DUALVISION: RGB-Infrared Multimodal Large Language Models for Robust Visual Reasoning

Abrar Majeedi[1], Zhiyuan Ruan[2], Ziyi Zhao[2], Hongcheng Wang[2], Jianglin Lu[3], Yin Li[1]
[1]University of Wisconsin-Madison, [2]Amazon, [3]Northeastern University

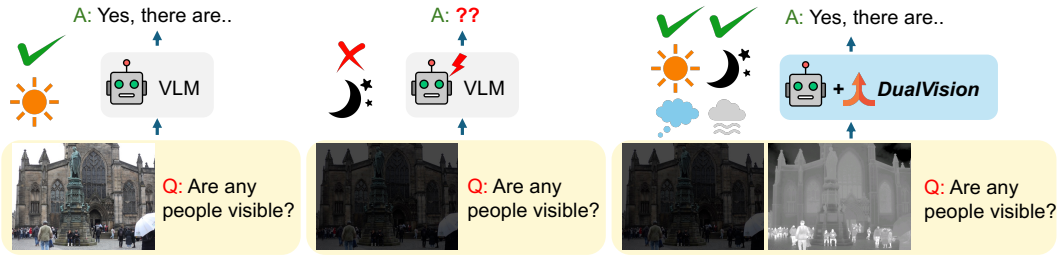

Figure 1. When visibility degrades in RGB images (*e.g.*, at night), MLLMs struggle to "see and reason", limiting their reliability in many real-world applications such as autonomous driving. By complementing RGB with infrared data, DUALVISION enables robust visual perception and reasoning while reducing computation by ∼75% compared to naïve fusion.

Abstract

Multimodal large language models (MLLMs) have achieved impressive performance on visual perception and reasoning tasks with RGB imagery, yet they remain fragile under common degradations, such as fog, blur, or low-light conditions. Infrared (IR) imaging, a well-established complement to RGB, offers inherent robustness in these conditions, but its integration into MLLMs remains underexplored. To bridge this gap, we propose DUALVISION*, a lightweight fusion module that efficiently incorporates IR–RGB information into MLLMs via patch-level localized cross-attention. To support training and evaluation and to facilitate future research, we also introduce* DV-204K*, a dataset of ∼25K publicly available aligned IR–RGB image pairs with* 204K *modality-specific QA annotations, and* DV-500*, a benchmark of* 500 *IR–RGB image pairs with 500 QA pairs designed for evaluating cross-modal reasoning. Leveraging these datasets, we benchmark both open- and closed-source MLLMs and demonstrate that* DUALVISION *delivers strong empirical performance under a wide range of visual degradations. Our code and dataset are available at https://abrarmajeedi.github.io/dualvision.*

1. Introduction

Multimodal large language models (MLLMs) [42] represent an increasingly important class of vision language models (VLMs) that connect visual and textual data through large language models (LLMs). MLLMs have achieved strong performance across many visual recognition tasks such as object recognition, visual grounding, and visual question answering, enabling broad applications in robotics [15], autonomous driving [7], and digital health [2]. Most existing MLLMs, however, rely exclusively on RGB imagery as their visual input, drawing from large-scale web datasets composed primarily of well-lit, natural scenes.

Although RGB imagery provides rich color and texture information and has driven impressive generalization in standard benchmarks, it exposes a critical weakness: their reliability drops sharply when inputs are degraded by adverse visual conditions [35]. This vulnerability stems from RGB's dependence on visible light and its susceptibility to optical distortions. Common examples include low-light environments, motion- or defocus-induced blur, and non-ideal weather such as rain or fog. These degradations are not rare anomalies but frequent realities in practical deployment, particularly in domains such as transportation, surveillance, and health, where robustness is paramount. For example, autonomous vehicles must sustain robust perception at night and in adverse weather, while home-based health monitoring systems must function effectively under poor lighting and motion blur.

Infrared (IR) imaging offers a valuable complement: by capturing electromagnetic radiation beyond the visible spectrum, IR can remain effective in darkness, fog, and other challenging environments faced by RGB imagery [10, 29]. However, IR imagery lacks the fine-grained appearance details and semantic richness that RGB captures

under favorable conditions. Fusing RGB and IR signals thus provides a promising pathway towards more robust visual perception and reasoning by leveraging their complementary strengths. This fusion has been widely explored in traditional vision tasks such as recognition, detection, and segmentation [40, 41, 43], and as a strategy for mitigating degradations through complementary sensing [28, 33]. However, its integration into MLLMs, particularly to overcome the limitations of RGB under *degraded visual conditions*, remains largely underexplored to date.

Developing MLLMs that can jointly comprehend RGB and IR data faces three key challenges. *First*, there is no principled design for a fusion mechanism that maintains spatial alignment between RGB and IR modalities while adaptively prioritizing the informative signals for MLLMs. *Second*, progress in IR-RGB perception is hindered by the scarcity of large-scale, semantically rich datasets. Existing datasets, often developed for detection or segmentation, tend to be narrow in scope, lack linguistic annotations, and are not aligned with the instruction-tuning paradigm that drives recent advances in MLLMs. *Third*, the absence of standardized IR-RGB benchmarks for vision language tasks leads to inconsistent evaluation, particularly under visual degradations, making it difficult to rigorously assess robustness across varying visual conditions.

In this paper, we address key challenges in developing IR-RGB MLLMs for robust visual reasoning. As illustrated in Fig. 1, we propose DUALVISION, a lightweight fusion approach that leverages multi-scale localized cross-attention to selectively route information between aligned IR and RGB tokens. Our design exploits the inherent spatial structure of visual data by employing progressively expanding local attention radii across different fusion levels. This hierarchical approach enables precise local correspondence at fine scales while capturing broader contextual relationships at coarser ones. The result is a unified IR-RGB representation that avoids the quadratic overhead of naïve concatenation, while the localized cross-attention mechanism provides inductive biases that strengthen cross-modal alignment, particularly under degraded RGB conditions.

Complementing our approach, we introduce DV-204K and DV-500, a new dataset suite designed to facilitate both instruction tuning and systematic evaluation of MLLM's robustness under visual degradations. Our datasets provide diverse, well-aligned IR-RGB image pairs with modality-aware question-answer pairs, allowing the study of both general reasoning and degradation-specific performance.

Our contributions are summarized as follows.

1. Our work is among the first to develop MLLMs that integrate RGB and IR modalities for robust visual reasoning under visual degradations (*e.g.*, blur, low-light, and fog).
2. We introduce DUALVISION, a lightweight IR-RGB fusion module with enhanced robustness to degradations, while remaining compatible with existing MLLMs.
3. To support training and evaluation of IR-RGB MLLMs, we create and release two datasets: 1) DV-204K: A dataset of ∼25K aligned IR–RGB image pairs with ∼204K modality-aware question-answer annotations designed for instruction tuning and 2) DV-500: A carefully curated evaluation benchmark featuring 500 IR-RGB image pairs with 500 associated QA pairs.
4. Leveraging our datasets and through extensive experiments, we demonstrate strong empirical results of DUALVISION under various degradations.

2. Related Work

VLMs and MLLMs. Modern VLMs align visual and textual modalities through contrastive and generative pre-training. CLIP [27] pioneers large-scale contrastive learning, enabling open-vocabulary recognition. Recent efforts, including ImageBind [9], LanguageBind [44] extend this paradigm to additional modalities including IR, depth and audio. However, these works focus on broad modality binding [9, 44], rather than more complex reasoning.

More recent MLLMs such as LLaVA [18], BLIP-2 [17], and Flamingo [1] integrate vision encoders with LLMs to support open-ended visual reasoning. Instruction tuning further improved alignment with human intent. InstructBLIP [8], MiniGPT-4 [45], and LLaVA-1.5 [19] demonstrated that curated instruction datasets enhance reasoning and zero-shot generalization. While most MLLMs rely solely on RGB imagery, a few very recent works, including IR-LLaVA [14] and IRGPT [6], have begun exploring IR-based MLLMs. However, these approaches operate on IR-only inputs and discard RGB, overlooking the complementary sensing capabilities of the two modalities. In contrast, we aim to develop IR–RGB MLLMs that fully exploit cross-modal synergy for robust visual reasoning.

Multimodal Fusion in MLLMs. Extending vision models beyond RGB requires effective fusion of complementary modalities. To integrate multiple modalities, recent MLLMs harness the flexibility of generic architecture, *e.g.*, Transformer [36]. In these models, signals from each modality are first tokenized using modality-specific encoders; the resulting tokens are then interleaved and passed to the Transformer, where the self-attention mechanism performs multimodal fusion [24].

However, self-attention scales quadratically with token count [36], thus rendering vanilla concatenation based multimodal fusion computationally expensive. Efficient variants such as local or sparse attention reduce cost but often weaken cross-modal alignment. Seminal works, such as Swin Transformer [20] mitigate this via windowed self-attention, though it remains intra-modal. Extensions like SwinFusion [22] achieve IR–RGB fusion for reconstruction tasks but are heavy and limited to low-level vision. More

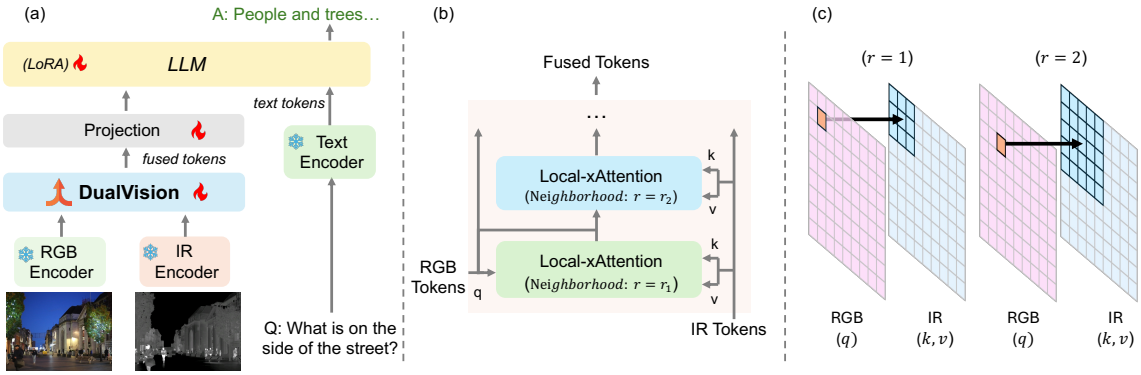

Figure 2. Overview of **DUALVISION**. **(a)** shows how DUALVISION integrates into a MLLM to fuse RGB and IR image tokens for robust visual reasoning. **(b)** illustrates the multi-scale localized cross-attention module, where RGB tokens serve as queries and IR tokens as keys and values. **(c)** visualizes the spatially aligned RGB–IR token grids over which localized cross-attention is performed.

recent multimodal systems explore lightweight and flexible fusion designs. PandaGPT [31] performs element-wise addition of visual and textual tokens for compact fusion, while our prior work [21] focuses on video-LLMs and employs cross-attention blocks to incorporate additional modalities (*e.g.*, audio or depth) without retraining the base architecture, enabling scalable multimodal integration.

Robustness to Visual Corruptions. The vulnerability of vision systems to common image corruptions has been systematically documented by [11], who introduced the ImageNet-C benchmark with 19 corruption types across four categories: noise, blur, weather, and digital distortions. Usama et al. [35] extended this analysis to VLMs, revealing vulnerability patterns across different tasks. To address such vulnerabilities, prior work has largely pursued two directions: degradation-aware processing and input restoration. Quality-aware networks [38] adapt feature extraction to estimate corruption levels and leverage that information during inference, while restoration-based methods [32] attempt to reconstruct clean inputs prior to inference. While effective for single-modality vision, these approaches add inference cost and fail to exploit complementary sensing.

IR-RGB for Robust Perception. Cross-modal sensing offers a promising approach to enhance robustness by combining complementary modalities. IR-RGB fusion has proven effective for object detection [39] and semantic segmentation [40], especially in autonomous driving [5], where thermal imaging compensates for RGB limitations in low-light conditions. Yet, this success has not translated to MLLMs, where the difficulty lies not only in perception but also in cross-modal reasoning and grounding.

A key barrier is the absence of suitable datasets. Existing vision-language benchmarks rely solely on RGB data, while robustness datasets [11, 35] test corruption within a single modality. Although IR-RGB datasets exist for detection and segmentation [13, 30], there is an unmet need for IR-RGB datasets with question-answer pairs necessary for multimodal reasoning evaluation.

To bridge the gaps, we propose DUALVISION, an effective and efficient IR-RGB fusion module designed for MLLMs. We also create new datasets for training and evaluating IR-RGB reasoning under challenging conditions.

3. Design of DUALVISION

DUALVISION, as shown in Fig. 2, presents a lightweight RGB-IR fusion module designed for MLLMs. Instead of interleaving tokens from IR and RGB, DUALVISION performs *multi-scale localized cross-attention*, allowing each RGB patch token to attend only to spatially corresponding IR regions. DUALVISION injects complementary IR cues where they are relevant, has low computational overhead, and remains compatible with many existing MLLMs.

Let $X_{\text{RGB}} \in \mathbb{R}^{H\times W\times 3}$ and $X_{\text{IR}} \in \mathbb{R}^{H\times W\times 1}$ denote spatially aligned RGB and IR inputs. A pre-trained vision encoder E then maps each input to a sequence of patch tokens:

$$Z^{\text{RGB}} = E(X_{\text{RGB}}), \quad Z^{\text{IR}} = E(X_{\text{IR}}), \tag{1}$$

with $Z^{\text{RGB}}, Z^{\text{IR}} \in \mathbb{R}^{N\times d}$. Each token corresponds to a patch feature that can be traced back to a specific location in the 2D image plane. With these patch coordinates, we define a local neighborhood for each RGB token and performs *local cross-attention* over the corresponding IR tokens.[1] As shown in Fig. 2 (c), each RGB token attends only to IR tokens within a radius-r region centered at its location, enforcing spatially aligned fusion. This design injects IR cues precisely where they are informative, while reducing the compute cost of global cross-modal interaction.

2D Local Cross-Attention. Let $\{z_u^{\text{RGB}}\}_{u=1}^N$ and $\{z_v^{\text{IR}}\}_{v=1}^N$ denote the RGB and IR token embeddings, and let $\{p_u^{\text{RGB}}\}$, $\{p_v^{\text{IR}}\} \subset \mathbb{Z}^2$ denote their patch coordinates. The neighborhood of an RGB token in the IR token grid is defined as

$$\mathcal{N}_r(u) = \left\{ v \in \{1,\dots,N\} : \rho\left(p_v^{\text{IR}}, \phi(p_u^{\text{RGB}})\right) \le r \right\}, \tag{2}$$

[1] A shared encoder implies implicit alignment, however, if separate encoders are employed, token grids can be aligned by interpolation.

where ρ is the 2D Euclidean distance, and ϕ maps RGB coordinates to the IR grid if resolutions differ.

Leveraging the neighborhood $\mathcal{N}_r(u)$, each RGB token serves as the query while keys and values are drawn from the IR tokens in its restricted neighborhood. We then compute the projected query, key, and value as:

$$q_u = z_u^{\text{RGB}} W_Q, \;\; K_u = Z^{\text{IR}}_{\mathcal{N}_r(u)} W_K, \;\; V_u = Z^{\text{IR}}_{\mathcal{N}_r(u)} W_V,$$

with $W_Q \in \mathbb{R}^{d\times d_k}$, $W_K \in \mathbb{R}^{d\times d_k}$, and $W_V \in \mathbb{R}^{d\times d_v}$.

Consequently, the attention weights are computed and applied to obtain the fused representations:

$$\begin{aligned} \alpha_u &= \text{softmax}\left(\frac{q_u K_u^\top}{\sqrt{d_k}}\right), \quad o_u = \alpha_u V_u, \\ \tilde{z}_u^{\text{RGB}} &= z_u^{\text{RGB}} + o_u W_O \end{aligned} \tag{3}$$

where $W_O \in \mathbb{R}^{d_v\times d}$. Stacking the fused tokens over all RGB tokens yields the fused representation $Z_{\text{fused}} \in \mathbb{R}^{N\times d}$.

Multi-Scale Cross-Attention. To capture interactions across local regions with varying size while preserving locality, we employ multiple 2D local cross-attention blocks sequentially with progressively increased radii $r_1 \le \cdots \le r_L$. This is initialized as $Z^{(0)} = Z^{\text{RGB}}$, and updated at each layer $\ell = 1, \ldots, L$, through local cross-attention:

$$\begin{aligned} \hat{Z}^{(\ell)} &= Z^{(\ell-1)} + \text{xAttn}_{r_\ell}\Big(\text{LN}\Big(Z^{(\ell-1)}\Big), Z^{\text{IR}}\Big), \\ Z^{(\ell)} &= \hat{Z}^{(\ell)} + \text{FFN}\Big(\text{LN}\Big(\hat{Z}^{(\ell)}\Big)\Big), \end{aligned} \tag{4}$$

where xAttn is the 2D local cross-attention in Eq. 3, and LN and FFN denote LayerNorm and MLP, respectively. The final fused representation is given by $Z_{\text{fused}} = Z^{(L)}$.

This hierarchical design, together with skip connections, preserves the local spatial alignment while allowing the model to have different receptive fields rather than being confined to a single arbitrarily fixed neighborhood. In practice, we rely on 3 sequential blocks with $r \in [1, 2, 3]$.

Interface with LLM. To interface the fused IR-RGB tokens Z_L with a LLM, we employ a linear projection P, trained jointly with the fusion module. During training, a LoRA adapter [12] is attached to the LLM and only its parameters Θ_{LoRA} are updated. At inference time, fused tokens Z_L are first projected and then concatenated with text tokens Q from a text query q. The resulting multimodal sequence is then passed to the LLM, which autoregressively generates the answer sequence:

$$y_{1:T} \sim \text{LLM}(P(Z_{\text{fused}}), Q). \tag{5}$$

Training Objective. Model training follows the standard next-token prediction loss over the answer sequence conditioned on both IR and RGB image and the text query q:

$$\mathcal{L}(\Theta) = -\Sigma_{t=1}^{T} \log p_\Theta\big(y_t \mid X_{\text{RGB}}, X_{\text{IR}}, q, y_{<t}\big), \tag{6}$$

Metric	Base Model	DUALVISION
Parameters (B)	∼7	+0.05 (+0.7%)
TFLOPs	9.24	+0.06 (+0.6%)

Table 1. **Compute cost and parameter count**: DUALVISION introduces negligible new parameters and computational cost, while maintaining competitive performance.

where $\Theta = \{\Theta_{\text{LoRA}}, P, \{W_Q, W_K, W_V, W_O\}_{1:L}\}$.

Data Augmentation. To enhance robustness in degraded conditions, we adopt a degradation-aware training scheme. Specifically, with probability p_d (=0.25 in our experiments), a degradation type $\delta \in \{\text{blur}, \text{darkness}, \text{fog}\}$ with severity $s \in \{\text{low}, \text{moderate}, \text{high}, \text{highest}\}$ is applied to the RGB image during training, *i.e.*, $X'_{\text{RGB}} = T_{\delta,s}(X_{\text{RGB}})$, which explicitly encourages the model to rely more on the IR features when RGB is unreliable.

Computational Efficiency. MLLMs with Transformers [36] have a quadratic cost to their input sequence length. RGB-IR concatenation yields $2N$ tokens with cost $\mathcal{O}((2N)^2) = \mathcal{O}(4N^2)$. In contrast, DUALVISION fuses both modalities into N tokens, reducing the cost to $\mathcal{O}(N^2)$, a 4× reduction. The fusion step uses local cross-attention with complexity $\mathcal{O}(Nw^2)$ for window size $w < N$, which is negligible compared to the repeated LLM attention blocks. This computational advantage is reflected in practice: as shown in Table 1, DUALVISION adds almost no parameters while substantially reducing TFLOPs, all while maintaining competitive performance. Implementation details are provided in the supplement.

4. DV-204K and DV-500 Datasets

To train and evaluate DUALVISION, we further introduce DV-204K and DV-500, two comprehensive resources for advancing multimodal understanding across infrared and RGB imagery. DV-204K is an instruction-tuning dataset of IR–RGB image pairs with fine-grained question–answer annotations, while DV-500 is a curated benchmark set for assessing modality-specific reasoning and robustness under controlled degradations. Importantly, DV-204K is created automatically using an agentic annotation framework.

4.1. Agentic Framework for IR-RGB Annotation

Building DUALVISION requires a way to generate rich, modality-grounded annotations associated with IR-RGB image pairs. These textual annotations must accurately reflect the unique characteristics of IR-RGB perception. A central obstacle is that IR imagery lacks large-scale captioning and QA resources, unlike the RGB domain where such datasets are abundant. Existing efforts [14] often sidestep this limitation by synthesizing IR data from RGB images and heuristically adapting RGB captions. However, this strategy relies on imagined thermal appearances rather than genuine measurements, leading to annotations that could

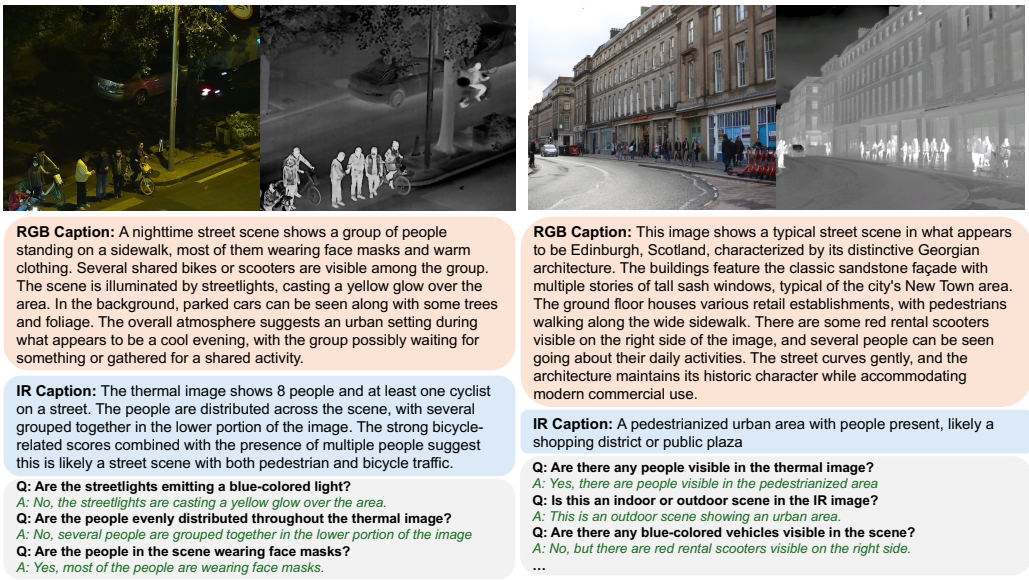

Figure 3. **RGB, IR Captions and QA Pairs.** We show examples of paired RGB–IR images along with their modality-specific captions and the corresponding question–answer pairs generated from those captions. The RGB caption includes richer scene details such as lighting, clothing, and context, while the IR caption reflects high level information like presence of people and overall scene type.

misrepresent IR-specific cues such as temperature gradients, emissivity patterns, and heat-based contrast. These systematic biases highlight the need for a more principled approach to annotation grounded in real IR data.

Agentic Framework for Captioning. To address this gap and inspired by [3], we introduce a modality-aware *agentic annotation framework* in which LLMs act as agents that iteratively generate and refine annotations under the supervision of a pretrained contrastive model. Crucially, unlike prior work that re-purposes RGB captions, our approach operates directly on real IR imagery, ensuring annotations reflect modality-specific content. The pipeline (illustrated in the Supplement), proceeds in three stages:

1. *Candidate Generation:* An LLM (Claude Sonnet 3.5 v2 [34]) generates a diverse set of candidate captions based on minimal input cues or object-detection labels when available.
2. *Contrastive Refinement:* A modality-specific contrastive model (IR LanguageBind [44]) evaluates the candidate text–IR image alignment and assigns similarity scores. These scores guide the LLM through multiple rounds (9 in our case) of iterative refinement in a closed-loop process, progressively improving accuracy and relevance.
3. *Final Selection:* Finally, the refined candidate captions, along with their contrastive scores, are synthesized by a stronger LLM (Claude Opus) into a final caption that best captures the semantics of the IR input image.

To strike a favorable balance between performance and compute, we employ the more efficient Claude Sonnet during the multi-iteration refinement loop and reserve Claude Opus for a single final reasoning pass. Empirically, this configuration provides a strong accuracy–efficiency trade-off. Using this framework, we automatically generate captions for ∼25K IR images in LLVIP [13] and HDRT [26].

	Score			
Metric	**Very Good**	**Good**	**Fair**	**Poor**
Accuracy	873 (51.1%)	609 (35.6%)	164 (9.6%)	64 (3.7%)
Detail	2 (0.1%)	807 (47.2%)	807 (47.2%)	94 (5.5%)

Table 2. **Quality of IR captions** is evaluated using an LLM-as-a-judge protocol, comparing IR captions against reference RGB captions from paired images. Most IR captions score *Good* or better in accuracy, while descriptive detail is lower, consistent with the reduced visual cues in infrared imagery.

Assessing the Quality of Generated Captions. We further evaluate the quality of captions from our agentic framework. Directly evaluating IR captions is challenging due to the lack of ground-truth benchmarks for infrared imagery. To address this, we apply an *LLM-as-a-judge* evaluation on a random subset of ∼1700 paired IR–RGB images.

This is achieved by the following steps. *First*, leveraging the strong RGB captioning of modern MLLMs, high-quality RGB reference captions are generated (using Claude Sonnet 3.5v2). *Second*, a judging LLM (a different instance of Sonnet 3.5v2) then scores the corresponding IR captions against these references while being instructed to account for modality-specific visibility differences. Each IR caption is rated along two dimensions: *Accuracy* (faithfulness to IR-visible content) and *Detail* (descriptive completeness).

As summarized in Table 2, IR captions demonstrate strong overall quality. Most (86.7%) achieve "Good" or "Very Good" ratings for accuracy, showing that they reliably capture IR scene content. About half reach similar

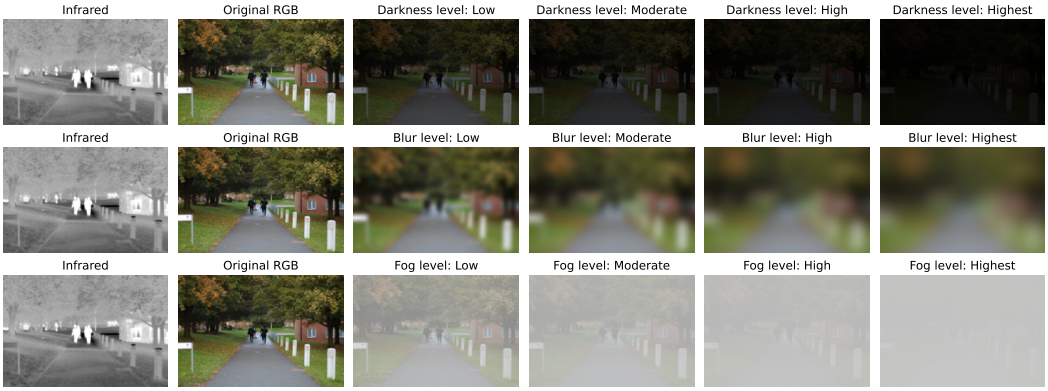

Figure 4. **RGB degradation conditions in DV-500.** Examples of the three corruption types: darkness, blur, and fog, applied at four severity levels each, with IR left unaltered. These controlled degradations enable systematic evaluation of the robustness of VLMs.

ratings for detail, as IR imagery inherently provides less fine-grained information than RGB. These results confirm that our framework produces informative and faithful IR descriptions directly from the IR images.

From Captions to QA Pairs. Finally, we convert each caption into 2-4 QA pairs using an LLM (Claude Sonnet 3.5 v2). Every QA pair is tagged with the modality (IR or RGB) that provides the key visual evidence, enabling modality-aware supervision. RGB questions target fine-grained appearance cues (*e.g.*, "*What color is the car?*"), while IR questions focus on properties reliably visible in infrared, such as object count or coarse scene layout (*e.g.*, "*How many people are visible?*"). Sample IR-RGB images, their captions and converted QA pairs are shown in Fig. 3.

4.2. DV-204K for Instruction Tuning

We applied our annotation pipeline to ∼9.5K aligned IR–RGB image pairs from the HDRT dataset [26] (reserving 500 pairs for testing) and an additional ∼15K pairs from the LLVIP [13] dataset. These two sources are complementary: HDRT captures diverse environmental settings, while LLVIP emphasizes low-light urban scenes.

Our DV-204K dataset contains ∼25K aligned IR–RGB pairs with ∼204K QA annotations, averaging 8.1 QA pairs per image. Following [19], QA pairs are formulated in an open-ended, instruction-tuning format (*e.g.*, "What color is the car?" → "The car is blue."). Importantly, questions are evenly divided between RGB- and IR- dependent cues, ensuring that learned models are exposed to both modality-specific reasoning and cross-modal alignment.

DV-204K *is intended for instruction tuning of MLLMs*, offering large-scale IR-RGB data for modality-aware reasoning and robust multimodal integration.

4.3. DV-500 for Evaluation

A key goal of IR–RGB fusion is maintaining accuracy when the RGB modality is degraded. To evaluate this behavior under controlled conditions, we introduce DV-500, a dataset of 500 aligned IR–RGB image pairs paired with 500 QA items. Each question is constructed to require complementary information from both modalities, semantic structure from IR and fine-grained appearance cues from RGB. This design allows us to assess if IR images can compensate for degraded RGB images, while still leveraging any remaining texture or color information.

For objective scoring, open-ended questions are converted into binary yes/no statements, following [3, 25]. Images are sampled from HDRT for high-quality RGB references and degradation is applied. To simulate visual degradation while preserving IR signals, we apply corruptions only to RGB images, since long-wave infrared sensors measure emitted thermal radiation and are largely unaffected by many visible-light degradations. Darkness does not alter thermal emission; optical blur primarily impacts visible-spectrum optics; and fog or haze scatters short wavelengths far more than long-wave IR. However, we acknowledge that this setup assumes idealized IR robustness under these conditions. We implement three corruption types (*darkness*, *blur*, and *fog*), each at four severity levels, plus a clean condition, yielding 13 evaluation categories (Fig. 4). Details are provided in the Supplement.

DV-500 *is intended for evaluation*, revealing how well models integrate complementary IR and RGB cues under diverse degradation types and severity levels.

5. Experiments and Results

We present a comprehensive evaluation of DUALVISION. We outline the baselines and evaluation metrics, followed by systematic ablations that examine the impact of IR-RGB fusion and the effects of model design choices. These analyses establish the final version of DUALVISION used for comparison with state-of-the-art MLLMs.

Baselines. We benchmark against both open- and closed-source MLLMs. Open-source baselines include LLaVA 1.5–7B [19], Qwen2-VL 7B [37], LLaVA-Next Interleave 7B [16], and LLaMA-4 Scout [23]. For completeness, we

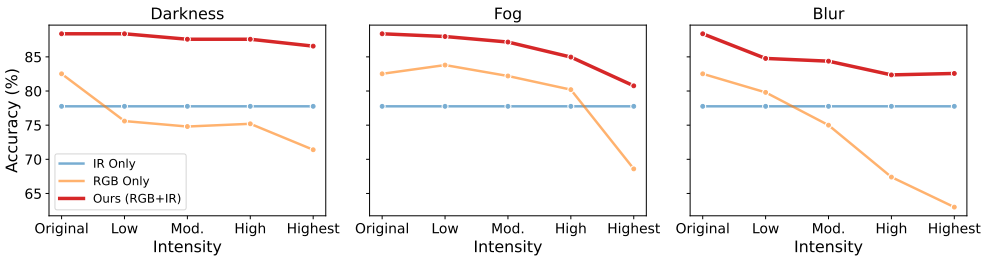

Figure 5. **Performance by Modalities (IR, RGB, RGB+IR)**. IR-only performance stays flat across degradations, RGB-only performance drops sharply as degradation severity increases, and RGB+IR provides the most robust results. Full results are in the Supplement.

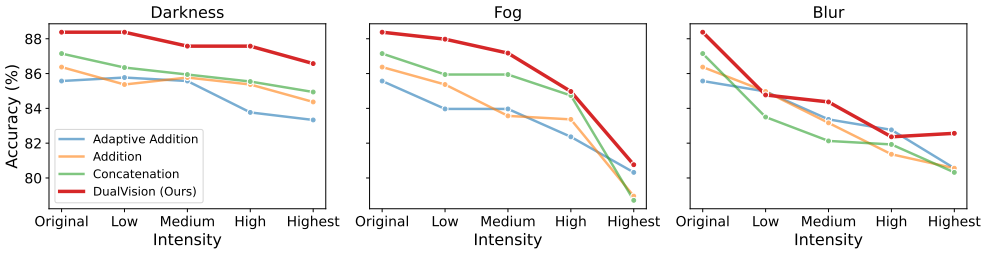

Figure 6. **Comparison of Fusion Methods.** We compare several fusion strategies: addition, adaptive addition, concatenation, and our DUALVISION. Note that the concatenation baseline is equivalent to finetuned LLaVA-1.5-7B. Full results can be found in the Supplement.

also evaluate some closed-source commercial systems such as Anthropic Claude Sonnet 3.5v2 and Claude Opus 4 [34]. All baselines are evaluated with RGB and IR images at a resolution of 336×336 to ensure a fair comparison.

To study the role of IR-RGB fusion, we evaluate LLaVA 1.5-7B variants using IR-only, RGB-only, and their combination under identical conditions. To assess the effectiveness of our fusion design, we implement several popular fusion strategies, each trained on DV-204K with identical hyperparameters and the same degradation-aware protocol. These include token-wise addition of embeddings, adaptive weighted addition (where learnable token weights are applied before summation), and the commonly used concatenation (interleaving tokens), followed by linear projection. All fusion variants share the LLaVA 1.5-7B backbone and the frozen CLIP ViT-$L/14$ encoder E, ensuring that performance differences arise solely from the fusion mechanism.

Training Details. Models are trained for 2 epochs on the DV-204K dataset, which contains both RGB- and IR-oriented QA pairs. Training is performed on 8× Nvidia A100 GPUs (80 GB RAM) with FlashAttention for memory efficiency, and DeepSpeed for further training optimization. We use a per-device batch size of 16 and learning rates of 10^{-4} for the fusion module and 10^{-5} for P, along with linear warm-up followed by cosine annealing. Under this setup, training completes in roughly one hour on DV-204K.

Metrics and Evaluation Protocol. Evaluation is conducted on DV-500. Since answers in DV-500 are binary (*i.e.*, yes/no), exact-match accuracy serves as the primary metric. We specifically analyze the setting where both modalities are provided and the question requires information from each. In this context, IR contributes stable semantic structure, while RGB supplies fine-grained texture and color cues. This design allows us to probe whether IR can compensate when RGB becomes degraded, while still leveraging whatever residual color or texture information remains in the corrupted RGB input. Following the DV-500 protocol, degradations applied to RGB images include blur, darkness, and fog, each at four severity levels, in addition to clean images. Results are stratified by corruption type and severity to enable fine-grained robustness assessment.

5.1. Design and Ablation Study

Effects of Modalities. We first analyze the contribution of each modality to overall robustness. The RGB-only model is evaluated using the pretrained LLaVA 1.5 7B weights without additional finetuning, while the IR-only variant is finetuned on DV-204K using only IR images for fairness. As in Fig. 5, single-modality models perform substantially worse under both clean and degraded settings. The IR-only model attains the lowest accuracy on DV-500, and its performance remains flat across degradations since no RGB input is available. The RGB-only model performs well on clean inputs but deteriorates rapidly as corruption severity increases. In contrast, integrating both modalities yields consistent improvements across all degradation types. The widening gap between the RGB-only model and our approach under degradation highlights the importance of multimodal fusion for robust perception.

Fusion Design. We compare DUALVISION with several established IR-RGB fusion methods to assess its effectiveness in multimodal integration. Specifically, we implement three canonical fusion strategies: token-wise addition, adaptive weighted addition, and concatenation, each finetuned on DV-204K with identical hyperparameters and the

Method	#Params	#Tokens	Original	Blur				Darkness				Fog			
				Low	Mod.	High	Highest	Low	Mod.	High	Highest	Low	Mod.	High	Highest
w/o Finetuning															
LLaVA 1.5–Text Only [19]	7B	–	48.8	–	–	–	–	–	–	–	–	–	–	–	
LLaVA 1.5 [19]	7B	2N	81.2	79.0	76.0	73.2	72.6	74.0	72.8	73.2	71.2	81.0	79.2	78.4	71.4
Qwen2-VL [37]	7B	2N	89.8	77.8	73.6	70.8	69.4	89.6	85.6	82.6	78.4	85.0	79.4	75.2	65.6
Qwen2.5-VL [4]	7B	2N	**90.2**	75.0	65.0	61.0	60.6	**89.4**	83.8	80.6	72.8	85.8	78.2	71.4	61.2
LLaVA-Next Interleave Qwen [16]	7B	2N	88.6	81.4	78.4	75.2	73.6	86.4	86.0	85.6	81.4	85.0	83.8	79.8	73.4
LLaMA-4 Scout [23]	17B	2N	85.8	71.6	62.6	59.2	57.6	79.4	71.6	66.4	56.6	73.4	63.6	58.0	54.0
Claude Opus 4 [34]	–	2N	86.6	67.0	60.8	60.0	58.8	83.6	72.4	64.6	57.0	74.2	63.8	59.2	56.2
Claude Sonnet-3.5 v2 [34]	–	2N	87.4	77.8	74.8	70.8	72.0	85.4	78.8	75.6	68.0	80.0	70.2	68.4	64.4
Finetuned															
LLaVA 1.5 [19]	7B	2N	87.15	83.50	82.13	81.93	80.32	86.35	85.94	85.54	84.94	85.94	85.94	84.74	78.71
DUALVISION (ours)	7B	**N**	88.38	**84.77**	**84.37**	**82.36**	**82.57**	88.38	**87.58**	**87.58**	**86.57**	**87.98**	**87.17**	**84.97**	**80.76**

Table 3. **Main results on DV-500**. We compare our method against strong open-source and commercial MLLM baselines under clean and corrupted conditions. Our approach achieves the highest overall accuracy and shows improved robustness against degradations.

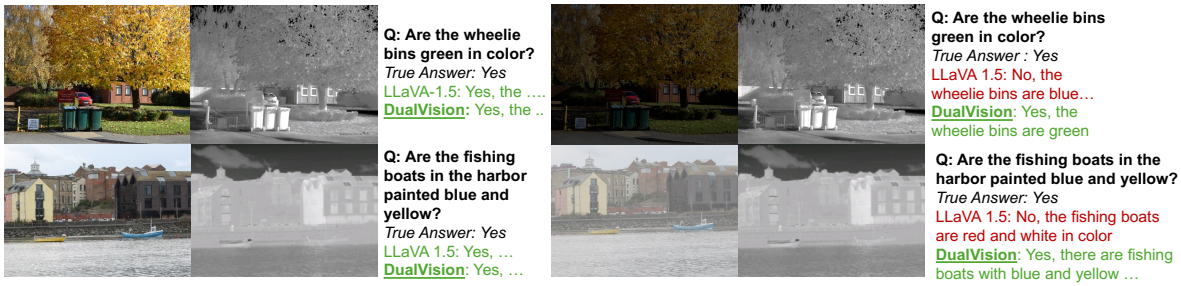

Figure 7. **Results on DV-500.** Both methods, finetuned on DV-204K, answer correctly with clean RGB–IR inputs (left). When the RGB is degraded (right), we can see DUALVISION remains robust, while finetuned LLaVA-1.5 shows reduced reliability.

same degradation-aware training protocol. All models share the LLaVA 1.5–7B backbone and frozen CLIP encoder E, ensuring that performance differences arise solely from the fusion mechanism. As illustrated in Fig. 6, DUALVISION delivers the best performance, winning $11/13$ evaluation settings over both clean and degraded RGB inputs. Detailed results can be found in the Supplement.

Additional Ablation Studies. We also conduct detailed analyses of both our attention design and degradation-aware training. More details are provided in the Supplement.

5.2. Comparison with Baselines

Table 3 summarizes results for the baselines, including open-source VLMs (LLaVA variants, Qwen2-VL) and large closed-source systems. As a sanity check, we also evaluate LLaVA 1.5 7B without providing any visual input, and it performs at chance level (48.8%), confirming dataset balance. When provided with RGB and IR inputs, all VLMs achieve strong performance on clean data, with Qwen2-VL, LLaVA-Next, and closed-source models consistently reaching the mid-to-high 80% range. This demonstrates that modern MLLMs can effectively interpret visual information under ideal conditions. However, performance drops sharply under degradations such as blur, darkness, or fog, with most baselines losing 15–20 percentage points in accuracy. Finetuning LLaVA 1.5 7B on DV-204K using our degradation-aware protocol substantially improves robustness, confirming the benefit of training under degraded visual conditions. Nonetheless, DUALVISION consistently outperforms this finetuned baseline and all other VLMs across most settings. These results highlight that while degradation-aware learning enhances general robustness, the fusion and attention mechanisms in DUALVISION remain crucial for achieving state-of-the-art multimodal resilience. Qualitative comparisons in Figure 7 further illustrate DUALVISION's superior ability to preserve semantic fidelity under challenging visual conditions.

6. Conclusion

This paper presents one of the first efforts to develop MLLMs that integrate IR and RGB images for robust visual perception and reasoning under various visual degradations. We introduced DUALVISION, a lightweight IR-RGB fusion module for MLLMs that performs multi-scale localized cross-attention, enabling effective interaction between modalities. Together with our DV-204K for instruction tuning and DV-500 for evaluation, DUALVISION establishes an effective solution for IR-RGB reasoning. Our experiments demonstrated that the performance of current MLLMs drops significantly under adverse degradations (*e.g.*, blur, low light, and fog), whereas DUALVISION delivers consistent improvements across all settings. Future work could investigate end-to-end finetuning, misalignment-tolerant fusion, and broader multimodal corruption benchmarks, along with real-world evaluations.

Acknowledgment. This material is partially based on work supported by the Army Research Office funded Assured Autonomy Innovation Institute (A2I2) under Contract number W911NF-2020-221, and the National Science Foundation under Grant Number CNS-2333491. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736. Curran Associates, Inc., 2022. 2

[2] Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505, 2024. 1

[3] Kumar Ashutosh, Yossi Gandelsman, Xinlei Chen, Ishan Misra, and Rohit Girdhar. LLMs can see and hear without any training. In *ICML*, 2025. 5, 6, 3

[4] 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 e-prints*, pages arXiv–2502, 2025. 8

[5] Martin Brenner, Napoleon H. Reyes, Teo Susnjak, and Andre L. C. Barczak. RGB-D and Thermal Sensor Fusion: A Systematic Literature Review. *IEEE Access*, 11:82410–82442, 2023. 3

[6] Zhe Cao, Jin Zhang, and Ruiheng Zhang. IRGPT: Understanding Real-world Infrared Image with Bi-cross-modal Curriculum on Large-scale Benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 166–176, 2025. 2

[7] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 958–979, 2024. 1

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in neural information processing systems*, pages 49250–49267, 2023. 2

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *CVPR*, 2023. 2

[10] Martin Grabner and Vaclav Kvicera. The wavelength dependent model of extinction in fog and haze for free space optical communication. *Optics Express*, 19(4):3379–3386, 2011. Publisher: Optica Publishing Group. 1

[11] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*, 2018. 3

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[13] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. 3, 5, 6

[14] Shixin Jiang, Zerui Chen, Jiafeng Liang, Yanyan Zhao, Ming Liu, and Bing Qin. Infrared-LLaVA: Enhancing Understanding of Infrared Images in Multi-Modal Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8573–8591, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2, 4

[15] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025. 1

[16] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models, 2024. arXiv:2407.07895 [cs]. 6, 8

[17] 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 *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. ISSN: 2640-3498. 2

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 2

[19] 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*, pages 26296–26306, 2024. 2, 6, 8, 3

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. ISSN: 2380-7504. 2

[21] Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, and Yin Li. PAVE: Patching and adapting video large lan-

guage models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3306–3317, 2025. 3

[22] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. Publisher: IEEE/CAA Journal of Automatica Sinica. 2

[23] Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025. 6, 8

[24] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention Bottlenecks for Multimodal Fusion. In *Advances in Neural Information Processing Systems*, pages 14200–14213. Curran Associates, Inc., 2021. 2

[25] Thuy Nguyen, Dang Nguyen, Hoang Nguyen, Thuan Luong, Long Hoang Dang, and Viet Dac Lai. OWLViz: An Open-World Benchmark for Visual Question Answering, 2025. arXiv:2503.07631 [cs]. 6

[26] Jingchao Peng, Thomas Bashford-Rogers, Francesco Banterle, Haitao Zhao, and Kurt Debattista. HDRT: A large-scale dataset for infrared-guided HDR imaging. *Information Fusion*, 120:103109, 2025. 5, 6

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. ISSN: 2640-3498. 2, 3

[28] Mani Ramanagopal, Sriram Narayanan, Aswin C Sankaranarayanan, and Srinivasa G Narasimhan. A theory of joint light and heat transport for lambertian scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11924–11933, 2024. 2

[29] Young-Sik Shin and Ayoung Kim. Sparse Depth Enhanced Direct Thermal-Infrared SLAM Beyond the Visible Spectrum. *IEEE Robotics and Automation Letters*, 4(3):2918–2925, 2019. 1

[30] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. PST900: RGB-Thermal Calibration, Dataset and Segmentation Network. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9441–9447, 2020. ISSN: 2577-087X. 3

[31] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One Model To Instruction-Follow Them All. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 11–23, Prague, Czech Republic, 2023. Association for Computational Linguistics. 3

[32] Shangquan Sun, Wenqi Ren, Tao Wang, and Xiaochun Cao. Rethinking image restoration for object detection. *Advances in Neural Information Processing Systems*, 35:4461–4474, 2022. 3

[33] Huixuan Tang, Xiaopeng Zhang, Shaojie Zhuo, Feng Chen, Kiriakos N Kutulakos, and Liang Shen. High resolution photography with an rgb-infrared camera. In *2015 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2015. 2

[34] Anthropic Team. Anthropic Claude LLMs, 2025. 5, 7, 8

[35] Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. Analysing the Robustness of Vision-Language-Models to Common Corruptions, 2025. arXiv:2504.13690 [cs]. 1, 3

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2, 4

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, 2024. arXiv:2409.12191 [cs]. 6, 8

[38] Qiangchang Wang and Guodong Guo. CQA-Face: Contrastive Quality-Aware Attentions for Face Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2504–2512, 2022. 3

[39] Qishun Wang, Zhengzheng Tu, Kunpeng Wang, Le Gu, and Chuanwang Guo. Mixture of Scale Experts for Alignment-free RGBT Video Object Detection and A Unified Benchmark, 2025. arXiv:2410.12143 [cs]. 3

[40] Yike Wang, Gongyang Li, and Zhi Liu. SGFNet: Semantic-Guided Fusion Network for RGB-Thermal Semantic Segmentation. *IEEE Trans. Cir. and Sys. for Video Technol.*, 33 (12):7737–7748, 2023. 2, 3

[41] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 2

[42] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403, 2024. 1

[43] Xingchen Zhang, Ping Ye, Henry Leung, Ke Gong, and Gang Xiao. Object fusion tracking based on visible and infrared images: A comprehensive review. *Information Fusion*, 63:166–187, 2020. 2

[44] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5, 3

[45] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 2

DUALVISION: RGB-Infrared Multimodal Large Language Models for Robust Visual Reasoning

Supplementary Material

This supplementary document provides expanded experimental results and analyses that complement the main paper. We present additional ablations that examine our design (Section A), provide further quantitative and qualitative results omitted from the main paper due to space limits (Section B), introduce details about the construction of our datasets (Sec. C), and describe implementation details of DUALVISION (Section D).

For sections, figures and equations, we use numbers (*e.g.*, Sec. 1) to refer to the main paper and capital letters (*e.g.*, Sec. A) to refer to this supplement.

A. Additional Ablations

We present ablation studies examining the attention design of DUALVISION as well as the contribution of degradation-aware training. These experiments highlight the effectiveness of our design choices and demonstrate their importance for achieving robust multimodal fusion.

Attention Design. We dissect the contributions of the central design choice *i.e.*, the attention mechanism design. Specifically, we compare global cross-attention, fixed-radius local attention (*i.e.*, $r = 1$), and our multi-scale local attention. As shown in Table A, the multi-scale variant achieves the best overall performance, outperforming alternatives in 8 of 12 degraded settings. Fixed-radius attention occasionally matches or slightly exceeds our approach on clean data (88.58% vs. 88.38%), suggesting that simpler fusion may suffice under ideal conditions. However, as degradations intensify, multi-scale attention consistently demonstrates stronger robustness, validating the hypothesis that localized interactions enable more effective cross-modal integration.

Degradation-Aware Training. We assess the role of training with stochastic degradations. Table B shows that degradation-aware training universally improves performance. Under severe blur, DUALVISION improves by +6.47% (82.57% vs. 76.10%); under the highest darkness level, by +4.81%; and under severe fog, also by +4.81%. These results demonstrate that exposure to corrupted inputs during training equips the model with more resilient fusion behaviors, enabling better generalization under adverse visual conditions.

B. Detailed Results

We provide the detailed numerical results corresponding to the modality ablation and fusion experiments shown in Figures 5 and 6 of the main paper. These tables list the exact accuracy values used to generate the plots, covering all degradation types and severity levels.

Figure A. Sample results of DUALVISION under degradations.

Effect of Modalities. Table C reports the detailed accuracy of the RGB-only, IR-only, and RGB–IR variants evaluated in Figure 5 of the main paper, covering all blur, darkness, and fog levels. The results detail how each model responds as degradations intensify, showing the characteristic drop in performance for RGB-only reasoning and the limited overall accuracy of IR-only predictions. In aggregate, the combined RGB–IR model maintains the strongest and most stable performance across the full degradation spectrum.

Fusion Design. Table D provides the full accuracy breakdown for the fusion mechanisms as illustrated in Figure 6 of the main paper, including simple addition, adaptive weighted addition, concatenation, and our proposed DUALVISION. The table shows how each method behaves under clean and progressively degraded RGB inputs, exposing where robustness differences emerge among the baselines. Across all corruption types and severity levels, DUALVISION achieves the highest overall performance among the evaluated fusion strategies.

Additional Qualitative Results. Figure A showcases the model's ability to produce more extended and detailed reasoning in its responses, while Figure B provides additional examples that highlight the robustness of DUALVISION across a range of degradations.

C. Dataset Details

We illustrate the agentic annotation framework and provide additional details on the degradation simulation procedures used throughout our experiments.

C.1. Agentic Framework for Captioning

Our three-stage annotation framework introduced in Section 4.1 of the main paper is further illustrated in Figure C. At each iteration, the LLM (Claude Sonnet 3.5v2) proposes caption candidates, receives similarity-based feedback via IR-CLIP, and produces improved captions, ultimately con-

Method	Wins	Original	Blur				Darkness				Fog			
			Low	Moderate	High	Highest	Low	Moderate	High	Highest	Low	Moderate	High	Highest
Global xAttention	1	88.18	83.94	81.96	81.76	80.96	87.37	87.58	**87.58**	85.57	86.17	86.37	84.37	78.92
Local xAttention (r=1)	5	**88.58**	**84.97**	83.37	**83.17**	81.96	**88.78**	**87.78**	87.17	85.97	87.17	86.17	84.17	80.56
DUALVISION (ours)	**8**	88.38	84.77	**84.37**	82.36	**82.57**	88.38	87.58	**87.58**	**86.57**	**87.98**	**87.17**	**84.97**	**80.76**

Table A. **Ablation on Attention Design**. Accuracy (%) is reported across corruptions on DV-500. All models use 3 blocks. *Global xAttn* uses full cross-attention; *Local xAttn ($r = 1$)* uses fixed local neighborhoods; DUALVISION applies multi-scale local xAttn ($r \in \{1, 2, 3\}$).

Method	Wins	Original	Blur				Darkness				Fog			
			Low	Moderate	High	Highest	Low	Moderate	High	Highest	Low	Moderate	High	Highest
DualVision (w/o Deg.-Aware Training)	0	88.18	84.17	79.56	76.91	76.10	86.97	85.57	84.17	81.76	86.57	83.97	81.76	75.95
DualVision (w/ Deg.-Aware Training)	**13**	**88.38**	**84.77**	**84.37**	**82.36**	**82.57**	**88.38**	**87.58**	**87.58**	**86.57**	**87.98**	**87.17**	**84.97**	**80.76**

Table B. **Ablation on Degradation-Aware Training.** Accuracy (%) is reported across corruption types and severities on DualVision-500.

Method	Original	Blur				Darkness				Fog			
		Low	Mod.	High	Highest	Low	Mod.	High	Highest	Low	Mod.	High	Highest
LLaVA 1.5 7B-IR Only (Finetuned)	77.76	-	-	-	-	-	-	-	-	-	-	-	-
LLaVA 1.5 7B-RGB Only (OOB)	82.52	79.80	75.00	67.40	63.00	75.60	74.80	75.20	71.40	83.80	82.20	80.20	68.60
DUALVISION (ours)	**88.38**	**84.77**	**84.37**	**82.36**	**82.57**	**88.38**	**87.58**	**87.58**	**86.57**	**87.98**	**87.17**	**84.97**	**80.76**

Table C. **Ablation on Different Modalities**. Accuracy (%) is reported across corruption types and severities on DualVision-500. All models use the LLaVA 1.5 7B [19] backbone.

Method	Wins	Original	Blur				Darkness				Fog			
			Low	Mod.	High	Highest	Low	Mod.	High	Highest	Low	Mod.	High	Highest
Addition	1	86.37	84.97	83.17	81.36	80.56	85.37	85.77	85.37	84.37	85.37	83.57	83.37	78.96
Adaptive Addition	2	85.57	**84.97**	83.37	**82.77**	80.56	85.77	85.57	83.77	83.33	83.97	83.97	82.36	80.32
Concatenation†	0	87.15	83.50	82.13	81.93	80.32	86.35	85.94	85.54	84.94	85.94	85.94	84.74	78.71
DUALVISION (ours)	**11**	**88.38**	84.77	**84.37**	82.36	**82.57**	**88.38**	**87.58**	**87.58**	**86.57**	**87.98**	**87.17**	**84.97**	**80.76**

Table D. **Ablation on Fusion Strategies**. Accuracy (%) is reported across corruption types and severities on DualVision-500. Note: All methods are finetuned with the same training data, settings as well as degradation aware training protocol. †Equivalent to LLaVA1.5 7B [19] finetuned on DV-204K with interleaved RGB and IR tokens.

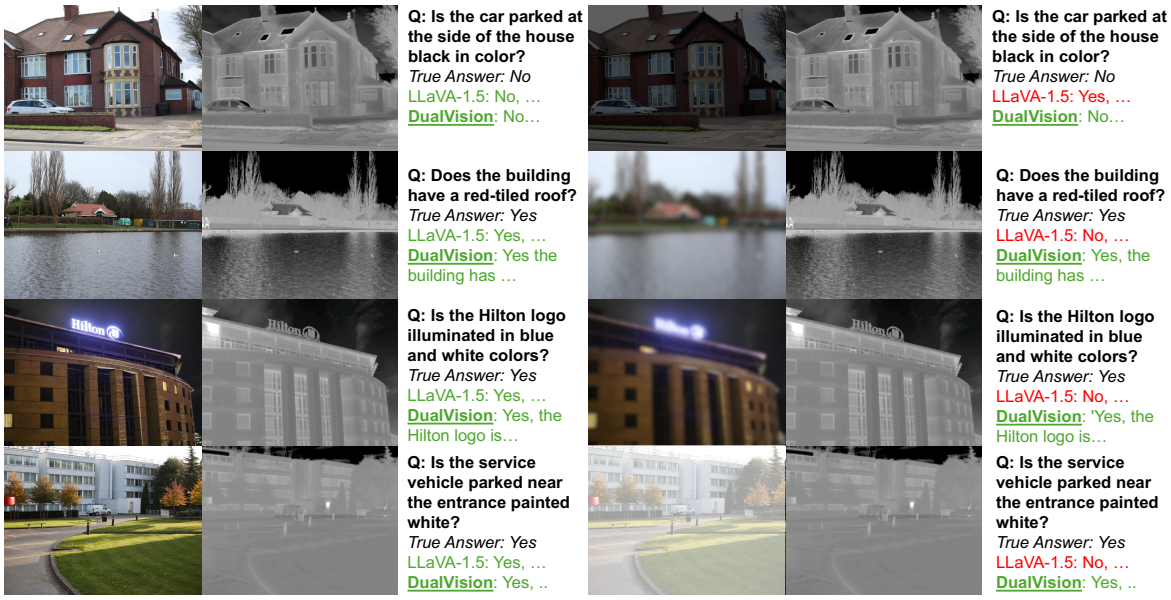

Figure B. **Sample results** (from DV-500) showing how DUALVISION maintains accurate predictions across different degradation types.

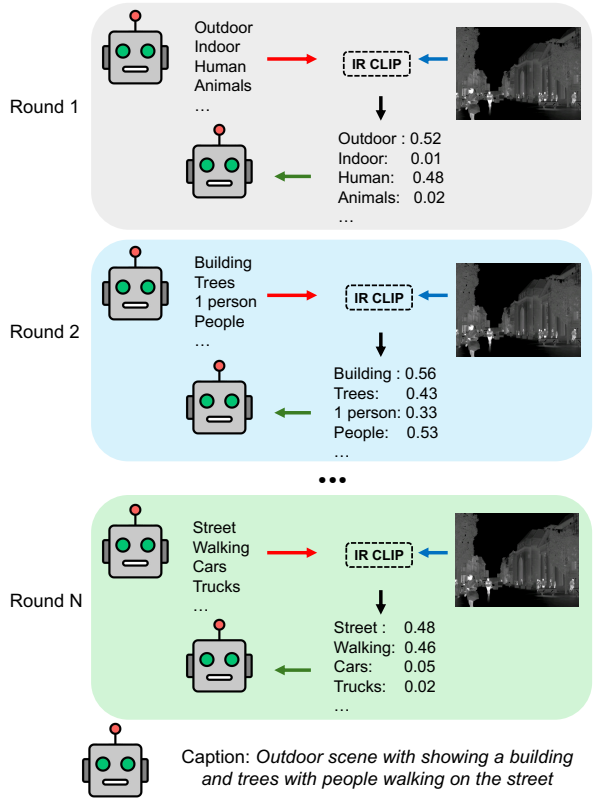

Figure C. **Our Agentic Framework** for captioning IR images. An LLM proposes and refines caption candidates while IR-CLIP [44] provides similarity-based supervision at each iteration.

verging to a final description selected by a stronger LLM (Claude Opus 4).

Further Discussion. While closely related to recent captioning method proposed in [3], our approach departs in two key ways. First, we forgo the fixed-prompt bootstrapping and aggressive truncation strategies (*e.g.*, seeding $\sim$30K prompts and retaining only the top-50 generations per step). Such strategies implicitly assume strong priors about the underlying image distribution—assumptions that may hold for curated RGB datasets but could break down for unlabeled, heterogeneous IR imagery. Second, rather than discarding low-scored candidates, we retain them as explicit hard negatives. Leveraging the longer context available in modern LLMs, our framework jointly conditions on both high- and low-quality captions, using low scores as counterfactual signals of what is not present in the IR image.

C.2. Degradations

To simulate real-world image degradations, we generated three types of altered inputs: blur, darkness, and fog. Blur was introduced by applying Gaussian smoothing with radii $\{0, 5, 10, 15, 20\}$, producing a controlled reduction of high-frequency detail. Darkness was simulated by scaling image brightness using factors $\{1.0, 0.45, 0.3, 0.2, 0.1\}$, where lower values correspond to reduced illumination. Fog effects were synthesized by blending the original image with a semi-transparent light-gray layer at intensities $\{0.0, 0.7, 0.85, 0.92, 0.97\}$, thereby decreasing contrast and diffusing edges. The selected parameter values for blur radius, brightness, and fog intensity were chosen based on qualitative visual inspection to ensure perceptually meaningful and progressively increasing degradation levels. Together, these degradations approximate common adverse conditions encountered in real outdoor environments.

D. Implementation Details

DUALVISION is implemented within the LLaVA [19] VLM backbone, where the vision encoder as well as the base language model remain frozen. Only the LLM LoRA adapters, the projection module P, and the fusion weights are updated. We make use of the pretrained CLIP ViT-$L/14$ [19, 27] as our frozen image encoder (E), to extract features from RGB and IR images resized to 336×336. With a patch size of 14×14, each image is thus represented by 576 tokens, each with an embedding size of 1024.