What You Have is What You Track: Adaptive and Robust Multimodal Tracking

Supplementary Material

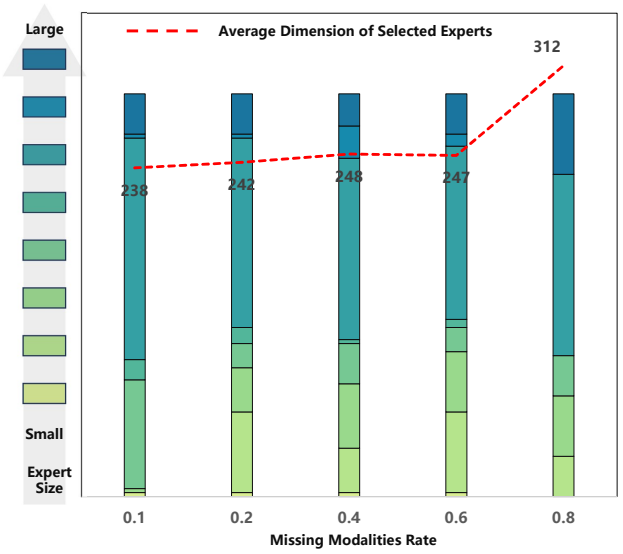

Figure A. Impact of missing modality rate on expert selection dynamics. The stacked bar graph illustrates the distribution of expert sizes selected by the HMoE across varying missing modality rates (x-axis). The red dashed line indicates the average dimension of the selected experts at each missing rate, showing an increasing trend as the missing rate rises.

1. More Details of the Encoder in FlexTrack

Inputs: For our video encoder, we select Fast-iTPN [5]. The video encoder takes two types of inputs: search region images from RGB, denoted as $S_{RGB} \in \mathbb{R}^{3\times H_x\times H_x}$, and corresponding modality data (Depth/Event/Thermal) represented as $S_X \in \mathbb{R}^{3\times H_x\times H_x}$. Additionally, the video clip consists of RGB video frames $V_{RGB} \in \mathbb{R}^{N\times H_z\times W_z}$ and X video frames $V_X \in \mathbb{R}^{N\times H_z\times W_z}$, where $N = 5$.

Embedding and Encoding: Initially, all images undergo downsampling through an embedding layer with a stride of 4. Following this step, the image tokens from RGB and X modalities are expanded and concatenated along the spatial dimension to form the following sets of tokens:

- RGB search region tokens, represented as T^s_{rgb}.
- Video clip tokens spanning from T^{c1}_{rgb} to T^{ci}_{rgb}.
- For the additional X modality, we obtain T^s_x, T^{c1}_x to T^{ci}_x, where $i \in \{1, \ldots, N\}$.

The encoder comprises 24 layers, differing from the standard Vision Transformer[1] in that not all layers pass through the same process. Notably, the resolution of all tokens is reduced during this process. This architecture enables the encoder to effectively capture both spatial and temporal features from the video clips, while also efficiently handling multiple modalities.

2. More Visualization and Analysis

To further investigate the selection patterns of different experts, we varied the masking ratio within a video to observe how the experts are chosen in Fig. A. As illustrated in the line graph, the average proportion of experts selected by the model changes under different missing modality rates. Moreover, we observed that as the missing ratio increases, the overall average size of the selected experts becomes larger. Notably, when the missing ratio increases from 0.6 to 0.8, the average hidden dimension of the selected experts experiences a substantial increase—from 247 to 312.

3. More Ablation Studies

We conduct systematic experiments to analyze the impact of expert quantity in our HMoE-Fuse architecture in Tab. A. While our main experiments employ 8 experts as the optimal configuration, we evaluate variations with 4 and 12 experts across multiple datasets to validate this design choice.

Table A. Performance comparison of different experts number.

	LasHeR	VisEvent	LasHeR$_{miss}$	VisEvent$_{miss}$
Ours	77.3	81.4	65.1	72.8
4	76.2	80.1	63.0	71.1
12	76.5	81.2	61.1	69.1

4. More Details of the Dataset

In this paper, we follow the dataset design in IPT[4]. But IPT on set the missing modality dataset in RGB-Thermal, including LasHeR [3] and RGBT234 [2]. We further keep the same design to RGB-Event with VisEvent [6] and RGB-Depth with DepthTrack [7]. As shown in Fig. B, the main missing types include long-time missing, switch missing, and random missing.

Table B. Statistics comparison among existing multimodal tracking datasets.

	RGBT234$_{miss}$	LasHeR$_{miss}$	VisEvent$_{miss}$	DepthTrack$_{miss}$
sequence number	234	245	320	50
Total Frames	116.7K	220.7K	157.8K	76.4K
Missing Frames	69.2K	133.2K	94.3K	44.4K

Figure B. Illustration of different missing patterns. A missing modality dataset mainly includes long-time missing, switch missing, and random missing as well as a combination of these three kinds of missing.

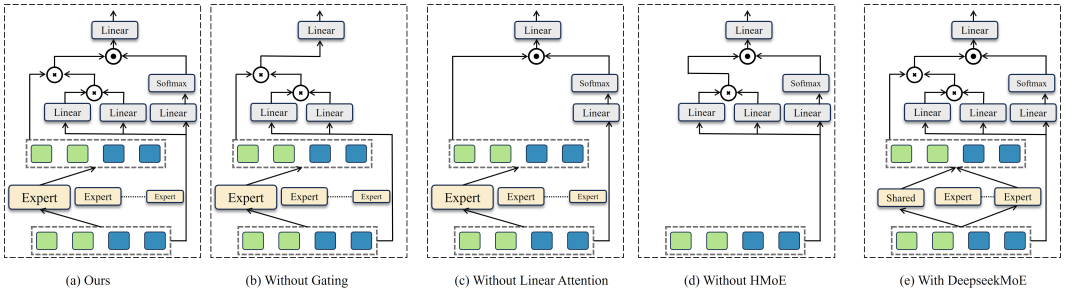

Figure C. Different Variant of Our HMoE-Fuse.

5. Different Variant on Ablation Study

To more accurately show the different MoE-Fuse variants, we have drawn the details in Fig. C.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *PR*, 96:106977, 2019. 1

[3] 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. 1

[4] Andong Lu, Chenglong Li, Jiacong Zhao, Jin Tang, and Bin Luo. Modality-missing rgbt tracking: Invertible prompt learning and high-quality benchmarks. *IJCV*, pages 1–21, 2024. 1

[5] Yunjie Tian, Lingxi Xie, Jihao Qiu, Jianbin Jiao, Yaowei Wang, Qi Tian, and Qixiang Ye. Fast-itpn: Integrally pre-trained transformer pyramid network with token migration. *TPAMI*, 2024. 1

[6] 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. 1

[7] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of rgbd tracking. In *ICCV*, 2021. 1