Hierarchical Spatiotemporal Transformers for Video Object Segmentation
-Supplementary Material-

In this *supplementary material*, we provide detailed explanations for the models used in the ablation studies and additional results for visual comparison.

1. Details on Ablation Studies

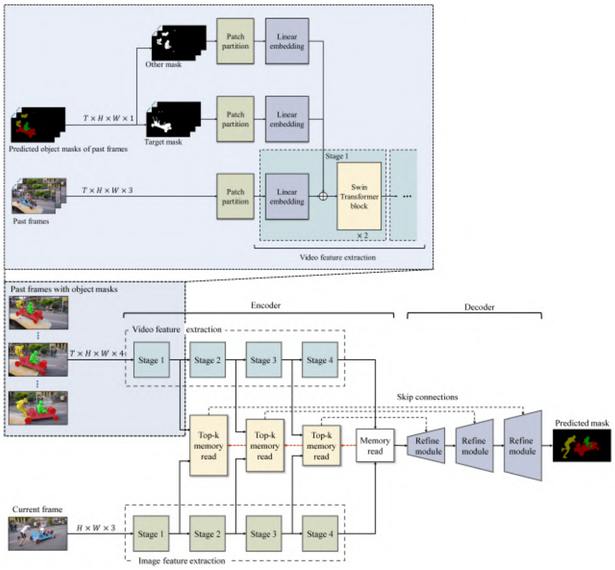

Figure S-1: Detailed illustration of HST. Note that our video Transformer takes past frames and their predicted masks as input, where the predicted masks are further divided into the target object masks and other masks.

Mask utilization. The video Transformer of HST takes the predicted mask of the other objects, called 'other mask,' as an additional input as shown in Figure S-1, following the common strategy [1, 2, 3, 4]. Note that the other mask is a single-channel binary map per frame (1: belongs to any other objects, 0: otherwise). Both the predicted mask of the target object and the other mask are fused into the image frame in the embedding space by passing through their respective linear embedding layers. This strategy can help HST find the target object outside the other objects as shown in Figures S-3 and S-4, leading to 1.8 % $\mathcal{J}\&\mathcal{F}$ improvement (See Table 4.4).

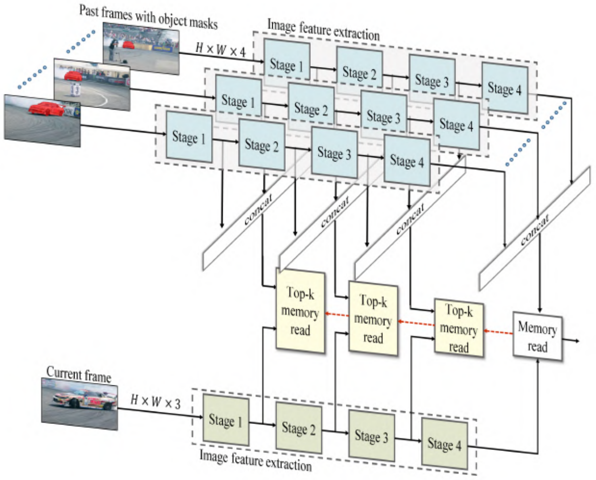

Figure S-2: An encoder architecture embodied with image Transformer only. Note that the encoder performs the memory read operations using only image features extracted from current and past frames.

Video Transformer. To demonstrate the effectiveness of using both image and video Transformers for spatiotemporal feature extraction, we built a model consisting of image Transformers only, as shown in Figure S-2. Note that the video feature $F^i_{video} \in \mathbb{R}^{T_i \times H_i \times W_i \times C_i}$ at the i-th stage is constructed by concatenating the spatial features $F^i_{image} \in \mathbb{R}^{H_i \times W_i \times C_i}$ obtained from the past T_i frames and their predicted object masks. In addition to the quantitative performance improvement that we demonstrated (See Table 4.5), we provide some results for visual comparison. As can be seen in Figures S-3 and S-4, both image and video Transformers play an essential role in VOS, especially for these difficult scenes containing occluded and moving objects.

Hierarchical memory read. In our implementation, HST processes multiple object segmentation in parallel, which is memory demanding but effective in reducing the processing time. However, in order to experiment with the model performing naive dense matching at all scales, we should process object segmentation sequentially to avoid memory

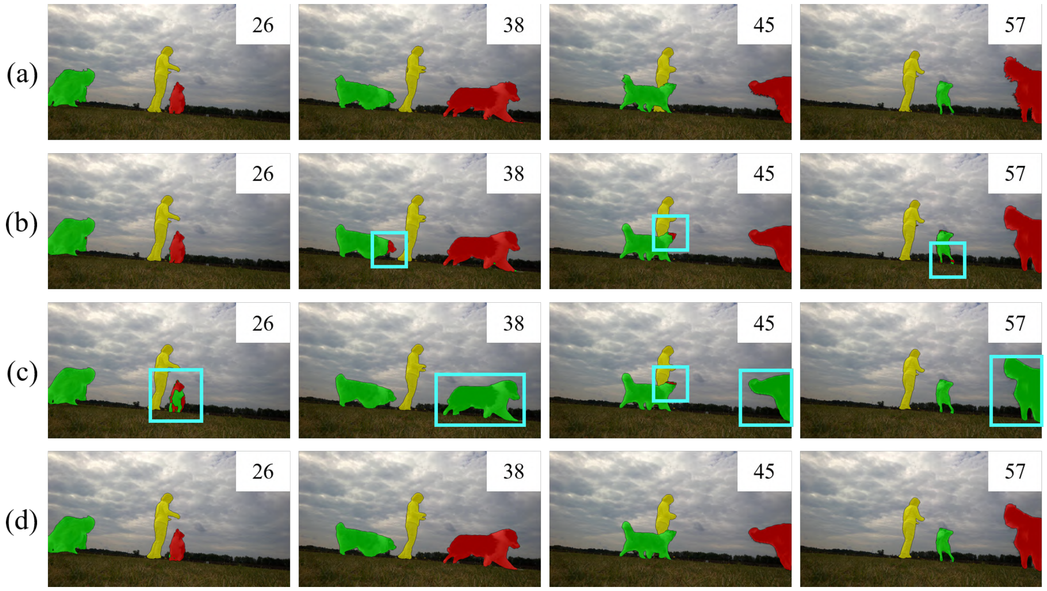

Figure S-3: Qualitative comparison of ablation studies: (a) Ground-truth, (b) result obtained w/o the other mask, (c) result obtained using image Transformer only, and (d) result of HST-B.

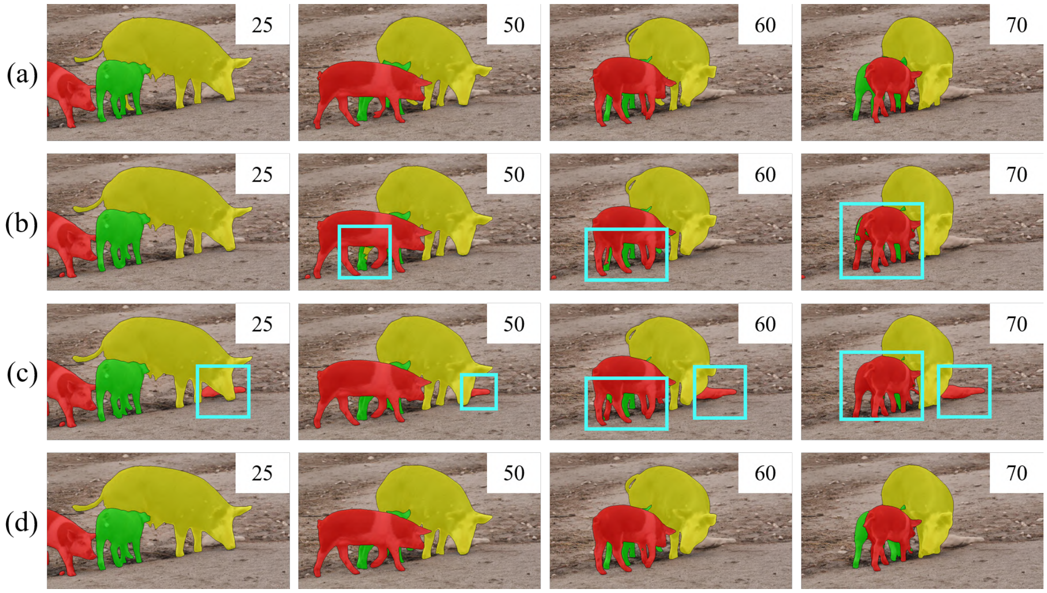

Figure S-4: Qualitative comparison of ablation studies: (a) Ground-truth, (b) result obtained w/o the other mask, (c) result obtained using image Transformer only, and (d) result of HST-B.

overflow, leading to the average 2.78 s processing time per frame (See Table 4.3).

2. More Qualitative Results

In Figures S-5 and S-6, we provide more results for visual comparison with the state-of-the-art methods: HMMN [2], STCN [4], and AOT [5].

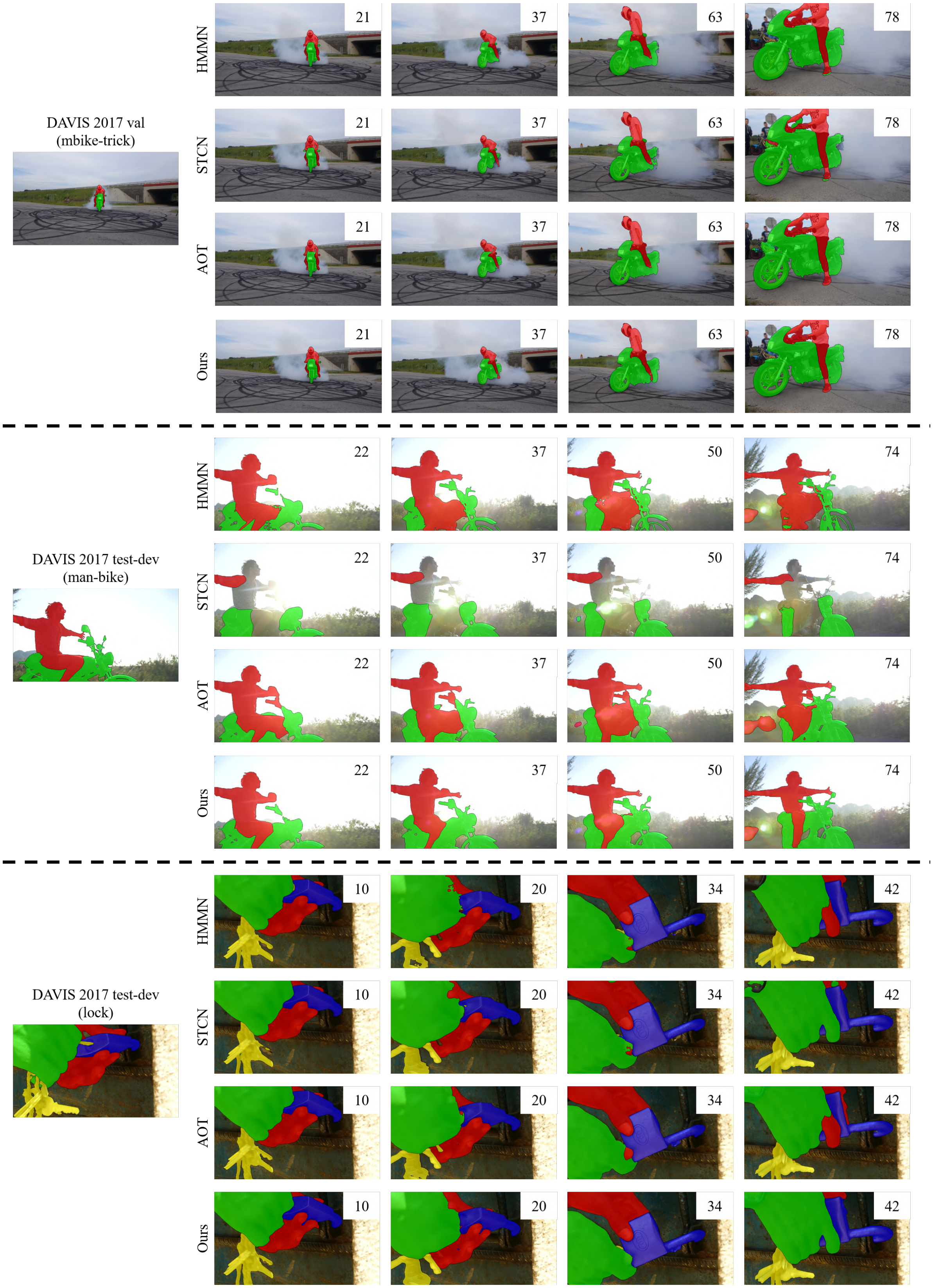

Figure S-5: Qualitative performance comparison of HST with HMMN [2], STCN [4], and AOT [5].

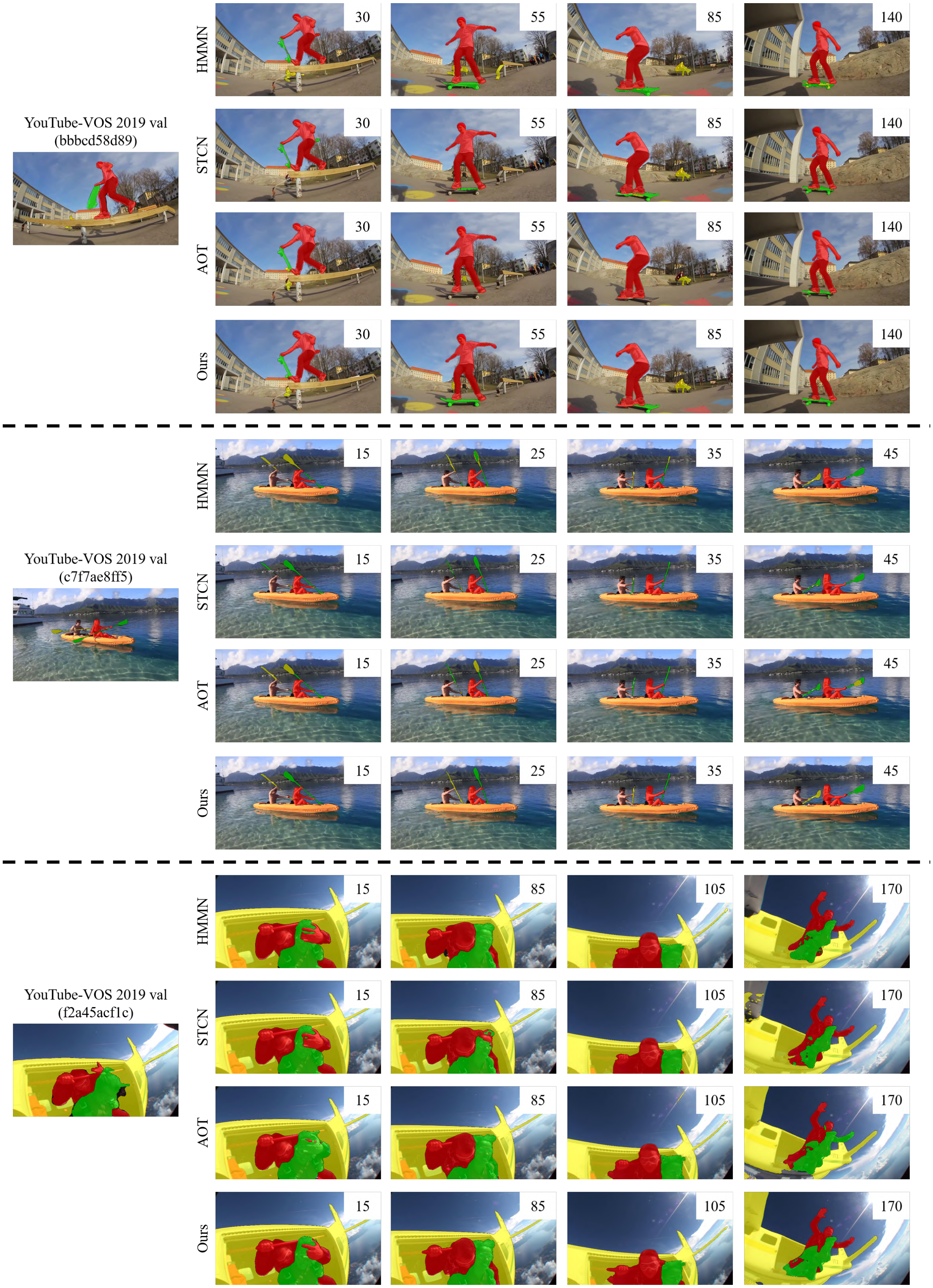

Figure S-6: Qualitative performance comparison of HST with HMMN [2], STCN [4], and AOT [5].

References

[1] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9226–9235. 1

[2] H. Seong, S. W. Oh, J.-Y. Lee, S. Lee, S. Lee, and E. Kim, "Hierarchical memory matching network for video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 889–12 898. 1, 2, 3, 4

[3] H. Seong, J. Hyun, and E. Kim, "Kernelized memory network for video object segmentation," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 629–645. 1

[4] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," in *Proceedings of the Advances in Neural Information Processing Systems*, 2021, pp. 11 781–11 794. 1, 2, 3, 4

[5] Z. Yang, Y. Wei, and Y. Yang, "Associating objects with transformers for video object segmentation," in *Proceedings of the Advances in Neural Information Processing Systems*, 2021, pp. 2491–2502. 2, 3, 4