. Grok-1.5 vision preview, 2024.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6585–6597, 2025.

Tianyi Xiong, Yi Ge, Ming Li, Zuolong Zhang, Pranav Kulkarni, Kaishen Wang, Qi He, Zeying Zhu, Chenxi Liu, Ruibo Chen, Tong Zheng, Yanshuo Chen, Xiyao Wang, Renrui Zhang, Wenhu Chen, and Heng Huang. Multi-crit: Benchmarking multimodal judges on pluralistic criteria-following, 2025a. URL https://arxiv.org/abs/2511.21662.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13618–13628, 2025b.

Tianyi Xiong, Shihao Wang, Guilin Liu, Yi Dong, Ming Li, Heng Huang, Jan Kautz, and Zhiding Yu. Phycritic: Multimodal critic models for physical ai. *arXiv preprint arXiv:2602.11124*, 2026.

Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. Shadowcast: Stealthy data poisoning attacks against vision-language models. *Advances in Neural Information Processing Systems*, 37:57733–57764, 2024.

Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24902–24912, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Tianpei Zhang, Jufeng Zhao, Yiming Zhu, and Guangmang Cui. Dual-domain perspective on degradation-aware fusion: A vlm-guided robust infrared and visible image fusion framework. *arXiv preprint arXiv:2509.05000*, 2025.

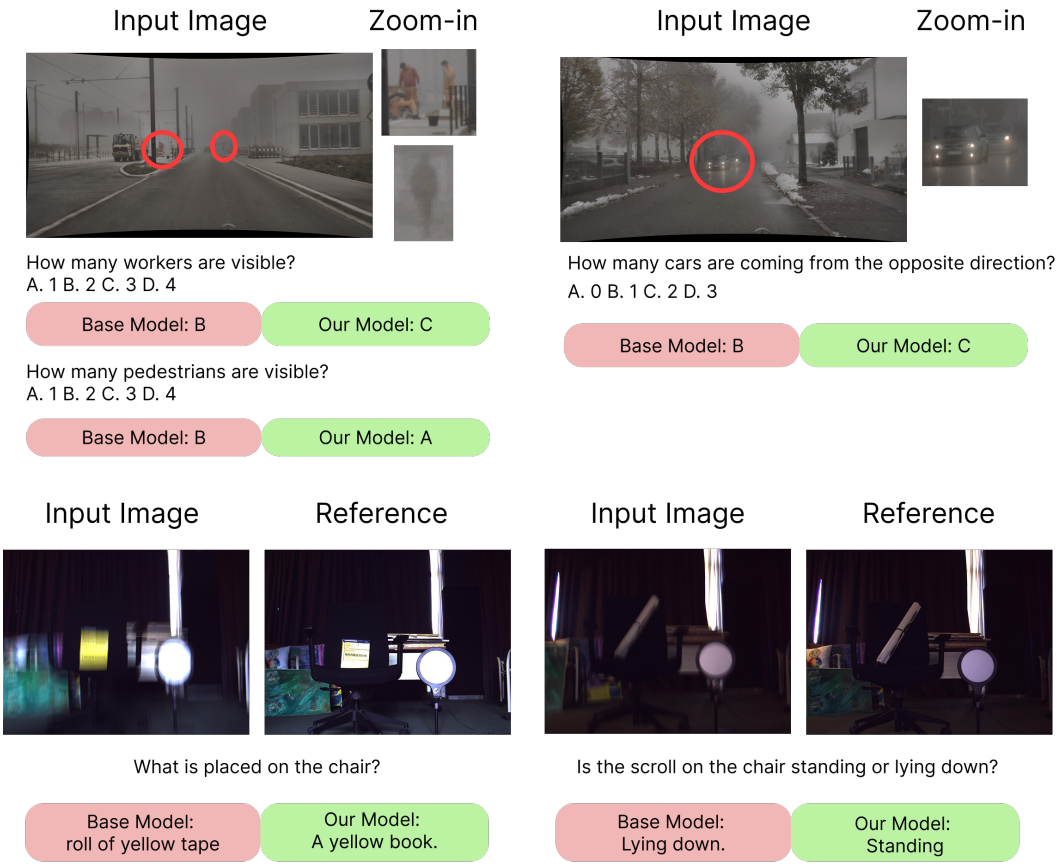


Figure 5: **Additional real-world Qualitative examples from Bijelic et al. (2020) and Li et al. (2025a)**. For each test case, we contrast the predicted outputs from the baseline Qwen-2.5 model (red) against our model (green). The comparisons demonstrate our model's enhanced robustness and accuracy when processing severe environmental visual degradations.

A Additional real world evaluation examples

In Fig. 5, we present additional qualitative examples from real-world degraded images in Li et al. (2025a) and Bijelic et al. (2020). Our method correctly identifies and counts objects under severe fog, rain, and motion blur, whereas the base model struggles under these conditions.

B More details on unified chain-of-thought

To encourage degradation-aware reasoning, we use a unified chain-of-thought format during training and inference. Concretely, we modify the system prompt to require the model to first identify the degradation category and estimate its severity before generating the final response. The system prompt is:

> You are an expert in analyzing degraded images. Upon receiving a new image, you must first reason inside `<think>` tags:
> 1. Identify degradation type.
> 2. Rate severity (1–10 from light to severe).
> Output your response after `</think>`.

Under this format, the model is required to begin its response with a structured reasoning prefix of the form

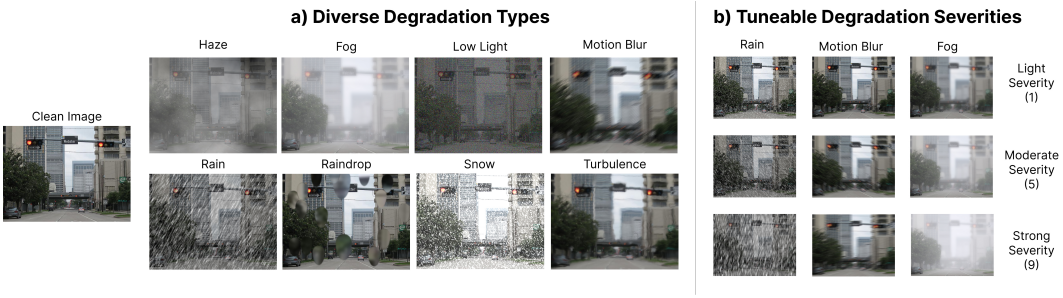


Figure 6: **Degradation Simulation Library**. We employ a physics-based simulation engine to generate a comprehensive set of weather conditions and degradations with adjustable severities. (a) Examples synthesized by our library across diverse degradation families, including adverse weather (fog, haze, rain streaks, adherent raindrops, snow), low light, motion blur, and atmospheric turbulence. (b) Each operator supports a set of severity parameters, which we discretize into 10 levels for dataset construction (examples shown at $s \in \{1, 5, 9\}$). This enables large-scale generation of degraded images with explicit supervision of degradation type t and severity s.

```
<think>Degradation type: <degradation_type>, Severity: <1-10></think>
```

before producing the final answer.

C More details on degradation simulation library

We provide more details on the physics-based simulation engine used to curate our training dataset (Fig. 6). Our library implements a diverse set of parametric degradations that model real-world optical and environmental phenomena. Each operator supports ten discretized severity levels, providing a smooth visual corruption curriculum for the model. **Haze** is generated by adding a procedural plasma-fractal overlay to the image, similar to Hendrycks & Dietterich (2019), followed by intensity renormalization. In practice, the simulator samples a fractal field and applies it as a veiling layer whose strength increases with severity. **Fog** is implemented as a depth-aware atmospheric scattering model. The code first estimates monocular relative depth using MiDaS Birkl et al. (2023) and then forms a transmission map $t = \exp(-\beta d)$, where farther regions receive stronger attenuation. The degraded image is synthesized by blending the clean image with a fixed atmospheric light color, and then applying an additional depth-weighted Gaussian blur so that distant regions become both lower contrast and less sharp. For **low-light** conditions, we apply a multi-stage degradation process involving brightness scaling, non-linear gamma adjustment, and additive Gaissoam noise. **Rain** is simulated with a noise-to-streak rendering pipeline. The simulator first samples sparse noise, thresholds it so only a subset of pixels survive, and then transforms that noise into elongated rain streaks using a rotated diagonal motion-blur kernel. The resulting streak map is alpha-composited back onto the image, while the background is slightly desaturated and attenuated to mimic reduced visibility under rainfall. **Raindrops** are treated separately from rain streaks. Instead of adding falling streaks, the simulator places discrete adherent water droplets directly on the lens/image plane Soboleva & Shipitko (2021). Each droplet is sampled with a random position and radius, an alpha map is generated for the droplet body and its darker boundary, and overlapping droplets are resolved with an explicit collision-handling step before compositing. As severity increases, both the droplet count and droplet size increase. We generate **snow** by sampling a random snow layer, enlarging it with zoom, thresholding it into visible flakes, and then applying motion blur to produce streaked or wind-swept snow particles. The snow layer is composited onto a partially desaturated/brightened image background, and a rotated copy is also added to increase particle diversity. As a result, snow combines both additive flakes and visibility reduction, rather than acting as pure white noise. Finally, for **Turbulence** we use the Phase-to-Space turbulence simulator of Mao et al. (2021). In this model, turbulence is

generated from turbulence-strength parameters and correlation statistics, producing both spatially varying blur and geometric warping.

All degradation operators are parametric and span a continuous range of intensities. For training, we discretize each degrdation into 10 manually designed severity levels, from weak but visible corruption at level 1 to severe perceptual degradation at level 10. Importantly, severity is not implemented as a fixed template: within each level, the simulator remains stochastic, so sampled corruptions still vary in their precise spatial patterns. This design ensures that severity increases degradation strength on average while preserving randomness, reducing simulator overfitting and providing explicit degradation-type and severity labels for downstream training.

During supervised fine tuning, we randomly sample a clean image/question pair and apply a randomly selected degradation with a randomly sampled severity level. For unified CoT training, we additionally provide the degradation type and severity in the prescribed output format, so the model learns to explicitly reason about both the corruption source and its strength. For the paired preference dataset in DPO, we feed both the clean and degraded images into the unified CoT model to construct preferred/rejected response pairs. The rejected response is obtained directly from the model's generated reasoning trace. For the preferred response, we edit the reasoning trace so that it correctly reflects the degradation type and severity associated with the degraded image.

D LLM disclosure

Our paper benchmarks and enhances the robustness of multimodal LLMs under degraded and adverse weather conditions. We use LLM-based models as baselines for evaluation and comparison. We also use closed-source GPT models as judges for more complex benchmarks, such as captioning. We additionally use LLMs to polish writing and brainstorm figure layouts.