RGBT-GroundBench: Visual Grounding Beyond RGB in Complex Real-World Scenarios

Tianyi Zhao[1*], Jiawen Xi[1*], Linhui Xiao[2], Junnan Li[1], Xue Yang[3], Maoxun Yuan[1†], and Xingxing Wei[1†]

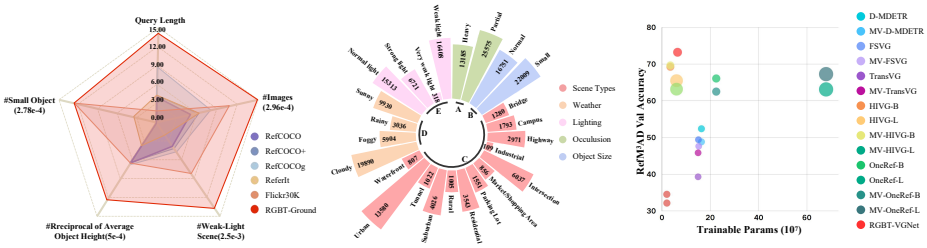
[1] Institute of Artificial Intelligence, State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2] Pengcheng Laboratory, Shenzhen, China
[3] Shanghai Jiao Tong University, Shanghai, China
{ty_zhao, xijiawen, li_jun_nan, mxyuan, xxwei}@buaa.edu.cn; xiaolinhui16@mails.ucas.ac.cn; yangxue-2019-sjtu@sjtu.edu.cn
Codes: https://github.com/crazyxiaoxi/RGBT-GroundBench
Dataset: https://huggingface.co/datasets/JiawenXi/RGBT-Ground-Dataset
* Equal contribution. † Corresponding author.

Abstract. Visual grounding (VG) localizes target objects in an image from natural-language expressions. In real-world perception, RGB cues often degrade under low illumination and adverse weather, making visual grounding substantially more challenging. However, existing VG benchmarks are largely RGB-only and provide limited, structured coverage of such conditions, hindering systematic robustness evaluation and cross-spectral comparison. We present **RGBT-GroundBench**, the first large-scale benchmark for RGB-Thermal (TIR) visual grounding in complex environments. It contains over 40K images (21,535 RGB-TIR pairs) and 38,760 object instances with referring expressions, bounding boxes, and fine-grained annotations at three levels: scene types, environmental conditions (illumination and weather), and object properties (size and occlusion). As a benchmark suite, RGBT-GroundBench provides not only curated RGB-TIR grounding annotations but also a unified evaluation protocol supporting RGB-only, TIR-only, and RGB+TIR inputs. Under this protocol, we benchmark 11 representative VG models across diverse scenes and environmental conditions. Our results show that grounding accuracy is strongly correlated with scene complexity, LoRA-based models are more robust in complex scenes, and low-illumination conditions cause significant performance degradation that has been rarely explored. Guided by these observations, we introduce **RGBT-VGNet**, a simple and reproducible reference baseline under the unified protocol, featuring Asymmetric Modality Adaptation, Language-Aware Visual Synergy, and Tri-Prior Fusion for reliability-aware RGB-TIR integration. Resources, annotations, code, checkpoints, and evaluation scripts have been publicly released.

Keywords: RGB-Thermal Visual Grounding · RGBT-GroundBench · RGBT-VGNet

Table 1: Comparison of RGBT-GroundBench with existing benchmarks. This table highlights key characteristics such as modality and the proportion of weak-light scenes and small objects.

Benchmark	Modality	Typical res.	#Instance	#Weak-light	#Small Object	Query Length	Data Source
Flickr30K [25]	RGB	500×375	276,000	2,588 (0.9%)	0 (0.0%)	1.59	Flickr [15], IAPR-TC [9]
ReferIt [15]	RGB	480×360	96,654	2,588 (2.77%)	0 (0.0%)	3.45	IAPR-TC [9]
RefCOCO [42]	RGB	640×480	50,000	3,068 (6.1%)	0 (0.0%)	3.49	MSCOCO [18]
RefCOCO+ [42]	RGB	640×480	49,856	1,415 (2.8%)	5,998 (12.0%)	3.58	MSCOCO [18]
RefCOCOg [24]	RGB	640×480	49,822	5,342 (10.7%)	22,609 (45.4%)	8.47	MSCOCO [18]
RGBT-GroundBench	RGB+TIR	640×512	38,760	16,726 (**43.2%**)	22,009 (**56.8%**)	14.24	FLIR [47], M^3FD [21], MFAD [12]



(a) Benchmark characteristics comparison (b) RGBT-GroundBench data distribution (c) Modal evaluation on RGBT-GroundBench

Fig. 1: Comparative overview of RGBT-GroundBench and evaluations of representative grounding methods. (a) Comparison with classical visual grounding benchmarks. (b) Distribution of lighting, occlusion, object, weather, and scene types. (c) Evaluation of representative methods and our RGBT-VGNet baseline on RGBT-GroundBench.

1 Introduction

Visual grounding (VG) aims to localize objects in an image from natural language expressions and is a fundamental capability for vision–language systems. It plays a key role in many applications, including human–robot interaction, embodied agents, autonomous driving, and assistive technologies. Over the past decade, visual grounding has advanced significantly with large-scale benchmarks such as RefCOCO/+/g [24, 42], ReferIt [15], and Flickr30K [25], and vision-language pretraining paradigms (*e.g.*, CLIP [26], BEIT3 [32], LLaVA [20], LLaVA-SP [22], etc.), enabling rapid progress in vision–language models and multimodal reasoning.

Despite remarkable progress in visual grounding, existing VG benchmarks still suffer from several critical limitations that hinder their applicability in real-world scenarios: 1) Limited Scene Complexity. Most benchmarks are built from datasets collected in clean environments, typically featuring centered and salient objects. Such settings fail to reflect the cluttered and dynamic nature of real-world scenes. 2) Insufficient Object Diversity. Small, distant, and occluded objects are underrepresented, making it difficult to evaluate model robustness in safety-critical domains such as autonomous driving and embodied perception. 3) RGB-Centric Modality Design. Current benchmarks rely almost exclusively on RGB imagery, which limits the evaluation of grounding models under conditions where visual cues are degraded.

Meanwhile, real-world perception often occurs under challenging conditions such as low illumination, nighttime scenes, adverse weather, or severe occlusion, where RGB sensors alone may provide unreliable visual cues and degrade

grounding accuracy. Thermal infrared (TIR) sensors can offer a complementary modality by capturing heat signatures and remaining robust under poor lighting conditions, suggesting potential benefits for grounding in such environments. Although RGB-Thermal fusion has been extensively studied in object detection [45,50,51] and semantic segmentation [43,49], its integration with language-guided perception remains largely unexplored. This motivates the study of visual grounding under RGB-thermal sensing.

To support research in this direction, we present **RGBT-GroundBench**, the first large-scale benchmark for RGB-Thermal visual grounding in complex real-world environments. As shown in Figure 1 and Table 1, RGBT-GroundBench includes a curated data resource of over 40K images (21,535 RGB-TIR pairs) and 38,760 object instances with fine-grained annotations at three levels: scene (13 scene types), environment (4 illumination and 4 weather conditions), and object (size and occlusion). Compared with previous benchmarks, it includes a substantially higher proportion of small objects (56.8%) and low-light instances (43.2%), as well as richer language descriptions with an average length of 14.24 words per query. These annotations enable systematic evaluation of grounding robustness across diverse real-world conditions. We hope that RGBT-GroundBench will serve as a comprehensive benchmark for RGB-Thermal visual grounding, providing not only curated RGB-TIR grounding resources but also standardized splits, fine-grained condition labels, and unified RGB/TIR/RGB+TIR evaluation settings for future research.

As part of RGBT-GroundBench, we construct a unified evaluation framework for the RGB-Thermal visual grounding task and conduct large-scale experiments to compare existing VG models under the proposed RGBT-GroundBench protocol. Specifically, we evaluate 11 visual grounding models under the 13 scenes, 4 weather conditions and 4 illumination levels. These models are evaluated with different input modalities including RGB-only, TIR-only and RGB+TIR. From the evaluation results, we find that: ***1)*** the localization accuracy of VG models is highly correlated with the scene complexity. ***2)*** LoRA-based VG models such as HiVG are more resistant in complex scenes. ***3)*** low-illumination scenes have a significant impact on the VG model performance, rarely explored before. ***More discussions are provided in Section 6.***

Guided by the above benchmark observations, we introduce RGBT-VGNet as a simple and competitive baseline under our unified RGBT grounding protocol. It includes three practical components: **AMA** (Asymmetric Modality Adaptation) follows our observation that LoRA-based grounding models are more robust in complex scenes. It can improve domain adaptation, extending RGB-pretrained VLM to cross-spectral inputs. Thus, we adopt an asymmetric design assigning higher adaptation capacity to the TIR branch to better align RGB-TIR representations. **TPF** (Tri-Prior Fusion) addresses the strong performance drop under low illumination. It performs illumination-aware fusion by adjusting the contributions of RGB and TIR cues with simple reliability priors. **LAVS** (Language-Aware Visual Synergy) is introduced to provide more informative cross-modal features for the TPF module. It uses a text-guided cross-modal in-

teraction mechanism, where textual features act as queries to select and align informative regions from each modality before fusion. Overall, our main contributions are as follows:

- **Comprehensive Benchmark.** We introduce **RGBT-GroundBench**, the first large-scale benchmark for RGB-Thermal visual grounding in complex real-world scenarios, with fine-grained annotations and a unified protocol for fair RGB/TIR/RGB+TIR evaluation.
- **Insightful Findings.** Through systematic benchmark-wide evaluation, we derive insightful findings into real-world cross-spectral grounding, highlighting: (i) scene-complexity sensitivity, (ii) LoRA robustness gains, (iii) low-light failure modes, and (iv) MLLM cross-spectral limitations.
- **Competitive Baseline.** We present RGBT-VGNet as a competitive baseline for benchmark validation, with practical components for modality adaptation and illumination-aware cross-modal interaction.
- **Release and Availability.** We have released RGBT-GroundBench resources, annotations, splits, code, checkpoints, and evaluation scripts. There are no hidden test sets or private servers and source datasets follow original licenses.

2 Related Works

2.1 Visual Grounding in the RGB Domain

With the rise of deep learning [10], visual grounding has rapidly advanced [1, 8, 16, 31, 39], driven by increasingly powerful visual and language representation models [5, 7, 29]. Early approaches such as ReSC [38] adopted CNN-based region-ranking pipelines. Later, transformer-based architectures, including TransVG [4], QRNet [41], D-MDETR [27], and FSVG [30], enabled end-to-end grounding with stronger global and contextual reasoning. The development of vision-language pre-training further boosted performance, as demonstrated by methods like CLIP-VG [37] and HiVG [35], which leverage large-scale image-text alignment to enhance cross-modal representations. Other lines of works explore attention-centric grounding. AttBalance [14] imposes and balances attention-driven training constraints to refine language-relevant regions, while F-LMM [33] and Localization Head [13] methods instead exploit intrinsic text-to-image attention in frozen MLLMs as grounding cues, either by translating attention to mask logits or by selecting a few localization heads for training-free localization. However, existing VG methods are mainly designed and evaluated on RGB-only visual input datasets, making them sensitive to illumination changes and adverse weather, and limiting robustness in challenging real-world conditions.

2.2 Visual Grounding Beyond the RGB Domain

Existing 2D visual grounding methods rely almost exclusively on RGB imagery, making them sensitive to illumination changes and adverse weather conditions.

Table 2: Dataset composition of RGBT-GroundBench, detailing the number of instances from each source and the split across Train, Val, Test, and specialized test subsets.

Benchmark	Sub-dataset	#Ins	Train	Val	Test	TestA	TestB	TestC	Data Source
RGBT-GroundBench	RefFLIR	9,712	7,000	608	2,104	837	640	968	FLIR [47]
	RefM^{3}FD	7,548	3,604	168	3,776	1,232	1,094	1,848	M^{3}FD [21]
	RefMFAD	21,500	16,000	1,256	4,244	789	2,452	2,544	MFAD [12]
	Total	38,760	26,604	2,032	10,124	2,858	4,186	5,360	ALL

Table 3: Lighting-weather cross statistics in RGBT-GroundBench for double-checking annotations; distributions follow illumination-weather relations.

Light-Weather	Foggy (FY)	Rainy (RY)	Sunny (SY)	Cloudy (CY)
Very Weak (VWL)	71(0.33%)	29(0.13%)	0(0.00%)	101(0.47%)
Weak (WL)	2,754(12.79%)	1,298(6.03%)	14(0.07%)	4,704(21.84%)
Normal (NL)	521(2.42%)	150(0.70%)	2,191(10.19%)	6,157(28.59%)
Strong (SL)	4(0.02%)	0(0.00%)	3,152(14.64%)	389(1.81%)

While some works explore extending VG beyond RGB, for example, by incorporating depth information in RGB-D settings [2, 19, 23], their formulations target 3D environments and are not directly applicable to 2D grounding. Moreover, depth sensing often degrades under low-light or adverse weather conditions, limiting its effectiveness as a complementary modality for robust 2D grounding. In contrast, thermal imaging is inherently illumination-invariant [44] and provides stable cues in nighttime, low-light, and adverse weather scenarios, motivating the exploration of RGB-Thermal (RGBT) visual grounding. Although RGBT perception has made notable progress in detection [6, 28, 46, 50, 51] and segmentation [17, 43, 48], RGBT visual grounding remains largely unexplored. This gap motivates the construction of our RGBT-GroundBench, which enables systematic evaluation of multi-sensor grounding models under diverse and challenging real-world conditions.

3 Benchmark: RGBT-GroundBench

3.1 Overview

RGBT-GroundBench is, to the best of our knowledge, the first large-scale RGB-Thermal visual grounding benchmark for complex real-world scenarios. It contains over 40K images (21,535 RGBT image pairs) and 38,760 object instances from diverse environments, with spatially aligned or weakly aligned RGB-TIR pairs. Each pair is annotated with high-quality natural language referring expressions, corresponding object bounding boxes, and fine-grained annotations. As illustrated in Figure 2, RGBT-GroundBench provides comprehensive annotations covering scene types, illumination conditions, weather variations, object sizes, and occlusion levels, factors that frequently occur in practical applications but remain largely underrepresented in existing benchmarks. These characteristics substantially enhance both the scenario realism and task difficulty of the dataset, thereby encouraging the development of more robust and generalizable visual grounding models.

Overall, RGBT-GroundBench introduces several key advances over existing visual grounding benchmarks:

- **Fine-grained multi-level annotations.** Covering scene-, environment-, and object-level attributes for comprehensive contextual understanding.
- **Off-central object distribution.** Objects appear in more diverse and realistic spatial positions, better reflecting real-world perception challenges.

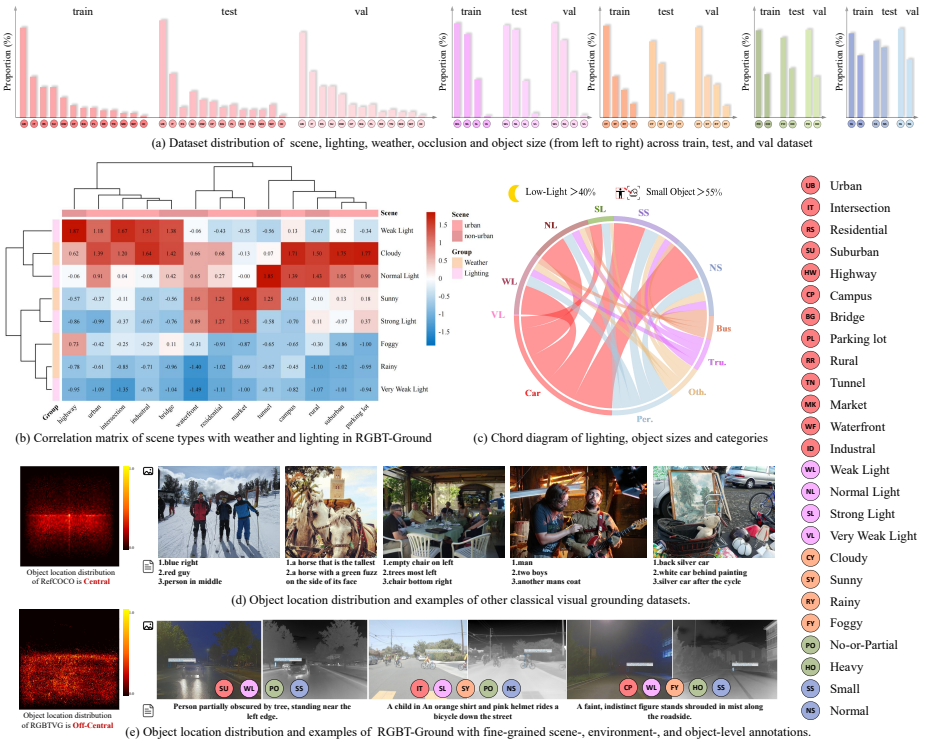


Fig. 2: Overview of RGBT-GroundBench. (a-c) show distributions and correlations of scene-, environment-, and object-level annotations; (d-e) show object-location distributions and example samples.

- **Increased diversity of challenging samples.** Including a higher proportion of nighttime, low-light, and long-range instances for robust evaluation (as shown in Table 1).
- **Paired multi-modal visual RGBT images.** Enabling reliable grounding across complex real-world scenarios.

3.2 Data Collection, Filtering and Captioning

Collection. We begin by collecting multi-modal object detection datasets [12, 21, 47] covering diverse real-world scenarios. As summarized in Tables 1 and 2, these data sources provide spatially aligned RGB and TIR image pairs, encompassing a wide spectrum of weather, lighting, and scene complexities. Each image pair is accompanied by object-level bounding box annotations, which ensure accurate multi-modal analysis and visual-language grounding.

Filtering. The collected data undergo a rigorous filtering process to ensure high quality and suitability for visual grounding research. We establish a comprehensive filtering pipeline covering three key aspects:

– Object visibility and scale: We exclude excessively small targets that occupy only a few pixels, as they provide insufficient semantic information for reliable annotation and meaningful textual descriptions.
– Cross-modality alignment accuracy: The image pair will be removed if it exhibit significant spatial misalignment between the RGB and TIR modalities.
– Category balance: Long-tailed or underrepresented categories (*e.g.*, *dog* in FLIR and *lamp* in M^3FD) are filtered out to alleviate class imbalance.

For the retained samples, we further select the largest instance of each object category per image pair to enhance annotation clarity and preserve cross-modality consistency. This filtering process ensures that RGBT-GroundBench faithfully captures the complexity, diversity, and challenges inherent in real-world deployment scenarios.

Captioning. As shown in Figure 3, we employ the Qwen3-VL to generate high-quality textual annotations in the captioning stage. Carefully designed prompts are used to produce two types of labels: (1) Object referring expressions, (2) Scene-level, environment-level, and object-level annotations. The specific prompt templates are provided in the *supplementary material*. After automatic captioning, a hierarchical random sampling strategy is applied for human verification and refinement to ensure annotation accuracy and consistency. As shown in Table 3, we further perform statistical validation to confirm the physical consistency of MLLM-generated labels, where lighting-weather distributions match real-world patterns. This dual-level annotation scheme enriches RGBT-GroundBench samples with comprehensive semantic and contextual information, thereby supporting robust visual grounding across diverse and challenging environments.

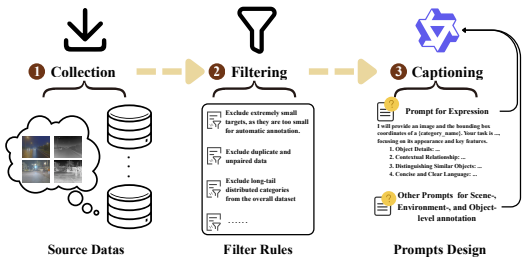


Fig. 3: The pipeline of RGBT-GroundBench data preparation. Full pipeline details are provided in the supplementary material.

Annotation Quality Control. After automatic captioning, we apply hierarchical random sampling across scene, illumination, weather, object size, and occlusion for manual verification. Annotators check hallucination, ambiguity, incorrect attributes, box-expression mismatch, and severe grammar issues, and correct them by rewrite/relabel or invalidation. An annotation is accepted only when the expression uniquely identifies one target and is consistent with the final box and attributes across modalities. Quality indicators are: 15% instances reviewed, 5% expressions edited, and 0.5% removed as invalid; in the reviewed subset, 95% of expressions correctly refer to the intended target. As an additional double-check, Table 3 shows plausible lighting-weather trends: sunny cases cluster in normal/strong light, while rainy/foggy cases cluster in weak/very-weak light, supporting fine-grained annotation consistency.

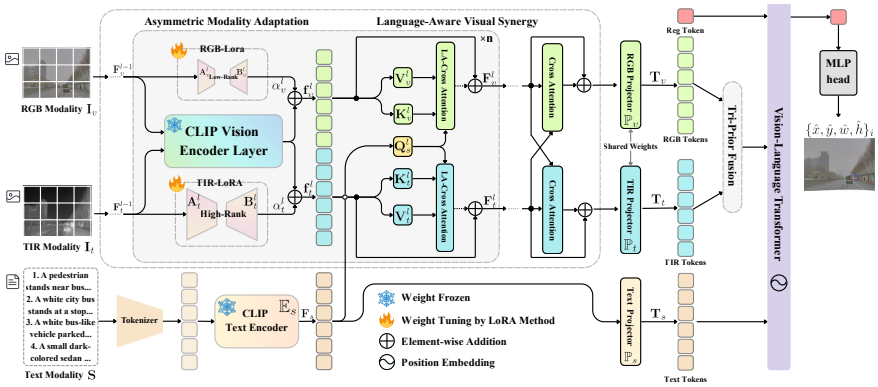


Fig. 4: Architecture of the reference baseline method RGBT-VGNet.

3.3 Evaluation Protocol

Subset Split. As summarized in Table 2, all instances are divided into train, val, and test subsets following the source data split. The test set is further partitioned into specialized subsets: **TestA**, covering normal-size (NS) targets under normal-light (NL) and strong-light (SL) conditions; **TestB**, focusing on nighttime scenarios with weak-light (WL) and very-weak-light (VL) environments; and **TestC**, containing small-size (SS) targets. Additional scene-, weather-, and occlusion-based splits and corresponding evaluation results will be provided in the supplementary material. This strategy enables fine-grained evaluation of model performance across diverse difficulty levels and environmental conditions.

Unified Evaluation Framework. As an integral part of RGBT-GroundBench, we build a unified implementation and evaluation infrastructure for benchmarking, rather than as a standalone model contribution. The unified evaluation framework adapts representative grounding models (TransVG [4], MMCA [40], D-MDETR [27], CLIP-VG [37], FSVG [30], AttBalance [14], HiVG [35], and OneRef [36]) to a single codebase supporting RGB-only, TIR-only, and RGB+TIR inputs. All methods follow the same data preprocessing, training routines, and evaluation metrics, enabling fair comparison across modalities and architectures. This benchmark infrastructure also serves as the code foundation for reproducible reporting and future benchmark-server integration.

Release and Licensing. All benchmark metadata resources are publicly released for direct use under repository licenses. Original source datasets used by RGBT-GroundBench remain subject to their own licenses and usage terms.

4 Reference Baseline Implementation: RGBT-VGNet

Overall architecture. This section presents implementation details of a reference baseline used for benchmark validation. Built on the unified evaluation framework in Sec. 3.3, RGBT-VGNet provides a transparent and reproducible reference point under a unified protocol. Given an RGB–TIR pair and a referring expression, the model localizes the described object by integrating complementary cues from the two modalities. Specifically, RGBT-VGNet is built upon the

CLIP model [26] and uses three practical components: ❶ Asymmetric Modality Adaptation (AMA) for modality-adaptive visual feature learning, ❷ Language-Aware Visual Synergy (LAVS) for semantics-aware cross-modal interaction, and ❸ Illumination-Local-Global Tri-Prior Fusion (TPF) for reliability-aware cross-spectral fusion.

Formally, given an RGB image I_v, a TIR image I_t, and a referring expression $\mathbf{S} = \{\mathbf{s}_i\}_{i=1}^{T}$, the visual features at layer l ($\mathbf{F}_v^l$ and $\mathbf{F}_t^l$) are extracted as:

$$\mathbf{f}_v^l = \begin{cases} \mathbb{E}_v^l(\mathbf{F}_v^{l-1}), & l > 1 \\ \mathbb{E}_v^l(\mathbf{I}_v), & l = 1 \end{cases}, \quad \mathbf{f}_t^l = \begin{cases} \mathbb{E}_t^l(\mathbf{F}_t^{l-1}), & l > 1 \\ \mathbb{E}_t^l(\mathbf{I}_t), & l = 1 \end{cases}, \tag{1}$$

where $\mathbb{E}_v$ and $\mathbb{E}_t$ denote the RGB and TIR CLIP vision encoders (shared parameters) with AMA modules, and $l \in \{1, \ldots, n\}$ is the encoder layer index. The intermediate RGB and TIR features $\mathbf{f}_v^l$ and $\mathbf{f}_t^l$ are then refined by the LAVS module to produce modality-enhanced features $\mathbf{F}_v^l$ and $\mathbf{F}_t^l$. The textual features are extracted as follows:

$$\mathbf{F}_s^n = \mathbb{E}_s(\mathbf{S}), \quad \mathbf{T}_s = \mathbb{P}_s\left(\mathbf{F}_s^n\right), \tag{2}$$

where $\mathbb{E}_s$ denotes the CLIP text encoder, and $\mathbb{P}_s$ is a linear projection that maps the last-layer textual embeddings into the grounding space.

All CLIP encoders remain frozen to preserve pretrained language-vision alignment. Extracted representations are aligned and refined by subsequent modules. The final input concatenates projected Fused Visual and Text tokens with a learnable regression token [Reg], which is then fed into the VL-Transformer [4]. A lightweight regression head predicts the bounding box from the [Reg] token:

$$\hat{\mathbf{B}} = \hat{x}, \hat{y}, \hat{w}, \hat{h}_i = \mathrm{MLP}([\mathrm{Reg}]). \tag{3}$$

❶ Asymmetric Modality Adaptation (AMA). Vision-language pretrained models such as CLIP have strong performance in the RGB domain, but their feature representations are biased toward RGB data used during pretraining, resulting in a large modality gap when extended to TIR modality. While LoRA-based [11] fine-tuning has been widely adopted to adapt frozen backbones to downstream tasks, applying identical LoRA configurations to different modalities overlooks their differences and the VLP models' bias. To address this limitation, we adopt an asymmetric LoRA configuration that assigns different adaptation capacities to the two visual modalities. Specifically, we perform low-rank decomposition on the attention projection weights of each vision encoder layer l:

$$\mathbf{W}_v^l = \mathbf{W}_{\mathbb{E}_v}^l + \alpha_v^l \mathbf{A}_v^l \mathbf{B}_v^l, \quad \mathbf{W}_t^l = \mathbf{W}_{\mathbb{E}_t}^l + \alpha_t^l \mathbf{A}_t^l \mathbf{B}_t^l, \tag{4}$$

where $\mathbf{W}_{\mathbb{E}_v}^l$ and $\mathbf{W}_{\mathbb{E}_t}^l$ are the frozen pre-trained CLIP weights (initialized from the same CLIP model), $\mathbf{A}_v^l \in \mathbb{R}^{d \times r_v}$, $\mathbf{A}_t^l \in \mathbb{R}^{d \times r_t}$, $\mathbf{B}_v^l \in \mathbb{R}^{r_v \times d}$, and $\mathbf{B}_t^l \in \mathbb{R}^{r_t \times d}$ are learnable low-rank matrices, and α_v^l, α_t^l are scaling factors. We set asymmetric ranks $r_v \leq r_t$ for RGB and TIR modalities, allowing the TIR branch to use

higher adaptation capacity when needed. We further adopt the hierarchical design [35] as a practical setting for progressive alignment from low-level structural to high-level semantic representations.

❷ **Language-Aware Visual Synergy (LAVS).** In the RGBT visual grounding task, the target described by the referring expression usually appears in both RGB and TIR images. Since grounding is inherently language-driven, we use a language-aware visual fusion module in which textual features act as semantic queries for cross-modal interaction. Given the encoder outputs $\mathbf{f}_v^l$ and $\mathbf{f}_t^l$ from layer l and the textual embedding $\mathbf{F}_s$, we first compute their Query-Key-Value projections and construct language-guided cross-modal attention:

$$\mathbf{A}^l_{\text{attn}_v} = \sigma\left(\frac{\mathbf{Q}_s^l(\mathbf{K}_v^l)^T}{\sqrt{d}}\right), \quad \mathbf{A}^l_{\text{attn}_t} = \sigma\left(\frac{\mathbf{Q}_s^l(\mathbf{K}_t^l)^T}{\sqrt{d}}\right), \tag{5}$$

where σ is Softmax, and $\mathbf{A}^l_{\text{attn}_v}, \mathbf{A}^l_{\text{attn}_t}$ denote the attention matrices.

$$\mathbf{F}_v^l = \mathbf{f}_v^l + (\mathbf{A}^l_{\text{attn}_v})^T(\mathbf{A}^l_{\text{attn}_v}\mathbf{V}_v^l), \quad \mathbf{F}_t^l = \mathbf{f}_t^l + (\mathbf{A}^l_{\text{attn}_t})^T(\mathbf{A}^l_{\text{attn}_t}\mathbf{V}_t^l). \tag{6}$$

Eq. 5 and 6 enable textual semantics to guide the integration of visual cues from each modality, determining both which features should be fused and to what extent, while preserving feature dimensional consistency via left multiplication with the transpose of $\mathbf{A}^l_{\text{attn}}$. After obtaining the text-queried visual features $\mathbf{F}_v^l$ and $\mathbf{F}_t^l$, we apply cross-attention $\mathbb{CA}$ between the two modalities:

$$\mathbf{T}_v^l = \mathbb{P}_v\left(\mathbf{F}_v^l + \mathbb{CA}(\mathbf{F}_v^l, \mathbf{F}_t^l)\right), \quad \mathbf{T}_t^l = \mathbb{P}_t\left(\mathbf{F}_t^l + \mathbb{CA}(\mathbf{F}_t^l, \mathbf{F}_v^l)\right) \tag{7}$$

where $\mathbb{P}_v$ and $\mathbb{P}_t$ denote RGB and TIR linear projections, respectively, mapping the output embeddings into a unified feature space for subsequent grounding.

❸ **Illumination-Local-Global Tri-Prior Fusion (TPF).** TPF is a lightweight heuristic fusion module in our reference baseline, designed to instantiate Insight 2 (cross-spectral complementarity) under a controlled protocol rather than to introduce standalone illumination modeling. It takes the two modality-enhanced outputs from Eq. 7, $\mathbf{T}_v^l$ and $\mathbf{T}_t^l$, as inputs. We first split each stream into a CLS token and patch tokens, and then flatten patch maps into token sets $\{\mathbf{t}_{v,p}^l\}_{p=1}^N$ and $\{\mathbf{t}_{t,p}^l\}_{p=1}^N$. The three priors are defined in two steps. We first compute the illumination prior: $i_p = [\mathcal{I}(\mathbf{I}_v)]_p, \quad q_p = 1 - |2i_p - 1|$, where $\mathcal{I}(\cdot)$ denotes grayscale conversion, pooling to the patch grid, and min-max normalization to $[0, 1]$; i_p is the normalized illumination intensity of the RGB patch, and q_p is the illumination quality that measures the proximity to the optimal illumination point (0.5). Then we compute local-global semantic priors:

$$\begin{aligned} g_p &= \sigma\big(\text{MLP}_{\text{tok}}([\hat{\mathbf{t}}_{v,p}^l, \hat{\mathbf{t}}_{t,p}^l, |\hat{\mathbf{t}}_{v,p}^l - \hat{\mathbf{t}}_{t,p}^l|])\big), \\ \mathbf{g}^{\text{ch}} &= \sigma\big(\text{MLP}_{\text{ch}}([\bar{\mathbf{t}}_v^l, \bar{\mathbf{t}}_t^l, |\bar{\mathbf{t}}_v^l - \bar{\mathbf{t}}_t^l|])\big), \qquad b = \text{mean}_c(\mathbf{g}^{\text{ch}}), \end{aligned} \tag{8}$$

where $\hat{\mathbf{t}}_{v,p}^l, \hat{\mathbf{t}}_{t,p}^l$ are layer-normalized patch features from Eq. 7, and $\bar{\mathbf{t}}_v^l, \bar{\mathbf{t}}_t^l$ are global pooled features. mean_c denotes channel-wise averaging. b is a global prior shared across all patches (broadcast over p). The RGB reliability weight is:

$$w_p = \sigma((i_p - 0.5)q_p + (g_p - 0.5) + (b - 0.5)), \tag{9}$$

where 0.5 is the neutral reliability point for each normalized prior; we use equal weighting to avoid extra hyperparameters. Finally, the fused patch token is:

$$\mathbf{z}_p^l = w_p \mathbf{t}_{v,p}^l + (1 - w_p)\mathbf{t}_{t,p}^l + \mathbf{g}^{\text{ch}} \odot (\mathbf{t}_{v,p}^l - \mathbf{t}_{t,p}^l), \tag{10}$$

where $\mathbf{g}^{\text{ch}}$ is the channel gate from Eq. 8. This design provides adaptive RGB–TIR fusion with illumination-aware reliability control and local-global semantic compensation in a compact form.

5 Experiments

5.1 Experimental Setup

Implementation Details. Our method is trained with a batch size of 8 using AdamW and a learning rate of 1×10^{-3}. Adapted baselines follow the default hyper-parameters in their original implementations where applicable. All experiments are trained for 120 epochs with the same input size (224×224) and the same data augmentation policy on RTX 4090 GPUs. More detailed implementation settings can be found in the open-source unified evaluation framework.
Evaluation Metrics. Following the previous visual grounding methods [3, 34], we adopt the **Acc@0.5** as the primary metric, measuring the localization accuracy when the intersection-over-union (IoU) between the predicted and ground-truth bounding boxes exceeds 0.5. In addition, we report results separately for val, test, testA, testB, and testC subsets (detailed in Section 3.3) to evaluate model robustness under diverse conditions. *Results for* ***Acc@0.7*** *are provided in the supplementary material for stricter performance validation.*

5.2 Evaluation on the RGBT-GroundBench

Pretrained model zero-shot transfer setting. All models are evaluated by directly loading pretrained weights without fine-tuning on the RGBT-GroundBench training split. As shown in Table 4, all methods exhibit low localization accuracy under both RGB and TIR modalities. This is mainly because the pretrained weights are trained on RefCOCO, ReferIt, and Flickr30K datasets, which contain limited scene complexity and lack environmental diversity, making zero-shot transfer difficult. Moreover, since these datasets contain only RGB modality, performance further degrades on TIR, where zero-shot results drop most drastically. Compared with subsequently trained models, zero-shot Acc@0.5 shows an average degradation of 30% on RGB and 50% on TIR modality, demonstrating that effective cross-modal adaptation is necessary for reliable RGBT visual grounding.
Single source dataset training setting. All models are trained under our unified framework with identical settings. To combat the deterioration of RGB-modality visual grounding in low-light conditions and small objects conditions, we introduce the TIR modality to provide illumination-invariant cues and thereby enhance overall robustness. In this setting, our RGBT-VGNet achieves highest

Table 4: Performance evaluation of current representative uni-modal visual grounding works and the baseline multi-modal RGBT visual grounding method on the RGBT-GroundBench. The "-'entries correspond to models that fail to produce stable bounding boxes. Complete evaluation results will be included in the supplementary materials.

Methods	Venue	Visual / Language Backbone	Visual Modality	RefFLIR					RefM^{3}FD					RefMFAD				
				val	test	testA	testB	testC	val	test	testA	testB	testC	val	test	testA	testB	testC
a. Zero-shot transfer from RGB-pretrained models:																		
CLIP-VG [37]	TMM'23	CLIP-B / CLIP-B	RGB	5.59	8.12	15.51	5.00	1.11	7.74	8.55	16.31	9.05	1.35	7.48	7.48	16.58	6.93	1.22
HiVG-B [35]	ACMMM'24	CLIP-B / CLIP-B	RGB	23.03	34.27	61.69	30.16	2.23	29.17	31.22	58.28	34.37	5.30	18.71	19.62	41.14	19.90	1.92
HiVG-L [35]	ACMMM'24	CLIP-L / CLIP-L	RGB	34.05	44.46	78.88	33.13	7.81	52.98	53.10	87.66	56.86	19.59	40.13	40.42	82.15	37.68	12.86
OneRef-B [36]	NeurIPS'24	BEIT3-B / BEIT3-B	RGB	32.89	41.46	77.68	31.56	3.55	33.93	39.65	74.11	48.63	3.19	29.70	30.05	70.63	28.14	2.86
OneRef-L [36]	NeurIPS'24	BEIT3-L / BEIT3-L	RGB	38.82	49.20	80.19	36.09	17.04	46.63	46.82	79.55	52.83	14.72	37.50	36.89	73.04	35.15	12.35
CLIP-VG [37]	TMM'23	CLIP-B / CLIP-B	TIR	4.77	8.03	14.92	6.41	0.61	2.38	6.59	9.33	11.15	0.76	4.54	4.26	9.49	4.04	5.49
HiVG-B [35]	ACMMM'24	CLIP-B / CLIP-B	TIR	10.86	12.10	20.64	15.94	0.81	7.74	11.49	18.34	18.19	0.97	6.29	7.36	17.97	6.44	0.51
HiVG-L [35]	ACMMM'24	CLIP-L / CLIP-L	TIR	18.26	23.80	43.56	24.06	0.00	24.40	28.89	55.36	36.47	2.92	23.73	23.11	56.33	20.39	2.82
OneRef-B [36]	NeurIPS'24	BEIT3-B / BEIT3-B	TIR	21.38	26.57	49.28	23.75	0.20	17.26	22.83	40.67	32.91	0.11	15.13	15.60	41.90	13.50	0.12
OneRef-L [36]	NeurIPS'24	BEIT3-L / BEIT3-L	TIR	21.55	27.00	46.42	26.56	3.65	23.21	23.65	38.15	33.55	4.17	17.52	17.81	41.14	16.39	2.43
b. In-domain training with uni-modal visual input (RGB or TIR):																		
TransVG [4]	ICCV'21	RN50+DETR / BERT-B	RGB	50.74	42.16	62.25	34.89	19.76	39.29	40.08	63.42	46.03	15.10	52.27	51.94	82.64	49.35	31.06
CLIP-VG [37]	TMM'23	CLIP-B / CLIP-B	RGB	43.68	42.77	67.14	33.80	15.81	32.14	33.76	58.80	41.83	5.39	36.52	37.47	73.13	34.92	13.16
D-MDETR [27]	TPAMI'24	CLIP-B / CLIP-B	RGB	57.07	49.25	70.17	39.22	26.77	52.38	50.19	71.67	57.04	26.08	60.59	59.39	86.71	55.79	40.98
D-MDETR [27]	TPAMI'24	RN50+DETR / BERT-B	RGB	49.67	44.28	67.14	34.38	20.95	-	-	-	-	-	51.99	51.18	82.07	48.65	30.41
MMCA [40]	ACMMM'24	RN50+DETR / BERT-B	RGB	50.16	46.57	71.48	35.63	21.50	-	-	-	-	-	47.85	46.52	76.71	43.84	26.24
FSVG [30]	ICME'25	RN50+DETR / BERT-B	RGB	56.09	45.25	67.14	36.88	20.55	49.40	49.39	73.86	54.65	23.70	59.71	58.65	86.24	55.34	39.81
AttBalance [14]	ACMMM'25	CLIP-B / CLIP-B	RGB	51.64	49.10	73.81	40.16	21.76	-	-	-	-	-	51.04	51.30	83.92	48.74	29.61
HiVG-B [35]	ACMMM'24	CLIP-B / CLIP-B	RGB	69.08	66.65	88.81	54.06	43.52	69.24	68.15	89.40	56.09	45.04	65.45	64.02	90.66	60.52	45.57
HiVG-L [35]	ACMMM'24	CLIP-L / CLIP-L	RGB	68.75	71.13	90.69	60.63	50.00	65.48	67.80	94.48	70.11	41.29	64.25	63.72	**91.52**	60.97	44.98
OneRef-B [36]	NeurIPS'24	BEIT3-B / BEIT3-B	RGB	63.82	61.69	84.73	52.34	35.90	66.07	66.66	**94.64**	70.93	39.07	62.34	60.61	90.76	57.63	40.86
OneRef-L [36]	NeurIPS'24	BEIT3-L / BEIT3-L	RGB	66.61	64.60	88.31	53.13	39.76	63.10	64.33	93.75	73.02	34.52	64.49	63.15	90.00	59.46	45.18
TransVG [4]	ICCV'21	RN50+DETR / BERT-B	TIR	49.92	42.63	62.37	35.67	21.18	45.83	47.54	69.72	55.58	22.51	52.79	52.33	83.92	49.84	31.37
CLIP-VG [37]	TMM'23	CLIP-B / CLIP-B	TIR	36.59	37.18	59.38	34.11	12.26	23.81	27.16	43.55	38.63	4.92	32.06	30.79	61.85	28.81	10.31
D-MDETR [27]	TPAMI'24	CLIP-B / CLIP-B	TIR	48.36	38.87	55.83	37.66	18.38	47.62	50.34	70.13	57.68	28.52	56.37	54.52	79.72	51.63	37.73
D-MDETR [27]	TPAMI'24	RN50+DETR / BERT-B	TIR	50.57	42.96	62.01	42.52	20.36	-	-	-	-	-	50.60	50.01	80.10	47.88	29.53
MMCA [40]	ACMMM'24	RN50+DETR / BERT-B	TIR	48.52	41.41	60.98	40.00	18.46	-	-	-	-	-	46.26	45.22	76.33	42.86	24.43
FSVG [30]	ICME'25	RN50+DETR / BERT-B	TIR	47.86	38.50	54.64	37.81	17.13	52.98	51.48	72.24	60.31	27.00	54.54	54.52	79.92	52.81	36.05
AttBalance [14]	ACMMM'25	CLIP-B / CLIP-B	TIR	43.75	43.39	66.83	39.22	15.92	33.93	37.63	64.44	45.80	8.77	45.30	45.45	78.61	43.35	22.43
HiVG-B [35]	ACMMM'24	CLIP-B / CLIP-B	TIR	68.75	64.07	81.55	66.09	42.41	68.09	61.77	78.21	68.31	40.59	63.22	62.63	87.50	60.85	45.69
HiVG-L [35]	ACMMM'24	CLIP-L / CLIP-L	TIR	65.79	66.71	85.44	68.44	43.91	61.31	65.12	87.34	73.95	40.69	59.87	58.87	85.95	57.18	39.33
OneRef-B [36]	NeurIPS'24	BEIT3-B / BEIT3-B	TIR	62.01	60.00	80.55	60.00	35.80	60.12	65.10	87.74	73.03	41.02	58.60	57.95	86.46	55.71	38.67
OneRef-L [36]	NeurIPS'24	BEIT3-L / BEIT3-L	TIR	64.14	61.17	81.15	60.47	37.53	65.48	68.30	90.26	75.32	45.29	62.34	60.09	85.70	57.54	42.67
c. In-domain training with multi-modal visual input (RGB+TIR):																		
MV-TransVG	ICCV'21	RN50+DETR / BERT-B	RGB+TIR	54.44	46.06	68.62	40.31	21.10	45.83	48.04	70.29	55.3	22.89	53.18	53.84	83.54	51.14	33.84
MV-CLIP-VG	TMM'23	CLIP-B / CLIP-B	RGB+TIR	45.57	46.01	72.62	41.56	16.06	34.52	38.92	63.42	50.05	10.01	47.02	47.52	82.76	45.07	23.96
MV-D-MDETR	TPAMI'24	CLIP-B / CLIP-B	RGB+TIR	55.26	47.42	68.33	41.25	23.38	48.81	46.50	65.91	53.82	24.19	58.76	57.31	83.84	53.79	39.30
MV-D-MDETR	TPAMI'24	RN50+DETR / BERT-B	RGB+TIR	55.01	47.84	68.10	43.77	24.52	44.64	45.62	67.23	53.88	21.81	54.73	54.13	82.89	52.00	34.24
MV-MMCA	ACMMM'24	RN50+DETR / BERT-B	RGB+TIR	54.93	48.97	72.43	42.34	22.62	46.43	47.83	72.16	56.76	20.45	53.62	54.41	84.41	51.47	34.16
MV-FSVG	ICME'25	RN50+DETR / BERT-B	RGB+TIR	54.61	47.96	68.57	40.16	24.29	47.62	46.32	64.61	54.74	25.00	58.44	58.60	85.86	55.18	40.05
MV-AttBalance	ACMMM'25	CLIP-B / CLIP-B	RGB+TIR	55.26	54.04	76.07	48.59	27.63	46.43	50.50	79.30	58.39	20.40	52.71	52.69	82.58	50.94	31.74
MV-HiVG-B	ACMMM'24	CLIP-B / CLIP-B	RGB+TIR	**75.33**	69.20	85.80	66.72	50.20	69.64	72.35	93.99	79.43	49.40	67.07	67.04	91.01	64.52	**51.53**
MV-HiVG-L	ACMMM'24	CLIP-L / CLIP-L	RGB+TIR	71.05	**72.50**	**91.07**	**70.17**	51.16	63.16	68.67	89.52	63.28	44.74	61.78	61.31	88.38	58.85	42.36
MV-OneRef-B	NeurIPS'24	BEIT3-B / BEIT3-B	RGB+TIR	63.82	62.05	86.19	57.34	35.43	62.50	63.98	89.20	69.34	37.93	61.62	59.97	89.25	56.85	41.58
MV-OneRef-L	NeurIPS'24	BEIT3-L / BEIT3-L	RGB+TIR	66.61	60.89	84.01	53.75	35.09	67.26	69.70	92.78	75.14	46.27	64.57	62.73	88.99	59.42	45.25
RGBT-VGNet	Ours	CLIP-B / CLIP-B	RGB+TIR	**75.33**	72.43	89.64	69.22	**52.53**	**73.21**	**74.89**	94.16	**80.75**	**55.52**	**68.31**	**67.58**	91.29	**64.80**	51.18

performance across all three sub-datasets, benefiting from the complementary information offered by RGB–TIR fusion. Uni-modality visual grounding baselines such as HiVG [35] and OneRef [36] also show competitive results within their respective settings. In addition, all models trained on the target dataset exhibit superior improvements over their zero-shot versions, confirming the inherent difficulty of the RGBT-GroundBench.

Multi-modal RGBT visual grounding. As shown in Table 4, multi-modal visual input consistently improves performance across all models, with an average improvement of approximately 10% in Acc@0.5 over uni-modal settings. All models exhibit noticeable improvements on testB and testC, demonstrating the advantages of multimodal visual input under small objects and low-light conditions. In particular, our RGBT-VGNet achieves the highest performance across all subsets, with Acc@0.5 exceeding 91%, 64%, and 49% on testA, testB, and testC, respectively, further validating its robustness and generalization under diverse lighting and object-size conditions. *More evaluation results are included in the supplementary materials.*

Table 5: Ablation of AMA, LAVS, and TPF in RGBT-VGNet. This table reports the effect of each practical component in the reference baseline under the same protocol. Best results are bold and underlined.

AMA	LAVS	TPF	RefFLIR					RefM^{3}FD					RefMFAD				
			val	test	testA	testB	testC	val	test	testA	testB	testC	val	test	testA	testB	testC
			55.42	46.19	67.42	41.56	22.81	54.16	57.57	78.89	65.63	34.46	53.50	54.63	81.77	51.91	36.62
✓			74.01	71.17	90.57	66.25	49.69	70.83	72.53	93.99	80.16	50.43	68.07	65.27	90.63	62.19	47.96
✓	✓		73.68	**72.65**	**91.31**	67.19	52.23	**73.21**	74.34	**94.72**	**81.93**	53.63	67.83	66.62	91.16	64.07	49.76
✓	✓	✓	**75.33**	72.43	89.64	**69.22**	**52.53**	**73.21**	**74.89**	94.16	80.75	**55.52**	**68.31**	**67.58**	**91.29**	**64.80**	**51.18**

Table 6: Comparison of different RGBT feature fusion method.

Fusion Method	Venue	RefFLIR					RefM^{3}FD					RefMFAD				
		val	test	testA	testB	testC	val	test	testA	testB	testC	val	test	testA	testB	testC
Naive Add	-	73.68	69.39	87.98	66.72	46.86	70.24	73.04	**94.40**	80.20	51.46	67.91	66.16	90.53	63.09	49.41
CMX	TITS'24	64.31	63.93	85.36	60.63	38.87	60.71	64.01	89.45	71.90	38.15	53.74	53.77	85.48	50.45	32.33
LIF	ICCV'25	74.34	71.39	**90.00**	67.50	49.70	69.64	72.96	93.99	80.47	51.30	67.59	66.33	90.28	63.66	49.96
TPF	**Ours**	**75.33**	**72.43**	89.64	**69.22**	**52.53**	**73.21**	**74.89**	94.16	**80.75**	**55.52**	**68.31**	**67.58**	**91.29**	**64.80**	**51.18**

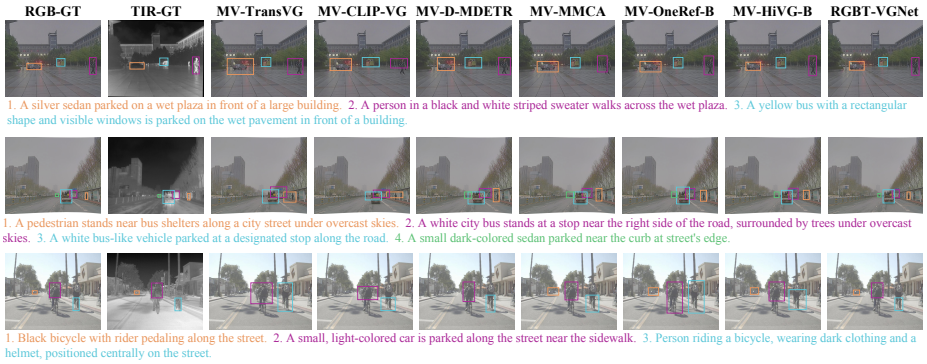


Fig. 5: Qualitative comparison of RGBT visual grounding on RGBT-GroundBench. Bounding box colors follow the legend below.

MLLM Evaluation. We additionally evaluate several representative multimodal large language models (MLLMs), including GLM-4.6V, Kimi-K2.5, and Qwen3.5-Plus, under both 224×224 and original-resolution settings. *The detailed experimental results and analysis are reported in the supplementary material.*

5.3 Ablation on RGBT-VGNet Components

Table 5 examines the effect of each component in the reference baseline. AMA brings the largest single-step gain over the no-module setting across all datasets and splits, indicating that modality adaptation is a prerequisite for stable RGB–TIR fusion. Adding LAVS on top of AMA further improves most subsets, with clear benefits on difficult splits (testB/testC), suggesting that language-guided cross-modal interaction improves robustness under weak visibility and small-target conditions. Adding TPF on top of AMA+LAVS yields additional gains on most hard subsets, especially RefFLIR-testB/testC, RefM^{3}FD-testC, and RefMFAD-testC, which is consistent with its illumination-aware design.

Table 6 compares different RGBT fusion strategies under the same evaluation protocol. The proposed TPF achieves the strongest overall performance,

obtaining the highest average accuracy across the three sub-datasets. Although simpler visual fusion methods such as Naive Add and LIF outperform TPF in a few specific cases, their performance is less consistent across datasets. In addition, the representative RGBT fusion baseline CMX shows a noticeable performance gap compared with the other approaches. Overall, the results indicate that the language-aware fusion design in TPF provides more robust cross-modal alignment, leading to stronger and more stable grounding performance.

5.4 Qualitative Comparisons and Analysis

The qualitative comparison of multi-modal RGBT visual grounding models under the RGBT-GroundBench protocol is shown in Figure 5. The objects are in various real-world conditions, including post-rain, foggy, and sunny weather, which presents challenges for models with different lighting, visibility, and object sizes. All results are evaluated using our framework adapted to the RGB–TIR modality. The results reveal that uni-modal visual grounding methods struggle with small or distant targets, often producing overly large or inaccurate boxes. In contrast, RGBT-VGNet consistently delivers precise and semantically consistent detections for humans and vehicles, effectively leveraging complementary cues from RGB and Thermal modalities under textual guidance. This demonstrates its robustness and generalization in various complex scenarios.

6 Discussion and Conclusion

In this paper, we present RGBT-GroundBench for RGB-Thermal visual grounding in complex real-world environments, which contains over 40K images (21,535 RGB-TIR pairs) and 38,760 object instances with fine-grained annotations in various complex scenes. By conducting large-scale experiments to evaluate 11 VG models under the proposed RGBT-GroundBench protocol, we draw some important findings, which are summarized below:

❶ **Scene Complexity Correlates with Grounding Accuracy.** VG models are stable in simpler environments (see Table 4-testA) but degrade in cluttered scenes (*Supplementary* Tables 11, 12, 13, 16 and Fig. 5), especially with distractors, occlusion, or small off-center targets. This suggests sensitivity to background complexity and object density, potentially amplified by dataset biases (e.g., salient/centered objects). Overall, realistic grounding demands stronger context modeling and more robust disambiguation under clutter and occlusion.

❷ **LoRA-based Adaptation Enhances Robustness.** LoRA-based grounding models show more stable behavior across diverse scenes (see Table 4 and Table 5). This suggests that parameter-efficient adaptation helps align pretrained vision-language representations with grounding objectives under domain shifts and challenging environments. The results highlight that adaptation strategies can play an important role in improving robustness beyond architectural design.

❸ **Low Illumination Exposes Robustness Gap.** Low-light conditions consistently degrade performance (see Table 4-testB and *Supplementary* Table 15), despite being underexplored in prior benchmarks. This suggests that

illumination variation is a major failure mode for models trained on well-lit RGB imagery. Thermal cues help under low illumination, but grounding remains difficult, especially for small targets or adverse weather (see Table 4-testC and *Supplementary* Tables 14 and 16).

❹ **MLLMs Struggle with Cross-Spectral Grounding.** Results from the MLLM evaluation (see *Supplementary* Tables 8 and 9) indicate that naively adding thermal inputs does not yield consistent gains in localization accuracy, and performance varies markedly with spatial resolution. This suggests that existing MLLMs still struggle to reliably fuse cross-spectral cues for precise spatial grounding.

Based on these observations, we propose RGBT-VGNet as a baseline model to achieve satisfactory performance in complex scenarios covered by RGBT-GroundBench. Overall, we hope our comprehensive benchmark, insightful findings, and competitive baseline method can be helpful for evaluating the performance of VG models in various complex scenes and improving their grounding accuracy in the future.

Acknowledgement

This work was supported in part by the Project of the National Natural Science Foundation of China under Grant 62576020, in part by the Fundamental Research Funds for the Central Universities and in part by the China Postdoctoral Science Foundation under Grant Number 2025M784289.

References

1. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV. pp. 104–120. Springer (2020) 4
2. Chen, Z., Hu, R., Chen, X., Nießner, M., Chang, A.X.: Unit3d: A unified transformer for 3d dense captioning and visual grounding. In: ICCV. pp. 18109–18119 (2023) 5
3. Chen, Z., Zhang, R., Song, Y., Wan, X., Li, G.: Advancing visual grounding with scene knowledge: Benchmark and method. In: CVPR. pp. 15039–15049 (2023) 11
4. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: ICCV. pp. 1769–1779 (2021) 4, 8, 9, 12, 23, 24, 25, 26, 27, 28
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019) 4
6. Do, D.P., Kim, T., Na, J., Kim, J., Lee, K., Cho, K., Hwang, W.: D3t: Distinctive dual-domain teacher zigzagging across rgb-thermal gap for domain-adaptive object detection. In: CVPR. pp. 23313–23322 (2024) 5
7. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4

8. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. NeurIPS **33**, 6616–6628 (2020) 4
9. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In: International workshop ontoImage. vol. 2 (2006) 2
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 4
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022) 9
12. Hu, K., He, Y., Li, Y., Zhao, J., Chen, S., Kang, Y.: Ei 2 det: Edge-guided illumination-aware interactive learning for visible-infrared object detection. TCSVT (2025) 2, 5, 6
13. Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your large vision-language model only needs a few attention heads for visual grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9339–9350 (2025) 4
14. Kang, W., Zhou, L., Wu, J., Sun, C., Yan, Y.: Visual grounding with attention-driven constraint balancing. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 1637–1645 (2025) 4, 8, 12, 23, 24, 25, 26, 27, 28
15. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP. pp. 787–798 (2014) 2
16. Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al.: mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In: EMNLP. pp. 7241–7259 (2022) 4
17. Li, G., Wang, Y., Liu, Z., Zhang, X., Zeng, D.: Rgb-t semantic segmentation with location, activation, and sharpening. TCSVT **33**(3), 1223–1235 (2022) 5
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014) 2
19. Liu, H., Lin, A., Han, X., Yang, L., Yu, Y., Cui, S.: Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In: CVPR. pp. 6032–6041 (2021) 5
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023) 2
21. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: CVPR. pp. 5802–5811 (2022) 2, 5, 6
22. Lou, H., Fan, C., Liu, Z., Wu, Y., Wang, X.: Llava-sp: Enhancing visual representation with visual spatial tokens for mllms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22014–22024 (2025) 2
23. Miyanishi, T., Azuma, D., Kurita, S., Kawanabe, M.: Cross3dvg: Cross-dataset 3d visual grounding on different rgb-d scans. In: 2024 International Conference on 3D Vision (3DV). pp. 717–727. IEEE (2024) 5
24. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: ECCV. pp. 792–807. Springer (2016) 2
25. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV. pp. 2641–2649 (2015) 2

26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICLR. pp. 8748–8763. PmLR (2021) 2, 9
27. Shi, F., Gao, R., Huang, W., Wang, L.: Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(2), 1181–1198 (2023) 4, 8, 12, 23, 24, 25, 26, 27, 28
28. Song, C., Li, H., Xu, T., Wu, X.J., Kittler, J.: Refinefuse: an end-to-end network for multi-scale refinement fusion of multi-modality images. Visual Intelligence **3**(1), 16 (2025) 5
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017) 4
30. Wang, J., Zhang, W., Huang, D., Wang, H., Zheng, Y.: A simple and better baseline for visual grounding. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025) 4, 8, 12, 23, 24, 25, 26, 27, 28
31. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: ICLR. pp. 23318–23340. PMLR (2022) 4
32. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: CVPR. pp. 19175–19186 (2023) 2
33. Wu, S., Jin, S., Zhang, W., Xu, L., Liu, W., Li, W., Loy, C.C.: F-lmm: Grounding frozen large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24710–24721 (2025) 4
34. Xiao, L., Yang, X., Lan, X., Wang, Y., Xu, C.: Towards visual grounding: A survey. arXiv preprint arXiv:2412.20206 (2024) 11
35. Xiao, L., Yang, X., Peng, F., Wang, Y., Xu, C.: Hivg: Hierarchical multimodal fine-grained modulation for visual grounding. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 5460–5469 (2024) 4, 8, 10, 12, 23, 24, 25, 26, 27, 28
36. Xiao, L., Yang, X., Peng, F., Wang, Y., Xu, C.: Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling. NeurIPS **37**, 139854–139885 (2024) 8, 12, 23, 24, 25, 26, 27, 28
37. Xiao, L., Yang, X., Peng, F., Yan, M., Wang, Y., Xu, C.: Clip-vg: Self-paced curriculum adapting of clip for visual grounding. TMM **26**, 4334–4347 (2023) 4, 8, 12, 23, 24, 25, 26, 27, 28
38. Yang, Z., Chen, T., Wang, L., Luo, J.: Improving one-stage visual grounding by recursive sub-query construction. In: ECCV. pp. 387–404. Springer (2020) 4
39. Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: Unitab: Unifying text and box outputs for grounded vision-language modeling. In: ECCV. pp. 521–539. Springer (2022) 4
40. Yao, R., Xiong, S., Zhao, Y., Rong, Y.: Visual grounding with multi-modal conditional adaptation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3877–3886 (2024) 8, 12, 23, 24, 25, 26, 27, 28
41. Ye, J., Tian, J., Yan, M., Yang, X., Wang, X., Zhang, J., He, L., Lin, X.: Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In: CVPR. pp. 15502–15512 (2022) 4
42. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV. pp. 69–85. Springer (2016) 2

43. Yuan, M., Cui, B., Zhao, T., Wang, J., Fu, S., Yang, X., Wei, X.: Unirgb-ir: A unified framework for visible-infrared semantic tasks via adapter tuning. In: ACM MM. pp. 2409–2418 (2025) 3, 5
44. Yuan, M., Meng, D., Xi, Z., Zhao, T., Zhao, S., Dai, Y., Wei, X.: Seeing through the noise: Improving infrared small target detection and segmentation from noise suppression perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27783–27792 (2026) 5
45. Yuan, M., Shi, X., Wang, N., Wang, Y., Wei, X.: Improving rgb-infrared object detection with cascade alignment-guided transformer. Information Fusion **105**, 102246 (2024) 3
46. Yuan, M., Wei, X.: C^2former: Calibrated and complementary transformer for rgb-infrared object detection. TGRS **62**, 1–12 (2024) 5
47. Zhang, H., Fromont, E., Lefevre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: ICIP. pp. 276–280. IEEE (2020) 2, 5, 6
48. Zhang, Q., Zhao, S., Luo, Y., Zhang, D., Huang, N., Han, J.: Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In: CVPR. pp. 2633–2642 (2021) 5
49. Zhao, S., Liu, Y., Jiao, Q., Zhang, Q., Han, J.: Mitigating modality discrepancies for rgb-t semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems **35**(7), 9380–9394 (2023) 3
50. Zhao, T., Liu, B., Gao, Y., Sun, Y., Yuan, M., Wei, X.: Rethinking multi-modal object detection from the perspective of mono-modality feature learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6364–6373 (2025) 3, 5
51. Zhao, T., Yuan, M., Jiang, F., Wang, N., Wei, X.: Removal then selection: A coarse-to-fine fusion perspective for rgb-infrared object detection. IEEE Transactions on Intelligent Transportation Systems (2025) 3, 5