When Transformers Meet Mamba: A Hybrid Transformer-Mamba Network for Video Object Detection

Supplementary Material

This supplementary material offers additional technical details, experimental settings, ablation studies and visualization detection results that could not be included in the main manuscript due to space limitations. The structure of this supplementary material is outlined as follows. First, we present more details of the Mamba entangled transformer decoder in Section A. Then, we provide more experimental settings in Section B. Next, we carry out more ablation studies in Section C. Finally, we present qualitative visualization results on the EPIC-KITCHENS-55 dataset in Section D.

A. More Details of Mamba Entangled Transformer Decoder

Figure A illustrates the diagram of the proposed Mamba entangled transformer (MaET) decoder. It is composed of four types of layers, formed by adding a multi-scale adaptive deformable attention layer and a cascaded bidirectional Mamba layer to the components of the standard transformer decoder. Specifically, the object queries are first processed by a multi-head self-attention layer to enable communication and interaction among themselves, allowing each object query to gather contextual information from other counterparts. Subsequently, the outputs of the multi-head self-attention layer, together with the spatial-temporal encoder features produced by TCBM, are fed into a multi-scale adaptive deformable attention layer to enhance object queries by gathering contextual information from encoder features. The formulation of the multi-scale adaptive deformable attention (MADAttn) can be defined as follows:

$$\begin{aligned}&\text{MADAttn}(z_q, \hat{p}_q, \{x^l\}_{l=1}^{L}) \\ &= \sum_{h=1}^{H} \boldsymbol{W}_h \Bigg[\sum_{l=1}^{L} \sum_{k=1}^{K_q} \mathcal{O}_{hlqk} \cdot \boldsymbol{W}'_h x^l \big(\psi_l(\hat{p}_q) + \Delta p_{hlqk}\big) \Bigg],\end{aligned} \tag{1}$$

where q represents the query element, and $z_q \in \mathbb{R}^D$ denotes the query feature with the dimension of D. $\hat{p}_q \in [0,1]^2$ denotes the normalized coordinates of the reference point for the query element q, in which $(0,0)$ and $(1,1)$ indicate the top-left and the bottom-right corners of frame images, respectively. x^l denotes the l-th scale feature map. h, l and k index the attention head, feature scale and sampling point, respectively. $\mathcal{O}_{hlqk} \in [0,1]$ represents the attention weight of the k-th sampling point in the l-th scale and the h-th attention head, and it is normalized by $\sum_{l=1}^{L} \sum_{k=1}^{K_q} \mathcal{O}_{hlqk} = 1$. $\boldsymbol{W}_h \in \mathbb{R}^{D \times D_v}$ and $\boldsymbol{W}'_h \in \mathbb{R}^{D_v \times D}$ denote the learnable projection weights, in which $D_v = D/H$. $\psi_l(\hat{p}_q)$ re-scales the normalized coordinates $\hat{p}_q$ to the input feature map of the l-th scale. $\triangle p_{hlqk}$ indicates the sampling offset of the k-th sampling point in the l-th scale and the h-th attention head. Since $\psi_l(\hat{p}_q) + \Delta p_{hlqk}$ is fractional, we employ the bilinear interpolation [2] to compute $x^l(\psi_l(\hat{p}_q) + \Delta p_{hlqk})$. H and L respectively denote the total number of attention heads and feature scales. K_q is the total number of sampling points for the query element q, and its calculation can refer to Equation (2) in the main manuscript.

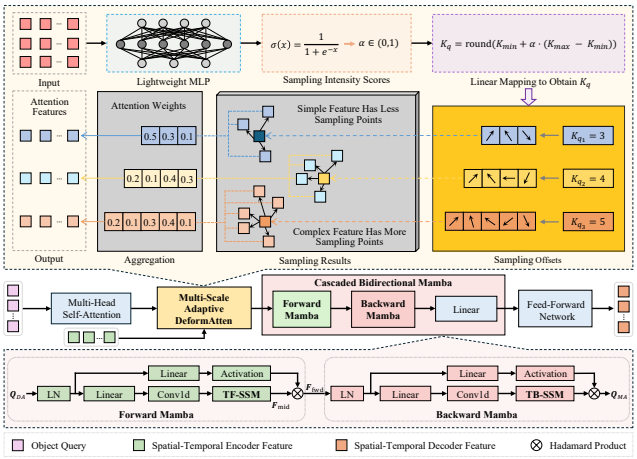

Figure A. Diagram of the proposed Mamba entangled transformer decoder. It is formed by adding a multi-scale adaptive deformable attention layer and a cascaded bidirectional Mamba layer to the components of the standard transformer decoder.

After that, a cascaded bidirectional Mamba layer is utilized to further refine the representations of object queries by integrating informative long-range contexts in both forward and backward directions. Specifically, the input features are first passed through a layer normalization (LN), after which they are split into two information flows. The first flow is processed by a linear projection (Linear) and a 1D convolution (Conv1d), followed by a temporal-prioritized forward state space model (TF-SSM) to capture forward long-range temporal dependencies. The second flow undergoes a linear projection and an activation function. Then, these two flows are combined through the Hadamard product to generate the forward feature representation $\boldsymbol{F}_{\text{fwd}}$:

$$\boldsymbol{F}_{\text{mid}} = \text{TF-SSM}(\text{Conv1d}(\text{Linear}(\text{LN}(\boldsymbol{\mathcal{Q}}_{DA})))), \tag{2}$$

$$\boldsymbol{F}_{\text{fwd}} = \boldsymbol{F}_{\text{mid}} \otimes \sigma(\text{Linear}(\text{LN}(\boldsymbol{\mathcal{Q}}_{DA}))), \tag{3}$$

where $\boldsymbol{\mathcal{Q}}_{DA}$ represents the output of MADAttn, $\otimes$ is the Hadamard product, and $\sigma(\cdot)$ is the activation function. Subsequently, we cascade the identical sequence of operations:

Table A. Analysis of the hyperparameters K_{min} and K_{max}. $*$ means the results obtained by Swin-B. Best results are marked in bold.

K_{min}	ImageNet VID		EPIC-KITCHENS-55	
	mAP (%)	mAP* (%)	mAP (%)	mAP* (%)
$K_{\text{min}} = 1$	87.0	91.4	44.5	50.3
$K_{\text{min}} = 2$	**87.9**	**92.1**	**45.1**	**50.8**
$K_{\text{min}} = 3$	87.5	91.6	44.8	50.4
K_{max}	ImageNet VID		EPIC-KITCHENS-55	
	mAP (%)	mAP* (%)	mAP (%)	mAP* (%)
$K_{\text{max}} = 7$	87.2	91.7	44.7	50.4
$K_{\text{max}} = 8$	**87.9**	**92.1**	**45.1**	**50.8**
$K_{\text{max}} = 9$	87.7	91.8	44.9	50.6

a layer normalization, a linear projection and a 1D convolution, followed by a temporal-prioritized backward SSM (TB-SSM) to capture backward long-range temporal dependencies in a reverse spatial order. The intuition behind TF-SSM and TB-SSM is that temporal continuity is more critical than spatial aggregation, and the bidirectional scan can capture richer temporal cues. Finally, the output $\mathcal{Q}_{MA}$ of the cascaded bidirectional Mamba is sent into a feed-forward network to yield spatial-temporal decoder features.

Since the two core components of the proposed MaET encoder, namely multi-scale adaptive deformable attention and cascaded bidirectional Mamba, share similar technical details with those described in Sections 3.2 and 3.3 in the main manuscript, we omit some redundant detailed descriptions in the main manuscript for the sake of brevity and to avoid unnecessary repetition.

B. More Experimental Settings

Datasets. Since the testing set of the ImageNet VID dataset [15] is not publicly available, we evaluate the performance of our TMambaDet on its validation set, following the widely adopted practice in existing video object detection methods [4, 8, 9, 16, 18–20]. We also adopt the EPIC-KITCHENS-55 dataset [3] as an additional benchmark for generalization evaluation. It consists of 55 hours of videos, documenting all daily activities of 32 participants in different kitchen scenarios. The EPIC-KITCHENS-55 training data for the object detection task contains 272 videos spanning 290 annotated object classes. Since the bounding box annotations for the EPIC-KITCHENS-55 testing set are not publicly available, we divide 272 training videos into 217 training and 55 validation videos, and conduct performance evaluation on the newly divided validation videos.

Implementation Details. We implement our TMambaDet on top of Deformable DETR [21], using ResNet-101 [10] and Swin transformer [11] as the backbones, which are initialized with pre-trained weights from the ImageNet dataset [5]. As a common protocol in [6, 7, 14, 17], we enlarge the feature resolution by adjusting the total stride of the Conv5 block in ResNet-101 from 32 to 16, and keep the receptive field size by setting the dilation of 3×3 convolutions in the Conv5 block to 2. The input frames are resized to hold a shorter size of 600 pixels for fair comparisons, following existing video object detection methods [1, 7, 12, 13, 20].

Table B. Analysis of the number of object queries for each frame on the ImageNet VID dataset.

# Queries	mAP (%)	Time (ms)	mAP* (%)	Time* (ms)
40	84.2	**18.6**	88.7	**36.7**
50	86.7	19.7	91.1	37.9
60	**87.9**	20.6	**92.1**	39.0
70	87.8	21.6	91.9	39.8
80	87.4	22.5	91.4	40.9

C. More Ablation Studies

In this section, we use ResNet-101 to conduct ablation studies on the ImageNet VID dataset, unless otherwise stated.

Analysis of the Hyperparameters K_{min} and K_{max}. We carry out extensive ablation studies to analyze the influence of the hyperparameters K_{min} and K_{max}, as illustrated in Table A. For K_{min}, the best performance is achieved when the value of K_{min} is set to 2, yielding 87.9% mAP and 92.1% mAP on the ImageNet VID dataset with ResNet-101 and Swin-B, respectively. However, increasing or decreasing K_{min} beyond this value slightly decreases performance. For K_{max}, the best performance is achieved when the value of K_{max} is set to 8, yielding 45.1% mAP and 50.8% mAP on the EPIC-KITCHENS-55 dataset with ResNet-101 and Swin-B, respectively. However, both smaller and larger values of K_{max} lead to minor performance drops.

Analysis of the Number of Object Queries. We carry out ablation experiments to comprehensively analyze and investigate the influence of the number of object queries for each frame on detection accuracy and runtime. As illustrated in Table B, we vary the number of object queries from 40 to 80, and the results indicate that the mAP improves consistently when the number of object queries increases and it tends to be optimal when using 60 object queries. However, when further enlarging the number of object queries from 60 to 80, the mAP is decreased by 0.5% (from 87.9% to 87.4%) and 0.7% (from 92.1% to 91.4%) with ResNet-101 and Swin-B, respectively. The reason is that excessive object queries possibly introduce redundant or low-confidence predictions, increasing the risk of false positives and duplicate detections, thus adversely affecting the final accuracy.

D. More Visualization Results

We visualize several detection results of our TMambaDet on the EPIC-KITCHENS-55 dataset in Figure B, where each detected object is marked with a colored bounding box and annotated with its class label and confidence score. From the results, we can observe that our TMambaDet is able to achieve reliable detection in more complex video scenarios, exhibiting accurate localization results and cor-

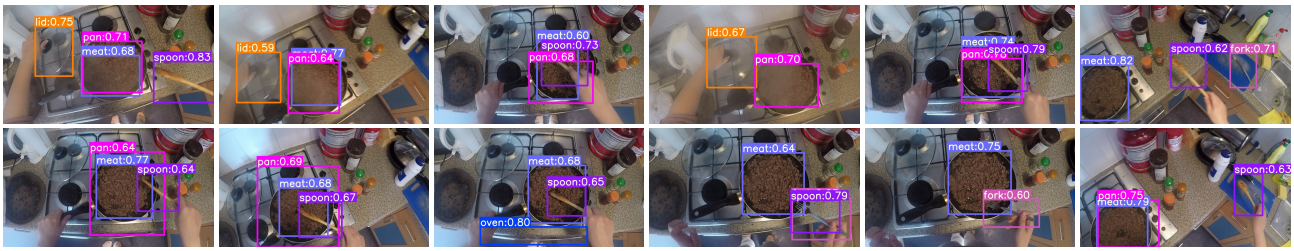

Figure B. Visualization detection results produced by our TMambaDet on the EPIC-KITCHENS-55 dataset.

rect classification results, which is consistent with the visualization results presented in the main manuscript, indicating the effectiveness and generalization of our TMambaDet.

References

[1] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10337–10346, 2020. 2

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 764–773, 2017. 1

[3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 720–736, 2018. 2

[4] Chaorui Deng, Da Chen, and Qi Wu. Identity-consistent aggregation for video object detection. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 13434–13444, 2023. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 248–255, 2009. 2

[6] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. Class-aware feature aggregation network for video object detection. *IEEE Trans. Circuits Syst. Video Technol. (TCSVT)*, 32(12):8165–8178, 2022. 2

[7] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 431–446, 2020. 2

[8] Khurram Azeem Hashmi, Talha Uddin Sheikh, Didier Stricker, and Muhammad Zeshan Afzal. Beyond boxes: Mask-guided spatio-temporal feature aggregation for video object detection. In *Proc. IEEE Win. Conf. App. Comput. Vis. (WACV)*, pages 8111–8122, 2025. 2

[9] Fei He, Naiyu Gao, Jian Jia, Xin Zhao, and Kaiqi Huang. Queryprop: Object query propagation for high-performance video object detection. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, pages 2620–2627, 2022. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, 2016. 2

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 10012–10022, 2021. 2

[12] Qiang Qi, Tianxiang Hou, Yang Lu, Yan Yan, and Hanzi Wang. Dgrnet: A dual-level graph relation network for video object detection. *IEEE Trans. Image Process. (TIP)*, 32: 4128–4141, 2023. 2

[13] Qiang Qi, Yan Yan, and Hanzi Wang. Class-aware dual-supervised aggregation network for video object detection. *IEEE Trans. Multimedia (TMM)*, 26:2109–2123, 2024. 2

[14] Qiang Qi, Hanzi Wang, Yan Yan, and Xuelong Li. Dgcnet: Dynamic graph contrastive network for video object detection. *IEEE Trans. Image Process. (TIP)*, 34:2269–2284, 2025. 2

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)*, 115:211–252, 2015. 2

[16] Yuheng Shi, Tong Zhang, and Xiaojie Guo. Practical video object detection via feature selection and aggregation. *Int. J. Comput. Vis. (IJCV)*, 134(95), 2026. 2

[17] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. Mamba: Multi-level aggregation via memory bank for video object detection. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, pages 2620–2627, 2021. 2

[18] Han Wang, Jun Tang, Xiaodong Liu, Shanyan Guan, Rong Xie, and Li Song. Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 732–747, 2022. 2

[19] Chao Xu, Jiangning Zhang, Mengmeng Wang, Guanzhong Tian, and Yong Liu. Multi-level spatial-temporal feature aggregation for video object detection. *IEEE Trans. Circuits Syst. Video Technol. (TCSVT)*, 32(11):7809–7820, 2022.

[20] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: End-to-end video object detection with spatial-temporal transformers. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 45(6):7853–7869, 2023. 2

[21] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, pages 1–16, 2021. 2