THERMO-VL: EXTENDING VISION–LANGUAGE MODELS TO THERMAL INFRARED PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) often fail under low illumination because their visual grounding is learned predominantly from RGB imagery, whereas thermal infrared preserves complementary scene structure when visible cues degrade. We present **Thermo-VL**, a language-conditioned RGB-thermal VLM that augments a frozen Molmo-7B backbone with a trainable thermal encoder and a text-guided dual-attention fusion module. Given aligned RGB tokens, thermal tokens, and prompt embeddings, the fusion module conditions thermal features on both language and RGB context, then injects a gated residual into the frozen RGB stream so thermal evidence can be incorporated without disrupting Molmo's pretrained RGB-language interface. We train the model with the standard language-modeling objective together with auxiliary alignment and regularization losses that improve cross-modal grounding and reduce over-reliance on RGB. We also introduce a pixel-aligned RGB-thermal instruction-tuning dataset and Thermo-VL-Bench, a manually screened RGB-thermal VQA benchmark for low-light and cross-spectrum reasoning. Experiments show improvements in thermal-only reasoning and substantial gains in RGB+thermal reasoning, highlighting the value of prompt-conditioned RGB-thermal fusion. Data and code will be made publicly available after the review process.

1 INTRODUCTION

Recent advances in vision-language models (VLMs) have dramatically improved open-ended image understanding, captioning, and visual question answering by combining large-scale multimodal pretraining with instruction tuning Deitke et al. (2025a); Liu et al. (2023); Huang et al. (2025). However, most current VLMs are trained almost exclusively on visible-spectrum RGB imagery and therefore inherit the failure modes of RGB sensing. Under low illumination, glare, haze, smoke, or other adverse conditions, RGB observations can lose critical scene evidence even when reliable reasoning is most needed.

Thermal infrared provides complementary information in precisely these settings. Because it captures radiometric structure rather than reflected visible light, thermal imagery can preserve salient cues such as humans, animals, and warm vehicles when visible features degrade Zhao et al. (2024). This mismatch makes RGB-only VLMs brittle in safety-critical scenarios such as nighttime driving, search-and-rescue, and low-visibility inspection.

A common workaround is to project infrared imagery into an RGB-compatible representation and then apply an RGB-trained perception model Shyam & Yoo (2024). However, translation-based pipelines introduce a generative bottleneck, may hallucinate visible-style details, and can suppress modality-specific information that is not naturally represented in RGB space. A more principled alternative is to integrate thermal observations directly inside the VLM. The challenge is to do so without destabilizing the pretrained RGB–language interface or requiring expensive retraining of the backbone. Progress is further limited by the lack of language-supervised RGB–thermal data: widely used paired datasets such as KAIST, FLIR-ADAS, and OSU/OTCBVS largely target detection or tracking and provide little to no VQA-style supervision.

In this work, we present **Thermo-VL**, a language-conditioned multispectral VLM built on Molmo-7B-D Deitke et al. (2025a). Our model combines a frozen RGB vision tower with a partially trainable thermal encoder and a text-guided dual-attention fusion module. Given aligned RGB tokens,

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

thermal tokens, and prompt embeddings, the fusion module first refines thermal features using both the question and RGB context, then injects the resulting information into the frozen RGB stream through a token-wise gated residual. This design makes fusion question-aware, preserves compatibility with Molmo's pretrained language stack, and supports parameter-efficient adaptation by updating only the last N_ℓ blocks of the thermal encoder together with the fusion parameters. The overall architecture is shown in Fig. 1(A), and the detailed fusion block is shown in Fig. 1(B).

To support instruction tuning and standardized evaluation in this setting, we also introduce a unified RGB–thermal data suite. We construct a pixel-aligned RGB–thermal instruction-tuning dataset with question–answer pairs spanning scene understanding, counting, spatial relations, and safety-relevant cues. We further introduce **Thermo-VL-Bench**, a manually screened RGB-thermal VQA benchmark designed for low-light and cross-spectrum reasoning.

Experiments on Thermo-VL-Bench and RGB-Th Bench show that explicit prompt-conditioned RGB–thermal fusion is especially helpful when both modalities are available. Relative to unmodified VLM baselines evaluated with paired inputs, **Thermo-VL** preserves RGB-only performance while improving thermal-only and RGB+thermal settings. These results suggest that thermal cues can complement RGB when fused inside the model rather than appended at the input level.

This paper makes the following contributions. **1. Language-conditioned multispectral VLM with question-aware fusion.** We introduce **Thermo-VL**, a Molmo-based RGB–thermal VLM that integrates paired RGB and thermal inputs through a text-guided dual-attention fusion module with token-wise gated residual injection, while keeping the RGB tower and language model frozen. **2. Parameter-efficient multispectral adaptation.** We fine-tune only the last N_ℓ blocks of the thermal encoder together with the thermal LayerNorm and fusion module. We further regularize training using token-level RGB–thermal alignment, visual–thermal–text contrastive supervision, gate-entropy regularization, and block-wise RGB masking. **3. Unified RGB–thermal language supervision.** We introduce a pixel-aligned RGB–thermal instruction-tuning dataset and Thermo-VL-Bench, a manually screened binary benchmark built from real aligned image pairs, to support training and initial evaluation in low-light cross-spectrum settings.

2 Related Work

Vision-language models have advanced rapidly, enabling open-ended captioning and visual question answering (VQA) on RGB images through instruction tuning on large-scale datasets such as PixMo Deitke et al. (2025a). Models like Molmo-7B, which **Thermo-VL** directly extends, fuse a frozen ViT vision tower with a language model connected via a projector, trained on detailed captions and synthetic QA pairs Deitke et al. (2025a). These models perform well in natural lighting but typically overlook wavelength-specific inputs, limiting performance in low-light or adverse weather.

Thermal and Infrared Language Understanding. Thermal and infrared language understanding remains underexplored, with limited prior work on direct thermal input processing for captioning and VQA Jiang et al. (2024); Moshtaghi et al. (2025). Existing approaches often translate thermal images to pseudo-RGB before applying RGB-trained VLMs, or use language models to mine captions from infrared imagery without native cross-spectrum fusion Jiang et al. (2024). These strategies do not directly align thermal observations with language and therefore remain limited in settings where visible-spectrum cues degrade.

RGB–Thermal Fusion Mechanisms. RGB–thermal fusion has been widely studied in detection and segmentation, where attention and gating modules dynamically combine visible and thermal cues El Ahmar et al. (2023); Sun et al. (2019); Xie et al. (2025); Wang et al. (2025); Li et al. (2025); Rathinam et al. (2024). Recent work has extended VLMs beyond conventional RGB perception using complementary sensing modalities. RE-VLM Liu et al. (2026) integrates RGB imagery with event streams for robust scene understanding under challenging illumination and motion. In contrast, our setting specifically investigates prompt-conditioned RGB-thermal fusion within a pretrained VLM while preserving its frozen RGB-language interface.

Visible-to-Thermal Translation. Visible-to-thermal translation has emerged as a useful tool for data augmentation and cross-spectrum synthesis Paranjape et al. (2026); Xiao et al. (2026); Berg et al. (2018); Li et al. (2023); Lee et al. (2026). While such methods can expand training data, they are not a substitute for native multispectral reasoning because the downstream model still needs to

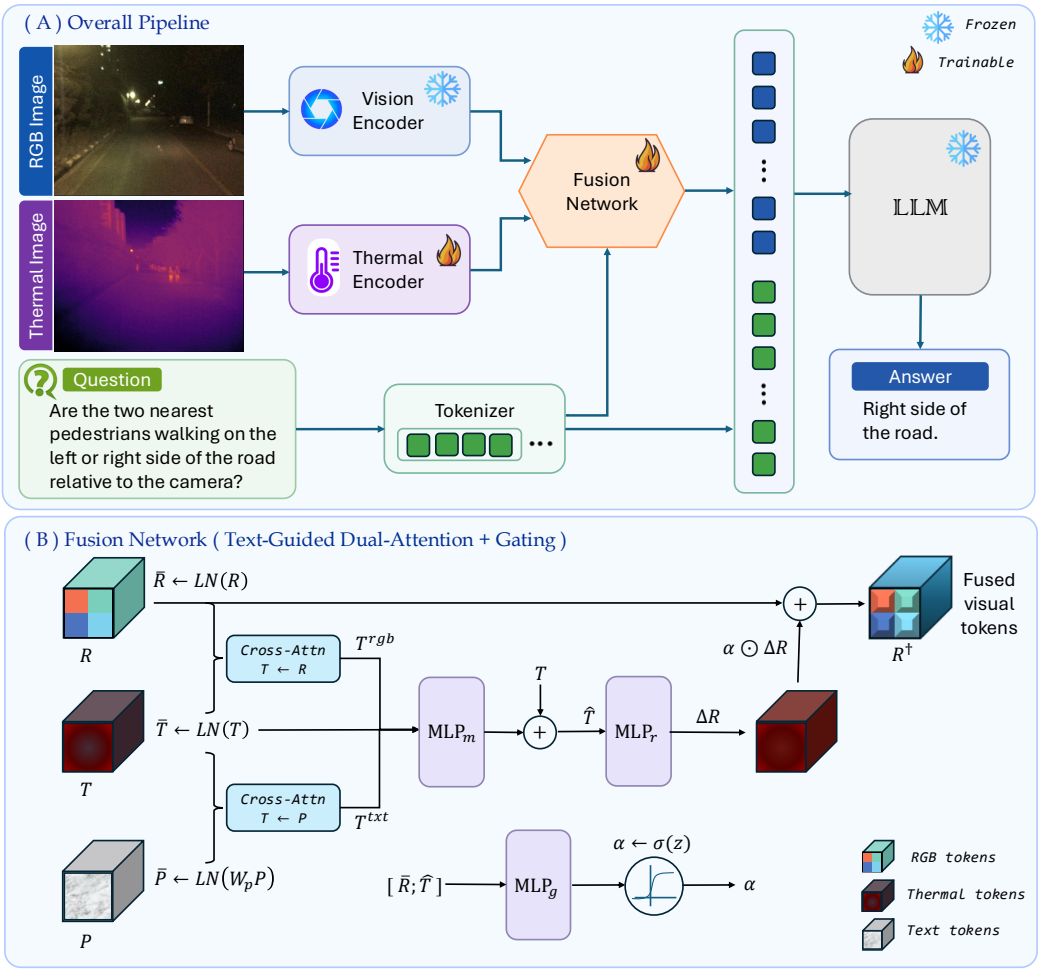


Figure 1: **Thermo-VL** overview. (A) Overall pipeline: a frozen RGB vision tower and a fine-tuned thermal encoder produce aligned patch tokens from paired RGB and thermal inputs. Prompt embeddings guide a text-aware fusion module, and the frozen language model processes the fused visual tokens to generate the final output. (B) Text-guided dual-attention fusion block: thermal tokens attend to both prompt embeddings and RGB tokens; the refined thermal representation predicts residual and token-wise gates, and the gated residual is injected into RGB tokens to form fused visual tokens $\mathbf{R}^{\dagger}$.

ground thermal evidence in language. In this work, we use synthetic thermal generation as augmentation only, while the proposed model reasons directly over paired RGB and thermal inputs.

RGB–Thermal Datasets. Public paired datasets such as LLVIP Jia et al. (2021), KAIST Hwang et al. (2015), FLIR Teledyne FLIR (2018), and OSU/OTCBVS Davis & Keck (2005) provide valuable cross-spectrum imagery, but most were created for detection, tracking, or recognition rather than language-supervised reasoning. Existing RGB–thermal VQA resources remain small and primarily evaluation-oriented. This lack of standardized training-scale language annotations motivates our dataset construction pipeline and benchmark design. Details are provided in Sec. 4.

3 METHOD

Overview: Our model, **Thermo-VL**, builds on the Molmo-7B-D backbone and augments its frozen RGB–language interface with a partially trainable thermal branch and a lightweight fusion module. Given a paired RGB image I^{rgb}, thermal image I^{th}, and prompt p, the model first encodes both modalities into spatially aligned patch tokens. The fusion module then uses the prompt and RGB

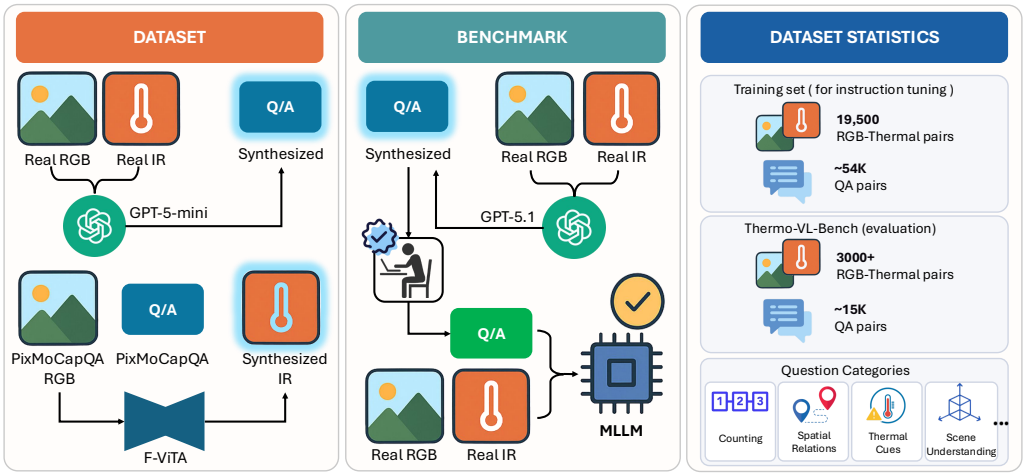


Figure 2: Construction of the RGB–thermal training data and Thermo-VL-Bench. For training, we pair real RGB images with real thermal images or synthetic thermal counterparts generated by F-ViTA and obtain QA supervision using GPT-5-mini or inherited PixMo-cap-qa annotations. For evaluation, candidate QA pairs are generated using GPT-5.1 from real aligned pairs and then manually verified.

context to refine thermal tokens and injects the resulting information into the frozen RGB stream through a gated residual. The fused visual tokens are finally passed to the frozen language model for autoregressive generation. Figure 1(A) summarizes the overall pipeline.

3.1 Visual Encoders and Input Preprocessing

Thermo-VL uses two aligned visual streams built on the Molmo-7B-D vision architecture. The RGB branch uses Molmo's pretrained vision tower ϕ_R and remains frozen throughout training. The thermal branch uses a ViT-L/14 encoder ϕ_T initialized from the MAE-pretrained checkpoint described in Appendix Sec. A.1. Because the RGB and thermal towers share the same patch embedding geometry and hidden dimension, both branches produce N spatially aligned patch tokens in $\mathbb{R}^{N\times d}$, enabling one-to-one token-level fusion without interpolation or an additional projection layer.

Molmo's vision encoder expects 3-channel inputs. Therefore, each single-channel thermal image is replicated across three channels before applying the standard Molmo image preprocessing pipeline. This preserves compatibility with the pretrained vision stack while keeping the thermal signal radiometrically single-channel. Thermal images shown in the figures are heatmap-colorized for visualization only and are not used as model inputs. During instruction tuning, we freeze the RGB encoder and update only the last N_ℓ transformer blocks of the thermal encoder together with the thermal LayerNorm LN_T and the fusion module. In all experiments, we set $N_\ell = 4$. This preserves low- and mid-level infrared structure learned during MAE pretraining while adapting higher-level thermal semantics to the downstream language objective.

3.2 Text-Guided Dual-Attention Fusion

As shown in Fig. 1(B), we introduce a text-guided dual-attention fusion block that refines thermal evidence before injecting it into the frozen RGB token stream. Let $\mathbf{R} \in \mathbb{R}^{N\times d}$, $\mathbf{T} \in \mathbb{R}^{N\times d}$, and $\mathbf{P} \in \mathbb{R}^{L\times d_p}$ denote RGB tokens, thermal tokens, and prompt embeddings. We normalize inputs and project prompts into the visual space as $\bar{\mathbf{R}} = \mathrm{LN}(\mathbf{R})$, $\bar{\mathbf{T}} = \mathrm{LN}(\mathbf{T})$, and $\bar{\mathbf{P}} = \mathrm{LN}(W_P\mathbf{P})$, where W_P maps prompts to dimension d.

We form two complementary thermal-context paths: a prompt-conditioned path $\mathbf{T}^{\mathrm{txt}} = \mathrm{MHA}(\bar{\mathbf{T}}, \bar{\mathbf{P}}, \bar{\mathbf{P}}; M_P)$ and an RGB-conditioned path $\mathbf{T}^{\mathrm{rgb}} = \mathrm{MHA}(\bar{\mathbf{T}}, \bar{\mathbf{R}}, \bar{\mathbf{R}})$. These are merged with the original thermal features to obtain $\hat{\mathbf{T}} = \mathrm{LN}\Big(\mathbf{T} + \mathrm{MLP}_m\big([\bar{\mathbf{T}}; \mathbf{T}^{\mathrm{txt}}; \mathbf{T}^{\mathrm{rgb}}]\big)\Big)$. Rather

than replacing RGB tokens, we predict a residual $\Delta\mathbf{R} = \mathrm{MLP}_r(\mathrm{LN}(\hat{\mathbf{T}}))$ and modulate it with token-wise gates $\boldsymbol{\alpha} = \sigma(\mathrm{MLP}_g(\mathrm{LN}([\bar{\mathbf{R}}; \hat{\mathbf{T}}])))$, where $\boldsymbol{\alpha} \in (0,1)^{N\times 1}$. The final fused tokens are $\mathbf{R}^\dagger = \mathbf{R} + \boldsymbol{\alpha} \odot \Delta\mathbf{R}$.

This design makes fusion query-conditional rather than a fixed preprocessing step. Thermal tokens interact with both prompt and RGB context, emphasizing task-relevant temperature cues. The gated residual preserves the frozen RGB interface of the pretrained VLM, stabilizing language grounding while enabling targeted thermal injection. Alignment and contrastive regularization organize RGB, thermal, and text features in a shared space, while gate entropy and block-wise RGB masking discourage over-reliance on the visible branch and reduce the risk of modality collapse, yielding a parameter-efficient architecture for robust cross-spectrum reasoning.

3.3 Training Objective

We train **Thermo-VL** while freezing the RGB vision encoder ϕ_R and the language model g_{LM}. The trainable parameters consist of the last N_ℓ blocks of the thermal encoder ϕ_T, the thermal LayerNorm LN_T, and the fusion module $\mathcal{F}_\theta$.

Block-wise RGB masking. To reduce over-reliance on the visible-spectrum branch, we randomly mask contiguous square regions whose union covers 10% of the RGB image area during training, while leaving the thermal input unchanged. Unlike token-level patch dropout, this augmentation operates in image space and acts as a coarse corruption augmentation for the visible branch. The masking is disabled at inference time.

Objective. Given fused tokens $\mathbf{R}^\dagger = \mathcal{F}_\theta(\mathbf{R}, \mathbf{T}, \mathbf{P})$, the frozen language model predicts $\hat{y} = g_{\mathrm{LM}}(\mathbf{R}^\dagger, \mathbf{P})$, and we optimize $\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{contr}}\mathcal{L}_{\mathrm{contr}} + \lambda_{\mathrm{gate}}\mathcal{L}_{\mathrm{gate}}$. The main task term is the autoregressive cross-entropy loss $\mathcal{L}_{\mathrm{LM}} = \mathrm{CE}(\hat{y}, y^\star)$. We apply weak token-correspondence regularization between spatially aligned RGB and thermal features: $\mathcal{L}_{\mathrm{align}} = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{R}_n - \mathbf{T}_n\|_2^2$. This term encourages feature correspondence rather than physical equivalence between RGB and thermal measurements. We use $\lambda_{\mathrm{align}} = 0.05$ to limit its influence relative to the language-modeling objective. Excessive alignment could suppress complementary modality-specific information. For global semantic consistency, we pool each stream as $\mathbf{p}_R = \mathrm{Pool}(\mathbf{R}^\dagger)$, $\mathbf{p}_T = \mathrm{Pool}(\mathbf{T})$, and $\mathbf{p}_P = \mathrm{Pool}_{M_P}(\mathbf{P})$, and define $\mathcal{L}_{\mathrm{contr}} = \frac{1}{2}\Big[\mathrm{NCE}(\mathbf{p}_R, \mathbf{p}_T; \tau) + \mathrm{NCE}(\mathbf{p}_R, \mathbf{p}_P; \tau)\Big]$, where NCE is InfoNCE with temperature τ and in-batch negatives. To avoid gate saturation, we maximize entropy $H(\alpha_n) = -\alpha_n \log \alpha_n - (1-\alpha_n)\log(1-\alpha_n)$ by minimizing $\mathcal{L}_{\mathrm{gate}} = -\frac{1}{N}\sum_{n=1}^{N} H(\alpha_n)$.

Optimization. We use AdamW with two parameter groups. Group-T contains the last N_ℓ blocks of the thermal encoder and is updated with learning rate η_{therm}. Group-F contains the fusion parameters and LN_T and is updated with learning rate η_{fusion}. This separation allows conservative adaptation of the pretrained thermal encoder while enabling faster learning in the newly initialized fusion parameters.

4 Datasets and Benchmark

To support multispectral instruction tuning and evaluation, we construct (i) a training set of aligned RGB-thermal question-answer pairs and (ii) **Thermo-VL-Bench**, a manually screened evaluation benchmark. The training data contains 15,919 (82%) real captured RGB-thermal pairs and 3,493 (18%) RGB images with F-ViTA-generated thermal counterparts. The benchmark uses only real, pixel-aligned RGB–thermal image pairs. Figure 2 summarizes the overall data pipeline.

Training Data Construction We collect aligned RGB–thermal image pairs from public datasets including KAIST Hwang et al. (2015), FLIR-ADAS Teledyne FLIR (2018), and OSU/OTCBVS Davis & Keck (2005). For real paired examples without language annotations, we use OpenAI GPT-5-mini to generate question–answer pairs grounded in the RGB image, the thermal image, or both modalities jointly. The prompts are designed to cover scene understanding, counting,

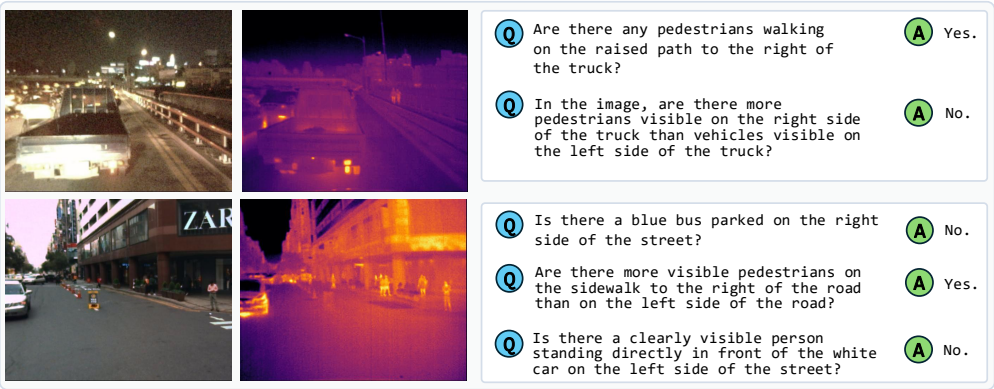


Figure 3: Representative samples from our evaluation benchmark. Each sample includes an RGB image (left), a thermal image (right), and corresponding question–answer pairs. Thermal images are displayed using heatmap colorization to enhance visual interpretability.

spatial relations, and safety-relevant cues that are often easier to identify in thermal imagery. This process yields instruction-tuning samples for RGB-only, thermal-only, and RGB+thermal reasoning from the same aligned scene.

Synthetic Thermal Augmentation. To increase data scale and linguistic diversity, we extend RGB-only images from PixMo-cap-qa Deitke et al. (2025a) with synthetic thermal counterparts. We first filter PixMo-cap-qa for natural scenes using CLIP Radford et al. (2021), then use F-ViTA Paranjape et al. (2026) to generate aligned thermal images. We retain the original QA supervision from PixMo-cap-qa and pair it with the aligned RGB–thermal inputs. These synthetic examples broaden the range of training scenes and question types while preserving the paired structure required for multispectral instruction tuning.

Thermo-VL-Bench. Given the limited availability of standardized RGB–thermal language benchmarks, we introduce **Thermo-VL-Bench**, a manually screened binary evaluation set built from real, pixel-aligned RGB–thermal image pairs. We first use OpenAI GPT-5.1 to propose candidate yes/no questions under three conditions: RGB-only, thermal-only, and RGB+thermal. We then manually screen each candidate for (i) answerability from the designated modality, (ii) consistency with the image pair, and (iii) absence of ambiguity or unintended reliance on the other modality. Approximately 20% of the candidate items are discarded during this process. Accepted items are kept verbatim after screening. As an additional quality check, three independent annotators audited a stratified sample of 100 benchmark questions while blinded to model predictions and the original generated answers. The audit achieved 91% raw agreement on answer correctness, with Fleiss' $\kappa = 0.81$ for answer correctness and $\kappa = 0.77$ for modality answerability; 91 of 100 items were retained unchanged. Full audit details are provided in Appendix A.7. We view Thermo-VL-Bench as an initial controlled evaluation resource rather than a fully human-authored benchmark.

Dataset Statistics. Our training set contains $\sim$ 19.5K pixel-aligned RGB–thermal image pairs and approximately 54K QA pairs. The questions cover a range of reasoning categories, including counting, spatial relations, safety-related cues visible in thermal imagery, and detailed scene understanding. Thermo-VL-Bench contains 3,148 RGB–thermal image pairs and approximately 15K QA pairs across RGB-only, thermal-only, and RGB+thermal settings. The training and evaluation are separated at the acquisition-sequence level; a duplicate and near-frame audit found no benchmark exposure during either MAE pretraining or instruction tuning (Appendix A.9). Table 1 compares our data with prior RGB–thermal benchmarks and highlights that our dataset supports both large-scale instruction tuning and standardized evaluation. Representative training examples and additional data visualizations are provided in Appendix Fig. 9 and Fig. 10.

Table 1: Comparison of existing RGB–thermal (and related) datasets and our proposed RGB–thermal VQA dataset. Prior RGB–thermal datasets (e.g., KAIST, FLIR-ADAS, OSU, NIRSCENE) do not provide question–answer supervision. RGB-Th-Bench offers an evaluation-only VQA benchmark, whereas our dataset supports both training (∼54K QA pairs) and evaluation (∼15K QA pairs).

Dataset	Modality	#Pairs	Task / Benchmark	#Train QA pairs	#Eval QA pairs	VQA Support
KAIST Hwang et al. (2015)	RGB + Th	∼95K	Detection / Tracking	✗	✗	✗
FLIR-ADAS Teledyne FLIR (2018)	RGB + Th	∼10K	Detection	✗	✗	✗
OSU (OTCBVS) Davis & Keck (2005)	RGB + Th	284	Detection / Tracking	✗	✗	✗
NIRSCENE Brown & Süsstrunk (2011)	RGB + NIR	477	Classification / Recognition	✗	✗	✗
RGB-Th-Bench Moshtaghi et al. (2025)	RGB + Th	29	VQA (evaluation only)	✗	1624	✓
Thermo-VL Data Suite (Ours)	RGB + Th	**22648**	**VQA (train + eval)**	**∼54K**	**∼15K**	✓

Table 2: **Thermo-VL-Bench results.** Accuracy (higher is better) under RGB-only, thermal-only, and RGB+thermal inputs. *Overall* is the QA-count-weighted average across the three subsets. Best in green, second in blue.

Model	LLM backbone	Vision backbone	RGB	IR	RGB+IR	Overall
`InternVL 2.5-8B` (IVL)	`InternLM2.5-7B`	`InternViT-300M-v2.5`	76.15	77.61	77.18	76.96
`Qwen2.5-VL-7B-Instruct` (QwenVL)	`Qwen2-7B`	`QwenViT`	83.02	86.31	76.25	82.45
`LLaVA-Next-Llama3-8B` (LlaVANext)	`Llama-3-8B`	`CLIP ViT-L/14`	89.97	78.67	85.89	84.73
`Molmo-7B-D-0924` (Molmo)	`Qwen2-7B`	`CLIP ViT-L/14`	88.67	81.20	75.79	82.51
`Thermo-VL` (**`Ours`**)	`Qwen2-7B`	`CLIP-ViT-L/14`	88.67	82.23	88.96	86.37

Table 3: **RGB-Th Bench results** Moshtaghi et al. (2025). Performance (higher is better) under RGB-Txt and RGB-Th-Txt; *Overall* averages across settings. Best in green, second in blue.

Model	LLM backbone	Vision backbone	RGB-Txt	RGB-Th-Txt	Overall
`InternVL 2.5-8B` (IVL)	`InternLM2.5-7B`	`InternViT-300M-v2.5`	79.68	63.31	71.50
`Qwen2.5-VL-7B-Instruct` (QwenVL)	`Qwen2.5-7B`	`QwenViT`	75.62	67.26	71.44
`LLaVA-Next-Mistral-7.6B` (LlaVANext)	`Mistral-7B`	`CLIP ViT-L/14`	77.83	52.82	65.33
`LLaVA-Next-Llama3-8B` (LlaVANext)	`Llama-3 8B`	`CLIP ViT-L/14`	77.22	58.63	67.92
`Molmo-7B-D-0924` (Molmo)	`Qwen2-7B`	`CLIP ViT-L/14`	77.25	60.84	69.05
`Thermo-VL` (**`Ours`**)	`Qwen2-7B`	`CLIP-ViT-L/14`	77.25	69.15	73.20

5 Experiments

5.1 Implementation Details

Unless otherwise stated, we initialize the thermal branch from the MAE-pretrained checkpoint described in Appendix Sec. A.1 and fine-tune **Thermo-VL** for 3 epochs on a single NVIDIA H100 80GB GPU with batch size 4. Under this setup, training takes approximately 34 hours. In all experiments, we update only the last $N_\ell = 4$ blocks of the thermal encoder together with the thermal LayerNorm and the fusion module, while keeping the RGB vision tower and language model frozen. Because Molmo's vision encoder expects 3-channel visual inputs, each single-channel thermal image is replicated across three channels before applying the standard Molmo image preprocessing pipeline. We use the same preprocessing convention at inference time and for external benchmark evaluation.

5.2 Evaluation Protocol

We evaluate on Thermo-VL-Bench under RGB-only, thermal-only, and RGB+thermal settings. For **Thermo-VL**, RGB-only evaluation bypasses the thermal branch, whereas thermal-only evaluation processes replicated three-channel thermal input through the visual path; no unreported modality is available. Baseline VLMs do not natively support RGB-thermal fusion, so paired RGB and thermal images are presented in a fixed order using the same prompt template. Their results therefore measure out-of-the-box paired-input adaptation rather than thermal-specific optimization.

Table 4: Ablation study of fusion design and training objective on the RGB+IR subset of Thermo-VL-Bench. Accuracy is reported on the manually screened QA pairs.

Variant	(RGB+IR)↑	Variant	(RGB+IR)↑
Full model		*Loss components*	
Full model	88.96	w/o $\mathcal{L}_{\text{align}}$	87.61
		w/o $\mathcal{L}_{\text{contr}}$	86.30
		w/o $\mathcal{L}_{\text{gate}}$	87.94
Fusion architecture		*Training strategy*	
w/o text-guided attention	82.34	w/o RGB block masking	88.11
w/o RGB-guided attention	85.06	Train all thermal blocks	86.54
w/o gated residual (Ungated residual fusion)	85.91	Train last $N_\ell = 2$ blocks	88.10

5.3 Main Results on Thermo-VL-Bench

Table 2 shows that **Thermo-VL** preserves Molmo's RGB-only performance, modestly improves thermal-only accuracy, and substantially improves RGB+thermal accuracy (88.96 vs. 75.79), achieving the highest overall score among the compared models.

To separate paired-data adaptation from architectural gains, we compare against a training-protocol-matched projected-residual baseline, obtaining $84.21 \pm 0.83\%$ versus $88.94 \pm 0.68\%$ for Thermo-VL across three seeds. On the same RGB+IR questions, zeroing thermal tokens after the thermal encoder reduces accuracy from 88.96% to 79.02%, while shuffling thermal inputs across scenes reduces it to 77.86%, indicating reliance on scene-consistent thermal evidence. Full controls and implementation details are provided in Appendix A.8.

5.4 Results on RGB-Th Bench

On the external RGB-Th Bench Moshtaghi et al. (2025), **Thermo-VL** preserves Molmo's RGB-Txt accuracy (77.25) while improving RGB-Th-Txt from 60.84 to 69.15 (Table 3), providing construction-independent evidence for improved joint RGB–thermal grounding.

5.5 Ablation Studies

We ablate the proposed fusion design and training objective to isolate the effect of each component. Unless otherwise noted, all ablations use the same training data, optimization settings, and evaluation protocol as the full model. We report accuracy on Thermo-VL-Bench, and analyze how prompt-guided attention, RGB-guided attention, auxiliary regularization, and partial thermal fine-tuning contribute to performance.

Table 4 isolates the effects of the proposed fusion architecture and training objective. Removing text-guided attention causes the largest drop, suggesting that conditioning fusion on the question is useful. Removing RGB-guided attention or the gated residual also hurts performance, which is consistent with the goal of preserving Molmo's pretrained RGB interface while incorporating thermal evidence. Among the auxiliary losses, $\mathcal{L}_{\text{contr}}$ contributes most to cross-modal consistency, while $\mathcal{L}_{\text{gate}}$ helps prevent collapse to a single modality. Block-wise RGB masking and partial thermal fine-tuning further improve robustness by discouraging RGB over-reliance and reducing overfitting in the thermal branch. We additionally evaluated an alternative text-conditioning strategy in which the fusion module uses frozen CLIP text embeddings instead of Molmo prompt embeddings; this variant underperformed the default design and is reported in Appendix Sec. A.3.

5.6 Qualitative Analysis

Figure 4 presents two qualitative comparisons between **Thermo-VL** and the RGB-only Molmo-7B-D-0924 baseline on challenging low-light scenes. In the left example, the task is detailed scene understanding. Molmo produces a plausible but generic description of a dark nighttime scene and misses structural cues. In contrast, **Thermo-VL** leverages complementary thermal evidence to recover the concrete walkway, safety railing, surrounding vegetation, and broader park-like layout,

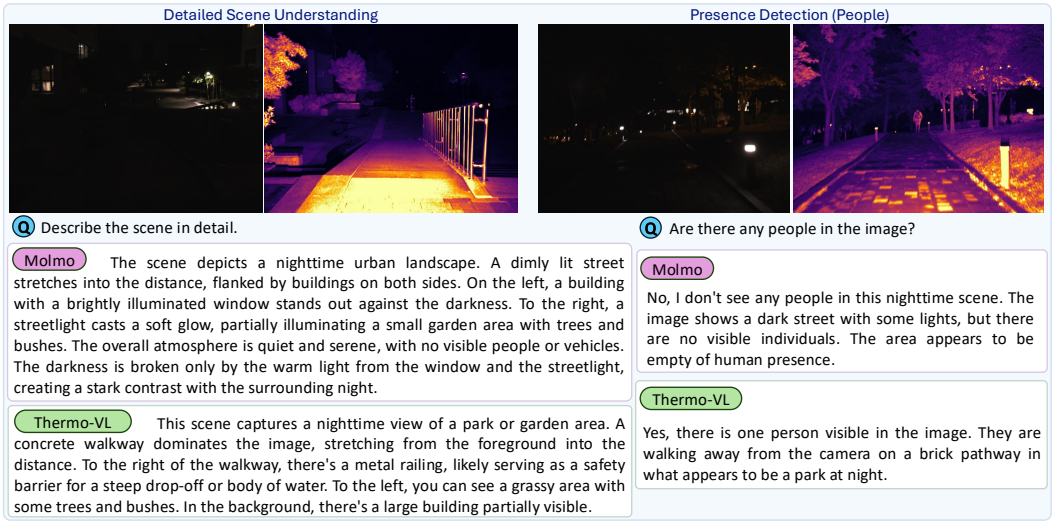


Figure 4: **Qualitative comparison in low light. Thermo-VL** improves detailed scene understanding (left) and correctly detects a person missed by Molmo (right). Thermal images are heatmap-colorized for visualization only.

yielding a more specific and spatially grounded response. In the right example, the task is presence detection. Molmo incorrectly concludes that the scene is empty, whereas **Thermo-VL** correctly identifies a person visible on the pathway. Together, these examples highlight two benefits of the proposed fusion design. First, RGB–thermal fusion improves fine-grained scene understanding when visible-spectrum content is heavily degraded. Second, it strengthens safety-relevant presence detection by injecting thermal evidence only where it is useful for the current question. These qualitative observations are consistent with the quantitative gains reported on Thermo-VL-Bench and RGB-Th Bench.

LIMITATIONS

Despite promising results, this work has several limitations. **Thermo-VL** assumes paired, spatially aligned RGB–thermal inputs and does not evaluate robustness to misalignment, missing modalities, or sensor shifts. The training data combines real paired imagery with synthetic thermal augmentation and model-generated QA supervision, which may introduce artifacts or bias. Thermo-VL-Bench is a manually screened, model-generated binary benchmark and should be viewed as an initial evaluation resource rather than a fully human-authored benchmark. Accordingly, we view **Thermo-VL** as a research prototype rather than a deployment-ready system, especially for safety-critical use.

6 CONCLUSION

We presented **Thermo-VL**, a language-conditioned multispectral VLM that integrates aligned RGB and thermal imagery through prompt-conditioned dual-attention fusion. By freezing the pretrained RGB–language backbone and adapting only a thermal encoder with a gated residual fusion module, **Thermo-VL** incorporates complementary thermal evidence while preserving RGB performance. We also introduced an RGB–thermal instruction-tuning dataset and Thermo-VL-Bench for low-light and cross-spectrum VQA. Results on Thermo-VL-Bench and RGB-Th Bench show improved thermal and RGB+thermal reasoning, highlighting native multispectral fusion as a promising direction for robust VLMs in visually degraded settings.

AI USE STATEMENT

Generative AI tools were used as part of the research methodology to generate question-answer supervision for the RGB-thermal instruction-tuning dataset and candidate question-answer pairs for Thermo-VL-Bench, as described in Section 4 and Appendix A.12. Generated benchmark items were manually screened, with an additional independent annotation audit described in Appendix A.7. Generative AI tools were also used to assist with literature search and writing refinement, including improving clarity and language. The authors reviewed and verified the AI-assisted literature search and all AI-assisted text. We take responsibility for the final content of this work, including all claims, citations, and AI-assisted material.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Amanda Berg, Jorgen Ahlberg, and Michael Felsberg. Generating visible spectrum images from thermal infrared. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1224–122409. IEEE, 2018.

M. Brown and S. Süsstrunk. Multispectral SIFT for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR11)*, pp. 177–184, Colorado Springs, June 2011.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

James W. Davis and Mark A. Keck. A two-stage approach to person detection in thermal imagery. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, January 2005. IEEE OTCBVS WS Series Bench.

James W. Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2-3):162–182, 2007. IEEE OTCBVS WS Series Bench.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025a.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025b.

Wassim El Ahmar, Yahya Massoud, Dhanvin Kolhatkar, Hamzah AlGhamdi, Mohammad Alja'Afreh, Riad Hammoud, and Robert Laganiere. Enhanced thermal-rgb fusion for robust object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 365–374, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, Xiaoqin Zhang, Ling Shao, Shijian Lu, and Dacheng Tao. Visual instruction tuning towards general-purpose multimodal large language model: A survey. *International Journal of Computer Vision*, 133(11):8151–8189, 2025.

Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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*, pp. 3496–3504, 2021.

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*, pp. 8573–8591, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Dong-Guw Lee, Tai Hyoung Rhee, Hyunsoo Jang, Young-Sik Shin, Ukcheol Shin, and Ayoung Kim. Thera: Thermal-aware visual-language prompting for controllable rgb-to-thermal infrared translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36803–36813, June 2026.

Gong Li, Guoyin Ren, Jingyu Wang, Mobing Zhi, Zhijie Yu, Bo Jiang, Haoliang Guan, and Qidan Guo. Multimodal fusion transformer network for multispectral pedestrian detection in low-light condition. *Scientific Reports*, 15(1):18778, 2025.

Yuchuan Li, Yoon Ko, and Wonsook Lee. A feasibility study on translation of rgb images to thermal images: Development of a machine learning algorithm. *SN Computer Science*, 4(5), 2023. doi: 10.1007/s42979-023-02040-4. URL https://dl.acm.org/doi/10.1007/s42979-023-02040-4.

Hanqing Liu, Mingjie Liu, Luoping Cui, Endian Lin, Donghong Jiang, and Chuang Zhu. Revlm: Event-augmented vision-language model for scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10377–10386, June 2026.

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

Mehdi Moshtaghi, Siavash H Khajavi, and Joni Pajarinen. Rgb-th-bench: A dense benchmark for visual-thermal understanding of vision language models. *arXiv preprint arXiv:2503.19654*, 2025.

Jay Nitin Paranjape, Celso M De Melo, and Vishal M Patel. F-vita: Foundation model guided visible-to-infrared translation. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5633–5642. IEEE, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Arunkumar Rathinam, Leo Pauly, Wassim Rharbaoui, Anis Kacem, Vincent Gaudillière, Djamila Aouada, et al. Hybrid attention for robust rgb-t pedestrian detection in real-world conditions. *IEEE Robotics and Automation Letters*, 2024.

Pranjay Shyam and HyunJin Yoo. Lightweight thermal super-resolution and object detection for robust perception in adverse weather conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7471–7482, 2024.

Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.

Teledyne FLIR. Teledyne FLIR thermal dataset for algorithm training. https://oem.flir.com/solutions/automotive/dataset/, 2018. Thermal and visible spectrum frames for ADAS and autonomous vehicle development.

Quan Wang, Ye Shen, Hao Wu, and Deji Chen. At-detr: A multispectral detector with adaptive feature fusion and scene enhancement for complex road scenes. In *2025 25th International Conference on Software Quality, Reliability and Security (QRS)*, pp. 623–633. IEEE, 2025.

Jiuhong Xiao, Roshan Nayak, Ning Zhang, Daniel Tortei, and Giuseppe Loianno. Thermalgen: Style-disentangled flow-based generative models for rgb-to-thermal image translation. *Advances in Neural Information Processing Systems*, 38:33111–33133, 2026.

Ruilin Xie, Sheng Jiang, Yiming Bie, and Miaolei Xia. Illumination-guided with crossmodal transformer fusion for rgb-t object detection. *Journal of Imaging Science & Technology*, 69(2):1, 2025.

Tianyi Zhao, Maoxun Yuan, Feng Jiang, Nan Wang, and Xingxing Wei. Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion. *arXiv preprint arXiv:2401.10731*, 8(9), 2024.

A Appendix

A.1 MAE encoder training

To obtain a robust thermal feature extractor, we pre-train a Masked Autoencoder (MAE) on a pooled multi-band infrared corpus (See figure 5) built from four public datasets: KAIST, FLIR-ADAS, OSU (OTCBVS), and NIRSCENE. KAIST provides aligned color–thermal pairs for pedestrian-centric driving scenes and explicitly characterizes thermal sensing in the long-wave infrared (LWIR) band (thermal band, 7.5–13 μm), while also contrasting it with near-infrared (NIR, 0.75–1.3 μm) imaging Hwang et al. (2015). FLIR-ADAS contributes large-scale annotated thermal/visible automotive data Teledyne FLIR (2018). The OSU OTCBVS benchmark adds additional thermal imagery and RGB–thermal collections recorded in outdoor pedestrian scenes Davis & Keck (2005); Davis & Sharma (2007). Finally, NIRSCENE (RGB–NIR) broadens the training distribution with near-infrared scene content Brown & Süsstrunk (2011). Training on this mixture encourages the encoder to learn modality-invariant structures (e.g., object boundaries and coarse geometry) that transfer across infrared bands and remain stable across diverse imaging conditions such as illumination, temperature variation, and sensor characteristics.

MAE pre-training objective We follow the asymmetric MAE design in which the encoder operates only on unmasked patches and a lightweight decoder reconstructs masked content in pixel space He et al. (2022). Each input image is resized to 224×224 and partitioned into non-overlapping 14×14 patches, producing $16 \times 16 = 256$ tokens. We randomly mask 75% of tokens per image and feed the remaining 25% to the encoder backbone (Molmo-7B-D-0924 ViT). The decoder receives the encoded visible tokens plus learned mask tokens (restored to original order using positional embeddings) and predicts all patch pixels. The reconstruction loss is computed only over masked patches:

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 , \tag{1}$$

where $\mathcal{M}$ is the set of masked patch indices and $\mathbf{x}_i$, $\hat{\mathbf{x}}_i$ are the ground-truth and reconstructed patch vectors, respectively He et al. (2022). After pre-training, the decoder is discarded and the encoder is used to initialize the thermal feature extractor.

A.2 LWIR-Only MAE Initialization

We compare the original LWIR+NIR MAE initialization with an LWIR-only initialization while keeping subsequent instruction tuning and evaluation unchanged.

Excluding NIR changes thermal-only accuracy by $+0.35$ percentage points and paired accuracy by -0.43 percentage points. These observed differences suggest limited sensitivity to NIR inclusion in this experiment. Without repeated-run uncertainty, they do not establish equivalence or a reliable advantage for either initialization.

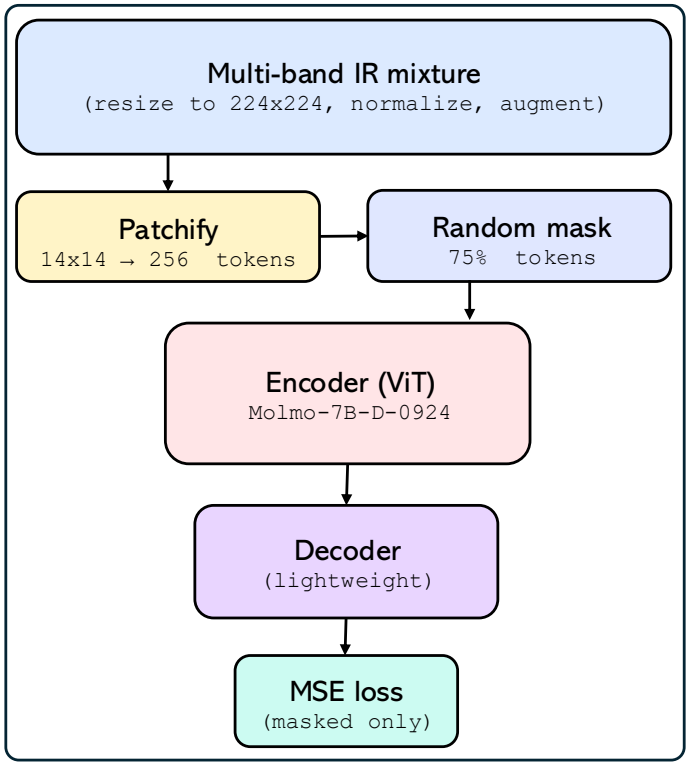


Figure 5: MAE encoder pre-training on a pooled infrared corpus (KAIST, FLIR-ADAS, OSU/OTCBVS, NIRSCENE). Inputs are resized to 224×224, patchified into 256 tokens, 75% randomly masked, encoded by a ViT backbone, and reconstructed by a lightweight decoder. MSE loss is applied on masked patches.

Table 5: Effect of excluding NIR imagery from MAE pretraining. Accuracy is reported on Thermo-VL-Bench.

MAE initialization	Thermal-only	RGB+thermal
LWIR+NIR	82.23	**88.96**
LWIR-only	**82.58**	88.53

A.3 Additional ablation: alternative text encoder

We additionally tested whether the fusion module benefits from external text representations by replacing Molmo prompt embeddings with features from a frozen CLIP text encoder. Concretely, we replace the prompt features used by the fusion module with CLIP text features followed by a learned projection into the visual space, while keeping the rest of the architecture unchanged. Table 6 shows that this variant performs substantially worse than using native Molmo prompt embeddings.

Table 6: Alternative text-conditioning ablation on Thermo-VL-Bench (RGB+IR). Replacing native Molmo prompt embeddings with frozen CLIP text features degrades performance.

Text-conditioning strategy	RGB+IR ↑
Molmo prompt embeddings (ours)	88.96
Frozen CLIP text encoder	83.10

We hypothesize that the drop arises from a representation mismatch: although CLIP text features provide a strong standalone text space, they are not naturally aligned with the token space used by

the frozen Molmo language model. As a result, question-conditioned fusion is weaker when CLIP text features replace the model's native prompt embeddings.

A.4 Algorithms for Training-VLM and Text-Guided Dual-Attention

Algorithm 1: Training Frozen-VLM Thermal Fusion with Text-Guided Dual-Attention

Input: $(I^{\mathrm{rgb}}, I^{\mathrm{th}}, p, y^\star),\ \phi_R,\ g_{\mathrm{LM}},\ \phi_T,\ \mathrm{LN}_T,\ \mathcal{F}_\theta$
Frozen modules: ϕ_R (RGB ViT encoder), g_{LM} (LLM head)
Trainable modules: ϕ_T (last-N_ℓ blocks only), $\mathrm{LN}_T,\ \mathcal{F}_\theta$
Hyperparameters: $\eta_{\mathrm{therm}},\ \eta_{\mathrm{fusion}},\ \mathrm{wd},\ \rho,\ \tau,\ \lambda_{\mathrm{align}},\ \lambda_{\mathrm{contr}},\ \lambda_{\mathrm{gate}}$
Output: $(\phi_T,\ \mathrm{LN}_T,\ \mathcal{F}_\theta)$

$\tilde{I}^{\mathrm{rgb}} \leftarrow \mathrm{BlockMask}_\rho(I^{\mathrm{rgb}})$ # apply contiguous RGB block masking during training
$\mathbf{R} \leftarrow \phi_R(\tilde{I}^{\mathrm{rgb}})$ # frozen RGB patch tokens from masked RGB input
$\mathbf{T} \leftarrow \mathrm{LN}_T\big(\phi_T(I^{\mathrm{th}})\big)$ # thermal patch tokens
$\mathbf{P} \leftarrow \mathrm{Embed}_{\mathrm{LLM}}(p)$ # prompt/token embeddings
$(\mathbf{R}^\dagger, \boldsymbol{\alpha}) \leftarrow \mathcal{F}_\theta(\mathbf{R}, \mathbf{T}, \mathbf{P})$ # fusion module; see Algorithm 2
$\hat{y} \leftarrow g_{\mathrm{LM}}(\mathbf{R}^\dagger, \mathbf{P})$ # frozen VLM forward
$\mathcal{L}_{\mathrm{LM}} \leftarrow \mathrm{CE}(\hat{y}, y^\star)$ # main task loss
$\mathcal{L}_{\mathrm{align}} \leftarrow \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{R}_n - \mathbf{T}_n\|_2^2$ # token-level RGB–thermal alignment
$\mathbf{p}_R \leftarrow \mathrm{Pool}(\mathbf{R}^\dagger), \quad \mathbf{p}_T \leftarrow \mathrm{Pool}(\mathbf{T}), \quad \mathbf{p}_P \leftarrow \mathrm{Pool}_{M_P}(\mathbf{P})$
$\mathcal{L}_{\mathrm{contr}} \leftarrow \frac{1}{2}\Big[\mathrm{NCE}(\mathbf{p}_R, \mathbf{p}_T; \tau) + \mathrm{NCE}(\mathbf{p}_R, \mathbf{p}_P; \tau)\Big]$ # contrastive regularization
$H(\boldsymbol{\alpha}) \leftarrow -\boldsymbol{\alpha}\log\boldsymbol{\alpha} - (1-\boldsymbol{\alpha})\log(1-\boldsymbol{\alpha})$
$\mathcal{L}_{\mathrm{gate}} \leftarrow -\frac{1}{N}\sum_{n=1}^{N} H(\alpha_n)$ # discourages saturated token gates
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{contr}}\mathcal{L}_{\mathrm{contr}} + \lambda_{\mathrm{gate}}\mathcal{L}_{\mathrm{gate}}$
Update Group-T: last-N_ℓ blocks of ϕ_T with AdamW$(\eta_{\mathrm{therm}}, \mathrm{wd})$
Update Group-F: $\{\mathcal{F}_\theta, \mathrm{LN}_T\}$ with AdamW$(\eta_{\mathrm{fusion}}, \mathrm{wd})$ # dual-group optimization
return $(\phi_T,\ \mathrm{LN}_T,\ \mathcal{F}_\theta)$

Algorithm 2: Text-Guided Dual-Attention Fusion and Gating Block

Input: $\mathbf{R} \in \mathbb{R}^{N\times d},\ \mathbf{T} \in \mathbb{R}^{N\times d},\ \mathbf{P} \in \mathbb{R}^{L\times d_p},\ M_P$
Trainable parameters: W_P (text-to-visual projection), MLP_m (multisource merger), MLP_r (RGB residual predictor), MLP_g (gate predictor)
Output: $\mathbf{R}^\dagger \in \mathbb{R}^{N\times d}$ (fused RGB tokens), $\boldsymbol{\alpha} \in (0,1)^{N\times 1}$ (token-wise gates)

$\mathbf{R} \leftarrow \mathrm{LN}(\mathbf{R}), \quad \mathbf{T} \leftarrow \mathrm{LN}(\mathbf{T}), \quad \mathbf{P} \leftarrow \mathrm{LN}(W_P\mathbf{P})$ # project text into visual feature space
$\mathbf{T}^{\mathrm{txt}} \leftarrow \mathrm{MHA}(\bar{\mathbf{T}}, \bar{\mathbf{P}}, \bar{\mathbf{P}}; M_P)$ # text-guided thermal attention
$\mathbf{T}^{\mathrm{rgb}} \leftarrow \mathrm{MHA}(\bar{\mathbf{T}}, \bar{\mathbf{R}}, \bar{\mathbf{R}})$ # RGB-guided thermal attention
$\hat{\mathbf{T}} \leftarrow \mathrm{LN}\Big(\mathbf{T} + \mathrm{MLP}_m\big([\bar{\mathbf{T}};\ \mathbf{T}^{\mathrm{txt}};\ \mathbf{T}^{\mathrm{rgb}}]\big)\Big)$ # merged thermal evidence
$\Delta\mathbf{R} \leftarrow \mathrm{MLP}_r\big(\mathrm{LN}(\hat{\mathbf{T}})\big)$ # candidate residual injected into RGB tokens
$\mathbf{z} \leftarrow \mathrm{MLP}_g\big(\mathrm{LN}([\bar{\mathbf{R}};\ \hat{\mathbf{T}}])\big)$ # gate logits from RGB/thermal context
$\boldsymbol{\alpha} \leftarrow \sigma(\mathbf{z})$ # token-wise soft gates
$\mathbf{R}^\dagger \leftarrow \mathbf{R} + \boldsymbol{\alpha} \odot \Delta\mathbf{R}$ # gated residual fusion
return $\mathbf{R}^\dagger,\ \boldsymbol{\alpha}$

A.5 Optimization and Hyperparameters

Table 7 summarizes the instruction-tuning configuration. AdamW uses separate parameter groups for the trainable thermal blocks and for the fusion module together with the thermal LayerNorm.

The original full-model experiment used one NVIDIA H100 80GB GPU and required approximately 34 hours. The repeated full-model and projected-residual baseline experiments used random seeds 42, 43, and 44.

Table 7: Instruction-tuning hyperparameters.

Setting	Value
Trainable thermal blocks	Last 4
Optimizer	AdamW
Thermal-block learning rate	5×10^{-5}
Fusion/thermal-LayerNorm learning rate	2×10^{-4}
Weight decay	0.1
Learning-rate schedule	Cosine
Warmup fraction	3%
Effective batch size	4
Training epochs	3
RGB block-mask ratio	0.10
λ_{align}	0.05
λ_{contr}	0.005
λ_{gate}	0.005
InfoNCE temperature τ	0.07
Seeds for repeated comparisons	42, 43, 44

A.6 Parameter Counts and Inference Efficiency

Table 8 reports trainable parameter counts and the computational measurements from the experiments. The projected-residual baseline contains fewer trainable parameters than **Thermo-VL** and is therefore not described as parameter-matched.

Table 8: Trainable parameters and reported inference measurements. The zero trainable parameter count for Molmo refers to its unmodified evaluation setting, not its total model size.

Model	Trainable params.	Latency/sample	Peak memory
Unmodified Molmo	0.000M	1.81 s	18.3 GB
Projected-residual baseline	62.979M	2.34 s	18.6 GB
Thermo-VL	79.779M	2.53 s	18.6 GB

A.7 Manual Screening and Independent Audit

Original screening. The generation pipeline produced approximately 19K candidate benchmark QA pairs. Four reviewers with postgraduate training in engineering screened the candidates for answer correctness, answerability from the designated modality, consistency with the aligned pair, ambiguity, and unintended reliance on another modality.

Approximately 80% of candidates were accepted verbatim and 20% were discarded. No candidates were manually rewritten during the original screening stage. The original procedure did not include independent overlapping annotations for every item. We therefore describe Thermo-VL-Bench as manually screened, rather than as a fully independently multi-annotated benchmark.

Independent audit. A separate stratified audit considered 100 questions spanning RGB-only, thermal-only, RGB+thermal, low-light, and thermal-intensity categories. Three annotators independently assessed each item while blinded to model predictions and the original generated answer. Table 9 summarizes the reported audit.

On the independently adjudicated subset, Thermo-VL achieves 83% accuracy, compared with 79% for the matched projected-residual baseline and 74% for unmodified Molmo.

The independent adjudication stage is distinct from the original generate-and-screen procedure.

Table 9: Independent annotation audit of 100 benchmark questions.

Statistic	Value
Raw agreement on answer correctness	91%
Fleiss' κ: answer correctness	0.81
Fleiss' κ: modality answerability	0.77
Items retained unchanged	91
Items corrected	6
Items removed	3

A.8 Training-Protocol-Matched Residual Baseline

The projected-residual baseline uses the same frozen Molmo RGB encoder and language model, MAE-initialized thermal encoder, last four trainable thermal blocks, thermal LayerNorm, instruction-tuning data, and optimization protocol as **Thermo-VL**.

Both attention paths and the token-wise gate are replaced by

$$\mathbf{R}' = \mathbf{R} + \mathrm{MLP}\left([\mathrm{LN}(\mathbf{R}); \mathrm{LN}(\mathbf{T})]\right).$$

All applicable auxiliary losses are retained. The gate-entropy loss is omitted because this baseline has no gate.

Table 10: Paired-input comparison on Thermo-VL-Bench. Repeated-run results are means over seeds 42, 43, and 44.

Model	Trainable parameters	Accuracy (%)
Unmodified Molmo	0.000M	75.79
Projected-residual baseline, three-seed mean	62.979M	84.21 ± 0.83
Thermo-VL, three-seed mean	79.779M	88.94 ± 0.68
Thermo-VL, original single run	79.779M	88.96

The projected-residual baseline improves over unmodified Molmo by 8.42 percentage points. The proposed model provides a further 4.73-point improvement when comparing the three-seed means. The first comparison reflects the combined effects of paired-data adaptation and the added thermal/residual pathway.

The second comparison supports the proposed fusion design under a shared training protocol. However, **Thermo-VL** contains 16.800M more trainable parameters, approximately 26.7% more than the residual baseline. This comparison therefore does not fully separate architectural structure from trainable capacity.

Thermal intervention controls. Because the RGB-only and RGB+thermal subsets contain different questions, their scores cannot directly quantify the contribution of thermal information. We therefore evaluate the same RGB+thermal questions while modifying only the thermal stream.

Table 11: Same-question thermal intervention on Thermo-VL-Bench.

Input condition	Accuracy (%)
Correct RGB + correct thermal	**88.96**
Correct RGB + zero thermal tokens	79.02
Correct RGB + shuffled thermal	77.86

For the zero-thermal condition, thermal tokens are set to zero after the thermal encoder, preventing encoder biases or an artificial black image from providing residual thermal information. Shuffled thermal inputs are drawn from different scenes while keeping the RGB image and question unchanged. The degradation under both interventions indicates that the model relies on scene-consistent thermal evidence rather than merely the presence of a second visual stream.

A.9 Sequence-Level Split and Leakage Audit

Because several source datasets contain sequential imagery, we additionally audited the splits using acquisition-sequence identifiers, frame timestamps or source indices where available, and exact SHA-256 hashes. The resulting separation is at the acquisition-sequence level. We found zero Thermo-VL-Bench images in MAE pretraining or instruction tuning, zero adjacent-frame or near-duplicate benchmark exposure, and zero overlap between RGB-Th Bench and either training stage.

A.10 Data provenance and licensing

Our images are sourced from public datasets and remain subject to the original licenses and terms of use. We do not claim ownership of the underlying images; attribution and licensing remain with the original dataset authors. As with the source datasets, the collected data may inherit annotation noise, sampling biases, and privacy-related limitations documented by the original releases.

Thermal-image interpretation. The source datasets provide processed thermal imagery rather than calibrated radiance or temperature measurements. Thermo-VL therefore does not estimate physical temperature or emissivity. Thermal-intensity questions are restricted to visually observable relative thermal cues and should not be interpreted as quantitative thermodynamic measurements.

Generation of questions and answers. Questions and reference answers are generated using OpenAI models. As a result, the generated text may contain residual hallucinations, omissions, or social biases. We apply filtering and, where noted, human verification, but the resulting dataset should not be treated as infallible ground truth.

Responsible use. Because the dataset combines reused imagery with model-generated language, downstream users should validate outputs before high-stakes deployment. Models trained on this data may reproduce artifacts from both the source datasets and the generation models.

A.11 Documentation for released assets

Released assets. This work introduces two new assets: (i) a pixel-aligned RGB–thermal instruction-tuning dataset with question–answer pairs, and (ii) Thermo-VL-Bench, a manually screened RGB–thermal evaluation benchmark. We also plan to release training and evaluation scripts, model configuration files, and instructions for reproducing the reported experiments.

Intended use. The new assets are intended for research on multispectral vision–language learning, instruction tuning, and benchmark evaluation under low-light and cross-spectrum conditions. They are not intended to serve as a stand-alone basis for operational or safety-critical decision-making.

A.12 Question-Answer Pair Generation

Figure 6 shows the prompt used with OpenAI GPT-5-mini to generate QA pairs for the training data. This prompt is applied only to a subset of the dataset, since part of the data is taken from the filtered PixMoCap-QA corpus, which already contains QA pairs.

Figure 7 shows the prompt used with OpenAI GPT-5.1 to generate QA pairs for the evaluation benchmark.

A.13 Data Statistics

A.13.1 Train Dataset Statistics

Figure 8 provides an overview of the training dataset. The left panel reports the most common question starters (first three tokens), showing that prompts are organized around a set of recurring linguistic patterns that reflect the dataset's task structure. The right panel plots answer frequency versus rank and highlights a characteristic head–tail profile: a compact set of high-frequency answers is complemented by a broad set of less frequent ones. Together, these statistics indicate that the

```
SYSTEM_PROMPT = (
    "You are a dataset builder for Visual Question Answering (VQA) using paired RGB and thermal infrared (IR) images. "
    "Generate high-quality QA pairs that can be answered from EITHER the IR image alone, or from BOTH the IR and RGB
    images together. "
    "Avoid questions that require RGB-only cues such as color names unless they can also be inferred with IR (e.g., heat
    patterns). "
    "Prefer concrete, verifiable questions. Mix simple factoid questions with harder reasoning ones (counting, spatial
    relations, "
    "temporal inference, safety-related cues visible in thermal, detailed captioning, etc.). Keep answers concise and
    unambiguous."
    "Keep NEGATIVE or UNANSWERABLE questions as well, they ask about objects or facts that are NOT visible in the image.
    For such questions, the answer must clearly say that the information is not present in the image."
)

USER_PROMPT = (
    "Create {n} QA pairs. Include a mix of SIMPLE and COMPLEX questions.\n"
    "- For at least half of the items, the question must be answerable using the IR image ALONE; mark these as
    modality_scope='IR'.\n"
    "- For the rest, the question must require BOTH images (fuse heat cues from IR with appearance/structure from RGB);
    mark as modality_scope='RGB+IR'.\n"
    "- Provide short, definite answers. If question is about detailed captioning, provide a longer answer if needed.\n"
    "Return strictly valid JSON per the provided schema."
).format(n=qas_per_pair)
```

Figure 6: Prompt used for QA generation for train data

```
SYSTEM_PROMPT = (
    "You are a dataset builder for Visual Question Answering (VQA) using paired RGB and thermal infrared (IR) images. "
    "For these test batches, ONLY generate Yes/No style questions whose correct answer is exactly 'Yes' or 'No'. "
    "Produce a balanced mixture of SIMPLE and COMPLEX Yes/No questions. SIMPLE should be directly verifiable from the
    image(s) with minimal reasoning; "
    "COMPLEX should require multi-step inference (counting comparisons, spatial reasoning, occlusion/visibility inference,
    cross-modal fusion, "
    "or safety/thermal reasoning). Each QA item must indicate whether it is answerable from the IR image ALONE
    (modality_scope='IR') or requires BOTH images (modality_scope='RGB+IR') or requires RGB only
    (modality_scope='RGB'). "
    "Avoid questions that *require* RGB-only cues such as precise color names unless RGB image has relevant cues.\n"
    "Keep the question text clear and focused so that the Yes/No answer is unambiguous."
)

USER_PROMPT = (
    "Create {n} QA pairs. ALL questions must be Yes/No questions whose answer is exactly 'Yes' or 'No'. Include a mix of
    SIMPLE and COMPLEX Yes/No questions.\n"
    "- SIMPLE: easily decided from the indicated modality with minimal inference.\n"
    "- COMPLEX: requires multi-step reasoning, cross-modal fusion, counting/threshold checks, or spatial inference.\n"
    "- Each answer must be exactly 'Yes' or 'No'. Keep questions concise and avoid ambiguity so the binary answer is
    clear.\n"
    "- As the images are pixel aligned, avoid questions specifing 'in the RGB image' or 'in the IR image' unless necessary.
    Refer as 'the image'.\n"
    "- Keep balanced between questions with positive and negative answers. Keep balance in modality_scope.\n"
    "Return strictly valid JSON per the provided schema."
).format(n=qas_per_pair)
```

Figure 7: Prompt used for QA generation for evaluation benchmark

training set contains both strong coverage of common cases and diverse, lower-frequency examples, which we account for in our training.

A.13.2 Evaluation data distribution

Table 12 summarizes the distribution of QA pairs categorized by modality, including RGB-only, Thermal-only, and combined RGB+Thermal questions.

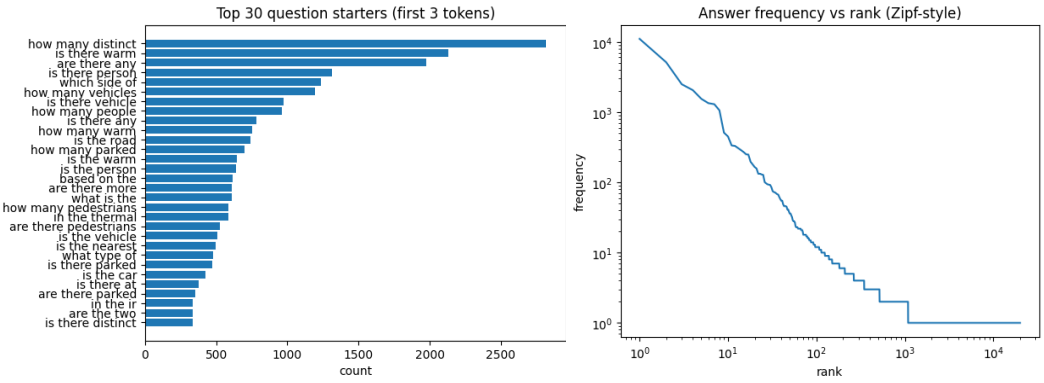


Figure 8: Data statistics for the training dataset. Left: top 30 question starters for the first 3 tokens. Right: answer frequency vs. rank.

Table 12: Distribution of question–answer (QA) pairs across modalities in the evaluation benchmark.

Answer Type	RGB	IR	RGB+IR
YES	42.60%	48.90%	50.12%
NO	57.40%	51.10%	49.88%

A.14 Additional Data Visualizations

A.14.1 Training data samples

Figure 9 shows representative examples from the training split, illustrating paired RGB (left) and heatmap-colorized thermal images (right) alongside the associated question–answer annotations.

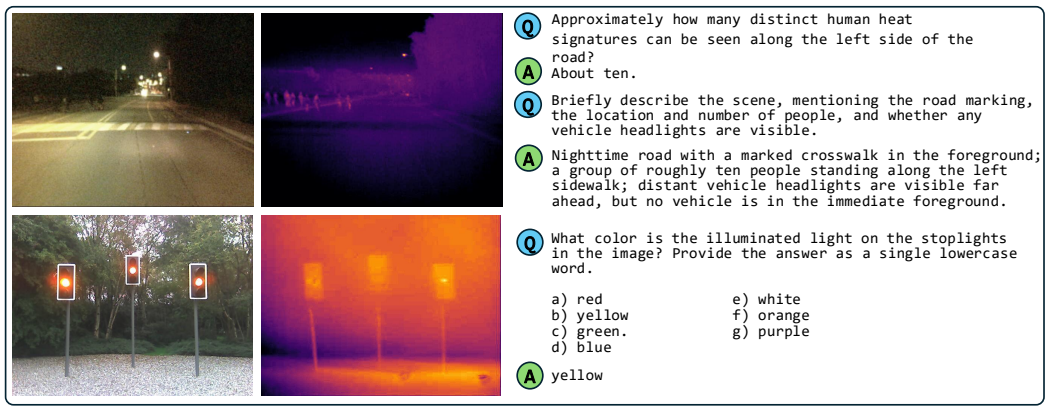


Figure 9: Representative samples from the training dataset. Each sample includes an RGB image (left), a single-channel thermal image visualized with a heatmap (right), and corresponding question–answer pairs. The heatmap colorization is for visualization only; model inputs use single-channel thermal images replicated to 3 channels.

A.14.2 Evaluation data samples

Figure 11 presents representative evaluation examples in the same format, highlighting how the model is assessed on RGB–thermal pairs with corresponding QA pairs.

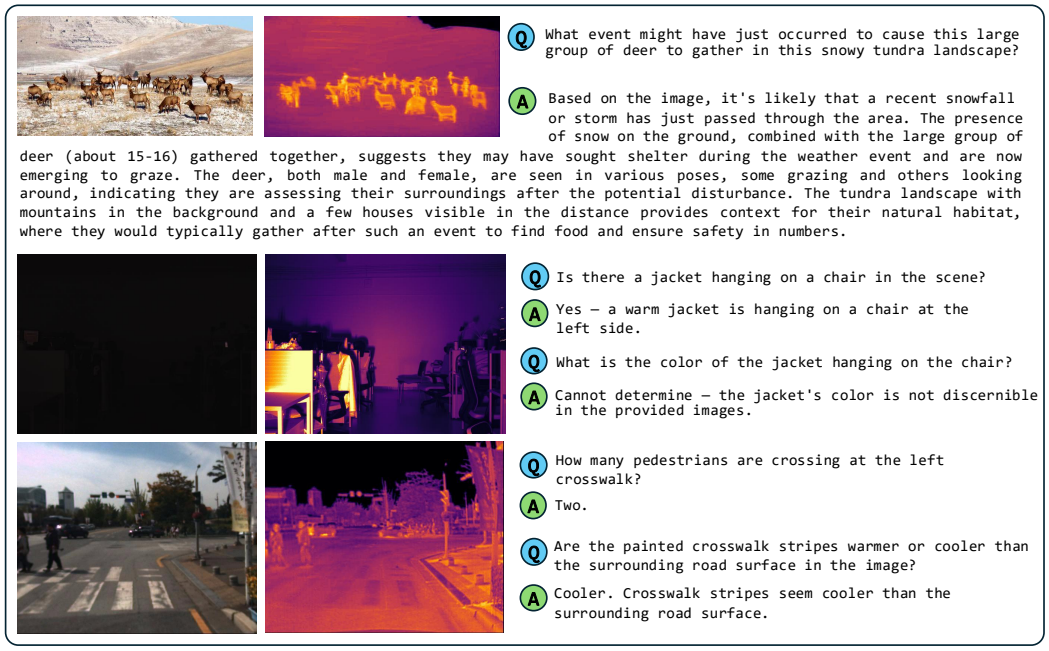


Figure 10: Additional samples from the training dataset. Each sample includes an RGB image (left), a thermal image (right), and corresponding question–answer pairs. Thermal images are displayed using heatmap colorization to enhance visual interpretability.

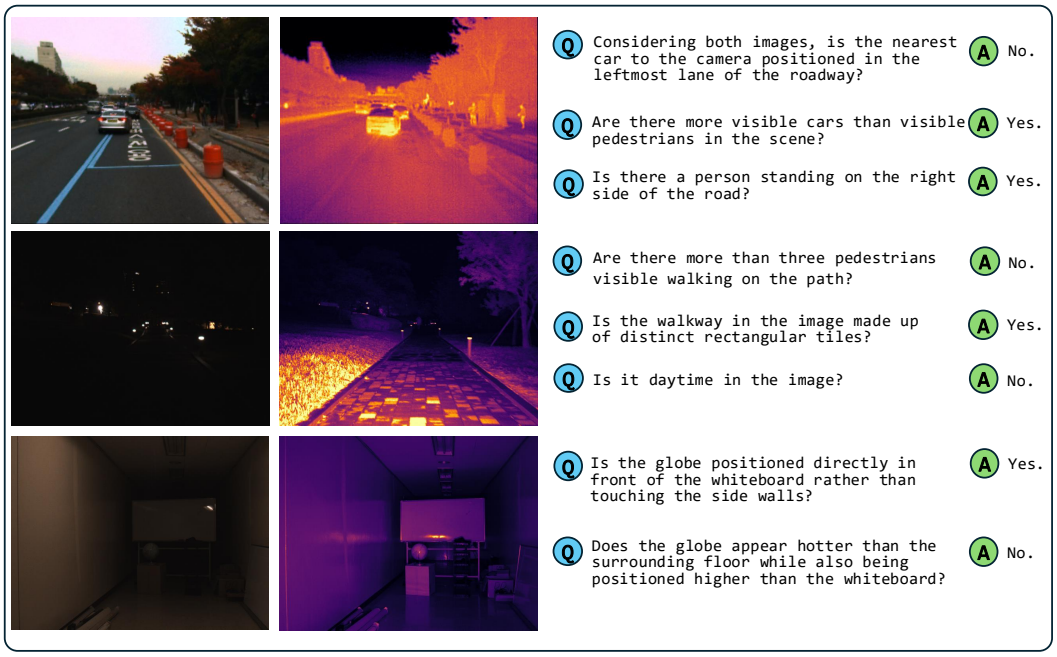


Figure 11: Representative samples from the evaluation dataset. Each sample includes an RGB image (left), a thermal image (right), and corresponding question–answer pairs. Thermal images are displayed using heatmap colorization to enhance visual interpretability.

A.15 BROADER IMPACT

This work improves multimodal reasoning in low-visibility conditions by combining RGB and thermal imagery within a language-guided framework. It enables applications such as nighttime scene understanding, search-and-rescue, and infrastructure inspection, where RGB-only systems often fail. However, these capabilities may be misused for surveillance or other dual-use scenarios, and errors could create false confidence in safety-critical settings.

The training data and benchmarks also inherit biases and imperfections from source datasets and model-generated annotations. We present this as a research prototype, not a stand-alone decision system, and recommend clear usage guidelines, respect for dataset licenses, and avoiding deployment in high-stakes settings without human oversight.