Exploring Modality-Aware Fusion and Decoupled Temporal Propagation for Multi-Modal Object Tracking

Shilei Wang[1], Pujian Lai[1], Dong Gao[1], Jifeng Ning[2], Gong Cheng[1*]

[1]School of Automation, Northwestern Polytechnical University
[2]College of Information Engineering, Northwest A&F University
{shileiwang, laipujian, 2019302284}@mail.nwpu.edu.cn, njf@nwsuaf.edu.cn, gcheng@nwpu.edu.cn

Abstract

Most existing multi-modal trackers adopt uniform fusion strategies, overlooking the inherent differences between modalities. Moreover, they propagate temporal information through mixed tokens, leading to entangled and less discriminative temporal representations. To address these limitations, we propose MDTrack, a novel framework for modality-aware fusion and decoupled temporal propagation in multi-modal object tracking. Specifically, for modality-aware fusion, we allocate dedicated experts to each modality (Infrared, Event, Depth, and RGB) to process their respective representations. The gating mechanism within the Mixture of Experts (MoE) then dynamically selects the optimal experts based on the input features, enabling adaptive and modality-specific fusion. For decoupled temporal propagation, we introduce two separate State Space Model (SSM) structures to independently store and update the hidden states h of the RGB and X-modal streams, effectively capturing their distinct temporal information. To ensure synergy between the two temporal representations, we incorporate a set of cross-attentions between the input features of the two SSMs, facilitating implicit information exchange. The resulting temporally enriched features are then integrated into the backbone via another set of cross-attention, enhancing MDTrack's ability to leverage temporal information. Extensive experiments demonstrate the effectiveness of our proposed method. Both MDTrack-S (Modality-Specific Training) and MDTrack-U (Unified-Modality Training) achieve state-of-the-art performance across five multi-modal tracking benchmarks.

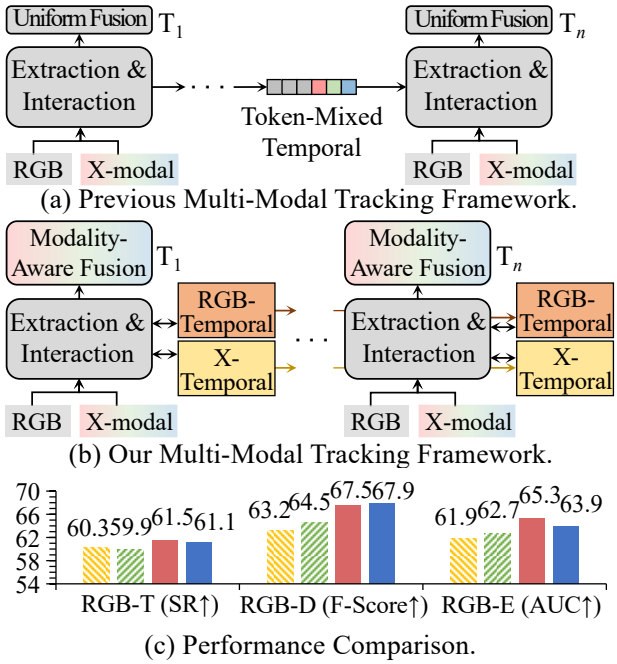

Figure 1: Overview of multi-modal tracking frameworks (a) and (b), with performance comparison (c) from left to right: STTrack, SUTrack, MDTrack-S, and MDTrack-U.

Introduction

Visual object tracking (VOT) is a fundamental task in computer vision that aims to continuously localize an object in a video based on its initial position and has been widely used in autonomous driving, robotics, surveillance, and augmented reality. While recent RGB-based trackers (Danelljan et al. 2019; Bhat et al. 2019; Yan et al. 2021a; Wang et al. 2024; Chen et al. 2021; Wang et al. 2021) have achieved impressive accuracy under normal conditions, they still struggle in challenging scenarios such as low illumination, motion blur, occlusion, and textureless backgrounds where appearance cues are unreliable. To address these limitations, multi-modal tracking has emerged as a promising paradigm by incorporating complementary sensor modalities such as infrared (IR), event, and depth data alongside RGB inputs (Hui et al. 2023; Cao et al. 2024; Hou et al. 2024; Wu et al. 2024). For instance, IR effectively captures thermal signatures under poor lighting, event cameras detect rapid motion changes with high temporal resolution, and depth sensors provide geometric structure invariant to appearance variations. Consequently, multi-modal tracking offers a powerful solution to overcoming the inherent weaknesses of RGB-based methods and achieving reliable performance across diverse real-world environments.

Despite these advances, existing state-of-the-art multi-modal trackers (Hu et al. 2025; Chen et al. 2025), as illus-

*Corresponding author

trated in Fig. 1(a), predominantly adopt a uniform fusion strategy that overlooks modality-specific differences. They use the same fusion module to integrate RGB+infrared (IR), RGB+event, or RGB+depth data, ignoring their distinct signal characteristics, noise patterns, and semantic properties. Under unified-modality training, this "one-size-fits-all" approach limits fusion adaptability and hinders the effective exploitation of each modality's distinct strengths, ultimately resulting in suboptimal tracking performance.

Moreover, for temporal modeling, these trackers typically propagate temporal features through mixed tokens, following the paradigm of RGB-only trackers (Zheng et al. 2024; Wang et al. 2024). However, this entangles heterogeneous temporal dynamics, as RGB streams encode appearance and texture changes while X-modal streams (IR, event, depth) capture thermal stability, polarity events, or geometric consistency. Mixing these distinct temporal cues within a single propagation path causes mutual interference and confounded representations, ultimately impeding robust tracking under challenging scene variations.

To address the limitations of existing multi-modal trackers, we propose MDTrack, which integrates modality-aware fusion with decoupled temporal propagation, as illustrated in Fig. 1(b). This design fully exploits the unique characteristics of each modality while preserving their distinct temporal dynamics, enabling robust multi-modal object tracking.

Specifically, MDTrack adopts a Mixture of Experts (MoE) (Shazeer et al. 2017; Fedus, Zoph, and Shazeer 2022) framework for modality-aware fusion, where dedicated experts are assigned to IR, Event, Depth, and RGB modalities. A gating mechanism dynamically selects appropriate experts based on input features, enabling effective modality-specific fusion. For decoupled temporal propagation, MDTrack employs two independent State Space Models (SSMs) (Gu and Dao 2023) to maintain and update the hidden states of the RGB and X-modal streams separately, thereby modeling their distinct temporal dynamics without interference. In addition, cross-attention is applied to the input features to facilitate implicit inter-stream information exchange. Together, these designs enhance inter-modal collaboration and temporal modeling, resulting in accurate and robust multi-modal tracking performance.

Performance comparisons on three multi-modal tracking tasks, as shown in Fig. 1(c), demonstrate that MDTrack significantly outperforms previous methods such as STTrack (Hu et al. 2025) and SUTrack (Chen et al. 2025).

The main contributions of this work are summarized as follows:

- We propose MDTrack, a novel multi-modal tracking paradigm that combines modality-aware fusion with decoupled temporal propagation to improve tracking robustness across diverse scenarios.
- We develop a modality-aware fusion based on an MoE, which dynamically selects dedicated experts for each modality to achieve effective cross-modal integration.
- We design a decoupled temporal propagation scheme that employs two independent SSMs for RGB and X-modal streams, allowing separate temporal dynamics modeling, while leveraging bidirectional cross-attention to achieve synchronized temporal reasoning and enriched temporal-contextual features.
- Extensive experiments on five mainstream multi-modal tracking benchmarks demonstrate that both MDTrack-S (Modality-Specific Training) and MDTrack-U (Unified-Modality Training) achieve state-of-the-art performance, validating the effectiveness and robustness of our proposed framework.

Related Work

Multi-Modal Object Tracking. Recent advances in RGB-based visual tracking (Danelljan et al. 2019; Ye et al. 2022; Yan et al. 2021a; Wang et al. 2024, 2025) have achieved remarkable performance with deep neural architectures. However, their robustness degrades in challenging scenarios such as low illumination, occlusion, and textureless regions, where appearance cues alone become unreliable. To address these limitations, multi-modal tracking leverages complementary sensory modalities, including infrared, depth and event. TBSI (Hui et al. 2023) enhances RGB-T tracking through temporal-bilateral semantic interactions, while DepthTrack (Yan et al. 2021b) and ProTrack (Yang et al. 2022) improve RGB-D tracking by integrating geometric and visual cues. More recently, unified multi-modal frameworks have emerged as promising solutions: ViPT (Zhu et al. 2023a) employs prompt-based fusion but underutilizes non-RGB information, whereas SDSTrack (Hou et al. 2024), BAT (Cao et al. 2024), and STTrack (Hu et al. 2025) adopt symmetric architectures with temporal modeling to enhance robustness.

Despite these advances, existing methods predominantly adopt uniform fusion strategies and mixed-token temporal propagation, overlooking modality-specific differences and leading to entangled temporal representations. This highlights the need for more flexible and adaptive frameworks capable of achieving modality-aware fusion while decoupling temporal modeling, motivating the design of our proposed MDTrack.

Mixture of Experts. The Mixture of Experts (MoE) method, introduced by Shazeer et al. (Shazeer et al. 2017), dynamically selects specialized experts to handle tasks, significantly increasing model capacity without adding computational complexity. MoE's core advantage is its ability to assign experts to different sub-tasks, improving efficiency while maintaining low cost. Lepikhin et al. (Lepikhin et al. 2020) and Fedus et al. (Fedus, Zoph, and Shazeer 2022) extended MoE to Transformer architectures, enabling larger-scale pre-trained models. MoE has been widely adopted in multi-modal learning. Ma et al. (Ma et al. 2018) proposed a multi-gated MoE method for multi-task learning, assigning experts to specific tasks. Mustafa et al. (Mustafa et al. 2022) demonstrated MoE's potential in vision-language model training, enhancing cross-modal alignment. These applications show that MoE excels not only in single-task optimization but also in multi-modal fusion tasks. In multi-modal image fusion, Zhu et al. (Zhu et al. 2024b) proposed TC-MoA, where expert modules serve as adapters to customize tasks

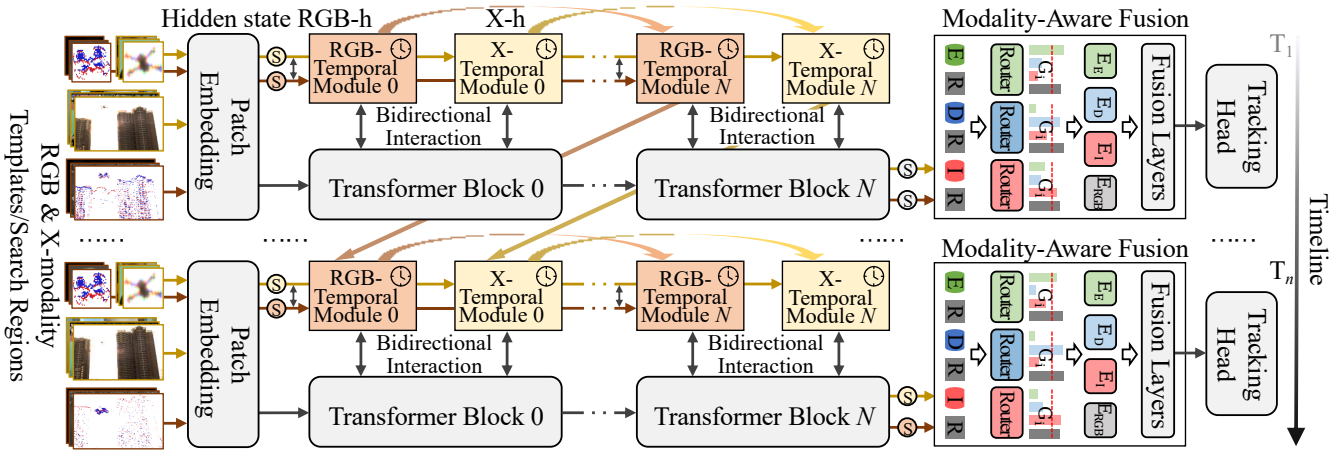

Figure 2: Overall tracking framework of MDTrack. Template and search region tokens from different modalities are concatenated and fed into the backbone. The search region tokens are then decoupled and stored separately for temporal propagation. Finally, the resulting features are fused by the modality-aware fusion module and passed to the tracking head for prediction.

without additional computational overhead. Building on this idea, MDTrack adopts an MoE-based fusion in which each modality is guided by a dedicated expert, enabling dynamic expert selection to improve fusion quality and tracking performance.

State Space Models. The structured State Space Model (SSM) family, especially Mamba introduced by Gu et al. (Gu and Dao 2023) based on S4, has recently drawn attention in vision due to its ability to model long-range dependencies with linear complexity. Following this, Vision Mamba (ViM) (Zhu et al. 2024a) used bidirectional blocks to efficiently adapt SSM to visual data, achieving strong performance while significantly reducing memory usage compared to traditional ViT (Dosovitskiy et al. 2021). VMamba (Liu et al. 2024) further refined this approach by incorporating a 2D scanning SSM, enabling spatially aware sequence modeling with high efficiency. In the single-RGB tracking domain, previous work MCITrack (Kang et al. 2025) employed Mamba SSMs within a HiViT backbone to model long-term sequence dependencies, enabling robust tracking over extended frames. This line of work validates the strong applicability of SSMs in capturing temporal context for visual tracking. Inspired by these advances, MDTrack integrates dual SSM modules to decouple and specialize temporal modeling for RGB and X-modal streams, benefiting from the temporal efficiency and expressiveness of Mamba-style dynamics.

Method

In this section, we first present an overview of the overall architecture of our proposed MDTrack. We then describe the decoupled temporal propagation module, which independently models the temporal dynamics of RGB and X-modal streams. Next, we introduce the modality-aware fusion strategy based on an MoE, which adaptively integrates cross-modal features. Finally, we detail the design of the tracking head.

Overview

The overall architecture of MDTrack is illustrated in Fig. 2. It comprises four key components: a backbone network for visual feature extraction, two decoupled temporal modules (RGB-temporal module and X-temporal module) for independent temporal propagation of RGB and X-modal streams, a modality-aware fusion module for modality-aware feature integration, and a tracking head for final target prediction.

MDTrack takes two video modalities as input, which jointly participate in the object tracking decision process. Specifically, for each modality, input frames are first transformed into template tokens ($\mathbf{Z}_{\text{RGB}}$, $\mathbf{Z}_{\text{X}}$) and search tokens ($\mathbf{S}_{\text{RGB}}$, $\mathbf{S}_{\text{X}}$) through patch embedding and positional encoding. These tokens are then concatenated along the spatial dimension and fed into the backbone to extract unified feature representations.

The backbone consists of N stacked blocks, each paired with its corresponding temporal module. The temporal modules interact bidirectionally with the backbone: temporal features are injected into the backbone to enhance feature accuracy, while the backbone features are used to update the hidden states within the temporal modules, capturing distinct temporal dynamics of each modality.

Subsequently, the extracted features are refined and fused by the modality-aware fusion module, which identifies the modality types and adaptively integrates them using dedicated experts. Finally, the fused features are passed into the tracking head to predict the target location in the current frame.

Decoupled Temporal Propagation

At each time step t, we first process the RGB and X-modal templates $\mathbf{Z}_{\text{RGB}} \in \mathbb{R}^{N\times 3\times H_z\times W_z}$ and $\mathbf{Z}_{\text{X}} \in \mathbb{R}^{N\times 3\times H_z\times W_z}$, as well as the corresponding search regions $\mathbf{S}_{\text{RGB}} \in \mathbb{R}^{N\times 3\times H_s\times W_s}$ and $\mathbf{S}_{\text{X}} \in \mathbb{R}^{N\times 3\times H_s\times W_s}$. Each input is divided into patches and transformed into token sequences us-

ing the patch embedding strategy of HiViT, which progressively downsamples the inputs to better preserve spatial information.

The embedded tokens from both modalities are then concatenated along the spatial dimension and fed into the HiViT backbone, which extracts multi-scale features and models their contextual relationships through its four-stage architecture. To capture temporal dynamics in a modality-specific manner, the search tokens of the RGB and X-modal streams are propagated through their respective temporal modules at each stage.

Specifically, within each stage, the search tokens from both modalities first undergo a bidirectional cross-attention operation to enable implicit feature exchange while maintaining their modality-specific representations. These updated tokens are then fed into their respective Mamba layers based on SSMs to encode and update distinct temporal information for each modality.

The SSM is formulated as a linear dynamical system inspired by continuous-time SSMs, defined as:

$$\begin{aligned} h' &= \mathbf{A}h + \mathbf{B}\mathbf{S}_i, \\ \mathbf{S}'_i &= \mathbf{C}h + \mathbf{D}\mathbf{S}_i, \end{aligned} \quad i \in \{\mathrm{RGB}, \mathrm{X}\}. \tag{1}$$

where h denotes the hidden state, $\mathbf{S}_i$ the input, and $\mathbf{S}'_i$ the output, with $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ being learnable system parameters. To implement SSM in a deep network, it is discretized using the zero-order hold method:

$$\bar{\mathbf{A}} = \exp(\Delta\mathbf{A}), \quad \bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \approx \Delta\mathbf{B}, \tag{2}$$

where Δ is the time scale parameter. The discretized SSM update equations are thus represented as:

$$\begin{aligned} h^t &= \bar{\mathbf{A}}h^{t-1} + \bar{\mathbf{B}}\mathbf{S}_i^t, \\ \mathbf{S}_i'^t &= \mathbf{C}h_t + \mathbf{D}\mathbf{S}_i^t, \end{aligned} \quad i \in \{\mathrm{RGB}, \mathrm{X}\}. \tag{3}$$

Here, h_{t-1} stores historical hidden states that carry crucial temporal context, h_t is the updated hidden state based on the current input $\mathbf{S}_i^t$, and $\mathbf{S}_i'^t$ integrates temporal information for downstream tasks. In our framework, modality-specific SSMs independently update the hidden states h_{RGB}^t and h_{X}^t, thereby preserving the unique temporal dynamics of each modality.

Subsequently, the search tokens carrying temporal information interact twice with the backbone features via cross-attention modules. This bidirectional interaction injects temporal context into the backbone representations to enhance feature accuracy, while also updating the search tokens with enriched spatial-semantic information.

At the next time step $t+1$, the updated hidden states h_{RGB}^t and h_{X}^t are propagated as h_{RGB}^{t+1} and h_{X}^{t+1}, enabling MDTrack to efficiently and decoupledly transmit long-term temporal information across video frames. This decoupled temporal modeling design ensures that the RGB and X-modal streams maintain their unique temporal dynamics without interference, resulting in more robust multi-modal tracking performance.

Modality-Aware Fusion Module

As illustrated in Fig. 2, the modality-aware fusion module consists of a modality expert library $\{\mathrm{E}_{\mathrm{RGB}}, \mathrm{E}_{\mathrm{T}}, \mathrm{E}_{\mathrm{E}}, \mathrm{E}_{\mathrm{D}}\}$, which contains modality-specific processing methods, a gating weight library $\{\mathbf{G}_{\mathrm{RGB}}, \mathbf{G}_{\mathrm{T}}, \mathbf{G}_{\mathrm{E}}, \mathbf{G}_{\mathrm{D}}\}$ to determine the contribution of each expert, and guided fusion weights $\mathbf{F}_i$. This module operates in two main stages: modality-specific expert selection and expert-guided fusion.

In the modality-specific expert selection stage, we first concatenate the multi-modal features to form a unified representation of the RGB and X-modality token pairs, enabling cross-modal interactions in subsequent processing. The router then generates the gating weights $\mathbf{G}_i$ for each modality expert based on the joint features. Specifically, the routing computation is defined as:

$$\begin{aligned} \mathbf{G}_i = &\mathrm{Softmax}(\mathrm{TopK}(\mathbf{S}_{\mathrm{RGBX}} \cdot \mathbf{W}_g + N(0,1) \\ &\cdot \mathrm{Softplus}(\mathbf{S}_{\mathrm{RGBX}} \cdot \mathbf{W}_{\mathrm{noise}}))), \end{aligned} \tag{4}$$

where $\mathrm{TopK}(\cdot)$ retains only the top K ($K = 2$) values, setting others to $-\infty$ so that their Softmax outputs become zero. $\mathbf{W}_{\mathrm{g}}$ and $\mathbf{W}_{\mathrm{noise}}$ are learnable parameters.

In the expert-guided fusion stage, each modality expert E_i and its associated routing output $\mathbf{G}_i$ are used to generate modality-aware fusion weights:

$$\mathbf{F}_{\mathrm{RGB}} = \mathrm{GAP}(\mathrm{Sigmoid}(\mathbf{G}_{\mathrm{RGB}} \cdot \mathrm{E}_{\mathrm{RGB}}(\mathbf{S}_{\mathrm{RGB}}))), \tag{5}$$

$$\mathbf{F}_{\mathrm{X}} = \mathrm{GAP}(\mathrm{Sigmoid}(\mathbf{G}_{\mathrm{X}} \cdot \mathrm{E}_{\mathrm{X}}(\mathbf{S}_{\mathrm{X}}))), \tag{6}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, and $\mathbf{G}_i$ is the routing value for the corresponding adapter. The resulting modality-customized weights $\mathbf{F}_i \in \mathbb{R}^{H \times W \times 1}$ range within $(0, 1)$ and serve as an importance assessment for each modality's information. We employ a load-balancing loss to regularize the routing process, thereby encouraging proper activation of experts.

Finally, these weights refine and fuse the modality features via element-wise multiplication and weighted summation:

$$\mathbf{S} = \lambda_{\mathrm{RGB}}(\mathbf{F}_{\mathrm{RGB}} \odot \mathbf{S}_{\mathrm{RGB}}) + \lambda_{\mathrm{X}}(\mathbf{F}_{\mathrm{X}} \odot \mathbf{S}_{\mathrm{X}}), \tag{7}$$

where both λ_{RGB} and λ_{X} are set to 0.5. This operation effectively removes redundant information while preserving complementary features, producing an adaptive and modality-aware fused representation for tracking.

Head and Loss Function

We adopt a prediction head architecture commonly used in recent Transformer-based trackers, with necessary adaptations for our multi-modal framework. Specifically, our tracking head comprises three parallel convolutional subnetworks dedicated to different prediction tasks. The first branch outputs the classification confidence map $\mathbf{P}_{\mathrm{S}} \in \mathbb{R}^{1 \times \frac{H}{16} \times \frac{W}{16}}$, indicating the likelihood of the target at each spatial location. The second branch predicts the target's width and height $\mathbf{P}_{\mathrm{B}} \in \mathbb{R}^{2 \times \frac{H}{16} \times \frac{W}{16}}$, while the third branch

	Method	Publication	LasHeR		RGBT234		DepthTrack			VOT-RGBD2022		
			Pr (↑)	AUC (↑)	MPR (↑)	MSR (↑)	Pr (↑)	Re (↑)	F-score (↑)	EAO (↑)	Acc. (↑)	Rob. (↑)
Modality-Specific Training	ViPT (Zhu et al. 2023a)	CVPR'23	65.1	52.5	83.5	61.7	59.2	59.6	59.4	72.1	81.5	87.1
	SPT (Zhu et al. 2023b)	AAAI'23	-	-	-	-	52.7	54.9	53.8	65.1	79.8	85.1
	Un-Track (Wu et al. 2024)	CVPR'24	66.7	53.6	83.7	61.8	61.3	61.0	61.2	72.1	81.5	87.1
	SDSTrack (Hou et al. 2024)	CVPR'24	66.5	53.1	84.8	62.5	61.9	60.9	61.4	72.8	81.2	88.3
	OneTracker (Hong et al. 2024)	CVPR'24	67.2	53.8	85.7	64.2	60.7	60.4	60.9	72.7	81.9	87.2
	TBSI (Hui et al. 2023)	CVPR'24	70.5	56.3	86.4	64.3	-	-	-	-	-	-
	TATrack (He et al. 2023)	AAAI'24	70.2	56.1	87.2	64.4	-	-	-	-	-	-
	BAT (Cao et al. 2024)	AAAI'24	70.2	56.3	86.8	64.1	-	-	-	-	-	-
	GMMT (Tang et al. 2024)	AAAI'24	70.7	56.6	87.9	64.7	-	-	-	-	-	-
	STTrack (Hu et al. 2025)	AAAI'25	76.0	60.3	89.8	66.7	63.3	63.4	63.2	77.6	82.5	93.7
	MDTrack-S	Ours	**76.5**	**61.4**	**93.0**	**70.5**	**67.5**	**67.5**	**67.5**	**79.7**	**83.6**	**94.8**
Unified-Modality Training	Stark (Yan et al. 2021a)	ICCV'21	41.8	33.3	67.7	49.6	39.7	40.6	38.8	44.5	71.4	59.8
	AiATrack (Gao et al. 2022)	ECCV'22	46.3	36.5	71.1	50.8	51.5	52.6	50.5	64.1	76.9	83.2
	OSTrack (Ye et al. 2022)	ECCV'22	53.0	42.2	75.5	56.9	56.9	58.2	55.7	66.6	80.8	81.4
	SeqTrack (Chen et al. 2023)	CVPR'23	58.2	44.1	80.6	59.9	59.0	60.0	58.0	67.9	80.2	84.6
	ViPT (Zhu et al. 2023a)	CVPR'23	60.8	49.0	-	-	56.1	56.2	56.0	-	-	-
	Un-Track (Wu et al. 2024)	CVPR'24	64.6	51.3	84.2	62.5	61.0	61.0	61.0	71.8	82.0	86.4
	SUTrack (Chen et al. 2025)	AAAI'25	74.5	59.9	92.2	69.5	65.1	65.7	64.5	76.5	82.8	91.8
	XTrack (Tan et al. 2025)	ICCV'25	73.1	58.7	87.8	65.4	65.4	64.3	64.8	74.0	82.8	88.9
	MDTrack-U	Ours	**76.3**	**61.1**	**92.6**	**70.6**	**68.1**	**67.6**	**67.9**	**80.0**	**83.5**	**95.1**

Table 1: Comparisons with state-of-the-arts on LasHeR, RGBT234, DepthTrack, and VOT-RGBD2022. The best results are highlighted with bold underlines, and the second-best results are shown in bold fonts.

estimates the bounding box offsets $\mathbf{P}_{\mathrm{O}} \in \mathbb{R}^{2\times\frac{H}{16}\times\frac{W}{16}}$ to refine localization precision.

During training, we formulate the objective as a weighted combination of classification, regression, and load-balancing losses. The classification branch is supervised with a focal loss to address class imbalance and emphasize hard examples. For the regression branches, we employ a hybrid loss comprising an ℓ_1 loss for bounding box parameter regression and a generalized IoU (GIoU) loss to improve overlap consistency with ground truth. Additionally, a load-balancing loss $\mathcal{L}_{\text{balance}}$ is introduced to regularize the routing process and encourage proper expert activation. Formally, the total loss for a training batch is defined as:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{l1}}\mathcal{L}_{\text{l1}} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{balance}}\mathcal{L}_{\text{balance}}, \quad (8)$$

where $\mathcal{L}$cls, $\mathcal{L}$l1, $\mathcal{L}$giou, and $\mathcal{L}$balance denote the classification, ℓ_1, GIoU, and load-balancing losses, respectively. The coefficients λ_{cls}, λ_{l1}, λ_{giou}, and λ_{balance} weight the contributions of the corresponding objectives during optimization.

Experiments

In this section, we first present the implementation details of the proposed MDTrack model, including its training pipeline and inference strategy. We then conduct comprehensive comparisons with state-of-the-art methods on various multi-modal tracking benchmarks to evaluate its effectiveness. Finally, we perform ablation studies on the key components of MDTrack to analyze their contributions to the overall tracking performance.

Implementation Details

Training. We adopt two training strategies for MDTrack. The first follows the conventional multi-modal tracking setup, where the model is trained using only one modality at a time (i.e., RGB+Depth, RGB+Thermal, or RGB+Event). The second strategy merges all modality datasets for joint training, enabling a single model to handle tracking tasks across all modalities. We initialize parts of our model parameters using the pretrained weights from the RGB tracker (Kang et al. 2025). For all training settings, the template size is set to 112×112 and the search region size to 224×224. Training is performed on four NVIDIA RTX 4090 GPUs, with modality-specific training running for 20 epochs and mixed-modality training running for 30 epochs. Each epoch consists of 60,000 sample pairs, with a batch size of 16. The AdamW optimizer is used with a learning rate of 5e-5.

Inference. During inference, the template and search region sizes remain consistent with those used during training. Temporal information is incrementally integrated into the tracking pipeline in a modality-decoupled manner. MDTrack-S and MDTrack-U achieve inference speed of approximately 25 frames per second (FPS) when tested on an NVIDIA RTX 4090 GPU.

Method	Modality-Specific Train						Unified-Modality Train					
	ViPT	Un-Track	SDSTrack	OneTrack	STTrack	MDTrack-S	SeqTrack	ViPT	Un-Track	SUTrack	XTrack	MDTrack-U
Precision (↑)	75.8	76.3	76.7	76.7	78.6	**82.2**	66.5	74.0	75.5	79.9	80.5	**81.3**
Success (↑)	59.2	59.7	59.7	60.8	61.9	**65.3**	50.4	57.9	58.9	62.7	63.3	**63.9**

Table 2: Precision and Success comparisons between Modality-Specific and Unified-Modality training on the VisEvent dataset. The best results are highlighted with bold underlines, and the second-best results are shown in bold fonts.

Comparison with State-of-the-Arts

We conducted a comprehensive comparison of MDTrack against recent state-of-the-art multi-modal trackers across five benchmark datasets. All experimental results are obtained by running MDTrack once on each dataset, following the standard evaluation approach for most tracking methods. As shown in Tab. 1 and Tab. 2, both MDTrack-S (the Modality-Specific Training variant) and MDTrack-U (the Unified-Modality Training variant) consistently achieved either the best or second-best results on all five datasets. Unlike Un-Track, which exhibits a noticeable performance gap between its two training modes, MDTrack-S and MDTrack-U deliver similarly strong performance, highlighting the effectiveness of the proposed modality-aware fusion design.

LasHeR. LasHeR is a large-scale RGBT tracking benchmark with 1,224 RGB-thermal video pairs and over 730K annotated frames (Li et al. 2021). On LasHeR, both MDTrack-S and MDTrack-U achieve state-of-the-art performance among their respective categories. MDTrack-S obtains 76.5% precision and 61.4% AUC, while MDTrack-U achieves 76.3% precision and 61.1% AUC, significantly outperforming all previous methods and setting new benchmarks in RGBT tracking.

RGBT234. RGBT234 contains 234 spatially aligned RGB-thermal video pairs with 234K annotated frames under diverse real-world conditions (Li et al. 2019). MDTrack-S delivers the best performance with an MPR of 93.0% and MSR of 70.5%, surpassing the previous best method STTrack by 3.2% and 3.8%, respectively. MDTrack-U also achieves strong results, reaching 92.6% MPR and 70.6% MSR, outperforming SUTrack (92.2% / 69.5%) and demonstrating excellent generalization under unified training.

DepthTrack. DepthTrack is the largest RGB-D tracking benchmark, consisting of 200 sequences across 90+ object classes and 40+ scenes (Yan et al. 2021b). MDTrack-S achieves balanced and robust performance with 67.5% precision, 67.5% recall, and 67.5% F1-score, improving upon the previous best (STTrack, 63.2% F1) by 4.3%. MDTrack-U further advances the performance, attaining 68.1% precision, 67.6% recall, and 67.9% F1-score, thereby setting new state-of-the-art results across all metrics on this dataset.

VOT-RGBD2022. VOT-RGBD2022 is a short-term RGB-D tracking benchmark featuring over 140 sequences and evaluated with EAO, accuracy, and robustness (Kristan et al. 2022). MDTrack-S achieves 79.7% EAO, 83.6% accuracy, and 94.8% robustness, ranking second among all methods. In contrast, MDTrack-U attains the highest EAO score of 80.0%, with 83.5% accuracy and 95.1% robustness, surpassing all existing trackers and demonstrating superior performance under unified training.

VisEvent. VisEvent is the first large-scale benchmark for RGB-event tracking, with 820 synchronized video pairs and 371K annotated frames (Wang et al. 2023). MDTrack-S achieves the highest Precision and Success of 82.2% and 65.3%, outperforming the previous best method STTrack by 3.6% in Precision and 3.4% in Success. Meanwhile, MDTrack-U obtains competitive performance with 81.3% Precision and 63.9% Success, ranking second overall and consistently outperforming other leading methods such as OneTrack, SUTrack, and Un-Track.

Ablation Studies

We conduct ablation studies on key components of MDTrack, including the modality-aware fusion module and the decoupled temporal propagation module.

Module Contribution Analysis. From Tab. 3, adding the decoupled temporal module improves performance to 60.2% (+1.7%) AUC on LasHeR, 67.6% (+1.8%) F-score on DepthTrack, and 63.3% (+1.1%) Success rate on VisEvent, achieving a mean gain of +1.5%. This demonstrates that decoupling temporal modeling effectively enhances target representation across diverse modalities. Incorporating the modality-aware fusion module achieves 59.6% (+1.1%) AUC on LasHeR, 66.3% (+0.5%) F-score on DepthTrack, and 62.8% (+0.6%) Success rate on VisEvent, with an average improvement of +0.7%. This validates the advantage of expert-based adaptive fusion in leveraging complementary cross-modal cues. Finally, combining the decoupled temporal module and modality-aware fusion module achieves the best results, reaching 61.1% (+2.6%) AUC on LasHeR, 67.9% (+2.1%) F-score on DepthTrack, and 63.9% (+1.7%) Success rate on VisEvent, leading to a mean gain of +2.1%. These results confirm that temporal decoupling and modality-aware fusion are complementary, jointly contributing to robust multi-modal tracking performance.

Temporal Propagation Analysis. We investigate three designs for temporal information propagation in MDTrack. As shown in Tab. 3, using the Token-based Temporal Module, which concatenates historical template and search region tokens at the beginning of the video and propagates them directly across frames, yields only marginal improvements (e.g., +0.5% on average). This limited gain is mainly due to the entanglement of temporal information between RGB and X modalities, leading to feature interference. Similarly, introducing the Temporal Module based on a single SSM (Gu and Dao 2023) for mixed temporal propaga-

Configuration	LasHeR	DepthTrack	VisEvent	Mean
Baseline	58.5	65.8	62.2	62.2
+ Token-based Temporal Module	59.4 (+0.9)	66.3 (+0.5)	62.3 (+0.1)	62.7 (+0.5)
+ Temporal Module	59.3 (+0.8)	66.6 (+0.8)	62.6 (+0.4)	62.8 (+0.6)
+ Decoupled Temporal Module	60.2 (+1.7)	67.6 (+1.8)	63.3 (+1.1)	63.7 (+1.5)
+ Fusion Module	58.8 (+0.3)	65.9 (+0.1)	62.4 (+0.2)	62.4 (+0.2)
+ Modality-Aware Fusion Module	59.6 (+1.1)	66.3 (+0.5)	62.8 (+0.6)	62.9 (+0.7)
+ Decoupled Temporal Module & Modality-Aware Fusion Module	**61.1 (+2.6)**	**67.9 (+2.1)**	**63.9 (+1.7)**	**64.3 (+2.1)**

Table 3: Ablation study of MDTrack on the LasHeR, DepthTrack, and VisEvent datasets. Each row illustrates a different module added to the baseline.

tion brings slightly higher improvements (+0.6% on average). However, the mixed design still cannot fully disentangle modality-specific temporal cues. In contrast, the decoupled temporal module, which employs separate SSMs for RGB and X modalities along with implicit cross-attention between them, achieves the most significant performance boost (+1.5% on average). This demonstrates that decoupling modality-specific temporal representations while enabling their interaction effectively enhances temporal modeling and overall tracking accuracy.

Modality-aware Fusion Analysis. We evaluate two fusion designs in MDTrack. As shown in Tab. 3, employing the Fusion Module, which directly combines RGB and X modality features after processing them with a unified expert structure, brings only limited improvements (+0.2% on average). This is because treating all modalities uniformly without considering their modality-specific differences leads to suboptimal fusion, as modality-specific cues may interfere with each other. In contrast, the modality-aware fusion Module leverages an MoE mechanism to adaptively select different experts and guidance strategies based on the specific modality task. This design yields a higher performance gain (+0.7% on average), demonstrating that adaptive expert selection effectively captures complementary information across modalities while suppressing irrelevant features. Overall, modality-aware fusion significantly enhances the robustness and generalization ability of our tracker in diverse multimodal scenarios.

Visual Comparison. As shown in Fig. 3, in the RGBT tracking scenario with numerous similar objects, MDTrack-S and MDTrack-U achieve stable tracking by effectively utilizing decoupled temporal propagation and modality-aware adaptive fusion, enabling them to accurately distinguish targets from distractors by leveraging both temporal cues and infrared information. In the RGBD tracking scenario, where the cup is partially occluded, MDTrack-S and MDTrack-U robustly localize the target by integrating depth information through their modality-specific experts while preserving temporal consistency via the decoupled temporal module. In the RGBE tracking scenario, characterized by fast-moving basketball players under dim lighting, MDTrack-S and MDTrack-U maintain accurate tracking by exploiting the high temporal resolution of event data alongside robust temporal modeling. These results demonstrate that the proposed decoupled temporal modeling and modality-aware fusion effectively improve tracking robustness and accuracy.

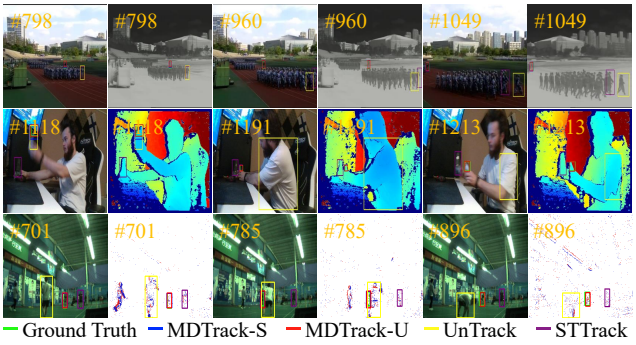

Figure 3: Visual comparisons of MDTrack-S and MDTrack-U with other multimodal trackers on the LasHeR, DepthTrack, and VisEvent datasets.

Conclusion

We present MDTrack, a novel multi-modal tracking framework that effectively addresses the challenges of modality heterogeneity and temporal entanglement. By integrating a modality-aware fusion mechanism based on an MoE with decoupled temporal propagation through dual structured SSMs, MDTrack captures modality-specific features and temporal dynamics while enabling their synergistic interaction. Extensive evaluations on five diverse benchmarks demonstrate that both modality-specific and unified training paradigms achieve state-of-the-art performance, underscoring the robustness and generalizability of our approach. MDTrack provides an effective approach for multi-modal tracking and offers valuable insights for future research on robust visual understanding using heterogeneous sensor data.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62376223 and 62476227, and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2025087, as well as the Fundamental Research Funds for the Central Universities.

References

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

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

Chen, X.; Kang, B.; Geng, W.; Zhu, J.; Liu, Y.; Wang, D.; and Lu, H. 2025. SUTrack: Towards Simple and Unified Single Object Tracking. In *AAAI*.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. Seqtrack: Sequence to sequence learning for visual object tracking. In *CVPR*, 14572–14581.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8126–8135.

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 1–22.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Gao, S.; Zhou, C.; Ma, C.; Wang, X.; and Yuan, J. 2022. AiATrack: Attention in Attention for Transformer Visual Tracking. In *European Conference on Computer Vision*, 146–164. Springer.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

He, K.; Zhang, C.; Xie, S.; Li, Z.; and Wang, Z. 2023. Target-Aware Tracking with Long-term Context Attention. *arXiv preprint arXiv:2302.13840*.

Hong, L.; Yan, S.; Zhang, R.; Li, W.; Zhou, X.; Guo, P.; Jiang, K.; Chen, Y.; Li, J.; Chen, Z.; et al. 2024. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *CVPR*, 19079–19091.

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

Hu, X.; Tai, Y.; Zhao, X.; Zhao, C.; Zhang, Z.; Li, J.; Zhong, B.; and Yang, J. 2025. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3581–3589.

Hui, T.; Xun, Z.; Peng, F.; Huang, J.; Wei, X.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2023. Bridging Search Region Interaction With Template for RGB-T Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13630–13639.

Kang, B.; Chen, X.; Lai, S.; Liu, Y.; Liu, Y.; and Wang, D. 2025. Exploring Enhanced Contextual Information for Video-Level Object Tracking. In *AAAI*.

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

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019. RGB-T object tracking: Benchmark and baseline. *Pattern Recog.*, 96: 1856–1864.

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2021. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE TIP*, 31: 392–404.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1930–1939.

Mustafa, B.; Riquelme, C.; Puigcerver, J.; Jenatton, R.; and Houlsby, N. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35: 9564–9576.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Tan, Y.; Wu, Z.; Fu, Y.; Zhou, Z.; Sun, G.; Ma, C.; Paudel, D. P.; Van Gool, L.; and Timofte, R. 2025. XTrack: Multimodal Training Boosts RGB-X Video Object Trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tang, Z.; Xu, T.; Zhu, X.; Wu, X.-J.; and Kittler, J. 2024. Generative-based Fusion Mechanism for Multi-Modal Tracking.

Wang, N.; Zhou, W.; Wang, J.; and Li, H. 2021. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1571–1580.

Wang, S.; Cheng, G.; Lai, P.; Gao, D.; and Han, J. 2025. Multi-State Tracker: Enhancing Efficient Object Tracking via Multi-State Specialization and Interaction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 4087–4096.

Wang, S.; Wang, Z.; Sun, Q.; Cheng, G.; and Ning, J. 2024. Modelling of Multiple Spatial-Temporal Relations for Robust Visual Object Tracking. *IEEE Transactions on Image Processing*.

Wang, X.; Li, J.; Zhu, L.; Zhang, Z.; Chen, Z.; Li, X.; Wang, Y.; Tian, Y.; and Wu, F. 2023. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE TCYB*, 54(3): 1997–2010.

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

Yan, B.; Peng, H.; Fu, J.; Wang, D.; and Lu, H. 2021a. Learning spatio-temporal transformer for visual tracking. In *ICCV*, 10448–10457.

Yan, S.; Yang, J.; Käpylä, J.; Zheng, F.; Leonardis, A.; and Kämäräinen, J.-K. 2021b. Depthtrack: Unveiling the power of rgbd tracking. In *ICCV*, 10725–10733.

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

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 341–357.

Zheng, Y.; Zhong, B.; Liang, Q.; Mo, Z.; Zhang, S.; and Li, X. 2024. Odtrack: Online dense temporal token learning for visual tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 7588–7596.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023a. Visual prompt multi-modal tracking. In *CVPR*, 9516–9526.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024a. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.

Zhu, P.; Sun, Y.; Cao, B.; and Hu, Q. 2024b. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7099–7108.

Zhu, X.-F.; Xu, T.; Tang, Z.; Wu, Z.; Liu, H.; Yang, X.; Wu, X.-J.; and Kittler, J. 2023b. RGBD1K: A large-scale dataset and benchmark for RGB-D object tracking. In *AAAI*, volume 37, 3870–3878.