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 [2] 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 [2] finetuned on DV-204K with interleaved RGB and IR tokens.

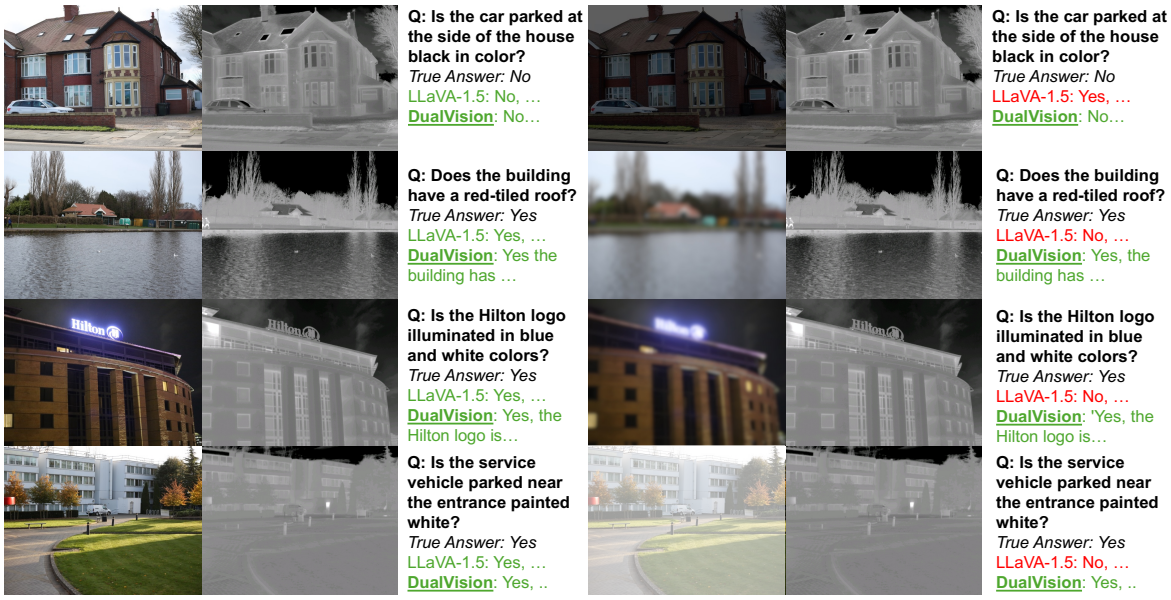

Figure B. **Sample results** (from DV-500) showing how DUALVISION maintains accurate predictions across different degradation types.

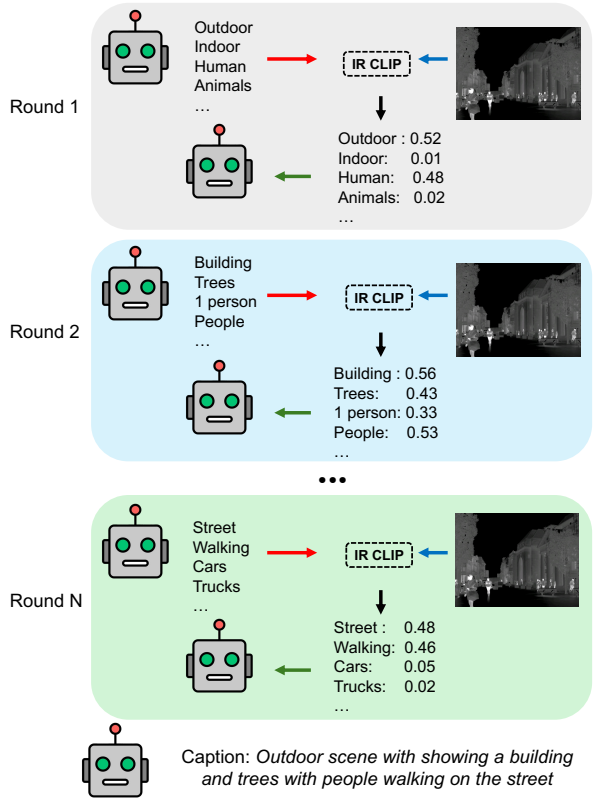

Figure C. **Our Agentic Framework** for captioning IR images. An LLM proposes and refines caption candidates while IR-CLIP [4] provides similarity-based supervision at each iteration.

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

Further Discussion. While closely related to recent captioning method proposed in [1], 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 [2] 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-$L/14$ [2, 3] 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.

References

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

[2] 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, 3

[3] 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. 3

[4] 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. 3