RT-RMOT: A Dataset and Framework for RGB-Thermal Referring Multi-Object Tracking

Yanqiu Yu[1*], Zhifan Jin[2*], Sijia Chen[1*], Tongfei Chu[1], En Yu[1], Liman Liu[2†], and Wenbing Tao[1†]

[1] Huazhong University of Science and Technology, China
{yanqiuyu6, sijiachen, tongfeichu, yuen, wenbingtao}@hust.edu.cn
[2] South-Central Minzu University, China
{zhifanjin, limanliu}@mail.scuec.edu.cn
https://github.com/yan-qiu-yu/RT-RMOT

Abstract. Referring Multi-Object Tracking has attracted increasing attention due to its human-friendly interactive characteristics, yet it exhibits limitations in low-visibility conditions, such as nighttime, smoke, and other challenging scenarios. To overcome this limitation, we propose a new **R**GB-**T**hermal **RMOT** task, named **RT-RMOT**, which aims to fuse RGB appearance features with the illumination robustness of the thermal modality to enable all-day referring multi-object tracking. To promote research on RT-RMOT, we construct the first **Ref**erring Multi-Object Tracking dataset under **R**GB-**T**hermal modality, named **RefRT**. It contains 388 language descriptions, 1,250 tracked targets, and 166,147 Language-RGB-Thermal (L-RGB-T) triplets. Furthermore, we propose **RTrack**, a framework built upon a multimodal large language model (MLLM) that integrates RGB, thermal, and textual features. Since the initial framework still leaves room for improvement, we introduce a Group Sequence Policy Optimization (GSPO) strategy to further exploit the model's potential. To alleviate training instability during RL fine-tuning, we introduce a Clipped Advantage Scaling (CAS) strategy to suppress gradient explosion. In addition, we design Structured Output Reward and Comprehensive Detection Reward to balance exploration and exploitation, thereby improving the completeness and accuracy of target perception. Extensive experiments on the RefRT dataset demonstrate the effectiveness of the proposed RTrack framework.

Keywords: Language-RGB-Thermal · Reinforcement Learning · Referring Multi-Object Tracking · Multi-modal Large Language Model

1 Introduction

With the recent advances in vision-language understanding, language-guided object detection has enabled objects to be localized according to textual descriptions [9]. Built upon this capability, Referring Multi-Object Tracking (RMOT)

* Authors contributed equally. † Corresponding author.

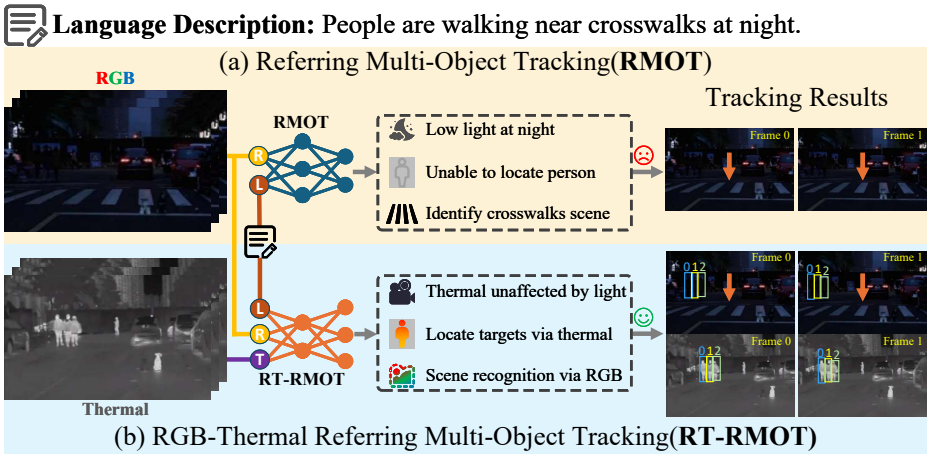


Fig. 1: The difference between RMOT and RT-RMOT. The RMOT model takes RGB images as input, but pedestrian positions cannot be reliably obtained from RGB alone, often leading to tracking failure. In contrast, the RT-RMOT model integrates information from three modalities: language, RGB, and thermal. Specifically, it utilizes thermal (T) images to capture pedestrian contours and positions, and combines this information with the crosswalk regions provided by RGB images, thereby enabling the model to accurately identify and track people according to the language description.

has emerged as a challenging and active research topic due to its human-friendly interactive characteristics. RMOT aims to continuously track specific targets according to language descriptions. Numerous RMOT approaches have appeared, including TransRMOT [26], TempRMOT [34], and CPAny [17].

However, current RMOT research is largely constrained to RGB observation modality. In low-visibility conditions, such as nighttime and smoke, the appearance information of targets often degrades severely, making it difficult for the model to establish reliable semantic associations between compromised visual features and language descriptions. RMOT task paradigm inherently limits the robustness of language-guided tracking systems in complex and dynamic environments, making it inadequate for all-day real-world applications. For example, in nighttime, tunnel, or smoky scenarios, RGB images often become dark or noisy and fail to provide clear contour information of the targets, making it difficult for the model to accurately localize targets.

To overcome this limitation, we propose a new task, **R**GB-**T**hermal **R**eferring **M**ulti-**O**bject **T**racking (**RT-RMOT**). This task requires the model to deeply fuse language, RGB, and thermal multimodal data, enabling robust all-day referring multi-object tracking even under low-visibility conditions such as nighttime and smoke. In this task, the language modality allows users to specify targets in a human-friendly manner, the RGB modality captures fine-grained appearance details to distinguish between targets, and the infrared modality provides contour information to facilitate accurate localization. Such a integrated multimodal

paradigm significantly enhances the reliability of referring multi-object tracking in complex environments. As shown in Fig. 1, for the language description "people walking near a crosswalk at night", RT-RMOT task recognizes the crosswalk using RGB images and detects pedestrians using thermal images, enabling accurate identification of the people mentioned in the language description.

To advance the research of RT-RMOT task, we construct the first **Ref**erring **R**GB-**T**hermal Multi-Object Tracking dataset, named **RefRT**. RefRT dataset consists of 388 high-quality language descriptions, 72 scenes, 1,250 referred targets and 166,147 language-RGB-thermal triplets, covering diverse scenarios such as campuses and urban areas, as well as imaging conditions including night, rain, and snow. It provides a high-quality and adaptable foundation for RMOT research under RGBT scenarios.

To address the challenges of multimodal feature fusion and effective information utilization in the RT-RMOT task, we propose a framework, termed **RTrack**. Centered on a Multimodal Large Language Model (MLLM) with strong cross-modal alignment and reasoning capabilities, RTrack achieves joint representation learning of language, RGB and thermal features for referring tracking. As the initial framework still leaves room for performance improvement, we further introduce a Group Sequence Policy Optimization (GSPO) [36] strategy to better unlock the model's potential. To ensure stable fine-tuning, we propose a novel Clipped Advantage Scaling (CAS) strategy to effectively suppress gradient fluctuations commonly observed in reinforcement learning. In addition, to balance exploration and exploitation, we carefully design Structured Output and Comprehensive Detection Reward that guides the model to maintain target perception integrity and spatial precision in challenging environments.

Experimental results demonstrate that RTrack significantly achieves state-of-the-art (SOTA) performance. Our method achieves the highest scores across five evaluation metrics: HOTA, DetA, AssA, DetRe, and AssRe. Among these, the HOTA metric provides a comprehensive assessment of both detection and tracking performance. Compared with other methods, our approach improves the HOTA metric by 6.84%, DetA by 9.8%, DetRe by 17.1%, and DetPR by 5.64%, demonstrating strong generalization capability on the RefRT dataset.

In summary, the main contributions are as follows:

- We propose a **R**GB-**T**hermal **R**eferring **M**ulti-**O**bject **T**racking (**RT-RMOT**) task, which aims to integrate the high-level semantic information from the language modality, RGB appearance features, and the contour cues from the thermal modality, thereby overcoming the perceptual limitations of traditional RMOT under low-visibility conditions such as nighttime and smoke, and enabling all-day referring multi-object tracking.
- We construct **RefRT**, the first dataset specifically designed for RT-RMOT task. It provides pixel-level alignment between RGB and thermal modalities, along with high-quality language descriptions. It contains 388 language descriptions, 1,250 referred targets, and 72 scenes.
- We propose **RTrack**, a unified framework guided by a Multimodal Large Language Model (MLLM) that integrates language, RGB and thermal in-

formation. By effectively exploiting the complementary cues of both the three modalities, RTrack achieves precise cross-modal fusion and all-day tracking. Extensive experiments on the RefRT dataset achieve state-of-the-art (SOTA) performance, highlighting the effectiveness of the framework.

2 Related Work

Referring Multi-Object Tracking. It aims to simultaneously track multiple language-specified objects in video. Traditional MOT methods, TLTDMOT [4], OVTR [29], RelationTrack [28], Motrv3 [16] provide effective trajectory modeling, relation reasoning, and open-vocabulary tracking paradigms. TransRMOT [26] introduces the RMOT task, proposing an end-to-end Transformer framework that handles cross-modal reference and temporal association. Subsequent works, iKUN [8], CRTrack [5], ReaMOT [6], CPAny [17] focus on optimizing cross-modal fusion, cross-view association, reasoning ability, and tracking paradigms within the dominant Transformer architecture. However, these methods often struggle with dynamic references involving temporal status changes (e.g., "the moving vehicles"). To enhance robustness against these challenges, TempRMOT [34] presents a query-driven Temporal Enhancement Module (TEM), which refines object queries using long-term spatiotemporal interaction with historical frames, significantly boosting temporal consistency.

RGB-Thermal Object Tracking. RGBT tracking integrates RGB and thermal infrared (T) data to improve robustness against challenges such as poor illumination and occlusion. Early methods, MANet [20], CMPP [30], MaCNet [25] relied on shallow feature fusion via handcrafted features or correlation filters. With the advent of deep learning, the focus shifted toward modality-adaptive representation [33], utilizing dual-stream networks and dynamic attention or distillation mechanisms to balance modality contributions. Most recently, Transformer architectures [27] have been adopted for their global modeling capacity to address modality heterogeneity and long-range dependencies. Notably, DeformCAT [11] introduces a deformable cross-attention mechanism that flexibly samples cross-modal features, significantly enhancing robustness to target shape and positional variations.

Multimodal Large Language Models. They have rapidly progressed from simple image-text alignment to complex multimodal perception and reasoning. Foundational work like CLIP [22] established contrastive learning for image-text alignment, enabling visual representations to be associated with natural language semantics. The shift toward reasoning was pioneered by LLaVA [19], which integrated Multimodal Large Language Models (MLLMs) via visual instruction tuning and enhanced their ability to follow language-based visual queries. Subsequent powerful models, including GPT-4 [1] and Gemini [24], demonstrated stronger unified perception, language understanding, and reasoning capabilities across diverse multimodal inputs. Recent efforts, such as Qwen2.5-VL [2] and InternVL2 [7], further focus on enhancing fine-grained visual grounding, video comprehension, and overall generalization capacity through efficient fusion

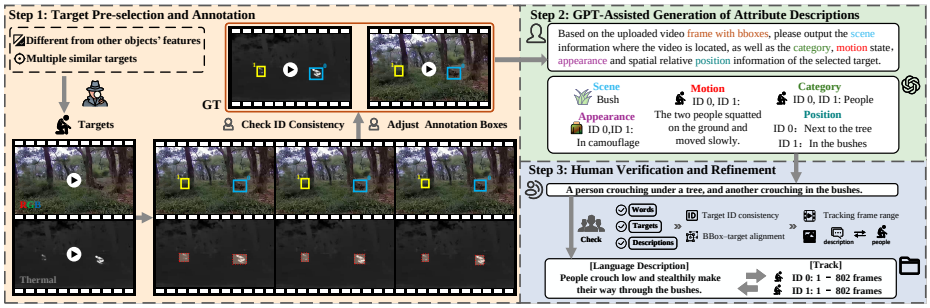


Fig. 2: Data annotation process. The annotation process consists of three steps: (1) Target Pre-selection and Annotation; (2) GPT-Assisted Generation of Attribute Descriptions; (3) Human Verification and Refinement. First, visually similar targets are selected, and bounding boxes are annotated across the entire sequence. Then, annotated key frames and task instructions are provided to GPT to analyze the target's category, scene, appearance, motion, and spatial position. Finally, GPT-generated attributes are integrated into initial descriptions, which is refined through multiple annotators' review to produce the final language descriptions.

and scaling strategies. These advances provide strong foundations for language-guided visual understanding, but adapting them to temporally consistent target localization and tracking remains a key challenge for RMOT.

3 Dataset

To advance research on the RGB-Thermal Referring Multi-Object Tracking (RT-RMOT) task, address the scarcity of datasets in this field, and evaluate model performance, we construct the RefRT dataset.

3.1 Dataset Collection

Our task requires the source dataset to have three core characteristics: (1) **Diverse Data Samples.** The dataset must include a variety of real-world scenes and object categories to provide sufficient and diverse samples for subsequent language annotations and support multi-object labeling. (2) **RGB-Thermal Spatial Alignment.** Each video frame must contain a pair of pixel-level aligned RGB and thermal infrare (T) images to facilitate effective information fusion across the RGB-T modalities. (3) **Continuous Video Sequences.** The original dataset should provide complete video streams containing continuous frames, in which the full motion trajectories of the targets can be observed throughout the frame sequence. After extensive evaluation across multiple datasets, we selected two public datasets, LasHeR [14] and VTUAV [32], that satisfy the aforementioned three criteria as the foundation of our dataset.

3.2 Dataset Annotation

To construct high-quality RGB-T modality data for referring multi-object tracking trajectories, we construct RefRT dataset based on two fundamental datasets using a GPT [1] assisted manual annotation strategy. The annotation process is depicted in Fig. 2. The detailed workflow is as follows:

Step 1: Target Pre-selection and Annotation. First, the annotator examines the aligned RGB-T video sequence and pre-selects several targets that share similar characteristics but are distinguishable from other objects. The annotator then labels bounding boxes for these targets across all video frames. During this process, the annotator ensures the accuracy of the bounding boxes and verifies the consistency of target IDs throughout the sequence.

Step 2: GPT-assisted Generation of Attribute Descriptions. We feed GPT with RGB-T aligned annotated frames containing bounding boxes and target IDs, along with structured instructions specifying the attributes to be analyzed. GPT then generates key attributes such as target category, appearance features, spatial location, and other visual cues.

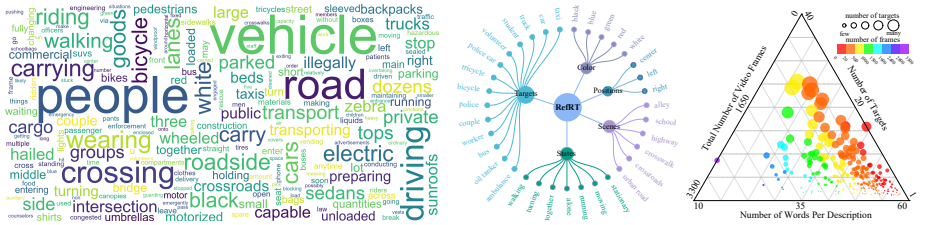
Step 3: Human Verification and Refinement. First, an annotator organizes the attributes generated by GPT into an initial language description. Then, multiple reviewers verify the spelling, the consistency and validity of the description, and the accuracy of bounding boxes, as well as the correctness of target IDs and the alignment of bounding boxes with their corresponding targets. Finally, they check the tracking frame ranges of the targets and ensure that the described targets match the language description. After this review process, we obtain the final language description, along with the target IDs and the corresponding video frame ranges for each target matching the description.

3.3 Dataset Split

To ensure that the model has sufficient data to learn the relationships between RGB and thermal modalities during training while retaining enough unseen samples for evaluation. We split all video sequences in the RefRT dataset into training and test sets at a ratio of 6:4. The language descriptions and annotated targets associated with each video sequence are assigned to the corresponding subset along with the sequence. This results in a training set containing 42 videos and a test set containing 30 videos.

3.4 Dataset Statistics

The RefRT dataset contains 72 scenes, 166,147 RGB-Thermal-Language triplets, 388 language descriptions, and 1,250 targets with language annotations. It features pixel-level RGB-thermal alignment, covers both standard and low visibility scenarios, and includes diverse targets such as pedestrians and vehicles. Below, we present a comprehensive analysis of the dataset.



(a) (b) (c)

Fig. 3: Visualization results on the RefRT dataset. (a) The word cloud of the RefRT dataset contains a rich set of keywords. (b) The RefRT dataset covers diverse scenes, targets and attributes. (c) We visualize all samples in the RefRT dataset from three perspectives: the number of targets, the number of video frames, and the number of words in each language description. Each circle represents a data sample. The distribution illustrates the diversity and complexity of the RefRT dataset.

Word Cloud. We perform a word cloud analysis of the RefRT dataset, and the results are shown in Fig. 3 (a). RefRT encompasses rich semantic information and provides detailed language descriptions through multi-dimensional descriptions, including category attributes, behavioral characteristics, and spatial locations. This highlights the diversity of the RefRT dataset.

RefRT Language Analysis. As shown in Fig. 3 (b), we conduct a detailed analysis of the language descriptions in the RefRT dataset. These descriptions cover common scenes such as urban areas, roads, and schools, demonstrating the dataset's strong real-world relevance. Moreover, to align with the RT-RMOT task, the language descriptions focus on targets capable of emitting thermal radiation, such as pedestrians, cars, and electric vehicles. The associated language descriptions capture behavioral cues, scene attributes, and inter-object interactions, further highlighting the dataset's rich semantic complexity.

Data Distribution Characteristics. Each data point in Fig. 3 (c) represents a semantic description, where the three axes denote word count, tracking frames, and target numbers. The figure shows that the RefRT dataset is broadly and evenly distributed across these dimensions. By including descriptions of varying lengths and complexities, RefRT meets model learning requirements. Variations in target instances challenge the model's tracking ability, while different frame spans evaluate its generalization under diverse temporal dynamics.

4 Methodology

To address this challenge of the RT-RMOT tasks, we propose RTrack, which leverages the perception capability of Multimodal Large Language Models (MLLMs) to detect and associate multiple targets.

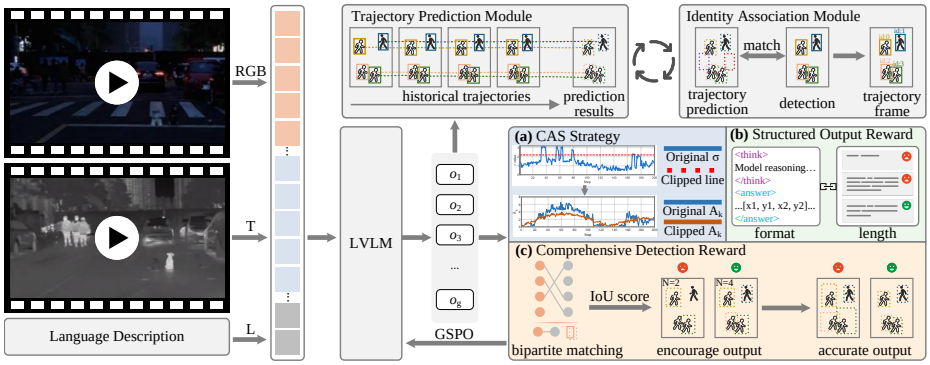


Fig. 4: Overall pipeline of RTrack. The RTrack framework consists of three modules: (1) Large-model perception module: performs inference detection of the target in the current frame based on the language description. (2) Trajectory Prediction Module: predicts the target box in the current frame based on the historical trajectory. (3) Identity Association Module: matches the trajectory box and the detection box to generate the target ID. Among them, the GSPO algorithm includes three aspects: (a) CAS strategy: constrains relative rewards to avoid gradient explosion. (b) Structured output reward: requires the model to format the output and limits the output length. (c) Comprehensive detection reward: completes the output encouragement and accurate output based on the IoU value.

4.1 Framework

We propose **RTrack**, a framework for the RGB-Thermal Referring Multi-Object Tracking (RT-RMOT) task. It leverages the strengths of **Multi-modal Large Language Models (MLLMs)** in RGBT modal fusion and semantic reasoning. Unlike traditional RMOT models that only process RGB images, RTrack can accurately locate and continuously track referred targets in RGBT scenarios. As illustrated in Fig. 4, RTrack consists of three main components: **Large-model Perception Module**, **Trajectory Prediction Module**, and **Identity Association Module**.

MLLM-based Perception Module. We adopt a Multimodal Large Language Model (MLLM) as the perception backbone. Compared with task-specific detectors, large-scale multimodal pretraining provides stronger cross-modal alignment and a more robust understanding of complex language expressions, which are crucial for jointly modeling language, RGB, and thermal inputs in RT-RMOT. Specifically, given an aligned pair of RGBT images $\{I^{\mathrm{V}}, I^{\mathrm{T}}\}$ and a language description L, the MLLM sequentially performs visual encoding E_v, text encoding E_l, modality fusion *Fuse*, cross-modal attention *CrossAttn*, and visual instruction reasoning *MLLM*, ultimately obtaining the spatial location of the target specified by the language description in the RGBT image.

Trajectory Prediction Module. This module dynamically models historical trajectory sequences to analyze motion trends and generate prior predictions of target positions in the current frame. It uses a Kalman Filter [12] to linearly

model the targets' states, including position and velocity, and performs tracking through a prediction-update loop. In each frame, the filter first predicts the target positions from the previous state and then updates the state using the detection results from the Perception Module.

Identity Association Module. This module maintains target identity consistency across consecutive frames and ensure trajectory continuity. Specifically, the system first generates candidate targets for the current frame based on the output of the Perception Module, and utilizes predicted bboxes from the Trajectory Prediction Module to construct an IoU-based cost matrix between the predicted and detection bboxes. On this basis, the Hungarian Matching Algorithm [13] is applied to achieve global optimal matching, thereby determining the identity correspondences between detection bboxes and historical trajectories. For successfully matched trajectories, the system updates their states using current detections; unmatched detections are initialized as new trajectories, while unmatched trajectories that exceed their predefined lifespan are considered lost and consequently terminated from tracking. The algorithm of the identity association module is illustrated in Algorithm 1.

4.2 RL Fine-tuning based on the GSPO algorithm

To further enhance cross-modal fusion capability, language understanding, and tracking performance of the RTrack framework on the RefRT dataset, we employ the Group Sequence Policy Optimization (GSPO) [36] algorithm to fine-tune the model using reinforcement learning. To enhance the stability of training and cross-modal detection capability, we make improvements in the GSPO framework from both the policy and reward dimensions.

Group Sequence Policy Optimization (GSPO) is a sequence-level policy optimization algorithm for the reinforcement learning stage of large language models. It treats the entire generated sequence as the optimization unit, aligning the granularity of policy updates with reward evaluation. For each input, the model generates multiple candidate sequences, computes their average reward as a baseline, and updates the policy by relative rewards. Unlike token-level methods (e.g., GRPO [23]), GSPO optimizes at the sequence level with an importance ratio and length-normalized clipping mechanism, effectively reducing training noise and instability. Its objective is formulated as:

$$\begin{aligned}\mathcal{J}_{\mathrm{GSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_k\}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{k=1}^{G}\min(s_1 A_k, s_2 A_k)\right] \\ - \beta_{\mathrm{KL}}\,\mathbb{E}_{x\sim\mathcal{D}}\left[D_{\mathrm{KL}}\big(\pi_\theta(\cdot|x)\,\|\,\pi_{\theta_{\mathrm{old}}}(\cdot|x)\big)\right]\end{aligned} \tag{1}$$

$$s_1(\theta) = \left(\frac{\pi_\theta(y_k \mid x)}{\pi_{\theta_{\mathrm{old}}}(y_k \mid x)}\right)^{\frac{1}{|y_k|}} \tag{2}$$

$$s_2 = \mathrm{clip}\left(s_1(\theta), 1+\varepsilon, 1-\varepsilon\right) \tag{3}$$

Algorithm 1 Identity Association Module

Input: Detection boxes $\mathcal{D}$; predicted trajectories $\mathcal{T}^{pred}$.
Parameter: Calculation operation $\mathcal{C}()$; Hungarian algorithm $\mathcal{H}()$; IoU cost matrix M_{iou}; matched/unmatched sets $\{\mathcal{T}^m, \mathcal{T}^u, \mathcal{D}^m, \mathcal{D}^u\}$; maximum lifespan δ_{max}.

/* IoU-based matching */
$M_{iou} \leftarrow \mathcal{C}(\mathcal{D}, \mathcal{T}^{pred})$
$(\mathcal{T}^m, \mathcal{T}^u, \mathcal{D}^m, \mathcal{D}^u) \leftarrow \mathcal{H}(M_{iou})$
/* Update matched trajectories */
for all $(\mathcal{T}_i^m, \mathcal{D}_i^m)$ **do**
    Update $\mathcal{T}_i^m$ with $\mathcal{D}_i^m$ at frame F_c
    $\mathcal{T}_i^m$.status $\leftarrow$ *active*; reset missing count
end for
/* Initialize new trajectories */
for all $\mathcal{D}_i^u$ **do**
    $\mathcal{T}_i^{new} \leftarrow \text{Init}(\mathcal{D}_i^u, F_c)$
    $\mathcal{T}_i^{new}$.status $\leftarrow$ *active*; add to $\mathcal{T}$
end for
/* Handle unmatched trajectories */
for all $\mathcal{T}_i^u$ **do**
    $\mathcal{T}_i^u$.missing_count $\leftarrow \mathcal{T}_i^u$.missing_count $+ 1$
    if $\mathcal{T}_i^u$.missing_count $> \delta_{max}$ **then**
        $\mathcal{T}_i^u$.status $\leftarrow$ *terminated*
    else
        $\mathcal{T}_i^u$.status $\leftarrow$ *temporary*
    end if
end for
return $\mathcal{T}$

Where ε and β_{KL} are hyperparameters, and A_k denotes the relative advantage of the k-th generated sequence, $x \sim \mathcal{D}$ denotes an input sample drawn from the data distribution $\mathcal{D}$, $\pi_\theta(\cdot \mid x)$ represents the current policy to be optimized, $y_k \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ is the k-th sequence generated by the old policy, G is the number of candidate sequences per input, and $D_{\text{KL}}(\pi_\theta \| \pi_{\theta_{\text{old}}})$ represents the KL divergence between the current and old policies.

Clipped Advantage Scaling (CAS) Strategy. To mitigate gradient explosion that may arise when the relative rewards are standardized, leading to abnormal amplification of the advantage term, especially when the reward distribution within the group becomes concentrated, we constructed a clipped, normalized advantage based on the CAS strategy. Specifically, given a set of scalar rewards $\{r_k\}_{k=1}^G$ with their group mean μ and standard deviation σ, the formula for calculating the new normalized advantage is as follows:

$$A_k \;=\; (r_k - \mu) \cdot \text{clip}\left(\frac{1}{\sigma},\, 0,\, \mathit{Scale}_{\max}\right) \tag{4}$$

where $\mathit{Scale}_{\max}$ controls the maximum global amplification factor, indicating the maximal multiplier of the advantage during advantage normalization.

Rule-based Reward Function. To effectively guide the model in generating structured and high-precision outputs, we comprehensively consider the reward rules of the task and design a rule-based composite reward mechanism consisting of two weighted parts, namely the structured output reward R_{str} and the comprehensive detection reward R_{ctr}.

Structured Output Reward. To standardize the output and balance exploration and exploitation, we introduce a structured output reward. The structured output reward consists of two parts: the format reward and the length reward, which are weighted and summed.

Specifically, the model's reasoning process must be enclosed within the `<think></think>` tags, and the final conclusion within the `<answer></answer>` tags. To ensure detectable outputs, the conclusion must include at least one coordinate set in the format `[x1,y1,x2,y2]`. If the output fully meets these formatting rules, $R_{\text{format}} = 1$; otherwise, $R_{\text{format}} = 0$. In addition, a sine-window–smoothed length reward $R_{\text{len}}(L) \in [0, 1]$ is applied to regularize response length.

$$s(x) = \min\left(\max\left(\sin^2\left(\frac{\pi}{2}x\right), 0\right), 1\right) \tag{5}$$

$$R_{\text{len}}(L) = s\left(\frac{L - L_{\min}}{L_{\text{low}} - L_{\min}}\right) \cdot s\left(\frac{L_{\max} - L}{L_{\max} - L_{\text{high}}}\right) \tag{6}$$

where L means the response length and is obtained by summing over the completion mask.

Comprehensive Detection Reward. To balance the comprehensiveness and accuracy of multi-object preception, we propose a reward based on the Intersection over Union (IoU). The reward comprises two components: **Output Encouragement Reward** and **Precision Detection Reward**. Given the detection set B_{det} and ground-truth set B_{gt}, we compute the IoU matrix after mapping all detected boxes to the same coordinate space as B_{gt}. A cost matrix is then constructed based on the IoU values, and the Hungarian algorithm is applied for optimal one-to-one matching. The weighted sum of matched IoUs, denoted as IoU_{score}, serves as the basis for subsequent reward computation.

To enhance the model's exploration ability and encourage it to identify as many potential targets related to the language instructions as possible, we designed **Output Encouragement Reward** that encourages the model to output more predictions. As the number of valid matched boxes detected by the model increases, its reward also increases, thereby guiding the model to identify and cover more target areas, striving to ensure that all ground truth targets are covered. This reward can be defined as:

$$R_{\text{OER}}(B_{\text{det}}, B_{\text{gt}}) = \alpha \cdot MatchedGT + \beta \cdot \text{IoU}_{\text{score}} \tag{7}$$

where $MatchedGT$ denotes the number of valid ground-truth boxes that are successfully matched with predictions.

In the later stages of training, to further improve the model's detection accuracy, we introduce the **Precision Detection Reward**. This reward not only

Table 1: Performance of many methods on the RefRT test set. ↑ indicates that higher scores are better. The best results are marked in **bold**. L represents Language.

Modality	Method	Venue	HOTA↑	DetA↑	AssA↑	DetRe↑	DetPr↑	AssRe↑	AssPr↑	LocA↑
L-RGB	TransRMOT [26]	CVPR 2023	8.69	2.57	29.96	3.01	14.46	30.73	85.49	**79.63**
	TempRMOT [34]	ArXiv 2024	8.19	1.86	36.23	2.04	16.68	39.28	75.39	77.48
	CRTracker [5]	AAAI 2025	9.30	2.37	37.01	3.81	5.83	40.10	67.48	73.25
	YOLOX+ByteTrack+iKUN [8]	CVPR 2024	2.32	0.29	19.86	0.29	12.71	21.18	61.45	69.70
	ReferGPT [3]	CVPR 2025	5.41	1.04	29.20	1.55	3.07	32.65	61.10	73.97
	DKGTrack [15]	ICCV 2025	3.39	0.46	25.38	0.49	6.06	26.04	80.48	76.14
	Qwen2.5-VL-3B [2]	ArXiv 2025	2.09	0.93	5.28	0.97	17.14	5.40	**87.46**	76.69
L-RGB-T	DeformCAT [11]+SORT+iKUN	IEEE TMM	2.03	0.41	11.25	0.77	0.87	12.07	47.65	62.61
	Unismot+iKUN [31]	PR 2025	1.95	0.29	14.34	0.31	3.98	15.41	65.48	70.86
	PFTrack+iKUN [37]	PR 2025	8.55	1.66	**45.92**	2.40	5.05	**49.15**	73.96	76.31
	MCTrack+iKUN [21]	TCSVT 2025	4.71	1.22	18.91	1.51	5.73	19.83	71.17	68.95
	Qwen2.5-VL-3B [2]	ArXiv 2025	4.98	2.59	10.19	3.05	14.29	10.65	83.40	75.52
	RTrack	**Ours**	**15.53**	**12.39**	20.79	**20.15**	**22.78**	22.02	81.99	75.53

requires the predicted bounding boxes to be spatially close to the ground truth boxes (i.e., having a high IoU), but also prevents the model from repeatedly outputting excessive predictions near high IoU regions to "boost" the reward. Additionally, this mechanism prevents the model from neglecting samples that are close to the ground truth but have slightly lower IoU values due to overemphasis on precision. The Precision Detection Reward can be defined as:

$$R_{\text{PDR}}(B_{\text{det}}, B_{\text{gt}}) = \frac{IoU_{\text{score}}}{(N_{\text{det}})^{\gamma}} + \frac{\lambda \cdot MatchedGT}{N_{\text{gt}}} \tag{8}$$

where IoU_{score} and $MatchedGT$ have the same meanings as defined in Eq. (7). Here, N_{det} and N_{gt} denote the numbers of predicted and ground-truth boxes.

5 Experiments

5.1 Implementation Details

Our RTrack framework is built upon a Multi-modal Large Language Model, using Qwen2.5-VL-3B as the baseline. In the RL stage, we apply LoRA fine-tuning with an initial learning rate of 1×10^{-5} and use 4 rollouts by default. We temporally and uniformly sample 10% of the frames from the training set, with detailed hyperparameter settings for Eqs. (4)–(8) provided in the supplementary material. We further measure the inference cost on an AMD EPYC 7K62 48-core processor with a single NVIDIA RTX 4090 GPU. Under this setting, RTrack runs at 3.82 FPS and consumes 7.54 GFLOPs per generated token.

5.2 Metrics

We evaluate our tracking framework following the RMOT community's evaluation protocol, using HOTA for comprehensive assessment of detection and tracking performance. In addition, DetA, DetPr, and DetRe measure detection quality. AssA, AssPr, and AssRe assess tracking consistency. LocA evaluates bounding box accuracy.

5.3 Quantitative Results

To evaluate the performance of RTrack on the RefRT dataset, we compare it with representative methods under both L-RGB and L-RGB-T settings. For L-RGB inputs, we include end-to-end RMOT methods, including TransRMOT [26], TempRMOT [34], CRTracker [5], ReferGPT [3], and DKGTrack [15], as well as the two-stage iKUN [8] pipeline. Since iKUN requires external detection and tracking modules, we combine it with YOLOX [10] and ByteTrack [35] to handle the diverse object categories in RefRT. These L-RGB methods cannot directly exploit thermal images; therefore, for the L-RGB-T setting, we construct RGBT tracker + iKUN baselines, where DeformCAT serves as the RGBT detector and Unismot, PFTrack, and MCTrack serve as RGBT trackers. We also report Qwen2.5-VL-3B [2] as a general MLLM baseline and include RTrack without CoT reasoning to assess the contribution of explicit reasoning. As shown in Tab. 1, RTrack achieves the best HOTA, DetA, DetRe, and DetPr scores among all methods, outperforming the strongest L-RGB baseline CRTracker by 6.23 HOTA points and the best RGBT tracker-based baseline PFTrack+iKUN by 6.98 HOTA points. The comparison indicates that simply coupling existing RGBT trackers with a language-guided RMOT module is insufficient for the RT-RMOT task, while the proposed MLLM-based framework better integrates language, RGB, and thermal cues. In addition, RTrack improves over RTrack (no-CoT) by 5.65 HOTA points, showing that structured reasoning further benefits referring target localization and tracking.

5.4 Ablation Study

To further verify the effectiveness of the proposed task, framework, and design mechanism, we conduct three groups of ablation experiments: multimodal input, MLLMs and the design mechanism, as shown in Tab. 2 and Tab. 3.

Validation of Multimodal Input Effectiveness. We compare the performance of the RTrack framework under different input modalities. Specifically, we conduct experiments with both L-RGB and Languge-RGB-Thermal inputs under zero-shot settings and after RL fine-tuning. As shown in Tab. 2, the proposed training strategy achieves significant performance improvements under both L-RGB and L-RGB-T input settings. Specifically, with L-RGB inputs, the trained model improves HOTA by 10.4% over the untrained version and yields consistent improvements across the five key metrics: DetA, AssA, DetRe, DetPr, and AssRe. With L-RGB-T inputs, the trained model improves HOTA by 10.55% and similarly achieves substantial gains on these five metrics. Furthermore, comparing the trained models reveals that using L-RGB-T as input leads to superior overall performance compared to L-RGB, achieving a 3.04% higher HOTA score and outperforming L-RGB on five detection and association metrics. These findings collectively demonstrate the importance of leveraging RGBT multimodal information to enhance the performance of the RTrack framework.

Analysis of Multi-modal Large Language Models. Under the same RTrack framework, we replace different vision-language models (LLaVA, LLaVA-NeXT, InternVL and Qwen2.5-VL) and conduct zero-shot inference comparisons

Table 2: Results of more experiments on the RefRT test set. ↑ indicates that higher scores are better. The best results are marked in **bold**.

Modality	Method	Size	RL Fine-tuning	HOTA↑	DetA↑	AssA↑	DetRe↑	DetPr↑	AssRe↑	AssPr↑	LocA↑
L-RGB	Qwen2.5-VL [2]	3B	✗	2.09	0.93	5.28	0.97	17.14	5.40	**87.46**	**76.69**
	RTrack	3B	✓	12.49	11.05	15.25	16.14	**23.95**	16.18	79.39	73.90
L-RGB-T	Qwen2.5-VL [2]	3B	✗	4.98	2.59	10.19	3.05	14.29	10.65	83.40	75.52
	RTrack	3B	✓	**15.53**	**12.39**	**20.79**	**20.15**	22.78	**22.02**	81.99	75.53

Table 3: Ablation study results of RTrack on the RefRT test set. Module A represent the untrained results, while Module B reflects the results after training. ↑ indicates higher is better. Best results are in **bold**.

Module	Method	Size	RL Fine-tuning	HOTA↑	DetA↑	AssA↑	DetRe↑	DetPr↑	AssRe↑	AssPr↑	LocA↑
A. MLLMs	LLaVA [19]	7B	✗	1.13	0.29	5.32	8.84	0.28	7.32	19.47	57.69
	LLaVA-NeXT [18]	8B	✗	2.99	0.84	13.35	5.70	0.96	17.84	38.32	60.17
	InternVL2 [7]	8B	✗	0.82	0.18	4.77	10.77	0.18	5.71	31.96	60.10
	Qwen2.5-VL [2]	3B	✗	4.98	2.59	10.19	3.05	14.29	10.65	83.40	75.52
B. Training Strategy	w/o CAS	3B	✓	5.27	3.93	7.37	4.32	**28.11**	7.43	**90.32**	**78.63**
	w/o Struct. Reward	3B	✓	11.24	9.47	14.25	16.81	17.10	14.79	86.08	77.20
	w/o Det. Reward	3B	✓	11.16	7.82	17.42	**28.44**	9.50	18.33	82.64	76.34
	RTrack (Ours)	3B	✓	**15.53**	**12.39**	**20.79**	20.15	22.78	**22.02**	81.99	75.53

to evaluate the differences among large vision-language models in multimodal perception and instruction understanding. As shown in section A of Tab. 3, we observe that although Qwen2.5-VL uses the 3B model, which has the smallest number of parameters among the compared models, it still achieves the best performance. This confirms that choosing Qwen2.5-VL as the baseline for RTrack provides better modality fusion and detection capabilities.

Analysis of Training Strategy Improvements. During training, we perform an ablation study on the RTrack optimization mechanism to evaluate the contributions of the improved GSPO algorithm and the multidimensional reward design. Each component's impact on training stability and performance is assessed by progressively removing it. As shown in section B of Tab. 3, the full model achieves the best detection performance, tracking performance, and multi-target matching accuracy. Specifically, the Cropping Advantage Scaling (CAS) strategy significantly improves the model's tracking performance by mitigating gradient fluctuations caused by group normalization; the Structured Output Reward plays a key role in constraining the rationality of output structures and maintaining stable output lengths; the Comprehensive Detection Reward improves the accuracy and completeness of multi-target detection. Overall, the designed RL fine-tuning strategy enhances the model's semantic understanding and referring multi-object tracking capabilities.

5.5 Visualization

To further verify the effectiveness of the proposed RTrack framework, we conduct qualitative visualization analysis. As shown in Fig. 5, RTrack accurately detects and continuously tracks targets following the given language descriptions. By fusing complementary RGB and thermal infrared (T) information, the

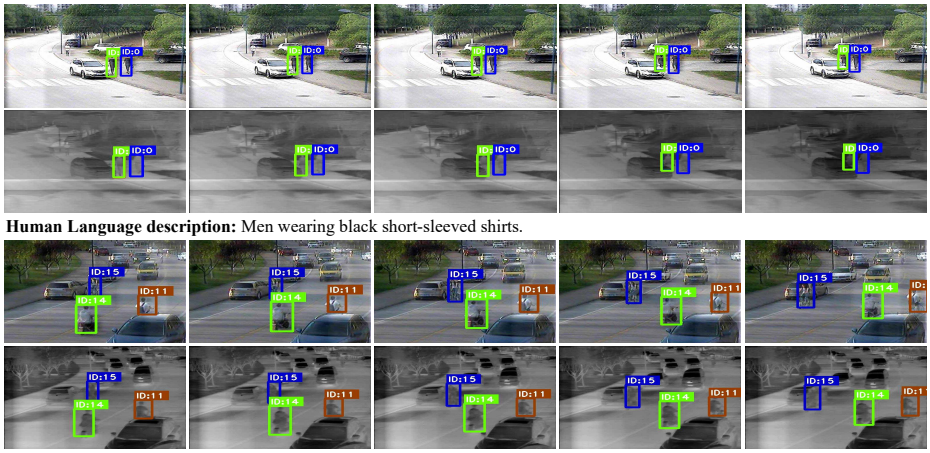


Fig. 5: Qualitative results of RTrack on the test set of RT-RMOT task.

model maintains stable performance under challenging conditions such as low illumination, occlusion, and complex backgrounds. These results highlight the advantages of multimodal feature fusion in improving semantic understanding.

6 Conclusion

This work proposes a new task, **R**GB**T** **R**eferring **M**ulti-**O**bject **T**racking (**RT-RMOT**), which aims to overcome the challenges of tracking under low-visibility conditions such as nighttime and smoke, enabling all-day referring multi-object tracking. To advance research, we construct the first **RefRT** dataset, featuring diverse scenarios and rich language descriptions, providing a benchmark for language-guided multi-object tracking in challenging environments. In addition, we propose the **RTrack** framework, which leverages MLLMs to enhance modality fusion and incorporates a CAS strategy to stabilize gradients, along with a Structured Output Reward and Comprehensive Detection Reward to balance exploration and accuracy. Experiments on RefRT dataset demonstrate that RTrack achieves **state-of-the-art (SOTA)** performance, providing a strong baseline for future research.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 62576144 and in part by Hubei Provincial Science and Technology Plan under Grant 2025BEB006.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023). https://doi.org/10.48550/arXiv.2303.08774
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025). https://doi.org/10.48550/arXiv.2502.13923
3. Chamiti, T., Di Bella, L., Munteanu, A., Deligiannis, N.: Refergpt: Towards zero-shot referring multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3888–3897 (2025)
4. Chen, S., Yu, E., Li, J., Tao, W.: Delving into the trajectory long-tail distribution for muti-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19341–19351 (2024)
5. Chen, S., Yu, E., Tao, W.: Cross-view referring multi-object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2204–2211 (2025)
6. Chen, S., Yu, Y., Yu, E., Tao, W.: Reamot: A benchmark and framework for reasoning-based multi-object tracking. arXiv preprint arXiv:2505.20381 (2025). https://doi.org/10.48550/arXiv.2505.20381
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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). https://doi.org/10.48550/arXiv.2312.14238
8. Du, Y., Lei, C., Zhao, Z., Su, F.: ikun: Speak to trackers without retraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19135–19144 (2024)
9. Fu, W., Li, J., Gao, B.B., Li, J., Lin, Y., Deng, H., Tao, W., Liu, Y., Wang, C.: Pet-dino: Unifying visual cues into grounding dino with prompt-enriched training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13039–13048 (2026)
10. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021). https://doi.org/10.48550/arXiv.2107.08430
11. Hu, Y., Chen, X., Wang, S., Liu, L., Shi, H., Fan, L., Tian, J., Liang, J.: Deformle cross-attention trnsformer for wekly aligned rgb–t pedestrin detection. IEEE Transactions on Multimedia (2025)
12. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960). https://doi.org/10.1115/1.3662552
13. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955). https://doi.org/10.1002/nav.3800020109
14. Li, C., Xue, W., Jia, Y., Qu, Z., Luo, B., Tang, J., Sun, D.: Lasher: A large-scale high-diversity benchmark for rgbt tracking. IEEE Transactions on Image Processing **31**, 392–404 (2021). https://doi.org/10.1109/TIP.2021.3130533
15. Li, G., Zhuang, S., Jian, Y., Yan, Y., Wang, H.: Language decoupling with fine-grained knowledge guidance for referring multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23626–23635 (2025)

16. Li, J., Yu, E., Chen, S., Tao, W.: Ovtr: End-to-end open-vocabulary multiple object tracking with transformer. arXiv preprint arXiv:2503.10616 (2025). https://doi.org/10.48550/arXiv.2503.10616
17. Li, W., Du, Y., Yin, Q., Zhao, Z., Su, F., Liu, D.: Cpany: Couple with any encoder to refer multi-object tracking. arXiv e-prints pp. arXiv–2503 (2025)
18. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llavanext: Improved reasoning, ocr, and world knowledge (2024)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023). https://doi.org/10.48550/arXiv.2304.08485
20. Long Li, C., Lu, A., Hua Zheng, A., Tu, Z., Tang, J.: Multi-adapter rgbt tracking. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019). https://doi.org/10.1109/ICCVW.2019.00279
21. Ma, J., Luo, H., Niu, S., Zhao, P., Liu, Y., Wei, Y., Zhang, J.: Multi-stage cross-modality feature interaction for rgb-thermal multi-object tracking. IEEE Transactions on Circuits and Systems for Video Technology (2025)
22. 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: International conference on machine learning. pp. 8748–8763. PmLR (2021). https://doi.org/10.48550/arXiv.2103.00020
23. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024). https://doi.org/10.48550/arXiv.2402.03300
24. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023). https://doi.org/10.48550/arXiv.2312.11805
25. Wang, C., Xu, C., Cui, Z., Zhou, L., Zhang, T., Zhang, X., Yang, J.: Cross-modal pattern-propagation for rgb-t tracking. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 7064–7073 (2020). https://doi.org/10.1109/CVPR42600.2020.00709
26. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring multi-object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14633–14642 (2023). https://doi.org/10.1109/CVPR52729.2023.01406
27. Xiao, Y., Zhao, J., Lu, A., Li, C., Yin, B., Lin, Y., Liu, C.: Cross-modulated attention transformer for rgbt tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8682–8690 (2025). https://doi.org/10.1609/aaai.v39i8.32938
28. Yu, E., Li, Z., Han, S., Wang, H.: Relationtrack: Relation-aware multiple object tracking with decoupled representation. arxiv 2021. arXiv preprint arXiv:2105.04322 (2021). https://doi.org/10.48550/arXiv.2105.04322
29. Yu, E., Wang, T., Li, Z., Zhang, Y., Zhang, X., Tao, W.: Motrv3: Release-fetch supervision for end-to-end multi-object tracking. arXiv preprint arXiv:2305.14298 (2023). https://doi.org/10.48550/arXiv.2305.14298
30. Zhang, H., Zhang, L., Zhuo, L., Zhang, J.: Object tracking in rgb-t videos using modal-aware attention network and competitive learning. Sensors **20**(2), 393 (2020). https://doi.org/10.3390/s20020393

31. Zhang, L., Wang, L., Wu, Y., Chen, M., Zheng, D., Cao, L., Zeng, B., Cai, Y.: Unirtl: A universal rgbt and low-light benchmark for object tracking. Pattern Recognition **158**, 110984 (2025). https://doi.org/10.1016/j.patcog.2024.110984
32. Zhang, P., Zhao, J., Wang, D., Lu, H., Ruan, X.: Visible-thermal uav tracking: A large-scale benchmark and new baseline. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8886–8895 (2022). https://doi.org/10.1109/CVPR52688.2022.00868
33. Zhang, T., Guo, H., Jiao, Q., Zhang, Q., Han, J.: Efficient rgb-t tracking via cross-modality distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5404–5413 (2023)
34. Zhang, Y., Wu, D., Han, W., Dong, X.: Bootstrapping referring multi-object tracking. arXiv preprint arXiv:2406.05039 (2024). https://doi.org/10.48550/arXiv.2406.05039
35. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European conference on computer vision. pp. 1–21. Springer (2022). https://doi.org/10.1007/978-3-031-20047-2_1
36. Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al.: Group sequence policy optimization. arXiv preprint arXiv:2507.18071 (2025). https://doi.org/10.48550/arXiv.2507.18071
37. Zhu, Y., Wang, Q., Li, C., Tang, J., Gu, C., Huang, Z.: Visible–thermal multiple object tracking: Large-scale video dataset and progressive fusion approach. Pattern Recognition **161**, 111330 (2025). https://doi.org/10.1016/j.patcog.2024.111330