PMPTrack: A Decoupled Two-Stage Progressive Modality Promotion Paradigm for RGB-T Tracking

Yuanhao Zheng
Wenzhou University
Wenzhou City, Zhejiang Province, China
23451350045@stu.wzu.edu.cn

Sixian Chan*
Zhejiang university of technology
Hangzhou City, Zhejiang Province, China
sxchan@zjut.edu.cn

Jie Hu
Wenzhou University
Wenzhou City, Zhejiang Province, China
20160204@wzu.edu.cn

Abstract

RGB-T tracking aims to accurately localize target objects by leveraging complementary information from both RGB and thermal-infrared modalities. A common belief is that a well-designed multimodal fusion strategy plays a critical role in improving tracking performance. However, the initially extracted raw features from each modality are often misaligned and noisy, greatly deteriorating the tracking performance. To address this issue, this paper proposes a new paradigm for RGB-T tracking, termed PMPTrack, which adopts a clearly decoupled two-stage progressive modality promotion framework prior to multimodal fusion. Through an elaborate modality enhancement process, we pursue more refined modality representations for subsequent fusion and downstream tasks, as naive multimodal fusion often overlooks important details. Specifically, we divide the process into two stages: the Token-Level Promotion (TLP) stage and the Feature-Level Promotion (FLP) stage. The TLP stage focuses on fine-grained target detail modeling at the token level, where tokens represent the fundamental units embedded and processed by Vision Transformer. A Context-aware Token Categorization (CTC) module is responsible for this fine-grained token modeling, with multiple Multimodal Synergistic Prompters (MSP) cooperating to facilitate the process. Following the TLP, the FLP stage performs coarse-grained refinement at a higher level of abstraction, emphasizing global enhancement based on deep-layer feature representations derived from token embeddings, where we utilize a Cross-modality Guided Enhancement Module (CGEM) to achieve this coarse-grained refinement. Finally, extensive experiments on LasHeR, RGBT234, and RGBT210 datasets validate the superiority of our PMPTrack, which outperforms previous state-of-the-art methods.

Keywords: ***RGB-T tracking, Decoupled Two-Stage, Modality Promotion, Token-Level, Feature-Level.***

1. Introduction

Visual object tracking [35, 37, 24] is a fundamental task in computer vision, aiming at estimating the position and shape of a target object in a video sequence based on its initial state. It is widely applied in various fields such as robotic vision [21], video surveillance [5], and autonomous driving [7]. While traditional RGB trackers perform well under normal conditions, they often struggle in challenging scenarios such as occlusion, low visibility, or fast motion. To address these limitations, integrating auxiliary modalities such as thermal infrared (TIR) data has proven effective for enhancing robustness in multimodal tracking [36]. With the increasing availability and maturity of TIR devices, RGB-T tracking has attracted growing attention by leveraging the complementary strengths of RGB and thermal modalities in recent years [10, 15, 22].

A widely held view in previous studies is that multimodal fusion plays a key role in building robust RGB-T trackers [23, 27, 39]. One of the earliest and most representative approaches, mfDiMP [32], performs pixel-level fusion by directly concatenating features from different modalities. However, this approach rudely merges raw inputs at shallow network layers, assigning equal weight to salient and low-frequency cues, which hinders the full exploitation of multimodal information and leads to performance degradation. Later, more refined feature-level fusion methods, such as APFNet [27], fuse information during the feature extraction phase and put greater emphasis on preserving semantic information. However, feature-level fusion at deeper network layers often requires multiple feature extractions,

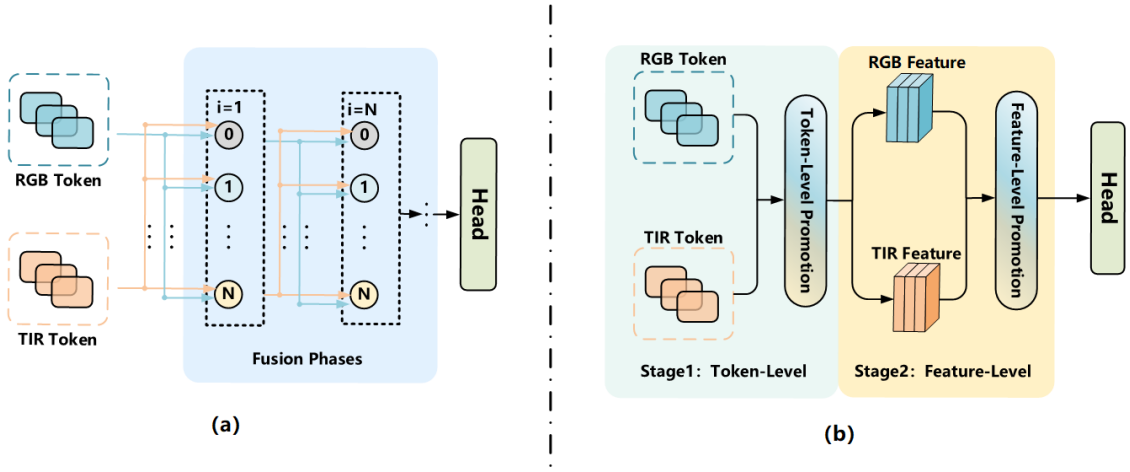

Figure 1: **Difference of RGB-T tracking paradigm. (a)** Traditional tracking based on redundant fusion structures. **(b)** Our proposed Two-stage Progressive Modality Promotion paradigm.

which dilutes fine-grained cues and impairs the learning of discriminative representations.

To address these issues, recent works [4, 18], as shown in Fig. 1 (a), have shifted focus to designing dynamic fusion strategies that adaptively select optimal approaches for different tracking scenarios. However, the use of complex fusion units often incurs additional computational overhead.

Despite advances in the design of multimodal fusion strategies, there is no conclusive evidence supporting the idea that an excellent RGB-T tracker can be achieved solely through fusion. Due to significant modality differences in resolution, noise, and other factors, naive fusion phase optimization does not fully take advantage of their potential. Thus, in this paper, we wonder: *Is it possible to explore a new modality promotion paradigm for RGB-T tracking, with the purpose of achieving more effective modality representation learning and enhancement prior to multimodal fusion or other downstream tasks?* Fortunately, the conceptual insights from prior work at both the pixel level (mfDiMP [32]) and the feature level (APFNet [27]) have inspired our study. Building on these foundations, this paper introduces a novel tracking paradigm, termed PMPTrack. As shown in Fig. 1 (b), instead of focusing on conventional multimodal fusion design, the core idea of PMPTrack is explicitly decoupled into two progressive modality promotion stages: Token-Level Promotion (TLP) and Feature-Level Promoton (FLP). The TLP focuses on fine-grained target detail modeling, capturing subtle and discriminative cues. Following this, the FLP emphasizes coarse-grained refinement at a higher abstraction level to enhance overall representations.

Specifically, for the TLP stage, we leverage Context-aware Token Categorization (CTC) to perform fine-grained object detail analysis, thereby enhancing the model's ability to understand target regions. Meanwhile, to transfer and amplify the subtle yet critical visual cues across modalities for a better inter-modality representation learning, we introduce the Multimodal Synergistic Prompter (MSP), which enables the model to more sensitively capture key information and achieve fine-grained inter-modal enhancement. For the FLP stage, we implement a Cross-modality Guided Enhancement Module (CGEM), which performs coarse-grained overall refinement at the feature level, operating on the deep-layer feature maps reshaped from the token representations produced by TLP, incorporating feature-level complementary information from the opposite modality to guide and enhance the representation capability of the current modality. Through this novel and progressive paradigm, our PMPTrack constructs a better representation learning from token details to feature abstraction, resulting in significant performance improvements in RGB-T tracking tasks.

In summary, our contributions are summarized as follows:

- PMPTrack, to our knowledge, is the first work to introduce such a decoupled two-stage progressive modality promotion paradigm, including the TLP stage and the FLP stage, for RGB-T tracking. Our analysis and validation demonstrate its theoretical soundness and exceptional performance.
- In the TLP stage, we focus on the fine-grained analysis and introduce the CTC and MSP to build token-level target modeling, enhancing detailed target learning.
- In the FLP stage, we progressively transition to coarse-grained refinement and design the CGEM to cross-modally enhance the overall representational capacity of the current modality by incorporating auxiliary information from the other modality at the feature level.
- Extensive experiments conducted on RGB-T tracking

datasets (LasHeR, RGBT210, RGBT234) have demonstrated the outstanding performance and remarkable tracking results of our proposed PMPTrack.

2. Related Work

RGB-T Object Tracking Conventional single-modality tracking methods [1, 2, 6] produce compelling results in normal conditions but tend to degrade significantly when faced with challenges such as low light, occlusion, or thermal cross. Given these challenges, RGB-T object tracking has attracted increasing attention, as its complementary modalities can enhance performance in scenarios where single-modality methods fall short. Recent state-of-the-art models, such as BAT [3], bring a bidirectional adapter to the prompt-based learning framework for cross-modality tracking. TBSI [10] uses a fused template with an attention mechanism to enable information exchange between RGB and TIR modalities. ViPT [37] presents a visual prompt-based multimodal tracking framework that adapts frozen pre-trained models by learning a few modality-related prompts, enhancing efficient tracking across various multimodal scenarios. Despite performance gains, these methods remain superficially confined to visual prompts or attention modules, leading to bottlenecks in exploring a more efficient and in-depth progressive paradigm for optimal modality extraction and enhancement.

Modality promotion for multimodal tracking The effective utilization of rich modality-specific information from different modalities is crucial for multimodal tasks. Modality-promotion-based tracking aims to enhance the representational capacity of each modality, facilitating more robust adaptation to diverse and complex tracking environments. Existing approaches like STMT [19], UnTrack [26], and MFGNet [25], enhance modality processing through various techniques like spatio-temporal fusion and attention mechanisms. However, these methods often rely on arbitrarily stacked modules without a coherent strategy, overlooking the potential of a progressive modality-promotion process. In this work, we move beyond conventional paradigms by exploring a progressive modality promotion design, offering a novel perspective for future research.

3. Preliminaries and Motivation for the Two-Stage Progressive Modality Promotion

Preliminary. We first select the classic single-stream RGB model, *i.e.*, OSTrack [30], as the pre-trained model. We extend the OSTrack into a dual-branch multimodal tracking model, which serves as the foundation for our PMPTrack. Its input consists of a pair of the template and search frames, which are divided into tokens by patch-embedded layers and then concatenated as follows:

$$\begin{aligned} \mathbf{H}^i_{\text{RGB}} &= \left[\mathbf{X}^i_{\text{RGB}}; \mathbf{Z}^i_{\text{RGB}}\right], \quad \mathbf{H}^i_{\text{RGB}} \in \mathbb{R}^{(N_z+N_x)\times C} \\ \mathbf{H}^i_{\text{TIR}} &= \left[\mathbf{X}^i_{\text{TIR}}; \mathbf{Z}^i_{\text{TIR}}\right], \quad \mathbf{H}^i_{\text{TIR}} \in \mathbb{R}^{(N_z+N_x)\times C} \end{aligned} \tag{1}$$

where, the input tokens $\mathbf{X}^i_{\text{RGB}}, \mathbf{X}^i_{\text{TIR}}$ and $\mathbf{Z}^i_{\text{RGB}}, \mathbf{Z}^i_{\text{TIR}}$ are the tokens from the i-th layer of the encoder. After concatenation, $\mathbf{H}^i_{\text{RGB}}$ and $\mathbf{H}^i_{\text{TIR}}$ are fed into the encoder. Each encoder layer updates the input tokens through multi-head attention (MHA) and a feed-forward network (FFN).

Why pursing a Two-Stage Progressive Modality Promotion Paradigm? To understand the rationale behind shifting from traditional multimodal fusion to a modality promotion paradigm, it is essential to delve into the connotation of a two-stage progressive modality promotion strategy. Existing approaches [14, 20] generally adhere to the prevailing practice of directly merging RGB and TIR modalities to mitigate modality discrepancies like misalignment and noise. However, such naive fusion often results in suboptimal information extraction and limited feature enhancement. Consequently, the model tends to rely only on superficial representations from each modality, rather than benefiting from a more structured and effective promotion process. In contrast, we argue that a clearly decoupled and progressive two-stage modality promotion—starting with fine-grained token-level modeling at the micro level and advancing to coarse-grained feature-level exploration at a deeper level—is crucial for fully exploiting rich modality-specific information and laying a solid foundation for fusion and other downstream tasks in multimodal visual tracking.

4. Method

4.1. Overview

Building upon the above analysis, in this section, we give a detailed description of our proposed PMPTrack. As shown in Fig.2, distinct from focusing on multimodal fusion design, our proposed PMPTrack adopts a more clearly decoupled two-stage strategy for progressive modality promotion prior to fusion: the TLP stage and the FLP stage.

Given an input token $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ embedded in the image, the TLP first develops a fine-grained token-level target detail model by performing CTC within each modality and then uses multiple MSP to facilitate the transfer of visual representation cues from the target across modalities. Subsequently, after the model has sufficiently captured fine-grained token-level target region details with the aid of positional encoding, the token representations are reshaped into deep-layer feature maps $\mathbf{F} \in \mathbb{R}^{b\times c\times h\times w}$. The FLP then exploits these coarse-grained representations through the CGEM module to achieve global feature-level refinement. Finally, the tracking head concatenates the RGB and TIR features and predicts the target's current state.

4.2. Token-Level Promotion

Given that the Vision Transformer (ViT) operates on visual inputs at the token level, we reasonably initiate our

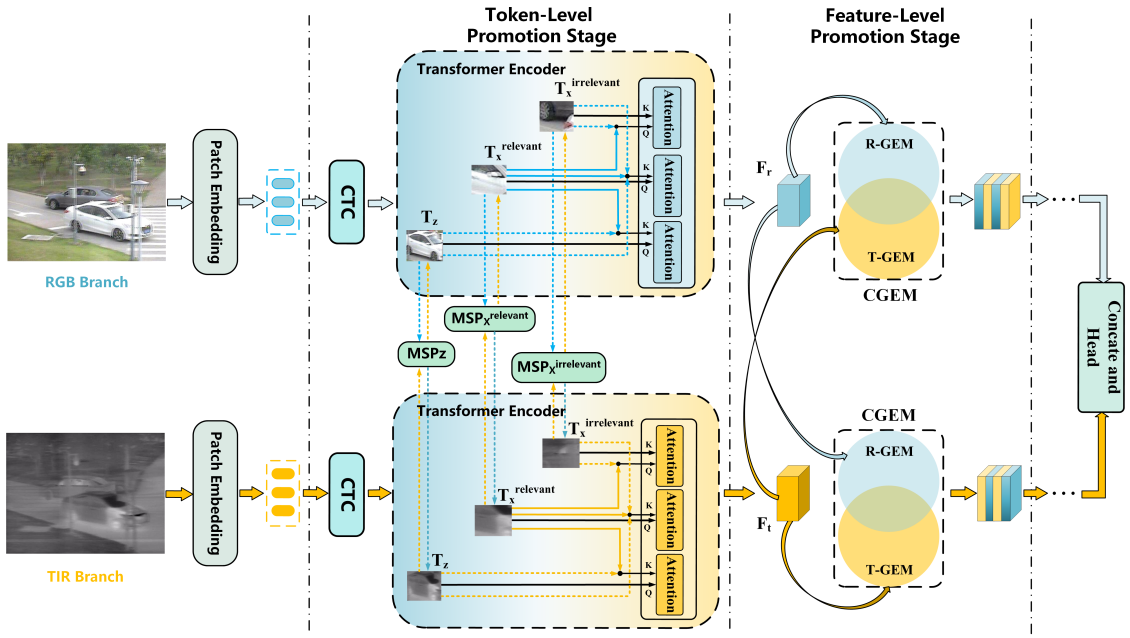

Figure 2: The overall framework of our proposed **PMPTrack**, including the TLP and FLP stages.

modality promotion from this fine-grained, micro-level representation. Accordingly, we design the TLP stage as the first phase, comprising two modules: the CTC and the MSP.

Context-aware Token Categorization. To better discriminate fine-grained details of target regions, we propose the CTC within the TLP, as shown in Fig. 2. We first divide all the tokens into three categories: T_x^{relevant}, $T_x^{\text{irrelevant}}$ and T_z, representing the template-relevant / irrelevant search tokens and template tokens, respectively. For clarity, we omit the RGB / TIR subscripts here.

For the above classification, we use a MLP to predict the category of each search token in real time:

$$P = \text{Softmax}\left(\text{MLP}\left([\text{MaxPool}(Z); X]\right)\right) \tag{2}$$

where Z and X are template and search tokens, respectively. We empirically set the threshold at 0.5, considering tokens with probabilities exceeding this value as relevant and those below it as irrelevant. The Softmax operation in Eq. 4.2, allowing the entire network—including token classification—to implicitly learn how to adaptively select appropriate search tokens for interaction with the template. This process is solely driven by gradients from the target localization loss and facilitates an end-to-end training.

Formally, the token interaction strategy in the i-th encoder layer can be formulated as follows:

$$\begin{aligned} q = k = v &= [X^i; Z^i], \\ [X'^i; Z'^i] &= [X^i; Z^i] + \text{MHA}(q, k, v), \\ [X^{(i+1)}; Z^{(i+1)}] &= [X'^i; Z'^i] + \text{FFN}([X'^i; Z'^i]) \end{aligned} \tag{3}$$

Algorithm 1: Token Interaction Strategy in the CTC

Input: Token sets: Template tokens T_z,
Template-Relevant search tokens T_x^{relevant},
Template-Irrelevant search tokens $T_x^{\text{irrelevant}}$
Output: Token interaction map for attention
foreach *token* $T_i \in T_z$ **do**
  Attend to tokens in $T_z \cup T_x^{\text{relevant}}$ via attention mechanism;
foreach *token* $T_i \in T_x^{relevant}$ **do**
  Attend to tokens in $T_z \cup T_x^{\text{relevant}} \cup T_x^{\text{irrelevant}}$ via attention mechanism;
foreach *token* $T_i \in T_x^{irrelevant}$ **do**
  Attend to tokens in $T_x^{\text{irrelevant}} \cup T_x^{\text{relevant}}$ via attention mechanism;
Block attention between T_z and $T_x^{\text{irrelevant}}$;
return Modified attention pattern

where, X^i, Z^i are the input search and template tokens to the i-th layer of the encoder respectively, with the RGB and TIR subscripts omitted for simplicity. We use q, k, and v to represent the query, key, and value. Since the template and search region tokens are jointly processed in the multi-head attention block, the modeling of both cross-modal and self-relations is seamlessly integrated into each encoder layer. The relation modeling rules for the three categories are defined as:

1. For category T_z, its tokens can aggregate information

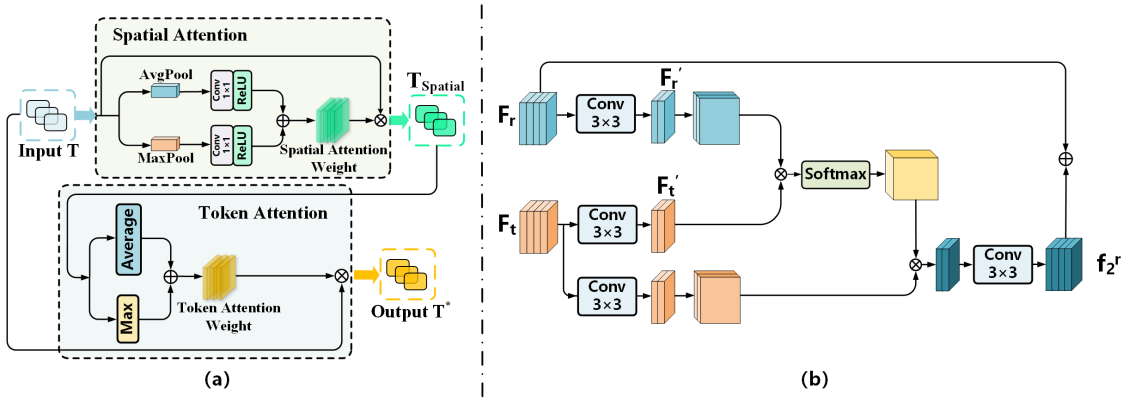

Figure 3: Conceptual illustration of the **MSP (a)** and **CGEM (b)**.

from the tokens in T_z and T_x^{relevant}:

$$\begin{aligned} q &= T_z = Z, \\ k &= v = [T_z; T_x^{\text{relevant}}] = [Z; T_x^{\text{relevant}}], \\ {T'}_z^i &= T_z^i + \text{MHA}(q, k, v) \end{aligned} \tag{4}$$

2. For category $T_x^{\text{irrelevant}}$, its tokens can aggregate information from the tokens in $T_x^{\text{irrelevant}}$ and T_x^{relevant}:

$$\begin{aligned} q &= T_x^{\text{irrelevant}}, \\ k &= v = [T_x^{\text{irrelevant}}; T_x^{\text{relevant}}] = X^i, \\ {T'}_x^{\text{irrelevant}} &= T_x^{\text{irrelevant}} + \text{MHA}(q, k, v) \end{aligned} \tag{5}$$

3. For category T_x^{relevant}, its tokens can aggregate information from all tokens:

$$\begin{aligned} q &= T_x^{\text{relevant}}, \\ k &= v = [T_z; T_x^{\text{irrelevant}}; T_x^{\text{relevant}}] = [Z; X], \\ {T'}_x^{\text{relevant}} &= T_x^{\text{relevant}} + \text{MHA}(q, k, v) \end{aligned} \tag{6}$$

Algorithm 1 provides the pseudocode of the attention interactions process among the token categories. Crucially, interactions between T_z and $T_x^{\text{irrelevant}}$ are explicitly blocked to suppress background noise, while still allowing search tokens in both T_x^{relevant} and $T_x^{\text{irrelevant}}$ to exchange information to obtain a more accurate object localization by learning the relationships between target-background search tokens.

Multimodal Synergistic Prompter. Building upon the CTC, we explicitly model intra-modality fine-grained target representations to achieve more accurate target discrimination. On this basis, to enable effective inter-modality representation learning, we further introduce the MSP. Specifically, each token category defined in the CTC is assigned a dedicated MSP: MSP_z, $\text{MSP}_x^{\text{relevant}}$ and $\text{MSP}_x^{\text{irrelevant}}$. As shown in Fig.3 (a), the input tokens T are firstly performed with average pooling and max pooling along the D-dimension to get token sequence $T_{\text{avr}}^S \in \mathbb{R}^{1\times D}$ and $T_{\text{max}}^S \in \mathbb{R}^{1\times D}$. Next, we apply the Spatial Attention by:

$$\begin{aligned} Att_{\text{spatial}} = \text{conv}_1\left(\text{ReLU}\left(\text{conv}_1\left(T_{\text{avr}}^S\right)\right)\right) + \\ \text{conv}_2\left(\text{ReLU}\left(\text{conv}_1\left(T_{\text{max}}^S\right)\right)\right) \end{aligned} \tag{7}$$

where conv_1 and conv_2 are two 1×1 convolutional layers used for dimensionality reduction and expansion, respectively. Subsequently, Att_{Spatial} is element-wise multiplied with the original tokens, yielding spatially reweighted tokens T_{Spatial}:

$$T_{\text{Spatial}} = T \cdot Att_{\text{Spatial}} \tag{8}$$

Next, the average and maximum values of T_{Spatial} are computed along the D-dimension to obtain the $T_{\text{avr}}^T \in \mathbb{R}^{1\times D}$ and $T_{\text{max}}^T \in \mathbb{R}^{1\times D}$. Then, another Token Attention is applied by:

$$Att_{\text{token}} = \text{conv}_3\left(\text{Concat}\left(T_{\text{avr}}^T, T_{\text{max}}^T\right)\right) \tag{9}$$

where conv_3 denotes a padded 7×7 convolution layer used to reduce dimensionality. Finally, $\text{Att}_{\text{token}}$ is element-wise multiplied with the original input to produce a token sequence reweighted by Token Attention:

$$T^* = T \cdot Att_{\text{token}} \tag{10}$$

So far, we have obtained T^* as the token after the Spatial Attention and Token Attention operations. Taking TIR as the modality being prompted as an example, the output of the MSP is as follows:

$$\text{MSP}^i = \text{MSA}(T_{\text{RGB}}^*) + T_{\text{TIR}}^* + \text{MSP}^{i-1} \tag{11}$$

where T_{RGB}^* and T_{TIR}^* represent the T^* obtained in Eq. 10. MSP^{i-1} is the output of the previous layer's MSP. When i=1 can ignore this input. $\text{MSA}(\cdot)$ refers to the Modality Spatial Amplification operation:

$$\text{MSA}(T^*) = \text{SmoothSoftmax}(\lambda, T^*) \tag{12}$$

where SmoothSoftmax refers to a λ-smoothed softmax function, and λ is set to 0.1, with a detailed analysis of its selection provided in the Sec. 5.3.4.

Method	Pub	Parameters M	LaSHeR PR	LaSHeR NPR	LaSHeR SR	RGBT234 PR	RGBT234 SR	RGBT210 PR	RGBT210 SR	Speed FPS
MFGNet [25]	TMM 2022	243.5	-	-	-	75.8	51.5	74.9	46.7	10.75
APFNet [27]	AAAI 2022	177.3	50.0	43.9	36.2	82.7	57.9	-	-	1.3
ProTrack [29]	ACM MM 2022	-	53.8	-	42.0	78.6	58.7	79.8	59.9	-
DMCNet [17]	TNNLS 2022	-	49.0	43.1	35.5	83.9	59.3	79.7	55.5	2.3
DFAT [22]	Inf fus 2023	173.6	44.6	40.0	33.6	75.8	55.2	74.5	52.8	22
CMD [33]	CVPR 2023	19.9	59.0	54.6	46.4	82.4	58.4	-	-	30
ViPT [38]	CVPR 2023	0.84	65.1	61.7	52.5	83.5	61.7	-	-	38.5
TBSI [10]	CVPR 2023	307	69.2	65.7	55.6	87.1	63.7	85.3	62.5	36.2
CAT++ [15]	TIP 2024	-	50.9	44.4	35.6	84.0	59.2	82.2	56.1	14
BAT [3]	AAAI 2024	0.96	70.2	66.0	56.3	86.8	64.1	84.8	62.2	35
TransAM [34]	TCSVT 2024	-	70.2	66.0	55.9	87.7	65.5	-	-	-
STMT [19]	TCSVT 2024	-	67.4	63.4	53.7	86.5	63.8	83.0	59.5	39.1
OneTracker [8]	CVPR 2024	2.8	67.2	-	53.8	85.7	64.2	-	-	-
Un-Track [26]	CVPR 2024	98.73	66.7	63.3	53.6	84.2	62.5	81.3	56.1	33.5
SDSTrack [9]	CVPR 2024	298.8	66.5	63.1	53.1	84.8	62.5	80.2	55.3	20.8
GMMT [23]	AAAI 2024	206.7	70.7	67.0	56.6	87.9	64.7	-	-	10.6
MFATrack [28]	PR 2025	2.9	63.3	-	49.8	81.7	57.5	-	-	25
FFE-BMFF [31]	Neurocomputing 2025	29.82	70.0	66.4	56.0	85.1	62.8	81.1	56.0	61
IPL [16]	IJCV 2025	183.6	69.4	65.6	55.3	88.3	65.7	86.7	63.2	28
PMPTrack (Ours)	CVM 2026	168.6	72.1	68.3	57.8	89.0	65.8	87.7	64.2	31

Table 1: Comparison with state-of-the-art methods on LasHeR, RGBT234 and RGBT210. Red: best results. Green: second-best. Yellow: third-best.

4.3. Feature-Level Promotion

Following the fine-grained modeling in the TLP, we progressively transition to the FLP for coarse-grained enhancement, which integrates the CGEM module to focus on deep-layer feature representation refinement.

Cross-modality Guided Enhancement Module. The CGEM aims to incorporate auxiliary feature-level cues from the complementary modality in a cross-modal manner, enhancing the representational capacity of the current main modality. As shown in Fig. 3 (b), we first hypothesize that the RGB as the main modality (R-GEM), the input feature maps $F_r \in \mathbb{R}^{b\times c\times h\times w}$ and $F_t \in \mathbb{R}^{b\times c\times h\times w}$ are reshaped from the tokens in the TLP stage. We first use a 1×1 convolution layer $\mathrm{CB}(\cdot)$ and a linear projection operation $\mathrm{v}(\cdot)$ to extract features and reduce the number of channels to half of the original, *i.e.*, $F_r' \in \mathbb{R}^{b\times \frac{c}{2}\times N}$ and $F_t' \in \mathbb{R}^{b\times \frac{c}{2}\times N}$. Next, the RGB feature goes through a series of procedures to obtain the finally enhanced feature map f_2^r represented as:

$$f_2^r = \mathrm{CB}'\left(\mathrm{v}'\left(\mathrm{S}\left(\mathrm{p}(F_r')\times F_t'\right)\times \mathrm{p}(F_t')\right)\right) + F_r \qquad (13)$$

where, $\mathrm{p}(\cdot)$ denotes the dimensional transformation operation ($\mathbb{R}^{b\times \frac{c}{2}\times h\times w} \rightarrow \mathbb{R}^{b\times N\times \frac{c}{2}}$), $\mathrm{S}(\cdot)$ denotes the Softmax operation , and $\mathrm{CB}'(\cdot)$ and $\mathrm{v}'(\cdot)$ are the inverse operations of $\mathrm{CB}(\cdot)$ and $\mathrm{v}(\cdot)$. To avoid introducing noise or interfering, we include a residual connection to preserve the feature integrity of the current modality.

By using F_t as the main guiding modality and F_t as the auxiliary modality, the coarse-grained complementary information from F_t is utilized to enhance F_r. Similarly, the TIR-Guided enhancement feature f_2^t can be expressed using the same principle as Eq. 13.

4.4. Prediction and Loss

We adopt the prediction head of OSTrack [30] directly as our prediction head, with detailed information available in OSTrack. The loss function is formulated as follows:

$$L = L_{\text{cls}} + \lambda_{\text{iou}} L_{\text{iou}} + \lambda_{L_1} L_1 \qquad (14)$$

where, L_{cls} refers to the weighted focal loss for training the classification branch, L_{iou} is used to optimize the overlap between predicted and ground truth bounding boxes, and L_1 is a measure of the average absolute difference between predicted values and true values. $\lambda_{\text{iou}} = 2$ and $\lambda_{L_1} = 5$ are two trade-off parameters.

5. Experiment

5.1. Experiment Setting and Implementation Details

LasHeR. LasHeR [12] is a large-scale RGB-T tracking benchmark with 1,224 pairs of aligned RGB-T video frames (730K frames). It includes manually annotated bounding boxes, covering diverse object categories, camera perspectives, and environmental factors such as season, weather, and day-night variations.

RGBT234. RGBT234 [11] dataset contains 234 RGB-T video pairs (234K frames), with the longest sequence having 8K frames, suitable for long-term tracking. The dataset is annotated with 12 attributes, which enable fine-grained analysis of tracking performance under various conditions.

RGBT210. RGBT210 [13] dataset contains 210 RGB-T video pairs, offering a valuable resource for RGB-T object tracking. Its

Variants	CTC	MSP	CGEM	LasHeR			RGBT234		RGBT210	
				PR	NPR	SR	PR	SR	PR	SR
Baseline				53.0	50.1	43.0	78.6	59.1	75.8	55.3
1	✓			66.1	64.0	54.7	84.9	63.7	84.1	61.9
2			✓	65.5	63.2	54.1	84.4	63.6	83.5	61.2
3	✓	✓		70.7	67.0	56.6	87.1	65.1	86.1	63.2
4	✓		✓	71.5	67.7	57.4	88.1	65.5	86.8	63.4
5	✓	✓	✓	72.1	68.3	57.8	89.0	65.8	87.7	64.2

(a) Ablation of the proposed components.

Layers			LasHeR		
4	7	10	PR	NPR	SR
			53.0	50.1	43.0
✓			69.0	65.8	53.7
		✓	65.5	63.2	51.4
✓	✓		70.7	67.0	56.6
✓		✓	71.5	67.7	57.4
✓	✓	✓	72.1	68.3	57.8

(b) Ablation of inserting layers.

Attribute	DAFNet	MANet	APFNet	TBSI	PMPTrack(Ours)
NO	90.0/63.6	88.7/64.6	93.4/66.4	96.1/72.8	96.8/73.3
PO	85.9/58.8	81.6/56.6	85.0/58.7	88.7/64.7	90.1/64.9
HO	68.6/45.9	68.9/46.5	72.9/49.0	81.5/58.6	82.0/59.2
LI	81.2/54.2	76.9/51.3	82.3/54.4	89.2/63.6	91.8/65.5
LR	81.8/53.8	75.7/51.5	82.9/54.8	85.1/60.0	86.8/61.7
TC	81.1/58.3	75.4/54.3	82.1/57.3	85.8/63.2	87.6/65.7
DEF	74.3/51.6	74.1/52.4	77.1/54.6	84.1/63.7	84.8/63.9
FM	74.0/46.5	69.4/44.9	78.2/49.2	81.4/58.7	81.0/58.2
SV	79.1/54.4	77.7/54.2	82.1/56.5	89.9/66.8	88.8/66.1
MB	70.8/50.0	72.6/51.6	72.8/53.0	88.1/64.9	91.6/66.3
CM	72.3/50.6	71.9/50.8	76.3/54.5	88.0/65.0	91.4/66.8
BC	79.1/49.3	73.9/48.6	80.6/52.4	83.4/57.8	85.6/58.9
ALL	79.6/54.4	77.7/53.9	82.7/57.9	87.1/63.7	89.0/65.8

(c) Attribute-based Tracking Results (PR/SR) on RGBT234.

Variants	LasHeR		
	PR	NPR	SR
Baseline	53.0	50.1	43.0
T_z	68.3	64.1	53.7
$T_x^{\text{irrelevant}}$	67.6	64.3	53.4
T_x^{relevant}	67.2	63.3	52.7
T_z+T_x^{relevant}	71.2	67.4	57.1
T_z+$T_x^{\text{irrelevant}}$	69.6	65.4	56.1
T_x^{relevant}+$T_x^{\text{irrelevant}}$	70.8	67.1	56.7
T_z+T_x^{relevant}+$T_x^{\text{irrelevant}}$	72.1	68.3	57.8

(d) Ablation of token interaction strategies in CTC.

Table 2: Ablation and Explorations on various datasets. Red: best results. Green: second-best. Yellow: third-best.

large scale, comprehensive annotations, and diverse attributes make it ideal for testing tracking algorithms under challenging conditions.

Following the evaluation standards adopted by most current RGB-T trackers [10, 23], for RGBT234 and RGBT210, we use Success Rate (SR) and Precision Rate (PR) as evaluation metrics. For the LasHeR, we extend the evaluation framework by adding an extra Normalized Precision Rate (NPR).

Implementation Details. Our PMPTrack is implemented on Python 3.8 and PyTorch 1.9.0. We train our model on four NVIDIA RTX 3090 GPUs with a batch size of 16 over 30 epochs on the LasHeR training set. Each epoch consists of 60k samples, with template and search region sizes set to 128×128 and 256×256, respectively. AdamW is used as the optimizer with a weight decay of 1e-4, and the learning rate is 1e-4, decaying by a factor of 10 after 10 epochs. Referring to TBSI [10], we use the pre-trained ViT weights from the SOT dataset.

5.2. Comparison with the State-of-the-Arts

Evaluation on LasHeR Dataset. LasHeR is considered more challenging than the RGBT210 and RGBT234 due to its inclusion of more extreme characteristics and attributes, making significant performance gaps in previous methods. When evaluating on this dataset, state-of-the-art methods such as IPL [16], GMMT [23], MFATrack [28], and BAT [3] fail to deliver satisfactory results. In Tab.1, we report the performance, model size and inference speed on the LasHeR dataset. It can be observed that GMMT [23] achieves the previous state-of-the-art performance with PR and NPR of 70.7%/67.0%, and SR of 56.6%. However, our PMPTrack achieves even greater performance improvements,

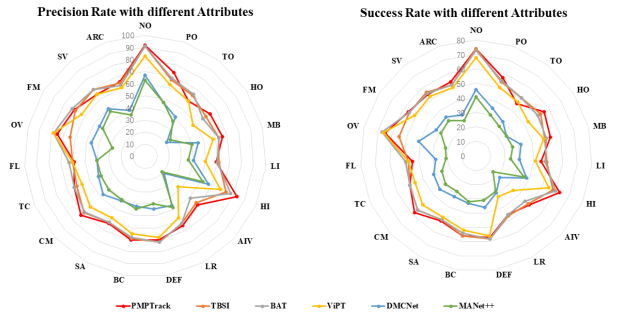

Figure 4: Comparison of PR and SR scores of the PMPTrack and other competing trackers under 19 attributes in the LasHeR dataset.

with PR of 72.1%, NPR of 68.3%, and SR of 57.8%, improves over existing methods by 1.4%, 1.3% and 1.2%, respectively.

To comprehensively analyze the robustness of our proposed PMPTrack, we compare its performance with previous methods on various challenge attributes on the LasHeR dataset, as shown in Fig. 4. Our PMPTrack achieves state-of-the-art performance in the majority of attributes. Specifically, in sequences like **PO** (Partial Occlusion), **HO** (Hyaline Occlusion), **MB** (Motion Blur), and **HI** (High Illumination), our method achieves the best results, indicating its effectiveness in accurately tracking targets even when they are partially or entirely occluded. Notably, it shows a precision improvement of 10.1% and a success improvement of 6.2% in **HI** (High Illumination).

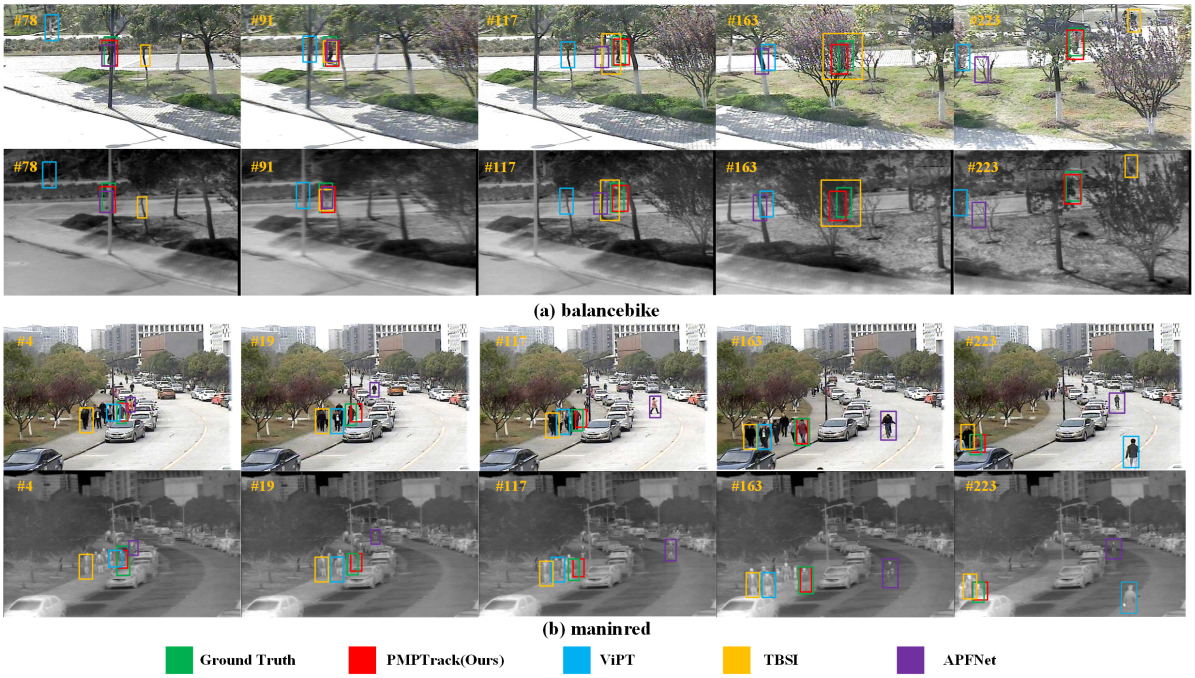

Figure 5: Visualizations of tracking results on two representative sequences from RGBT234.

Setting	Speed	LasHeR			RGBT234		RGBT210	
λ	(FPS)	PR	NPR	SR	PR	SR	PR	SR
0.01	31	70.9	66.6	57.2	88.2	64.7	86.6	63.0
0.03	30.3	71.0	66.8	56.9	87.5	64.5	87.0	63.3
0.05	30.5	70.4	65.9	56.8	87.1	64.1	86.1	62.5
0.07	30	69.6	65.3	55.7	86.1	63.4	85.5	62.3
0.1	31	72.1	68.3	57.8	89.0	65.8	87.7	64.2
0.13	30.8	70.8	66.9	57.1	88.1	65.1	86.8	63.4
0.15	30.7	68.6	65.3	55.2	86.0	63.6	85.3	62.1
0.17	31.2	69.4	65.6	55.9	87.3	64.3	86.5	62.9
0.2	31.5	69.7	66.0	55.9	87.5	64.8	86.7	63.1

Table 3: Ablation Studies on the Hyper Parameter λ. Red: best results. Green: second-best. Yellow: third-best.

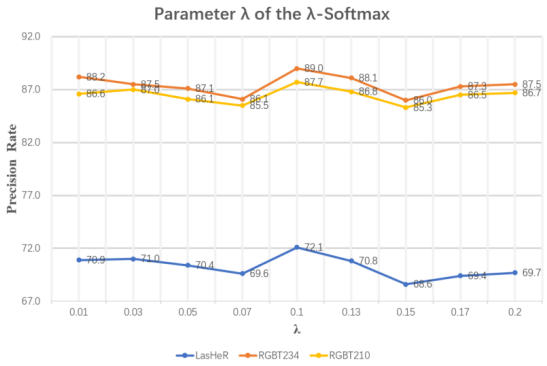

Figure 6: Ablation analysis of the Hyper Parameter λ on LasHeR, RGBT234 and RGBT210 datasets.

Evaluation on RGBT234 Dataset. As shown in Tab.1, it is clear that our tracking method surpassed all previous state-of-the-art trackers on the RGBT234 dataset, achieving a PR of 89.0% and an SR of 65.8%. Siamese-based trackers, such as DFAT (75.8/55.2), and VGG-M-based trackers, such as CAT++ (84.0/59.2), perform worse than our method. Notably, our method surpasses even state-of-the-art trackers built on ViT. Specifically, it outperforms the MFATrack [28] by 7.3%/8.3%, GMMT [23] by 1.1%/1.1%, SDSTrack [9] by 4.2%/3.3%, and TBSI [10] by 1.9%/2.1%, all robust trackers considered in recent years.

Meanwhile, we further evaluate the performance of the PMPTrack across various 12 challenging attributes on the RGBT234 dataset. As shown in Tab.2c, the PMPTrack outperforms other methods on most attributes. Specifically, attributes like **CM** (Camera Moving), **DEF** (Deformation), and **BC** (Background Clutter) can cause significant appearance changes, while **HO** (Hyaline Occlusion), **TC** (Thermal Crossover), and **LI** (Low Illumination) introduce substantial modality differences, which pose significant challenges that test the tracker's robustness. Despite these challenges, the PMPTrack consistently achieves the highest performance, validating the effectiveness of our progressive two-stage modality promotion paradigm, which explores and constructs an outstanding modality representation learning from token level to feature level, significantly exploiting rich modality-specific information. Furthermore, we provide qualitative visualizations of successful tracking cases on RGBT234, as illustrated in Fig. 5.

Evaluation on RGBT210 Dataset. As shown in Tab.1, it is clear that our PMPTrack outperforms all the state-of-the-art trackers in the RGBT210 dataset. For example, compared to TBSI and STMT [19], our PMPTrack achieves an improvement of 2.4%/4.7% in PR and 1.7%/4.7% in SR.

Variants	LasHeR			RGBT234		RGBT210	
	PR	NPR	SR	PR	SR	PR	SR
Baseline	53.0	50.1	43.0	78.6	59.1	75.8	55.3
w/o Token-Level	68.8	65.1	55.1	86.0	64.0	85.3	62.7
w/o Feature-Level	69.2	66.4	56.3	87.1	64.6	86.1	63.2
Feature-Level→Token-Level	67.9	65.1	54.0	84.9	62.8	84.1	62.1
Full Model	72.1	68.3	57.8	89.0	65.8	87.7	64.2

Table 4: In-depth analyses of the proposed two-stage progressive paradigm. Red: best results. Green: second-best. Yellow: third-best.

Variants	LasHeR			RGBT234		RGBT210	
	PR	NPR	SR	PR	SR	PR	SR
MSP_z	68.3	64.1	53.7	84.2	62.0	83.8	61.3
$\text{MSP}_{\tilde{z}}^{\text{irrelevant}}$	67.6	64.3	53.4	84.7	62.2	84.1	61.7
$\text{MSP}_{\tilde{z}}^{\text{relevant}}$	67.2	63.3	52.7	83.9	61.8	83.1	60.8
MSP_z+$\text{MSP}_{\tilde{z}}^{\text{irrelevant}}$	68.6	64.7	54.5	86.0	63.3	85.1	62.3
MSP_z+$\text{MSP}_{\tilde{z}}^{\text{relevant}}$	70.2	66.4	56.8	87.8	65.2	86.2	63.1
$\text{MSP}_{\tilde{z}}^{\text{relevant}}$+$\text{MSP}_{\tilde{z}}^{\text{irrelevant}}$	69.8	65.9	56.3	86.1	63.5	85.3	62.6
MSP_z +$\text{MSP}_{\tilde{z}}^{\text{relevant}}$+$\text{MSP}_{\tilde{z}}^{\text{irrelevant}}$	72.1	68.3	57.8	89.0	65.8	87.7	64.2

Table 5: Quantitative comparison between different combinations of MSP. Red: best results. Green: second-best. Yellow: third-best.

5.3. Ablation Studies

5.3.1 Ablation of the proposed components

In Tab.2a, we analyze the effectiveness of the CTC, MSP, and CGEM. Our baseline model is built upon the OSTrack [30], which is originally a single-modal RGB tracker. As shown in Tab.2a, we intend to gradually add these components to the baseline. To be specific, in **Variant 1**, the CTC is first introduced, which achieves a substantial improvement of 11.1% over the baseline at NPR on LasHeR. Notably, the addition of MSP and CGEM plays a crucial role in enhancing model performance, as evidenced by the overall performance of **Variant 3** and **Variant 4** in comparison to **Variant 1**, clearly highlighting the benefits of incorporating MSP and CGEM. Finally, **Variant 5**, our full model, achieves the best performance.

5.3.2 Inserting layers of the two-stage paradigm

We evaluate different inserting layers of our proposed two-stage paradigm and summarize the experimental results in Tab. 2b. Firstly, inserting the two-stage promotion in the 4-th layer of the ViT backbone yields a large performance boost against our baseline model (PR ↑ 16), showing the importance of the proposed modality promotion paradigm. Then, performance is further improved when inserted into the 7-th and 10-th layers (PR ↑3.1). Marginal improvements are found by inserting more layers, so we adopt the setting of the 4-th, 7-th and 10-th layers as our final model.

5.3.3 Analysis of the token interaction strategies in the CTC

To thoroughly evaluate the proposed CTC, we perform an ablation study on LasHeR to examine different token interaction strategies, as shown in Tab.2d. Removing T_z causes a noticeable drop in PR (72.1 → 70.8, ↓1.3). Excluding T_z^{relevant} leads to a larger PR decline (72.1 → 69.6, ↓2.5), while $T_z^{\text{irrelevant}}$ also has a smaller yet non-negligible impact. These results underscore the significant contribution of all three categories in the CTC.

5.3.4 Role of Hyper Parameter λ

Specifically, the SmoothSoftmax function used in the Modality Spatial Amplification (MSA) operation in Eq. 12 is defined as below:

$$\text{SmoothSoftmax}(\lambda, T^*) = \frac{\exp(\lambda \cdot T^*)}{\sum \exp(\lambda \cdot T^*)} \tag{15}$$

Thus, We further validate the impact of the hyper parameter λ described in Eq. 12. As shown in Tab. 3 and Fig. 6, λ vary within the range of 0.01 to 0.2. The optimal setting of λ is found to be $\lambda = 0.1$, yielding superior performance compared to all other settings. This suggests that setting the smoothing factor λ to 0.1 promotes better generalization by reducing overfitting, improving calibration, and enhancing the learned embeddings. Notably, the runtime remains largely unaffected by the choice of λ, as the smoothing factor is intended to regularize model learning rather than impact computational complexity.

5.3.5 In-depth analyses of Two-Stage Progressive Modality Promotion Paradigm

We conducte in-depth ablation experiments to evaluate the influence of the proposed two-stage modality promotion paradigm. As shown in Tab.4, where **w/o** denotes "without". Removing the TLP results in a more significant overall performance decreased (PR ↓ 3.3 in LasHeR) compared to removing the FLP (PR ↓2.9 on LasHeR), indicating that fine-grained modeling is more critical than coarse-grained refinement. Furthermore, reversing the order of the two stages results in a more substantial performance drop (PR ↓4.1 on RGBT234). These results validate the effectiveness and rigor of our progressive modeling strategy, which strictly transition from fine-grained to coarse-grained representations.

5.3.6 Analysis of the proposed MSP

To access the contribution of MSP, we conducted detailed ablation experiments in Tab. 5. To find the most efficient combination, we tested seven combinations of the three types of MSP we proposed, as shown in Tab.5. We first performed experiments by selecting each category of MSP individually, then tested their pairwise combinations, finally incorporated all three into the model. Excluding $\text{MSP}_z^{\text{relevant}}$ causes a noticeable performance drop in PR scores (↓3.5 on LasHeR, ↓3.0 on RGBT234 and ↓2.6 on RGBT210), highlighting its critical role in transferring fine-grained cross-modality cues for target representation. Removing T_z leads to a slighter PR decline (↓2.3 on LasHeR, ↓2.9 on RGBT234 and ↓2.4 on RGBT210), while $T_z^{\text{irrelevant}}$ also contributes positively, albeit to a lesser extent. Overall, the best performance is achieved when all three prompters are jointly employed. The experimental results demonstrate the indispensable role of the proposed MSP in the model's fine-grained processing.

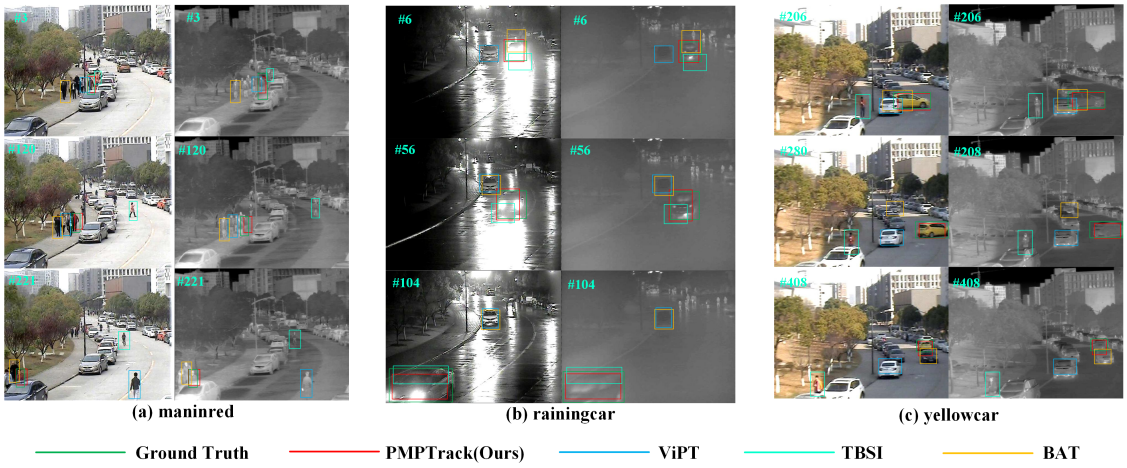

Figure 7: Additional qualitative results analysis for more challenging scenarios on the RGBT234.

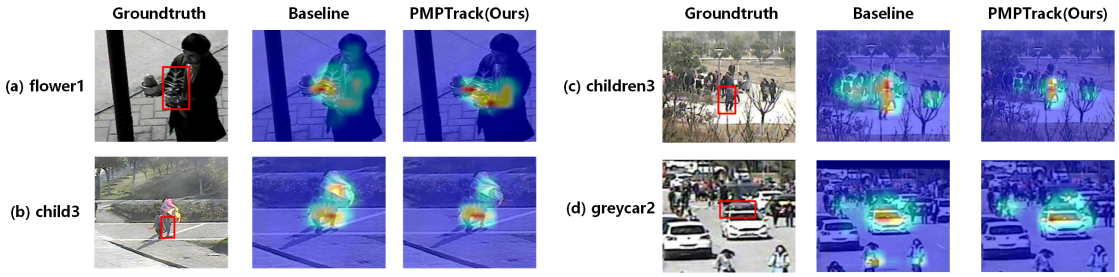

Figure 8: More visualization for the **attention maps** of baseline (OSTrack) and our proposed PMPTrack on the RGBT234.

Setting	LasHeR			RGBT234		RGBT210	
threshold	PR	NPR	SR	PR	SR	PR	SR
0.3	67.6	64.5	54.2	84.2	61.7	82.6	61.0
0.35	69.6	65.3	55.7	86.1	63.4	85.5	62.3
0.4	70.8	66.6	56.6	87.3	64.4	86.2	62.8
0.45	71.0	66.8	56.9	87.5	64.5	87.0	63.3
0.5	72.1	68.3	57.8	89.0	65.8	87.7	64.2
0.55	71.5	67.2	57.0	88.3	65.2	86.7	63.1
0.6	69.7	66.0	55.9	87.3	64.8	86.7	63.1
0.65	70.1	65.8	56.1	87.4	64.7	86.8	62.9
0.7	68.6	65.3	55.2	86.0	63.6	85.3	62.1

Table 6: Ablation Studies on different threshold settings in CTC. Red: best results. Green: second-best. Yellow: third-best.

5.3.7 Ablation of the threshold settings in CTC

In Eq. 4.2, the threshold is empirically set to 0.5. To further validate the impact of different threshold configurations on the CTC module, we conduct additional experiments where the threshold value is varied within the range of 0.3 to 0.7. As presented in Tab. 6, the threshold of 0.5 yields the optimal performance, outperforming all other candidate values. We conjecture that this threshold setting enables the token classification module to effectively mitigate the interference caused by abundant background clutter in the search region, thereby avoiding the mixing of target and distractor features induced by inappropriate cross-interaction and simplifying the target identification process.

5.3.8 More Visualization

We provide additional qualitative comparisons on the RGBT234, with heat map visualizations presented in Fig. 8. When tracking the object "flower" in Fig. 8(a), although the baseline (OSTrack) can attends to the location of object, it also matches with some additionally regions for the target object, whereas our method solely fixes to the flower, validating the effectiveness of our fine-to-coarse modality promotion strategy. Moreover, for some extraordinary conditions, like Fig. 8(d), where the target object "gray car" is heavily occluded by a white car in front, PMPTrack still manages to correctly localize the target. These visualizations further validate our motivation: the proposed two-stage progressive modality promotion paradigm effectively captures and exploits rich modality-specific information, from fine-grained modeling to coarse-grained

Backbone	LasHeR			RGBT234		RGBT210	
Architecture	PR	NPR	SR	PR	SR	PR	SR
Tiny-224	68.6	65.5	55.0	85.2	62.7	83.6	62.0
Base-224	72.1	68.3	57.8	89.0	65.8	87.7	64.2
Base-384	72.9	69.0	58.6	89.8	66.3	88.5	65.1
Large-224	73.7	69.8	59.3	90.5	67.0	89.3	66.0

Table 7: Ablation Studies of different backbone architectures. Red: best results. Green: second-best. Yellow: third-best.

refinement, thereby providing a solid foundation for robust multi-modal fusion in visual tracking.

In Fig. 7, we conduct additional qualitative results analysis for more challenging scenarios. In sequence (b), the target *rainingcar* is situated in a rainy night environment, where the vehicle lights severely interfere with the discrimination of the target's outer contour. Compared with other methods, the target contour in our method is effectively reconstructed and accurately located. In sequence (c), the target *yellowcar* is occluded by environmental distractors, leading to a significant decline in target discriminability. However, the proposed method still achieves superior performance in complex interference scenarios under harsh weather conditions. Experimental results demonstrate that the method proposed in this paper can fully exploit the deep complementary relationships among modalities, thereby exhibiting stronger robustness and generalization capability in the multi-modal object tracking task.

5.3.9 Ablation of different backbone architectures

In Tab. 7, we conduct an ablation study to evaluate the impact of different backbone architectures on the performance of our proposed PMPTrack, with Base-224 set as our default configuration. Specifically, on the LasHeR dataset, the Large-224 backbone achieves the highest PR of 73.7%, outperforming the default Base-224 by 1.6%; Base-384 follows with a PR of 72.9%, which is 0.8% higher than Base-224. While Base-384 and Large-224 yield marginal performance gains across all datasets, their substantially higher resource requirements on hardware devices make them less practical for real-world deployment. By contrast, the default Base-224 backbone achieves a favorable balance between tracking accuracy and computational efficiency, delivering competitive performance (e.g., PR of 72.1% on LasHeR, 89.0% on RGBT234, and 87.7% on RGBT210) with moderate resource consumption. Thus, Base-224 is identified as the most ideal choice for our tracking framework.

6. Conclusion

In this paper, we present a novel perspective on RGB-T tracking by proposing PMPTrack, a decoupled two-stage progressive modality promotion framework that prioritizes modality enhancement before fusion. Our PMPTrack consists of two logically decoupled stages: the TLP stage and the FLP stage. The TLP stage focuses on fine-grained detail modeling, employing the CTC and MSP to enhance the model's ability to discern target regions. Building upon the TLP, the FLP stage performs coarse-grained refinement, emphasizing global enhancement by using the CGEM. Both comparative and ablation experiments demonstrate the effectiveness of our proposed method, highlighting its potential as a promising direction for future research in multimodal visual tracking.

Limitation and Transferability. The main limitation of this work is that the memory usage of PMPTrack is non-trivial (shown in Tab. 1) and maybe unaffordable to some low-RAM devices. In future work, we plan to explore replacing the existing backbone with more lightweight and memory-efficient architectures, such as compact CNNs or efficient Transformer variants, to further reduce memory consumption while maintaining competitive tracking performance. This direction is expected to improve the deployability of the proposed method on resource-constrained platforms.

Regarding transferability, as discussed in Sec. 1 and 3, we argue that the key challenge in multimodal visual tracking is not multimodal fusion itself but effective modality promotion beforehand. To this end, we believe the proposed architecture is inherently flexible and holds potential for extension to other multimodal scenarios, such as RGB-depth and RGB-event tracking—directions we plan to explore in future work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 61906168, 62201400 and 62272267); the Fundamental Research Funds for the Provincial Universities of Zhejiang(Grant No.RF-A2024013); Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020023, LZ23F020001); Construction of Hubei Provincial Key Laboratory for Intelligent Visual Monitoring of Hydropower Projects (Grant No. 2022SDSJ01), the Hangzhou AI major scientific and technological innovation project (Grant No. 2022AIZD0061).

References

[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *ECCVW*, pages 850–865. Springer, 2016.

[2] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte. Learning discriminative model prediction for tracking. In *ICCV*, pages 6182–6191, 2019.

[3] B. Cao, J. Guo, P. Zhu, and Q. Hu. Bi-directional adapter for multimodal tracking. In *AAAI*, volume 38, pages 927–935, 2024.

[4] L. Chen, Y. Huang, H. Li, Z. Zhou, and Z. He. Simplifying cross-modal interaction via modality-shared features for rgbt tracking. In *ACM MM*, pages 1573–1582, 2024.

[5] L. Cheng, J. Wang, and Y. Li. Vitrack: Efficient tracking on the edge for commodity video surveillance systems. *TPDS*, 33(3):723–735, 2021.

[6] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, pages 4660–4669, 2019.

[7] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, pages 9710–9719, 2021.

[8] L. Hong, S. Yan, R. Zhang, W. Li, X. Zhou, P. Guo, K. Jiang, Y. Chen, J. Li, Z. Chen, et al. Onetracker: Unifying visual

object tracking with foundation models and efficient tuning. In *CVPR*, pages 19079–19091, 2024.

[9] X. Hou, J. Xing, Y. Qian, Y. Guo, S. Xin, J. Chen, K. Tang, M. Wang, Z. Jiang, L. Liu, et al. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *CVPR*, pages 26551–26561, 2024.

[10] T. Hui, Z. Xun, F. Peng, J. Huang, X. Wei, X. Wei, J. Dai, J. Han, and S. Liu. Bridging search region interaction with template for rgb-t tracking. In *CVPR*, pages 13630–13639, 2023.

[11] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019.

[12] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, and D. Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2021.

[13] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1856–1864, 2017.

[14] J. Liu, G. Wu, Z. Liu, L. Ma, R. Liu, and X. Fan. Where elegance meets precision: Towards a compact, automatic, and flexible framework for multi-modality image fusion and applications. In K. Larson, editor, *IJCAI*, pages 1110–1118, 8 2024.

[15] L. Liu, C. Li, Y. Xiao, R. Ruan, and M. Fan. Rgbt tracking via challenge-based appearance disentanglement and interaction. *TIP*, 2024.

[16] A. Lu, C. Li, J. Zhao, J. Tang, and B. Luo. Modality-missing rgbt tracking: Invertible prompt learning and high-quality benchmarks. *International Journal of Computer Vision*, 133(5):2599–2619, 2025.

[17] A. Lu, C. Qian, C. Li, J. Tang, and L. Wang. Duality-gated mutual condition network for rgbt tracking. *TNNLS*, 2022.

[18] A. Lu, W. Wang, C. Li, J. Tang, and B. Luo. After: Attention-based fusion router for rgbt tracking, 2024.

[19] D. Sun, Y. Pan, A. Lu, C. Li, and B. Luo. Transformer rgbt tracking with spatio-temporal multimodal tokens. *arXiv preprint arXiv:2401.01674*, 2024.

[20] Y. Sun, B. Cao, P. Zhu, and Q. Hu. Dynamic brightness adaptation for robust multi-modal image fusion. In K. Larson, editor, *IJCAI*, pages 1317–1325, 8 2024.

[21] Y. K. Tan, K. M. Chin, T. S. H. Ting, Y. H. Goh, and T. H. Chiew. Research on yolov8 application in bolt and nut detection for robotic arm vision. In *2024 16th International Conference on Knowledge and Smart Technology (KST)*, pages 126–131. IEEE, 2024.

[22] Z. Tang, T. Xu, H. Li, X.-J. Wu, X. Zhu, and J. Kittler. Exploring fusion strategies for accurate rgbt visual object tracking. *Information Fusion*, 99:101881, 2023.

[23] Z. Tang, T. Xu, X. Wu, X.-F. Zhu, and J. Kittler. Generative-based fusion mechanism for multi-modal tracking. In *AAAI*, volume 38, pages 5189–5197, 2024.

[24] J. Tao, S. Chan, Z. Shi, C. Bai, and S. Chen. Foctrack: Focus attention for visual tracking. *Pattern Recognition*, 160:111128, 2025.

[25] X. Wang, X. Shu, S. Zhang, B. Jiang, Y. Wang, Y. Tian, and F. Wu. Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking. *TMM*, 25:4335–4348, 2022.

[26] Z. Wu, J. Zheng, X. Ren, F.-A. Vasluianu, C. Ma, D. P. Paudel, L. Van Gool, and R. Timofte. Single-model and any-modality for video object tracking. In *CVPR*, pages 19156–19166, 2024.

[27] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang. Attribute-based progressive fusion network for rgbt tracking. In *AAAI*, volume 36, pages 2831–2838, 2022.

[28] H. Xing, W. Wei, L. Zhang, and Y. Zhang. Multi-scale feature extraction and fusion with attention interaction for rgb-t tracking. *Pattern Recognition*, 157:110917, 2025.

[29] J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song. Prompting for multi-modal tracking. In *ACM MM*, pages 3492–3500, 2022.

[30] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022.

[31] J. Zhang, J. Yang, Z. Liu, and J. Wang. Rgbt tracking via frequency-aware feature enhancement and unidirectional mixed attention. *Neurocomputing*, 616:128908, 2025.

[32] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. Van De Weijer, and F. Shahbaz Khan. Multi-modal fusion for end-to-end rgb-t tracking. In *ICCVW*, pages 0–0, 2019.

[33] T. Zhang, H. Guo, Q. Jiao, Q. Zhang, and J. Han. Efficient rgb-t tracking via cross-modality distillation. In *CVPR*, pages 5404–5413, 2023.

[34] T. Zhang, X. He, Q. Jiao, Q. Zhang, and J. Han. Amnet: Learning to align multi-modality for rgb-t tracking. *TCSVT*, 2024.

[35] H. Zhao, D. Wang, and H. Lu. Representation learning for visual object tracking by masked appearance transfer. In *CVPR*, pages 18696–18705, 2023.

[36] J. Zhao, X. Zhang, and P. Zhang. A unified approach for tracking uavs in infrared. In *ICCV*, pages 1213–1222, 2021.

[37] J. Zhu, Z. Chen, Z. Hao, S. Chang, L. Zhang, D. Wang, H. Lu, B. Luo, J.-Y. He, J.-P. Lan, et al. Tracking anything in high quality. *arXiv preprint arXiv:2307.13974*, 2023.

[38] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9516–9526, 2023.

[39] Y. Zhu, C. Li, J. Tang, B. Luo, and L. Wang. Rgbt tracking by trident fusion network. *TCSVT*, 32(2):579–592, 2021.