2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

DMTrack: Spatio-Temporal Multimodal Tracking via Dual-Adapter

Weihong Li[1,2], Shaohua Dong[3], Haonan Lu[4], Yanhao Zhang[4], Heng Fan[3,†], Libo Zhang[1,2,†,*]

***Abstract*— In this paper, we explore adapter tuning and introduce a novel dual-adapter architecture for spatio-temporal multimodal tracking, dubbed DMTrack. The key of our DMTrack lies in two simple yet effective modules, including a spatio-temporal modality adapter (STMA) and a progressive modality complementary adapter (PMCA) module. The former, applied to each modality alone, aims to adjust spatio-temporal features extracted from a frozen backbone by self-prompting, which to some extent can bridge the gap between different modalities and thus allows better cross-modality fusion. The latter seeks to facilitate cross-modality prompting progressively with two specially designed pixel-wise shallow and deep adapters. The shallow adapter employs shared parameters between the two modalities, aiming to bridge the information flow between the two modality branches, thereby laying the foundation for following modality fusion, while the deep adapter modulates the preliminarily fused information flow with pixel-wise inner-modal attention and further generates modality-aware prompts through pixel-wise inter-modal attention. With such designs, DMTrack achieves promising spatio-temporal multimodal tracking performance with merely 0.93M trainable parameters. Extensive experiments on five benchmarks demonstrate that DMTrack achieves state-of-the-art results. Our code and models will be available at https://github.com/Nightwatch-Fox11/DMTrack.**

I. INTRODUCTION

Over the past decades, visual object tracking has played a vital role in computer vision. The remarkable surge of excellent tracking frameworks [1]–[6] has boosted numerous real-world applications [7]–[10]. Despite the promising performance achieved by fine-tuning on large-scale benchmarks [11]–[14], RGB-based object tracking still fails to handle "corner scenarios" under open-world settings, such as extreme illumination and occlusion of similar distractors. Therefore, multimodal tracking is emerging as a pivotal catalyst for advancing more robust tracking performance.

Due to the limited scale of downstream training data [15]–[17], dominant multimodal trackers typically leverage the power of foundation models pre-trained on RGB sequences. To handle this issue, researchers explore parameter-efficient training approaches for multimodal tracking. As demonstrated in Fig. 1 (a), by introducing only a few trainable parameters, some methods [18]–[20] have pioneered the use of parameter-efficient fine-tuning (PEFT) techniques (*e.g.*, prompt tuning, adapter tuning, *etc.*) to adapt RGB-based foundational trackers for multimodal tracking tasks, sparking a trend of PEFT in this field. Recent efforts [21], [22] have further explored LoRA [23] techniques in pursuit of unified multimodal tracking. However, these attempts still adopt an image-level tracking paradigm that relies on a fixed initial template frame and only model spatial relationships, thus limiting their ability to handle complicated situations with significant target appearance variations.

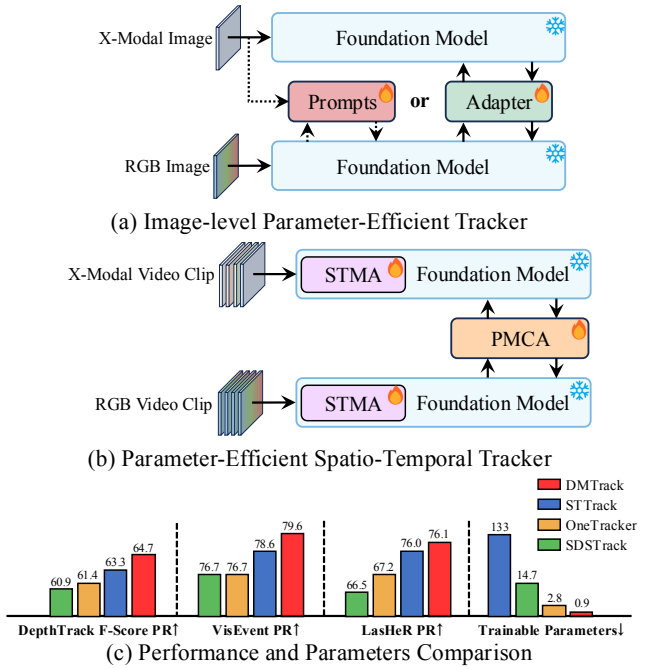


Fig. 1. Comparison of existing unified multimodal trackers and our proposed DMTrack in frameworks (a)-(b) and performance (c). *Best viewed in color for all figures in this paper.*

Conversely, some trackers [24], [25] begin to explore spatio-temporal multimodal tracking through fully fine-tuning on Mamba [26]-based architectures and incorporate global interaction between video streams from different modalities to jointly model spatio-temporal contexts. Although the incorporation of temporal information leads to performance gains, it also introduces a large number of trainable parameters and computational demands, resulting in high memory costs.

To mitigate these limitations, we propose a novel multimodal tracker, dubbed DMTrack, toward parameter-efficient spatio-temporal tracking. In contrast to existing non-temporal parameter-efficient multimodal trackers, we present the first attempt to extend PEFT to joint spatio-temporal context modeling. As shown in Fig. 1 (b), DMTrack freezes the entire foundation model and employs two separate branches to process different modalities. Each branch first performs

†Equal advising and co-last authors.
*Corresponding author: libo@iscas.ac.cn
[1]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
[2]Institute of Software, Chinese Academy of Sciences
[3]Department of Computer Science and Engineering, University of North Texas
[4]OPPO AI Center

pixel-wise inner-modality spatio-temporal modeling in a self-prompting manner, then progressively injects cross-modal complementary prompts, enriched with spatio-temporal cues, into the other modality branch on a per-pixel basis. All learned prompts are built upon the parameters of the foundation model. Specifically, 1) For inner-modality spatio-temporal information incorporation, we adopt a simple template memory bank without temporal propagation to establish temporal relationships efficiently, and we design an STMA that enhances the spatio-temporal feature within the modality-specific template memory while simultaneously reducing the gap between modalities; 2) For inter-modality prompts generation, we propose a PCMA module that facilitates cross-modal interactions with linear complexity. The PCMA module features twin adapters: the shallow adapter establishes bidirectional cross-modal feature alignment via dense connections, while the deep adapter employs pixel-wise attention to refine fused representations and incorporate complementary modality guidance simultaneously.

We summarize our **contributions** as follows:

(1) We present DMTrack, a parameter-efficient framework that adapts pre-trained image-level RGB-based trackers for robust video-level multimodal tracking by integrating dual spatio-temporal adapter modules; (2) DMTrack performs cost-effective modeling of inner-modality spatio-temporal correlation and further reduces computational expenses by progressively generating cross-modal prompts on a pixel-wise basis; (3) To the best of our knowledge, we are the first to leverage adapters to explore spatio-temporal contextual modeling for multimodal tracking. By incorporating only 0.93M trainable parameters (accounting for 0.9% of the total), DMTrack converges to optimal performance within a 5-hour training; (4) Extensive experiments demonstrate that DMTrack achieves state-of-the-art performance across five prevailing benchmark datasets, including DepthTrack, VOT-RGBD2022, VisEvent, LasHeR, and RGBT234.

II. Related Works

A. Multimodal Tracking

Recent RGB-based tracking methods [2]–[4] have achieved promising results on large-scale datasets [12]–[14]. However, despite the strong temporal mechanisms employed, single-modal tracking paradigms still struggle to tackle real-world challenges such as extreme illumination variations. As a result, multimodal trackers, which introduce auxiliary modalities to complement RGB, have gained significant attention. ViPT [19], as an early method, injects auxiliary modalities cues into the RGB information stream with a prompt-tuning architecture. BAT [20] introduces a bidirectional adapter that enables reciprocal interaction between the auxiliary modality and RGB. Although both methods leverage PEFT techniques to reduce training costs, they fail to account for the temporal information. MambaVT [24] and STTrack [25] jointly model spatio-temporal information by global interaction of video streams from different modalities with Mamba [26] architecture. Despite their reasonable performance, current spatio-temporal tracking methods rely on full fine-tuning strategies and global cross-modal interaction between video streams, thus suffering from prohibitive memory and computational demands. In this study, we pioneer a modality-specific adapter design for self-prompting spatio-temporal context in multimodal tracking. With such designs, we reduce the inherent gap between modalities for the following cross-modal prompts generation and avoid expensive global interactions among video tokens from two modalities.

B. Parameter-Efficient Tuning

Different from full fine-tuning, PEFT has recently garnered significant attention due to its ability to substantially reduce the number of trainable parameters, offering an efficient approach to leverage pre-trained models. Originally developed for NLP [27], PEFT has since been adapted and applied to a variety of vision tasks [18]–[20]. Some works [28]–[30] begin to adapt large pre-trained image models (*i.e.*, CLIP [31]) for video downstream tasks. AIM [28] proposed a joint spatio-temporal adaptation method to fine-tune pre-trained vision transformers. ST-Adapter [29] introduced a parameter-efficient space-time adapter that effectively unleashes the power of CLIP for video understanding. Meanwhile, with the advent of ProTrack [18], prompt-tuning was first applied to the tracking domain. Moreover, BAT and ViPT explore the potential of freezing the parameters of image-level trackers while incorporating various spatial adapters or prompts for multimodal tracking. Different from previous parameter-efficient trackers, we introduce spatio-temporal adapters to the multimodal tracking field for jointly modeling inner-modal spatio-temporal correlation, which to our knowledge has not been studied before. In addition to the STMA design, we incorporate pixel-wise attention mechanisms into the adapter architecture to generate modality-aware prompts for inter-modality interaction.

C. Multimodal Fusion

Multimodal fusion serves as a fundamental component in various perception tasks. In autonomous driving, existing transformer-based methods like TransFuser [32] and TriTransNet [33] achieve cross-modal interaction through self-attention mechanisms, but they suffer from quadratic complexity that limits computational efficiency. Recent advances in efficient fusion strategies reveal two promising directions: TokenFusion [34] enhances feature selectivity through dynamic token exchange between modalities, while GeminiFusion [35] introduces lightweight pixel-wise attention for multimodal semantic segmentation. Building upon these developments, we present a novel PMCA module that progressively integrates cross-modal complementary information through a twin adapter design. Our architecture features: a) A shallow bidirectional bridge adapter that synchronously aligns feature representations between modalities through shared dense connection layers, and b) A deep refinement adapter that employs a pixel-wise attention mechanism to modulate preliminary fused modality flow while iteratively injecting complementary guidance from the alternate modality. This dual-stage adaptation enables the progressive in-

corporation of cross-modal cues through parameter-efficient operations while preserving modality-specific characteristics

III. Methodology

In this section, DMTrack is presented step by step. First, we formulate the pipeline of video-level multimodal tracking. Next, we present an STMA designed for inner-modal spatio-temporal context self-prompting, followed by the introduction of a PMCA module that progressively generates cross-modal prompts on a pixel-wise basis. Finally, we introduce the prediction head and training objective function.

A. Video-Level Multi-modal Tracking

In contrast to image-level paradigms that rely on a single template image and a single search image as input, we construct a template memory bank $M \in \mathbb{R}^{T\times 3\times H_z\times W_z}$ using historical frames. This memory bank, combined with a search frame $X \in \mathbb{R}^{3\times H_x\times W_x}$, forms our input, thereby lifting the foundation model to the video level. As illustrated in Fig. 2, our framework processes dual-modality video streams $\{Z^1_{RGB}, Z^2_{RGB}, \ldots, Z^k_{RGB}, X_{RGB}\}$ and $\{Z^1_{XM}, Z^2_{XM}, \ldots, Z^k_{XM}, X_{XM}\}$, which are temporally synchronized and spatially aligned. The core operation within the frozen transformer layers of each modality branch can be formulated as follows:

$$\begin{aligned} Y_{RGB} &= \text{Attn}([Z^1_{RGB}, Z^2_{RGB}, ..., Z^k_{RGB}, X_{RGB}]) \\ Y_{XM} &= \text{Attn}([Z^1_{XM}, Z^2_{XM}, ..., Z^k_{XM}, X_{XM}]) \end{aligned} \tag{1}$$

where XM denotes the X modality (Thermal, Event, and Depth). k is the length of the memory bank. By employing a uniform interval sampling strategy for frame selection, our method enables robust temporal information modeling while maintaining a uniform number of sampled frames. We avoid temporal propagation when incorporating temporal context, as it may lead to overfitting given the limited scale of multimodal training data. With such designs, we simplify the video-level tracking pipeline, significantly reducing memory consumption during training and demonstrating that the memory bank is sufficient to provide robust spatio-temporal cues.

B. Spatio-Temporal Modality Adapter

Previous spatio-temporal trackers have predominantly followed a brute-force paradigm, relying on global cross-modal interactions through full fine-tuning of entire networks. While such approaches achieve moderate performance given sufficient computational and parametric budgets, they suffer from inefficiency and suboptimal performance by neglecting the inherent modality gap between heterogeneous modality video streams. For instance, event video frames exhibit sparse spatio-temporal distribution due to their asynchronous triggering mechanism, while RGB video frames contain dense spatio-temporal variations with continuous photometric changes. To address this limitation, we propose an STMA that dynamically learns spatio-temporal cues for each modality branch with modality-specific parameters. Designed in a modular fashion, STMA is integrated in the front of a transformer block, enabling parameter-efficient spatio-temporal self-prompting that reduces the gap between the two modalities in the high-dimensional feature space. As shown in Fig. 3, for the input of each modality denoted as $X \in \mathbb{R}^{B\times N\times C}$, we split it into the search part and template part after the down-projection:

$$\begin{aligned} X_{\text{down}} &= XW_{\text{down}} + b_{\text{down}} \\ X_x &= X_{\text{down}}[:, T\cdot N_x :] \\ X_z &= X_{\text{down}}[:, : T\cdot N_z] \end{aligned} \tag{2}$$

where N_x and N_z represent the length of search and template tokens, respectively. T is the size of the template memory bank. After we reshape X_z from $X_z \in \mathbb{R}^{B\times(N_z\cdot d)\times T}$ to $X_z \in \mathbb{R}^{(B\cdot N_z)\times d\times T}$, we perform the following operations:

$$X'_z = X_z + \text{Conv1d}(X_z) \tag{3}$$

where Conv1D denotes the 1D-convolution for spatio-temporal reasoning operating on the temporal dimension we introduce. It is noteworthy that after applying Conv1D, the X'_z will be reshaped back from $X'_z \in \mathbb{R}^{(B\cdot N_z)\times d\times T}$ to $X'_z \in \mathbb{R}^{B\times(N_z\cdot d)\times T}$. Finally, X'_z is concatenated with X_x followed by the up-projection:

$$\begin{aligned} X'_{\text{down}} &= \text{Concat}(X'_z, X_x) \\ X_{\text{up}} &= X'_{\text{down}}W_{\text{up}} + b_{\text{up}} \end{aligned} \tag{4}$$

Consequently, the STMA enjoys high efficiency and effectiveness in spatio-temporal modeling while merely incorporating tiny extra (0.6%) parameters.

C. Progressive Modality Complementary Adapter

The paradigm of generating complementary prompts for the other modality through pixel-wise operations has demonstrated promising results [19], [20]. Unlike BAT, which applies an identical processing strategy after both the MHA (multi-head attention) and MLP components within each ViT block, our proposed PMCA explicitly considers the difference in information density between these two stages. Leveraging this observation, PMCA introduces a progressive adaptation strategy composed of two complementary components: a shallow adapter and a deep adapter. Specifically, we adopt bi-directional adapter from BAT as our shallow adapter, which establishes inter-modal connectivity via parameter-shared transformations, creating a foundational feature bridge between each modality branch. On top of this, the deep adapter refines the fused features through dual pixel-wise attention mechanisms: intra-modal attention for feature recalibration and inter-modal attention for modality-aware prompting to guide cross-modal adaptation.

Shallow Adapter. As illustrated in Fig. 4, the shallow adapter includes a down-projection fully connected (FC) layer, an up-projection FC layer, and a linear FC layer. Formally, the shallow adapter can be expressed as:

$$\begin{aligned} Y_{RGB\to X} &= ((X_{RGB}W_{\text{down}})W_{\text{mid}})W_{\text{up}} \\ Y_{X\to RGB} &= ((X_X W_{\text{down}})W_{\text{mid}})W_{\text{up}} \end{aligned} \tag{5}$$

where X_{RGB} and X_X are the input tokens of RGB and X modality. Similar to the STMA, the shallow adapter employs

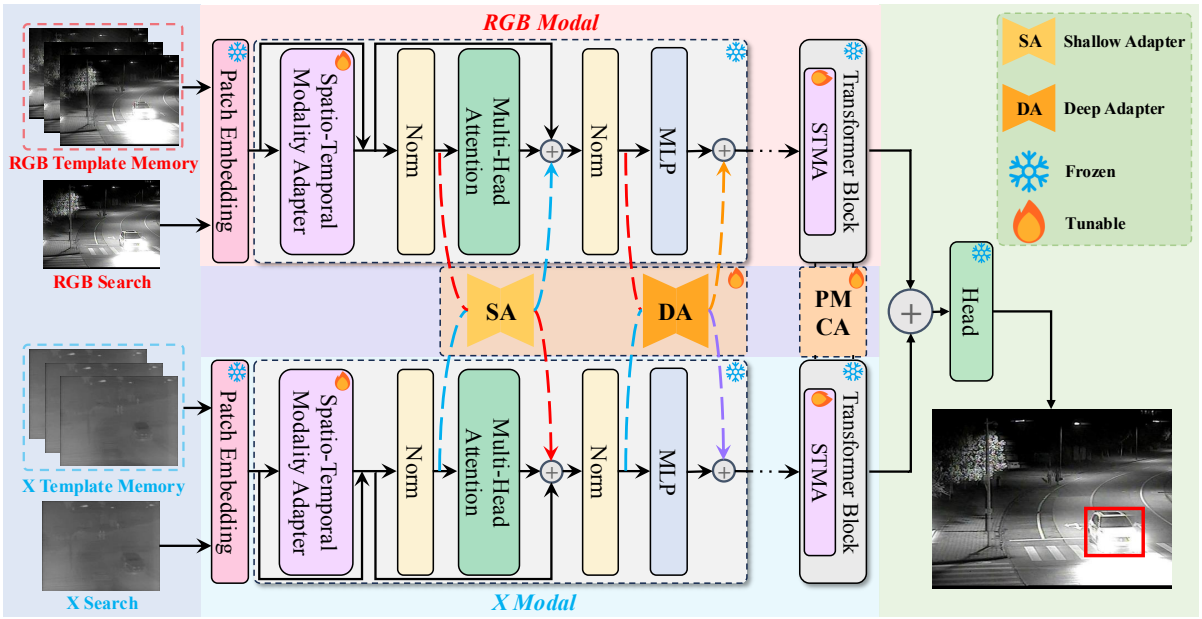


Fig. 2. Overview of the proposed DMTrack. We first tokenize the template and search frames from each modality, then concatenate the resulting token sequences and process them through the frozen transformer architecture. Within each block structure, the STMA remains the only trainable component, specifically designed to produce self-prompts that encode intra-modal spatio-temporal relationships. The PMCA module bridges two processing branches through a twin-adapter architecture, where a shallow adapter and a deep adapter progressively synthesize inter-modal complementary prompts.

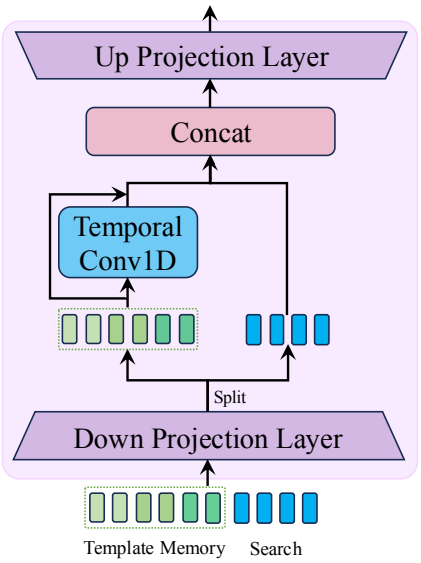


Fig. 3. Detailed design of STMA. In STMA, the temporal context is extracted from Template Memory via a 1D convolutional layer.

a modular design and is integrated into the MHA stage. Since it serves as a foundational feature bridge between each modality branch, the weights are shared across different modality streams. Finally, the complementary information is merged into the other modality stream via element-wise addition. With such a simple but effective design, we establish initial cross-modal correspondences.

Deep Adapter. Building upon the preliminary cross-modal interaction introduced by the shallow adapter, the deep adapter further leverages a pixel-wise MHA mechanism to generate modality-aware complementary prompts. As illustrated in Fig. 5, given the input of RGB and X modality, *i.e.*, $X_{RGB} \in \mathbb{R}^{B\times N\times C}$ and $X_X \in \mathbb{R}^{B\times N\times C}$, we first project them to lower-dimensional of d. Considering the differences between modalities, we adopt a lightweight gating unit to compute relation-aware scores of X modality and RGB as:

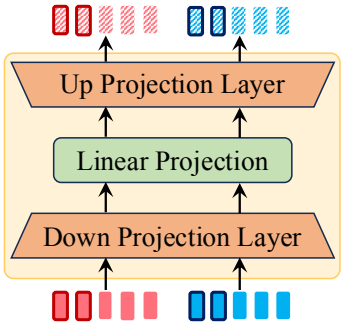


Fig. 4. Detailed design of Shallow Adapter. Multimodal input flows are processed through three FC layers to generate foundational cross-modal complementary prompts, which are subsequently supplied to another modality branch.

$$\begin{aligned} Score_{RGB\to X} &= \text{softmax}(\text{Concat}(X_X, X_{RGB})W_{gate}) \\ Score_{X\to RGB} &= \text{softmax}(\text{Concat}(X_{RGB}, X_X)W_{gate}) \end{aligned} \tag{6}$$

where W_{gate} is the weight of the linear-projection. To prevent the bias introduced by query and key containing information from the same modality, we inject a layer-

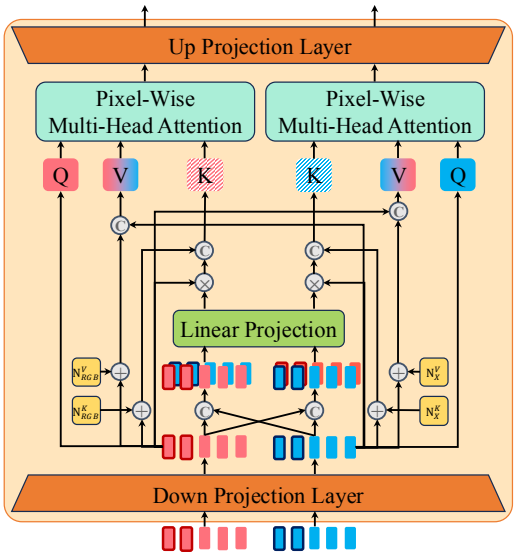


Fig. 5. Detailed design of Deep Adapter. In deep adapter, we construct both Key and Value using dual modalities, enabling pixel-wise attention to simultaneously refine intra-modal representations while adaptively fusing cross-modal information.

adaptive noise when computing the key and value as:

$$\begin{aligned} Q_{RGB} &= X_{RGB} \\ K_{RGB} &= \text{Concat}(X_{RGB} + N^k_{RGB},\ X_{RGB} \odot Score_{X\to RGB}) \\ V_{RGB} &= \text{Concat}(X_{RGB} + N^v_{RGB},\ X_X) \\ Q_X &= X_X \\ K_X &= \text{Concat}(X_X + N^k_X,\ X_X \odot Score_{RGB\to X}) \\ V_X &= \text{Concat}(X_X + N^v_X,\ X_{RGB}), \end{aligned} \tag{7}$$

where N^k_X, N^v_X, N^k_{RGB}, N^v_{RGB} are the learnable noise embeddings, and $\odot$ indicates the element-wise multiplication. As shown in Eq. 7, we integrate self-attention with cross-attention in the deep adapter. This dual mechanism simultaneously captures intra-modal dependencies and inter-modal interactions, thereby producing modality-aware complementary prompts. The attention is computed as:

$$\begin{aligned} \text{P}_{RGB} &= \text{PW-MHA}(Q_{RGB},\ K_{RGB},\ V_{RGB}) \\ \text{P}_X &= \text{PW-MHA}(Q_X, K_X, V_X) \end{aligned} \tag{8}$$

where PW-MHA denotes the pixel-wise MHA. P_{RGB} and P_X represent the modality-aware complementary cues for the RGB and X branches, respectively. It is noteworthy that the core PW-MHA mechanism employs modality-specific parameters. With such designs, we explicitly account for the complementarity of patches at corresponding spatial positions across different modalities. By employing a balanced combination of pixel-wise intra-modal self-attention and inter-modal cross-attention, we generate robust completion cues with minimal computational and parametric overhead. Finally, the P_{RGB} and P_X are projected back to the original dimension and merged into each modality stream via element-wise addition.

D. Head and Objective Loss

Following prevailing methodologies [1] in visual tracking, our framework features a fully convolutional network-based prediction head. The overall loss function is

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_G \mathcal{L}_{\text{GIoU}} + \lambda_l \mathcal{L}_l, \tag{9}$$

where $\lambda_G = 2$ and $\lambda_l = 5$ are the regularization parameters.

IV. Experiments

In this section, we first provide a detailed description of the experimental setup. Next, we compare DMTrack with other state-of-the-art (SOTA) methods across several benchmark datasets. Finally, the ablation study and qualitative comparison are presented.

A. Implementation Details

Training. As a unified multimodal tracking framework, we present a versatile RGB-X tracker that flexibly addresses a range of tasks, including RGB-T, RGB-D, and RGB-E tracking. The training process leverages the LasHeR, DepthTrack, and VisEvent datasets. DMTrack is implemented in Python 3.8 using PyTorch 2.2.2 and trained on four NVIDIA RTX 3090 GPUs over 60 epochs, with each epoch comprising 60,000 sample pairs. The total batch size is set at 64. The search and template images are resized to 256 $\times$ 256 and 128 $\times$ 128, respectively. We employ AdamW [36] optimizer with a weight decay of 1e-4 and initialize the learning rate at 4e-4, reducing it by 10% during the final 20% of the epochs.

Inference. In line with our training configuration, we integrate multiple template memory frames sampled at equal intervals into our tracker during inference. Evaluated on an NVIDIA RTX 3090 GPU, the tracker operates at approximately 39.21 frames per second (FPS).

B. Comparison with State-of-the-Arts

DepthTrack. DepthTrack is a long-term RGB-D tracking benchmark with an average sequence length of 1,473 frames. It includes 200 sequences across 40 scenes and 90 target objects. As shown in Table. I, our DMTrack achieves SOTA results, with an F-score of 64.7%, recall of 64.8%, and precision of 64.7%.

VOT-RGBD2022. VOT-RGBD2022 consists of 127 short-term RGB-D sequences and evaluates tracker performance with Accuracy, Robustness, and Expected Average Overlap (EAO). As demonstrated in Table. II, DMTrack achieves an EAO score of 79.4%, accuracy of 83.7%, and robustness of 94.3%, outperforming the previous SOTA tracker STTrack.

VisEvent. VisEvent, as a large-scale RGB-E dataset, comprises 500 training and 320 testing video sequences. As reported in Table. III, our DMTrack achieves SOTA performance with an AUC of 62.4% and a precision of 79.6%.

LasHeR. LasHeR is a large-scale RGB-T tracking benchmark, consisting of 1,224 aligned sequences. As shown in Table. IV, our DMTrack achieves a success rate (SR) of 60.3% and a precision rate (PR) of 76.1%, outperforming the previous SOTA tracker STTrack by 0.1% in PR. This

TABLE I

OVERALL PERFORMANCE ON DEPTHTRACK TEST SET [16].

	OSTrack [1]	DeT [16]	SPT [37]	ProTrack [18]	ViPT [19]	OneTracker [22]	UnTrack [21]	SDSTrack [38]	SeqTrackv2 [39]	STTrack [25]	**DMTrack (Ours)**
F-score(↑)	0.529	0.532	0.578	0.578	0.594	0.609	0.612	0.614	0.632	**0.633**	**0.647**
Recall(↑)	0.522	0.506	0.538	0.573	0.596	0.604	0.610	0.609	0.634	**0.634**	**0.648**
Precision(↑)	0.536	0.560	0.527	0.583	0.592	0.607	0.613	0.619	0.629	**0.632**	**0.647**

TABLE II

OVERALL PERFORMANCE ON VOT-RGBD2022 [40].

	KeepTrack [41]	STARK-RGBD [42]	SPT [37]	ProTrack [18]	DeT [16]	OSTrack [1]	SBT-RGBD [43]	ViPT [19]	UnTrack [21]	OneTracker [22]	SDSTrack [38]	SeqTrackv2 [39]	STTrack [25]	**DMTrack (Ours)**
EAO(↑)	0.606	0.647	0.651	0.651	0.657	0.676	0.708	0.721	0.718	0.727	0.728	0.744	**0.776**	**0.794**
Accuracy(↑)	0.753	0.803	0.798	0.801	0.760	0.803	0.809	0.815	0.820	0.819	0.812	0.815	**0.825**	**0.837**
Robustness(↑)	0.739	0.798	0.851	0.802	0.845	0.833	0.864	0.871	0.864	0.872	0.883	0.910	**0.937**	**0.943**

TABLE III

OVERALL PERFORMANCE ON VISEVENT [17] TEST SET.

	LTMU_E [44]	ProTrack [18]	TransT_E [45]	SiamRCNN_E [46]	OSTrack [1]	UnTrack [21]	ViPT [19]	SDSTrack [38]	OneTrack [22]	SeqTrackV2 [39]	STTrack [25]	**DMTrack (Ours)**
AUC(↑)	45.9	47.1	47.4	49.9	53.4	58.9	59.2	59.7	60.8	61.2	**61.9**	**62.4**
PR(↑)	65.9	63.2	65.0	65.9	69.5	75.5	75.8	76.7	76.7	78.2	**78.6**	**79.6**

TABLE IV

OVERALL PERFORMANCE ON LASHER [15] TEST SET.

	ProTrack [18]	OSTrack [1]	ViPT [19]	SDSTrack [38]	UnTrack [21]	OneTracker [22]	CAFormer [47]	SeqTrackv2 [39]	TATrack [48]	TBSI [49]	BAT [20]	GMMT [50]	STTrack [25]	**DMTrack (Ours)**
PR(↑)	53.8	52.5	65.1	66.5	66.7	67.2	70.0	70.4	70.2	70.5	70.2	70.7	**76.0**	**76.1**
SR(↑)	42.0	41.2	52.5	53.1	53.6	53.8	55.6	55.8	56.1	56.3	56.3	56.6	**60.3**	**60.3**

TABLE V

OVERALL PERFORMANCE ON RGBT234 [51].

	ProTrack [18]	OSTrack [1]	ViPT [19]	SDSTrack [38]	UnTrack [21]	OneTracker [22]	CAFormer [47]	SeqTrackv2 [39]	TATrack [48]	TBSI [49]	BAT [20]	GMMT [50]	STTrack [25]	**DMTrack (Ours)**
MPR(↑)	79.5	72.9	83.5	84.8	83.7	85.7	88.3	88.0	87.2	86.4	86.8	87.9	**89.8**	**90.3**
MSR(↑)	59.9	54.9	61.7	62.5	61.8	64.2	66.4	64.7	64.4	64.3	64.1	64.7	**66.7**	**65.7**

TABLE VI

ABLATION OF VARIOUS COMPONENTS. EACH ROW IS THE BASELINE MINUS SOME DMTRACK COMPONENT. 'Δ' DENOTES THE AVERAGED PERFORMANCE CHANGE.

Model Variants	LasHeR	Visevent	DepthTrack	Δ
DMTrack	60.3	62.4	64.7	-
w/o STMA	58.7	62.0	64.5	-0.73
w/o STMA & Memory Bank	56.5	60.3	61.4	-3.07
w/o Shallow Adapter	59.5	62.1	62.4	-1.13
w/o Deep Adapter	58.5	62.3	62.6	-1.33

TABLE VII

ABLATION STUDY ON THE SIZE OF THE TEMPLATE MEMORY BANK. GRAY DENOTES OUR FINAL CONFIGURATION.

Memory size	LasHeR	VisEvent	DepthTrack
2	59.4	61.7	64.7
3	60.3	62.4	63.0
4	60.0	62.2	64.4

highlights the effectiveness of continuous spatio-temporal thermal information modeling of DMTrack.

RGBT234. RGBT234, extended from RGBT210 [52] dataset, incorporates a broader range of environmental challenges, consisting of 234 aligned RGBT sequences. As shown in Table. V, DMTrack achieves the highest MPR score of 90.3%, exhibiting very competitive performance.

C. Ablation Study

Component Analysis. In Table. VI, comprehensive ablation studies are conducted to analyze key components of DMTrack. We select AUC in LasHeR, PR in DepthTrack, and AUC in VisEvent as the evaluation metrics. From the results, we observed that the incorporation of temporal information is the most critical factor for performance improvements. When both the memory bank and STMA are removed from the model, DMTrack is reduced to a non-temporal tracker, resulting in the most significant performance degradation. The incorporation of STMA, built upon the memory bank, yields substantial benefits, which demonstrates its ability to facilitate the model in learning the appearance evolution of the target in the memory bank. Additionally, the results reveal that either the absence of basic modality complementary prompts (resulting in blocked bidirectional information flow) or the lack of modality-aware complementary prompts leads to severe performance degradation.

Memory Bank Size. In DMTrack, a key design is the incorporation of a memory bank comprised of historical frames. The historical states provide critical cues of target changes and motion trajectories. The memory bank size represents the length of the temporal information we maintain. In multi-modal tasks, different modalities exhibit varying sensitivities to temporal information. Excessive temporal information can introduce disruptive noise, increasing the learning burden for the model. Therefore, as shown in Table. VII, we explore the

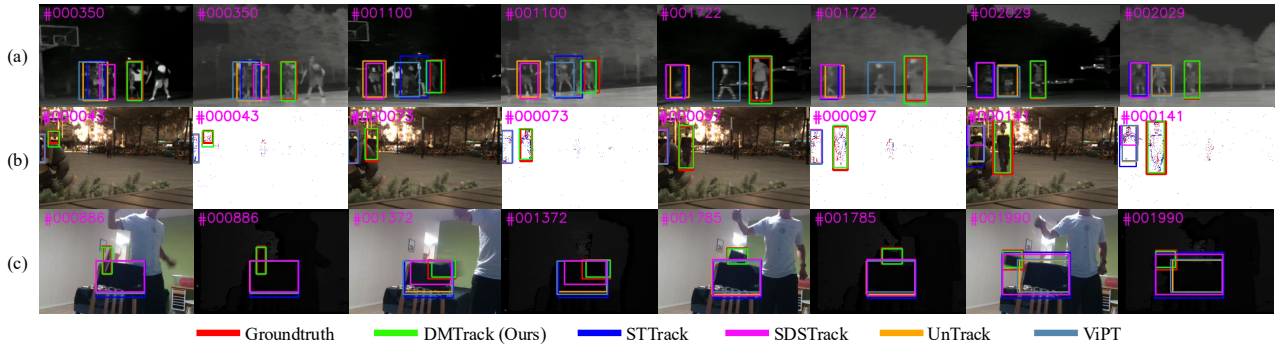
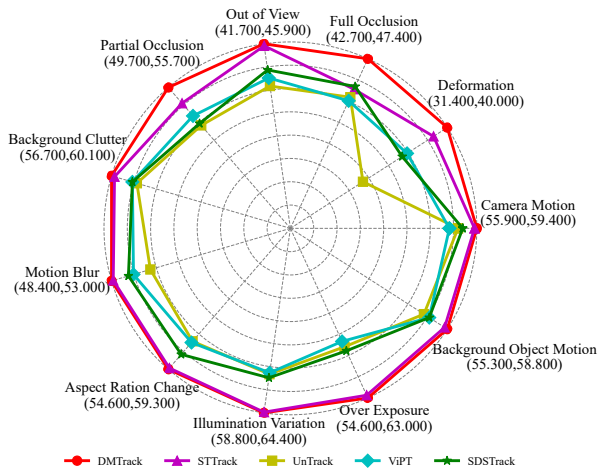
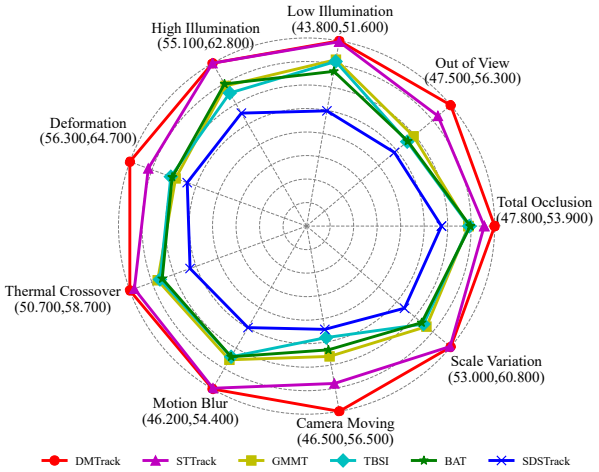


Fig. 6. Qualitative comparison with SOTA unified multimodal trackers across three challenging scenarios: (a) nighttime crowded environments, (b) severe occlusion, and (c) similar distractors. DMTrack demonstrates accuracy and temporal consistency via effective spatio-temporal modeling capabilities.



(a) Attributes within VisEvent.



(b) Attributes within LasHeR.

Fig. 7. Comprehensive comparison between DMTrack and SOTA trackers under challenging attributes within VisEvent (a) and LasHeR (b).

TABLE VIII

ABLATION STUDY ON HIDDEN STATES SIZE AND MODALITY SHARING IN STMA. GRAY DENOTES OUR FINAL CONFIGURATION.

Method	LasHeR	Visevent	DepthTrack
DMTrack	60.3	62.4	64.7
Modality Shared	59.0	62.2	62.0
8 hidden states	60.0	62.1	64.6
12 hidden states	59.9	**62.4**	**64.7**
16 hidden states	**60.3**	62.0	62.5

optimal memory bank size for each modality.

Ablation of STMA. STMA is a critical component of DMTrack, responsible for facilitating the capture of inner-modal spatio-temporal cues. Therefore, we conduct ablation studies on whether parameters are shared across modalities and the size of the hidden states. The results are presented in Table. VIII. We found that when the spatio-temporal information across the two modality video streams is modeled using shared parameters, performance significantly degrades. This supports our hypothesis that video streams from different modalities exhibit distinctly different spatio-temporal information densities, and thus, separate parameters should be employed. We further investigated the optimal hidden state size of STMA for every modality.

D. Visualization and Analysis

Qualitative Comparison. To further show tracking performance, we qualitatively compare DMTrack with four other SOTA multimodal trackers in Fig. 6. Leveraging historical memory and progressive cross-modal prompts, DMTrack addresses a range of challenges such as motion blur and severe occlusion, thereby achieving robust tracking performance.

Attribute-based Performance. Leveraging the rich attribute annotations provided by the VisEvent and LasHeR datasets, we select multiple representative attributes from each benchmark to analyze the performance of our method across various scenarios. As depicted in Fig. 7(b) and Fig. 7(a), DMTrack outperforms previous SOTA trackers on all attributes. The results compellingly demonstrate that our method achieves exceptional robustness across a wide range of challenging scenarios.

V. CONCLUSION

In this work, we present DMTrack, a parameter-efficient spatio-temporal tracking framework that incorporates two novel components: (1) The Spatio-Temporal Modality Adapter, which dynamically self-prompts modality-specific spatio-temporal cues through lightweight history template adaptation, and (2) The Progressive Modality Complementary Adapter module, which facilitates progressive cross-modal prompting via efficient pixel-wise operations. Experiments show that DMTrack is highly effective, achieving SOTA performance across multiple datasets.

ACKNOWLEDGMENT

Libo Zhang is supported by National Natural Science Foundation of China (No. 62476266). Heng Fan is not supported by any fund for this work.

REFERENCES

[1] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, "Joint feature learning and relation modeling for tracking: A one-stream framework," in *ECCV*. Springer, 2022, pp. 341–357.

[2] Y. Bai, Z. Zhao, Y. Gong, and X. Wei, "Artrackv2: Prompting autoregressive tracker where to look and how to describe," in *CVPR*, 2024, pp. 19 048–19 057.

[3] Y. Zheng, B. Zhong, Q. Liang, Z. Mo, S. Zhang, and X. Li, "Odtrack: Online dense temporal token learning for visual tracking," in *AAAI*, vol. 38, 2024, pp. 7588–7596.

[4] B. Kang, X. Chen, S. Lai, Y. Liu, Y. Liu, and D. Wang, "Exploring enhanced contextual information for video-level object tracking," in *AAAI*, 2025.

[5] L. Lin, H. Fan, Z. Zhang, Y. Xu, and H. Ling, "Swintrack: A simple and strong baseline for transformer tracking," in *NeurIPS*, vol. 35, 2022, pp. 16 743–16 754.

[6] L. Lin, H. Fan, Z. Zhang, Y. Wang, Y. Xu, and H. Ling, "Tracking meets lora: Faster training, larger model, stronger performance," in *ECCV*. Springer, 2024, pp. 300–318.

[7] F. Zhang, H. Peng, L. Yu, Y. Zhao, and B. Chen, "Dual-modality space-time memory network for rgbt tracking," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[8] L. Liu, J. Xing, H. Ai, and X. Ruan, "Hand posture recognition using finger geometric feature," in *ICPR*. IEEE, 2012, pp. 565–568.

[9] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE TIP*, vol. 13, pp. 1304–1318, 2004.

[10] J. Xing, H. Ai, and S. Lao, "Multiple human tracking based on multi-view upper-body detection and discriminative learning," in *ICPR*. IEEE, 2010, pp. 1698–1701.

[11] L. Peng, J. Gao, X. Liu, W. Li, S. Dong, Z. Zhang, H. Fan, and L. Zhang, "Vasttrack: Vast category visual object tracking," in *NeurIPS*, vol. 37, 2024, pp. 130 797–130 818.

[12] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "Lasot: A high-quality benchmark for large-scale single object tracking," in *CVPR*, 2019, pp. 5374–5383.

[13] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE TPAMI*, vol. 43, pp. 1562–1577, 2019.

[14] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *ECCV*, 2018, pp. 300–317.

[15] 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 TIP*, vol. 31, pp. 392–404, 2021.

[16] S. Yan, J. Yang, J. Käpylä, F. Zheng, A. Leonardis, and J.-K. Kämäräinen, "Depthtrack: Unveiling the power of rgbd tracking," in *ICCV*, 2021, pp. 10 725–10 733.

[17] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu, "Visevent: Reliable object tracking via collaboration of frame and event flows," *IEEE Transactions on Cybernetics*, 2023.

[18] J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song, "Prompting for multi-modal tracking," in *ACM MM*, 2022, pp. 3492–3500.

[19] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu, "Visual prompt multi-modal tracking," in *CVPR*, 2023, pp. 9516–9526.

[20] B. Cao, J. Guo, P. Zhu, and Q. Hu, "Bi-directional adapter for multimodal tracking," in *AAAI*, vol. 38, 2024, pp. 927–935.

[21] 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*, 2024, pp. 19 156–19 166.

[22] 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*, 2024, pp. 19 079–19 091.

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[24] S. Lai, C. Liu, J. Zhu, B. Kang, Y. Liu, D. Wang, and H. Lu, "Mambavt: Spatio-temporal contextual modeling for robust rgb-t tracking," *IEEE TCSVT*, 2025.

[25] H. Xiantao, T. Ying, Z. Xu, Z. Chen, Z. Zhenyu, L. Jun, Z. Bineng, and Y. Jian, "Exploiting multimodal spatial-temporal patterns for video object tracking," in *AAAI*, 2025.

[26] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv*, 2023.

[27] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[28] T. Yang, Y. Zhu, Y. Xie, A. Zhang, C. Chen, and M. Li, "Aim: Adapting image models for efficient video action recognition," *arXiv*, 2023.

[29] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li, "St-adapter: Parameter-efficient image-to-video transfer learning," *NeurIPS*, vol. 35, pp. 26 462–26 477, 2022.

[30] M. Wang, J. Xing, B. Jiang, J. Chen, J. Mei, X. Zuo, G. Dai, J. Wang, and Y. Liu, "A multimodal, multi-task adapting framework for video action recognition," in *AAAI*, vol. 38, 2024, pp. 5517–5525.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021, pp. 8748–8763.

[32] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *CVPR*, 2021, pp. 7077–7087.

[33] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network," in *ACM MM*, 2021, pp. 4481–4490.

[34] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *CVPR*, 2022, pp. 12 186–12 195.

[35] D. Jia, J. Guo, K. Han, H. Wu, C. Zhang, C. Xu, and X. Chen, "Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer," *arXiv*, 2024.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2018.

[37] X.-F. Zhu, T. Xu, Z. Tang, Z. Wu, H. Liu, X. Yang, X.-J. Wu, and J. Kittler, "Rgbd1k: A large-scale dataset and benchmark for rgb-d object tracking," in *AAAI*, vol. 37, 2023, pp. 3870–3878.

[38] 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*, 2024, pp. 26 551–26 561.

[39] X. Chen, B. Kang, J. Zhu, D. Wang, H. Peng, and H. Lu, "Unified sequence-to-sequence learning for single-and multi-modal visual object tracking," *arXiv*, 2023.

[40] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, H. J. Chang, M. Danelljan, L. Č. Zajc, A. Lukežič *et al.*, "The tenth visual object tracking vot2022 challenge results," in *ECCV*. Springer, 2022, pp. 431–460.

[41] C. Mayer, M. Danelljan, D. P. Paudel, and L. Van Gool, "Learning target candidate association to keep track of what not to track," in *CVPR*, 2021, pp. 13 444–13 454.

[42] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *CVPR*, 2021, pp. 10 448–10 457.

[43] F. Xie, C. Wang, G. Wang, Y. Cao, W. Yang, and W. Zeng, "Correlation-aware deep tracking," in *CVPR*, 2022, pp. 8751–8760.

[44] K. Dai, Y. Zhang, D. Wang, J. Li, H. Lu, and X. Yang, "High-performance long-term tracking with meta-updater," in *CVPR*, 2020, pp. 6298–6307.

[45] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *CVPR*, 2021, pp. 8126–8135.

[46] P. Voigtlaender, J. Luiten, P. H. Torr, and B. Leibe, "Siam r-cnn: Visual tracking by re-detection," in *CVPR*, 2020, pp. 6578–6588.

[47] Y. Xiao, J. Zhao, A. Lu, C. Li, Y. Lin, B. Yin, and C. Liu, "Cross-modulated attention transformer for rgbt tracking," *arXiv*, 2024.

[48] H. Wang, X. Liu, Y. Li, M. Sun, D. Yuan, and J. Liu, "Temporal adaptive rgbt tracking with modality prompt," in *AAAI*, vol. 38, 2024, pp. 5436–5444.

[49] 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*, 2023, pp. 13 630–13 639.

[50] Z. Tang, T. Xu, X. Wu, X.-F. Zhu, and J. Kittler, "Generative-based fusion mechanism for multi-modal tracking," in *AAAI*, vol. 38, 2024, pp. 5189–5197.

[51] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: Benchmark and baseline," *PR*, vol. 96, p. 106977, 2019.

[52] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for rgb-t object tracking," in *ACM MM*, 2017, pp. 1856–1864.