

XTrack: Multimodal Training Boosts RGB-X Video Object Trackers

Supplementary Material

In the supplementary material, we first provide more implementation details of our work in Sec. 1. We conduct extensive ablation studies to validate the effectiveness of key components, as detailed in Section 2. In Sec. 3, we provide a head-to-head comparison against SOTA methods across all datasets for ease of review. Qualitative comparison can be found in the **accompanying video**.

1. Implementation Details

In the main manuscript, we focus on detailing the Mixture of Modal Experts framework. For the implementation of XTrack, we adopt a pipeline consistent with existing Dual-Transformer models [1, 4]. Our model takes as input the search and template regions from both the RGB and sub-modality streams. The input sizes for the search and template regions are 256×256 and 128×128 , respectively.

Patch embedding: To achieve this, our XTrack starts by taking a pair of downstream modal features and a pair of RGB features as input, comprising template RGB and modal input $T_{rgb}^I, T_x^I \in \mathbb{R}^{3 \times H_t \times W_t}$ and search region input $S_{rgb}^I, S_x^I \in \mathbb{R}^{3 \times H_s \times W_s}$. These patches are initially segmented and subsequently flattened into sequences of patch vectors, specifically $T_{rgb}^p, T_x^p \in \mathbb{R}^{N_t \times (3 \cdot P^2)}$ and $S_{rgb}^p, S_x^p \in \mathbb{R}^{N_s \times (3 \cdot P^2)}$, where $P \times P$ defines the spatial extent of each individual patch. The quantities N_t and N_s , calculated as $N_t = \frac{H_t W_t}{P^2}$ and $N_s = \frac{H_s W_s}{P^2}$, respectively, signify the count of patches pertaining to the template and search regions from the modal input. Here, we set P as 16. Following this transformation, a learnable linear projection layer $proj$ is utilized to map the patch sequences into a D -dimensional latent space. These sequences are then concatenated, to meet the tracking specificity, and form T_{rgb} and T_x for subsequent processing and analysis:

$$T_{rgb} = \text{concat}(\text{proj}(T_{rgb}^p), \text{proj}(S_{rgb}^p)). \quad (1)$$

$$T_x = \text{concat}(\text{proj}(T_x^p), \text{proj}(S_x^p)). \quad (2)$$

After passing through patch embedding projection layer $proj$, we add T_{rgb} with position embedding.

MeME: After patch embedding, we employ MeME to facilitate multimodal interactions. As illustrated in the Fig. A, the process begins with a down projection that reduces the feature dimensionality. The down-projected features are then fed into a MoE module with shared experts. Here, the RGB features and the features from downstream modalities are integrated via a prompting mechanism. Finally, the

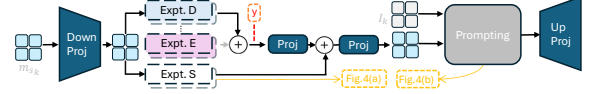


Figure A. The detail architecture of the MeME.

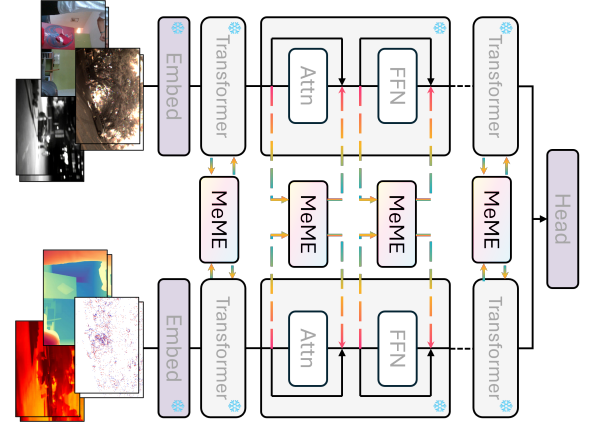


Figure B. The detail architecture of our XTrack.

Table A. The influence of Laplacian init.

Laplacian init	DepthTrack			LasHeR		VisEvent	
	F-score	Re	Pr	Pr	Sr	Pr	Sr
✓	61.2	61.5	61.4	68.8	54.8	77.0	60.1
	61.5	62.0	61.8	69.1	55.7	77.5	60.9

Table B. Multimodal Training Benefits.

E	D	T	VisEvent		DepthTrack			LasHeR	
			Pr	Sr	F-score	Re	Pr	Pr	Sr
			69.5	53.4	52.9	52.2	53.6	51.5	41.2
✓			76.3	59.9	45.5	45.8	45.7	58.1	45.1
	✓		65.8	48.3	60.1	61.0	60.6	50.2	39.3
		✓	67.0	48.4	48.8	45.2	46.9	68.4	55.0
✓	✓		76.6	60.2	60.8	59.5	60.1	58.1	45.2
✓		✓	76.6	60.3	48.4	48.1	48.3	68.5	55.0
	✓	✓	69.9	52.5	61.0	60.9	61.0	69.0	55.5
✓	✓	✓	77.5	60.9	61.5	62.0	61.8	69.1	55.7

fused features undergo an up projection to restore their dimensionality. Notably, all these operations are performed in a low-rank space, ensuring that the fusion process remains highly efficient.

Box Head: After passing through multiple layers of transformer blocks and MeME, we sum the final layer outputs, T_{rgb}^{l+1} and T_x^{l+1} , and feed the tokens into the box head. To ensure fair comparison with current multimodal trackers [2, 3], the box head we used is kept consistent with theirs [14].

Table C. State-of-the-art comparison on DepthTrack[12] for RGB-Depth, LasHeR[8] for RGB-Thermal, and VisEvent[10] for RGB-Event. The best two results are shown in **red** and **blue** fonts.

Method	DepthTrack [12]			VOT-RGBD2022 [5]			LasHeR [8]		RGBT234 [7]		VisEvent [10]	
	F-score(\uparrow)	Re(\uparrow)	Pr(\uparrow)	EAO(\uparrow)	Acc.(\uparrow)	Rob.(\uparrow)	Pr(\uparrow)	Sr(\uparrow)	MPR(\uparrow)	MSR(\uparrow)	Pr(\uparrow)	Sr(\uparrow)
XTrack-L	64.8	64.3	65.4	74.0	82.8	88.9	73.1	58.7	87.8	65.4	80.5	63.3
XTrack-B	61.8	62.0	61.5	74.0	82.1	88.8	69.1	55.7	87.4	64.9	77.5	60.9
Current Unified Models with Modality-Specific Branches												
UnTrack [11]	61.0	61.0	61.0	72.1	82.0	86.9	64.6	51.3	84.2	62.5	75.5	58.9
SDSTrack [3]	61.9	60.9	61.4	72.8	81.2	88.3	66.5	53.1	84.8	62.5	76.7	59.7
OneTracker [2]	60.9	60.4	60.7	72.7	81.9	87.2	67.2	53.8	85.7	64.2	76.7	60.8
ViPT [15]	59.4	59.6	59.2	72.1	81.5	87.1	65.1	52.5	83.5	61.7	75.8	59.2
ProTrack [13]	57.8	57.3	58.3	65.1	80.1	80.2	53.8	42.0	79.5	59.9	63.2	47.1

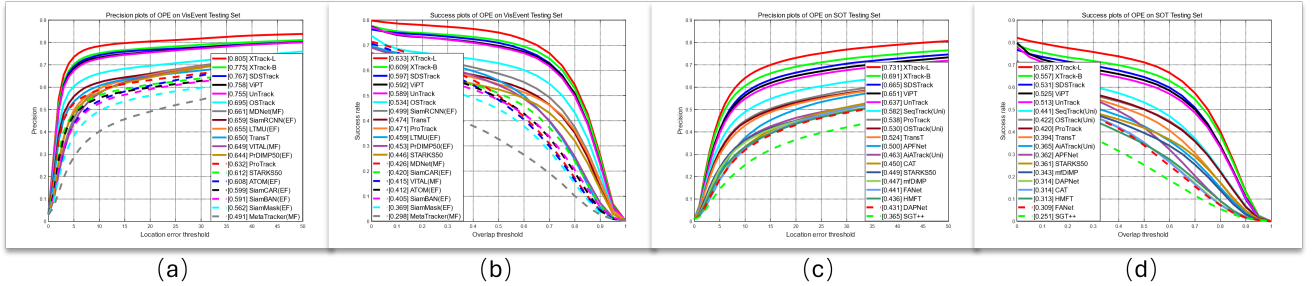


Figure C. (a) and (b) depict the precision and success plots for VisEvent[10]. (c) and (d) illustrate the precision and success plots for LasHeR[8].

2. Ablation Studies and Discussion

In this section, we present additional ablation studies on our network design. First, in Tab. A, we evaluate the effectiveness of Laplacian initialization for the shared expert. The results demonstrate that this initialization improves performance without incurring any additional computational cost. This finding supports our decision to use edge embedding as a common feature representation across all modalities.

Second, in Tab. B, we present a comprehensive study on the benefits of multimodal training by evaluating different combinations of training data. The results show that training with all modalities consistently improves performance compared to training with only two or single modalities.

It is worth noting that this paper primarily addresses the challenge of knowledge transfer within the multimodal training framework. Another factor that could influence model performance is data imbalance. For example, depth data differs in scale compared to thermal and event data. While addressing data imbalance is an intriguing avenue for future exploration, it lies beyond the scope of this work and remains a direction for our future research.

3. Comparison Against SOTA

Comparison with multimodal tracker: In the manuscript, we present the quantitative performance of our models across all tracking datasets, grouped by modality domains.

For clarity, we have consolidated these results to highlight our cross-domain advantages over SOTA unified models in Tab. C. Our models establish SOTA performance across all modality domains with a clear and significant margin. Specifically, XTrack-L achieves notable gains: +3% F-score on DepthTrack, +7% Precision on LasHeR, and +4% Precision on VisEvent.

Notably, our parameters are shared across all modalities, unlike existing unified models that include modality-specific parameters, which remain inactive if the corresponding modality is irrelevant. By training with multimodal data, even when only one modality is available at a time, our model effectively facilitates knowledge transfer, leveraging multimodal training to enhance the performance of each modality-specific tracker.

We can also observe that existing models often favor a single modality, achieving SOTA performance in one specific domain, such as SDSTrack on Depth, and OneTracker on LasHeR, etc. Our model, however, delivers more balanced performances, setting clear SOTA records across all modalities. In Fig. C we show the precision and success plot on Event and Thermal benchmarks. Our method outperforms existing counterparts by large margin.

More datasets: To validate the generalization capability of our model, we further incorporated two additional datasets: RGBT210 (an RGB-Thermal dataset) [6] and COESOT (an RGB-Event dataset) [9]. Notably, our model was not trained

Table D. Results on more datasets against modality-tailored SOTA models. RGBT210[6] for RGB-Thermal object tracking, and COESOT[9] for RGB-Event object tracking. We report Pr for these two datasets.

Tracker	ViPT	SDSTrack	Ours
RGBT210	80.0	81.2	85.5
COESOT	73.8	74.2	74.8

Table E. Complexity and Performance Metrics. Note that SOTA methods require sensor-specific param. sets*, where we achieve the unification. D./T./E. refer to DepthTrack, LasHeR and VisEvent dataset. We report F-score for D and Pr for T, E.

Tracker	Param.	FPS	D.	T.	E.
ViPT	93.0M*	25.2	59.4	65.1	75.8
SDSTrack	107.8M*	21.3	61.4	66.5	76.7
Ours	97.8M	22.1	61.5	69.1	77.5

on these datasets but only evaluated on them in a zero-shot cross-dataset testing manner. As demonstrated in the results, our tracker achieves competitive performance on these new datasets, exhibiting superior generalization capability compared to existing methods [3, 15].

4. Efficiency Analysis

We further compare the number of structural parameters and FPS of different models in Tab. E, and the experimental results show that our model has the best performance while maintaining the real-time inference speed. In addition, our model only needs to be trained once to achieve uniform target tracking in three different modal scenarios.

References

- [1] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bi-directional adapter for multimodal tracking. In *AAAI*, 2024. 1
- [2] Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, et al. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *CVPR*, 2024. 1, 2
- [3] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, et al. Sdsttrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *CVPR*, 2024. 1, 2, 3
- [4] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *CVPR*, 2023. 1
- [5] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, et al. The tenth visual object tracking vot2022 challenge results. In *ECCVW*, 2023. 2
- [6] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 2, 3
- [7] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *PR*, 96:106977, 2019. 2
- [8] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *TIP*, 31:392–404, 2021. 2
- [9] Chuanming Tang, Xiao Wang, Ju Huang, Bo Jiang, Lin Zhu, Jianlin Zhang, Yaowei Wang, and Yonghong Tian. Revisiting color-event based tracking: A unified network, dataset, and metric. *arXiv preprint arXiv:2211.11010*, 2022. 2, 3
- [10] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. VisEvent: Reliable object tracking via collaboration of frame and event flows. *TCYB*, pages 1–14, 2023. 2
- [11] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *CVPR*, 2024. 2
- [12] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of rgb-d tracking. In *ICCV*, 2021. 2
- [13] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *ACMMM*, 2022. 2
- [14] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 1
- [15] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *CVPR*, 2023. 2, 3