RAGTrack: Language-aware RGBT Tracking with Retrieval-Augmented Generation

Supplementary Material

This supplementary material provides additional implementation details, experimental analyses, and visualizations for our proposed method. We begin by detailing the pipeline for generating high-quality textual descriptions in RGB-Thermal (RGBT) tracking benchmarks (§A). We then present the architecture of our prediction head (§B) and elaborate on the evaluation metrics used in our experiments (§C). Next, we outline implementation details including data augmentation and parameter settings (§D). Furthermore, we conduct extensive ablation studies to analyze key components of our method, including fusion position selection and robustness to missing text (§E). Finally, we show additional visualizations that demonstrate the effectiveness of our dynamic token selection mechanism and present qualitative results under various challenging scenarios (§F). Collectively, these contents offer deeper insights into the design choices of RAGTrack, further validating the overall performance of our framework.

A. Textual Description Generation

We generate high-quality textual descriptions for RGBT tracking through a two-stage pipeline, ensuring semantic consistency and minimal hallucinations.

Step 1: Initial Description Generation. We utilize Multimodal Large Language Models (MLLMs) to automatically generate initial descriptions for the object of tracking in each frame. The specific prompt used is:

> *"Describe the object located in the image at <box> (x, y, x+w, y+h) </box>. Focus on distinctive visual features, motion patterns, and key identifiers to distinguish it from background elements and distractors. Keep the description in a continuous sentence under 20 words. Avoid mentioning bounding boxes or coordinates. Do not use parentheses for explanations."*

This approach efficiently produces informative descriptions at scale, overcoming the cost of manual annotation.

Step 2: Description Refinement. The initial descriptions from MLLMs may contain inaccuracies or hallucinations. To ensure quality, we perform a refinement step using the following prompt:

> *"Correcting the textual description of the tracking object. Ensure the final output is a continuous sentence under 20 words, logically coherent, does not mention bounding boxes, or coordinates terms, and does not use parentheses for explanations. Do not introduce new details. Output only the integrated description without any additional text. Textual description: [Initial description]"*

Finally, human experts review the refined descriptions to correct any remaining issues. This includes rectifying hallucinations, fixing grammatical errors, removing mixed-language content or garbled text, and ensuring descriptions accurately reflect the visual target.

Annotation Statistics. Using this pipeline, we annotate the entire LasHeR training set, which comprises 979 sequences. This process yields a total of 514,081 textual descriptions, with one description provided for each frame, to support model training. For evaluation, we annotate only the first frame across the test set of LasHeR and all sequences of the GTOT, RGBT210 and RGBT234 benchmarks. This results in a collection of 739 high-quality textual descriptions, which are used to assess the textual reasoning capability of trackers during inference.

B. Prediction Head

The enhanced features of search region are fed into the prediction head to produce tracking results in the form of bounding boxes $[x, y, w, h]$. The prediction head generates three outputs: (1) a target classification score map $\mathbf{I} \in [0,1]^{H_F \times W_F}$ indicating presence probabilities, (2) a spatial offset map $\mathbf{G} \in [0,1]^{2 \times H_F \times W_F}$ compensating for discretization errors, and (3) a normalized bounding box size map $\mathbf{J} \in [0,1]^{2 \times H_F \times W_F}$ representing target width and height. Here, H_F and W_F denote the height and width of the feature map. The final bounding box is constructed at the position (i^*, j^*) with maximum classification score by combining the corresponding predictions:

$$\begin{aligned} x &= i^* + \mathbf{G}(0, i^*, j^*), \quad & y &= j^* + \mathbf{G}(1, i^*, j^*), \\ w &= \mathbf{J}(0, i^*, j^*), & h &= \mathbf{J}(1, i^*, j^*). \end{aligned} \tag{1}$$

C. Evaluation Metrics

We employ Precision Rate (PR) and Success Rate (SR) as our primary evaluation metrics. On the LasHeR benchmark, we additionally use the Normalized Precision Rate (NPR) to address scale variations. For GTOT and RGBT234, we instead report Maximum Precision Rate (MPR) and Maximum Success Rate (MSR) due to modality annotation misalignment. PR quantifies the accuracy of target localization

Table 1. Comparison of different fusion positions.

Fusion Positions	MPR↑	MSR↑
[1, 2, 3, 4]	92.4	67.5
[11, 12, 13, 14]	92.8	67.9
[21, 22, 23, 24]	93.1	68.7
[6, 12, 18, 24]	**93.8**	**69.5**

as the proportion of frames where the predicted center position lies within a predefined distance threshold from the ground truth. SR measures the bounding box overlap, computed as the percentage of frames where the Intersection over Union (IoU) between the predicted and ground truth boxes surpasses a given threshold. NPR extends PR by normalizing the precision based on target size, providing fair comparison across scale variations. MPR and MSR address annotation inconsistencies by reporting the maximum performance across modalities, ensuring equitable evaluation when RGB and TIR annotations are misaligned.

D. Additional Implementation Details

During training, we apply standard data augmentation techniques [1, 4] to the training samples from each sequence, including rotation, translation, and grayscale transformation. During inference, the update threshold for multi-modal references is set to 0.65, with an update interval of 5 frames. Following common practice [2, 3], the backbone achieves a tracking speed of 24.3 FPS on a NVIDIA V100 GPU, with a computational cost of 62.7G FLOPs.

E. More Ablation Studies

This section presents further ablation studies to examine the impact of individual components in our method. Detailed analysis and discussions are provided as follows.

Selection of Fusion Positions in ATF. To evaluate the impact of fusion locations in Adaptive Token Fusion (ATF), we conduct an ablation study by applying cross-modal fusion at different layers of the Multi-modal Transformer Encoder (MTE). As shown in Tab. 1, fusing at Layers 6, 12, 18, and 24 achieves the best performance with 93.8% MPR and 69.5% MSR on RGBT234. Shallow-layer fusion (1-4) captures low-level features but yields suboptimal performance due to limited semantic information. Mid-layer fusion (11–14) improves semantic understanding but still lacks comprehensive representation. While deep-layer fusion (21–24) retains high-level context, it misses fine-grained spatial details. Our design progressively integrates features across multiple stages, effectively combining spatial details with semantic abstractions. The results confirm that cross-layer fusion is essential for robust tracking.

Robustness to Missing Text. To evaluate the robustness of our method under incomplete language guidance, we conduct an ablation study by randomly masking the input text during inference. As shown in the Fig. 1, RAGTrack maintains strong performance even when the first-frame text is partially or fully absent. The performance remains nearly unchanged with 0% to 60% of text missing. This demonstrates that our model effectively addresses absent language cues through its Retrieval-Augmented Generation (RAG) mechanism and context-aware reasoning. Even when all text is unavailable, our method still achieves competitive results of 92.9% MPR and 68.8% MSR on RGBT234, surpassing leading methods [3]. This highlights the capacity of RAGTrack to robustly utilize visual features when textual input is unavailable.

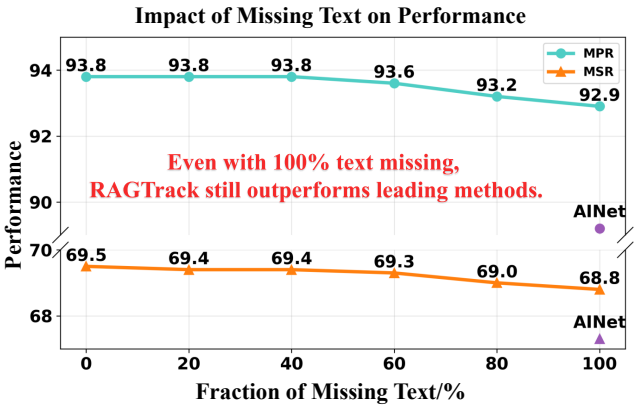


Figure 1. Comparison with different fractions of missing text.

F. More Visualizations

This section provides additional visual analysis to further validate the effectiveness of our method through the following examples and discussions.

More Discussions of Dynamic Token Selection in ATF. To better understand the behavior of our dynamic token selection mechanism in ATF, we provide visualizations of the selected tokens on the LasHeR benchmark. The visualization shows that the mechanism effectively focuses on target regions while suppressing background distractions. As shown in Fig. 2, the retained tokens mainly cover the target area, while the discarded tokens correspond primarily to background regions and distractors. This demonstrates that our attention-based selection identifies semantically important regions guided by the textual descriptions. Compared to processing all tokens equally, our method reduces unnecessary token processing while maintaining critical target information. This selective processing allows the model to concentrate its reasoning capacity on the most informative image regions. These results provide clear evidence that our dynamic token selection addresses search redundancies, enabling more efficient and accurate RGBT tracking.

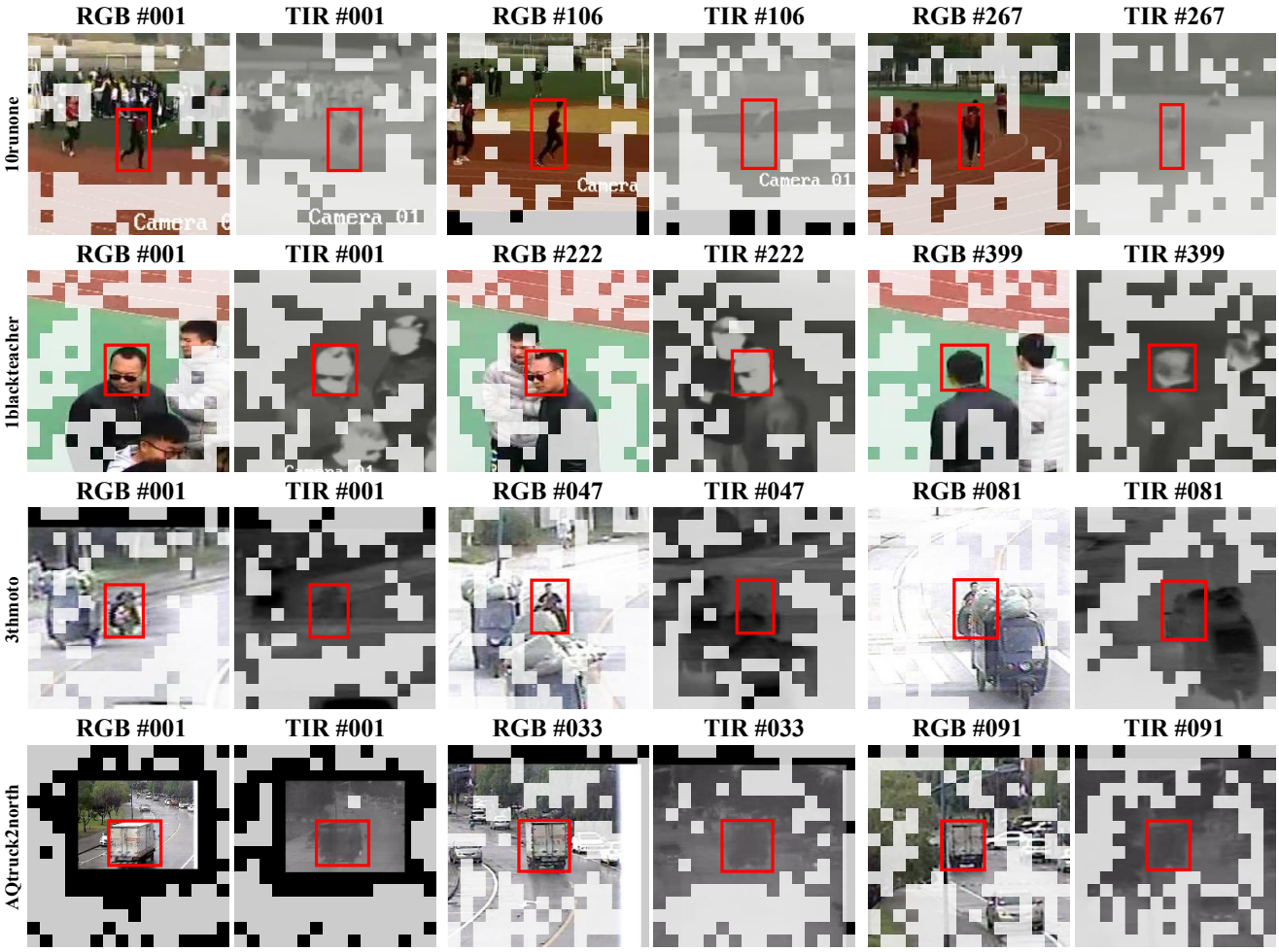


Figure 2. Visualization of dynamic token selection in ATF.

More Qualitative Results. Fig. 3 and Fig. 4 present comprehensive qualitative comparisons of RAGTrack on challenging sequences from the RGBT234 benchmark. The visualization highlights the capacity of our method to maintain precise tracking through dynamic language reasoning across diverse scenarios. Each sequence illustrates the tracking results, evolving textual descriptions and corresponding per-frame IoU curves. These results demonstrate how our framework effectively handles several challenging situations. Through context-aware reasoning and historical knowledge retrieval, RAGTrack successfully distinguishes similar-appearance targets in Fig. 3 (a) and Fig. 4 (a). The method resolves ambiguous target references in Fig. 3 (b) by leveraging linguistic guidance to maintain tracking consistency. During occlusion shown in Fig. 3 (c) and Fig. 4 (b), the visual-language unified modeling preserves target identity despite severe appearance changes. Additionally, as evidenced in Fig. 4 (c), our method overcomes insufficient visual cues through adaptive fusion of complementary multi-modal features. The stable IoU curves across challenging sequences confirm the robustness of our method in addressing complex tracking difficulties.

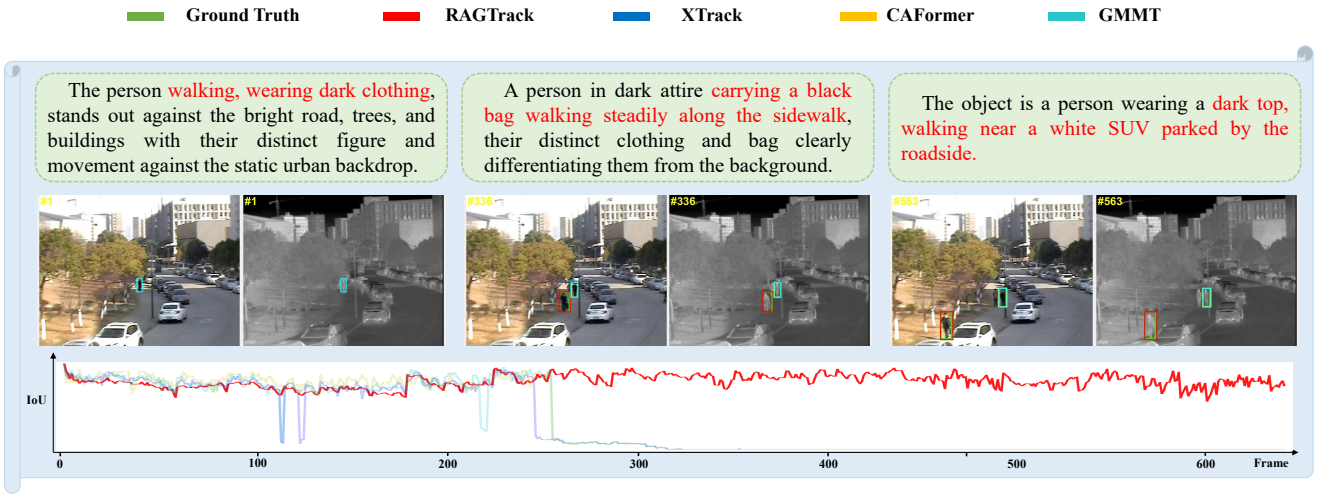
References

[1] Zong Ke, Yuqing Cao, Zhenrui Chen, Yuchen Yin, Shouchao He, and Yu Cheng. Early warning of cryptocurrency reversal risks via multi-source data. *FRL*, page 107890, 2025. 2

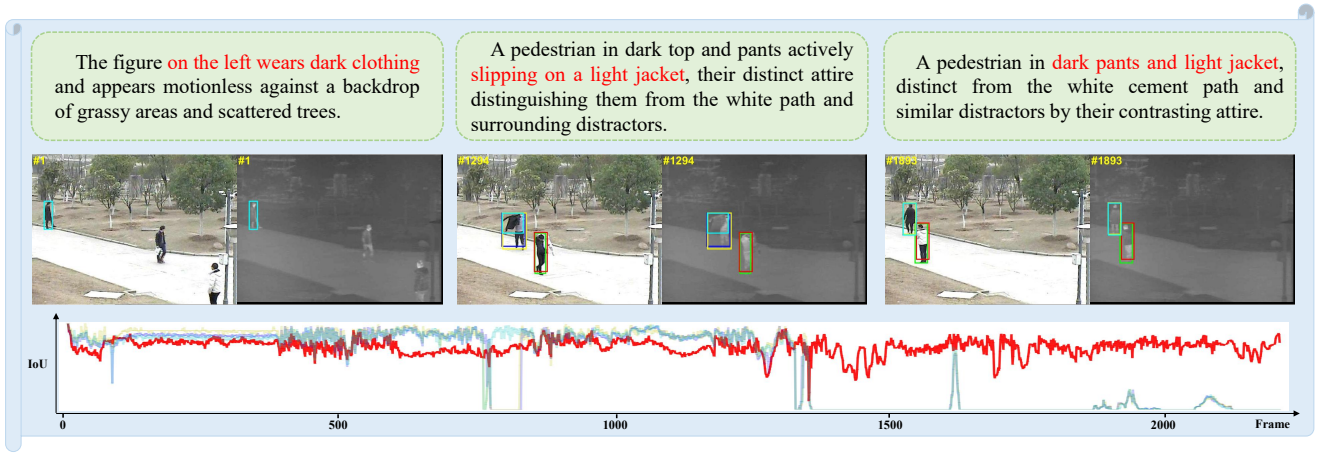
[2] Hao Li, Yuhao Wang, Xiantao Hu, Wenning Hao, Pingping Zhang, Dong Wang, and Huchuan Lu. Cadtrack: Learning contextual aggregation with deformable alignment for robust rgbt tracking. *arXiv preprint arXiv:2511.17967*, 2025. 2

[3] Andong Lu, Wanyu Wang, Chenglong Li, Jin Tang, and Bin Luo. RGBT tracking via all-layer multimodal interactions with progressive fusion Mamba. In *AAAI*, pages 5793–5801, 2025. 2

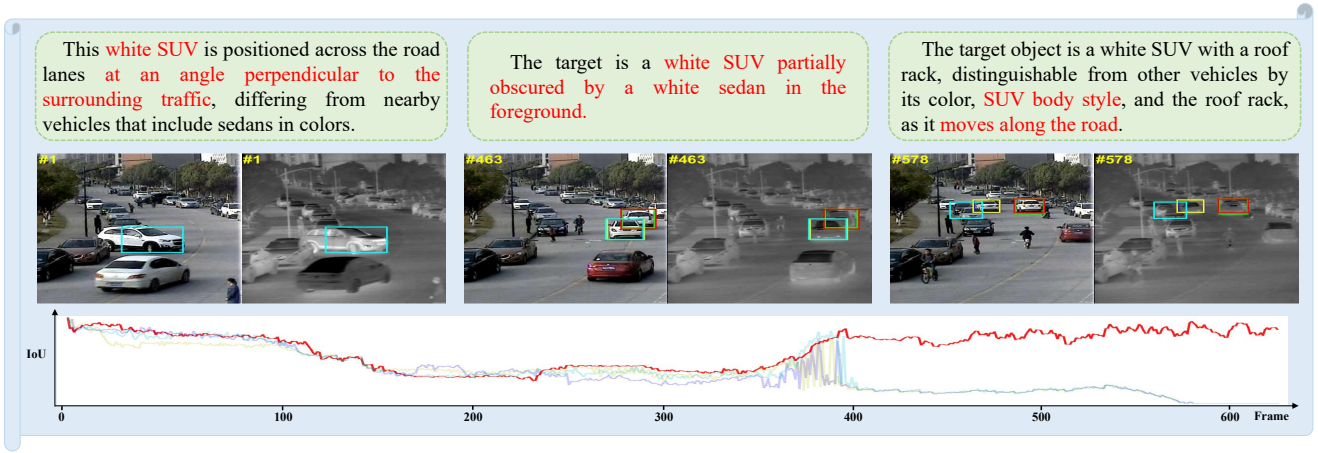
[4] Yongqi Shan, Yunzhi Zhuge, and Huchuan Lu. Hdvs: semi-supervised semantic segmentation via heterogeneous dual-branch voting supervision. *VI*, 4(1):4, 2026. 2



(a) walkingman1



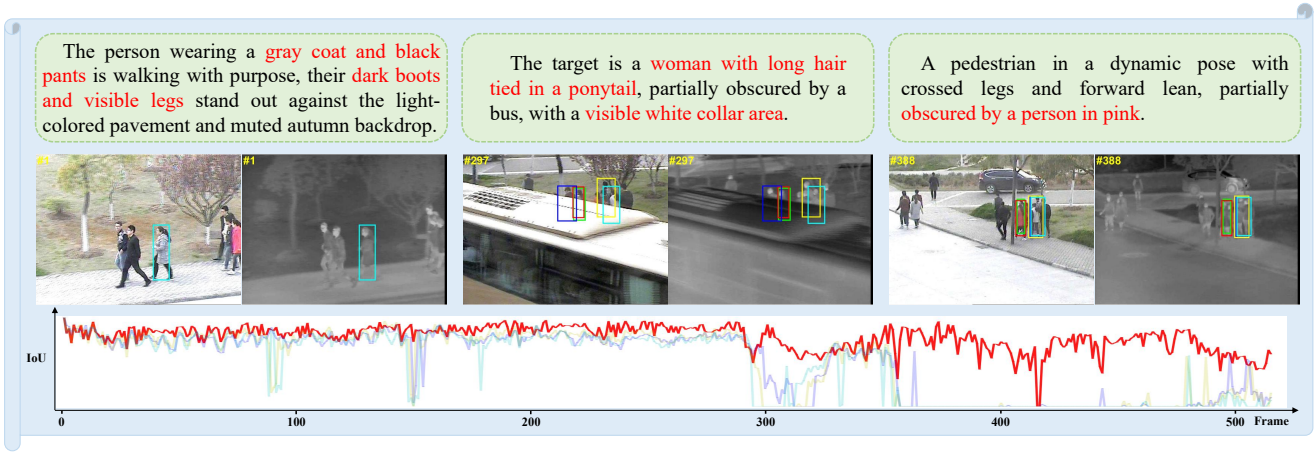
(b) tree5



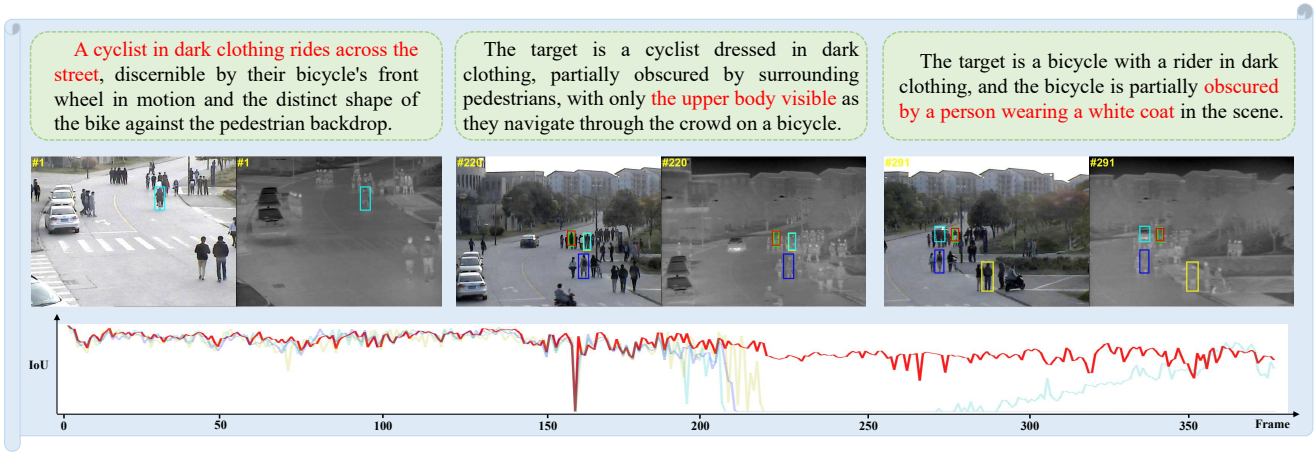
(c) whitesuv

Figure 3. Qualitative results on the RGBT234 benchmark.

(a) baketballwaliking



(b) walkingman41



(c) walking40

Figure 4. More qualitative results on the RGBT234 benchmark.