SOC: Semantic-Assisted Object Cluster for Referring Video Object Segmentation Supplementary Material

Anonymous Author(s)
Affiliation
Address
`email`

1 Dataset Details

The A2D-Sentences dataset contains 3782 videos and each video has 3-5 annotated segmentation masks and JHMDB-Sentences totally comprises 928 videos, each of which is associated with a text description. For the large-scale dataset, Ref-YouTube-VOS has 3978 videos with about 15K text descriptions. The Ref-DAVIS17 contains 90 videos with 1,544 expressions, including 60 and 30 videos for training and validation respectively.

2 Performance on JHMDB-Sentences

We also compare our SOC with existing methods on JHMDB-Sentences [2] and the results are shown in Table 1. Following ReferFormer [8], we directly report the results utilizing the models trained on A2D-Sentences without finetune. It can be seen that our method achieves new state-of-the-art performance with different backbone and training settings. Compared to other benchmarks, the performance gains on this dataset are relatively small. This can be attributed to JHMDB's imprecise annotations generated by coarse human puppet model.

Method	Backbone	Precision					mAP	IoU	
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9		Overall	Mean
Hu *et al.* [3]	VGG-16	63.3	35.0	8.5	0.2	0.0	17.8	54.6	52.8
Gavrilyuk *et al.* [2]	I3D	69.9	46.0	17.3	1.4	0.0	23.3	54.1	54.2
CMSA + CFSA [9]	ResNet-101	76.4	62.5	38.9	9.0	0.1	-	62.8	58.1
ACAN [7]	I3D	75.6	56.4	28.7	3.4	0.0	28.9	57.6	58.4
CMPC-V [5]	I3D	81.3	65.7	37.1	7.0	0.0	34.2	61.6	61.7
ClawCraneNet [4]	ResNet-50/101	88.0	79.6	56.6	14.7	0.2	-	64.4	65.6
MTTR [1]	Video-Swin-T	93.9	85.2	61.6	16.6	0.1	39.2	70.1	69.8
ReferFormer [8]	Video-Swin-T	93.3	84.2	61.4	16.4	0.3	39.1	70.0	69.3
SOC(Ours)	Video-Swin-T	**94.7**	**86.4**	**62.7**	**17.9**	0.1	**39.7**	**70.7**	**70.1**
With Image Pretrain									
ReferFormer	Video-Swin-T	95.8	89.3	66.8	18.9	0.2	42.2	71.9	71.0
ReferFormer	Video-Swin-B	96.2	90.2	70.2	21.0	0.3	43.7	73.0	71.8
SOC(Ours)	Video-Swin-T	**96.3**	**88.7**	**67.2**	**19.6**	0.1	**42.7**	**72.7**	**71.6**
SOC(Ours)	Video-Swin-B	**96.9**	**91.4**	**71.1**	**21.3**	0.1	**44.6**	**73.6**	**72.3**

Table 1: Comparison with the state-of-the-art methods on JHMDB-Sentences.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

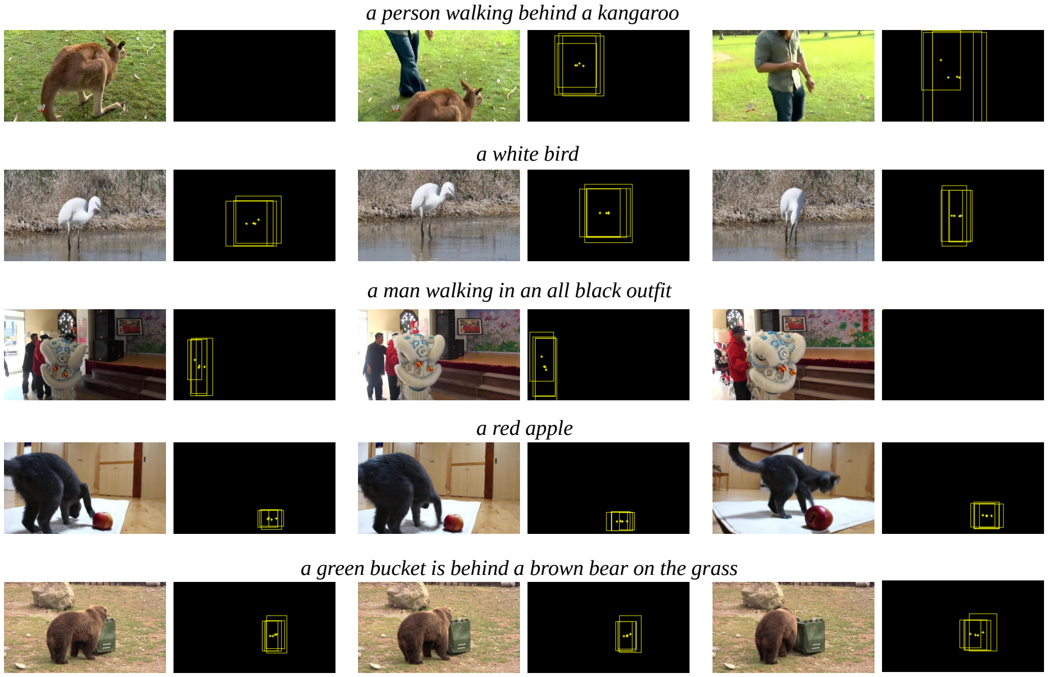

a person walking behind a kangaroo
a white bird
a man walking in an all black outfit
a red apple
a green bucket is behind a brown bear on the grass

Figure 1: Visualization of the frame-level object query

3 More Implementation Details

Training Settings Our models are trained with the AdamW optimizer using Pytorch. The weight decay is 1×10^{-4}. The batch size is set to 56 for pretraining and 8 for main training. The models are trained for 30 epochs. The initial learning rate is set to 1×10^{-4} for Ref-YouTube-VOS and RefCOCO/+/g, 5×10^{-5} for A2D-Sentences. The learning rate decays by 10 for the backbone network. During training, we apply RandomResize and Horizontal Flip for data augmentation. Specifically, all frames are downsampled to 360×640 for Ref-YouTube-VOS and RefCOCO/+/g, 320×576 for A2D-Sentences.

Inference Settings During inference, the input videos are downsampled to 320×576 for A2D-Sentences dataset and 360p for other datasets. We directly output the segmentation masks without any post-process.

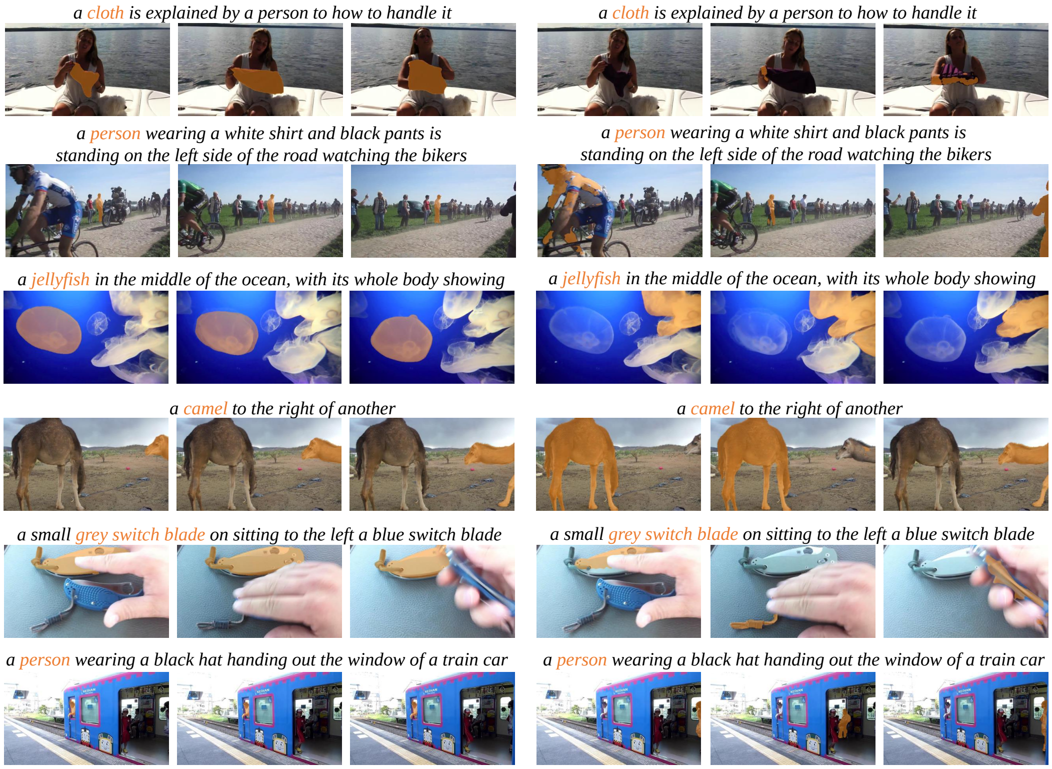
4 Additional Qualitative Results

4.1 Query Visualization

To demonstrate that the frame-level query embeddings can represent the referred object in a specific frame, we visualize the predicted bounding boxes corresponding to the query embeddings. As illustrated in Fig. 1, the majority of queries focus on regions of the referred object as expected. This indicates that the compact frame-level query embeddings indeed reflect object information and subsequent video-level object cluster is performing temporal interaction for referred objects.

4.2 Segmentation Stability Visualization

The benchmark performance and IoU variance analysis in the main paper have proven the effectiveness and stability of our method. Here we incorporate visual comparisons to further validate the segmentation stability of our model. In Fig. 2 (a), benefiting from the global object view, SOC is capable of tracking the referred object across frames in coherence. On the contrary, ReferFormer [8], the existing state-of-the-art method, may generate segmentation masks with high degree of variance, indicating that the frame-based paradigm fails to accurately understand the state of the object in the context of the entire video (see in Fig. 2 (b)).


a cloth is explained by a person to how to handle it
a cloth is explained by a person to how to handle it
a person wearing a white shirt and black pants is standing on the left side of the road watching the bikers
a person wearing a white shirt and black pants is standing on the left side of the road watching the bikers
a jellyfish in the middle of the ocean, with its whole body showing
a jellyfish in the middle of the ocean, with its whole body showing
a camel to the right of another
a camel to the right of another
a small grey switch blade on sitting to the left a blue switch blade
a small grey switch blade on sitting to the left a blue switch blade
a person wearing a black hat handing out the window of a train car
a person wearing a black hat handing out the window of a train car

(a) **(b)**

Figure 2: Visualization comparisons of segmentation stability between our SOC and existing state-of-the-art method ReferFormer [8]. (a) and (b) denote our SOC and ReferFormer, respectively.

4.3 Adaptability for Texts Describing Temporal Variation

Figure 6 in the main paper has shown some results to demonstrate that our SOC can better handle descriptions that focus on temporal variation. Here we provide more cases to demonstrate the adaptability of our method to such text descriptions. Fig. 4 and Fig. 5 show the segmentation results of our SOC and ReferFormer, where (a) indicates the segmentation results by SOC and (b) represents the results by ReferFormer[8].

5 Comprehensive Evaluation

We comprehensively measure our method by different perspectives, *e.g.*, performance, inference speed and computation cost under fair comparison. It is noted that the horizontal axis of Fig. 3 denotes performance on Ref-YouTube-VOS, vertical axis is FPS and the radius of the circle represents the relative FLOPs. Compared with ReferFormer [8] (blue ●), Our method (red ●) achieves superior performance with faster inference speed and less computation cost. Although MTTR [1] (orange ●) [1] has the lowest FLOPs, the lack of elaborate multi-modal fusion and temporal interaction significantly degrade the segmentation accuracy. In contrast, our method leverages video-level multi-modal understanding, which brings a significant increase in performance with only minimal computational costs.

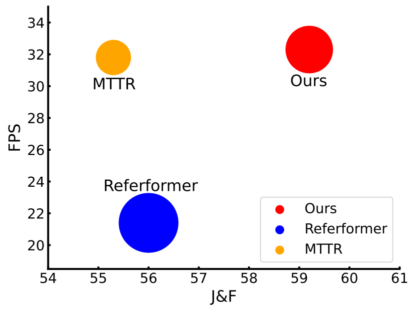


Figure 3: Performance *vs* Inference Speed *vs* Computation Cost.

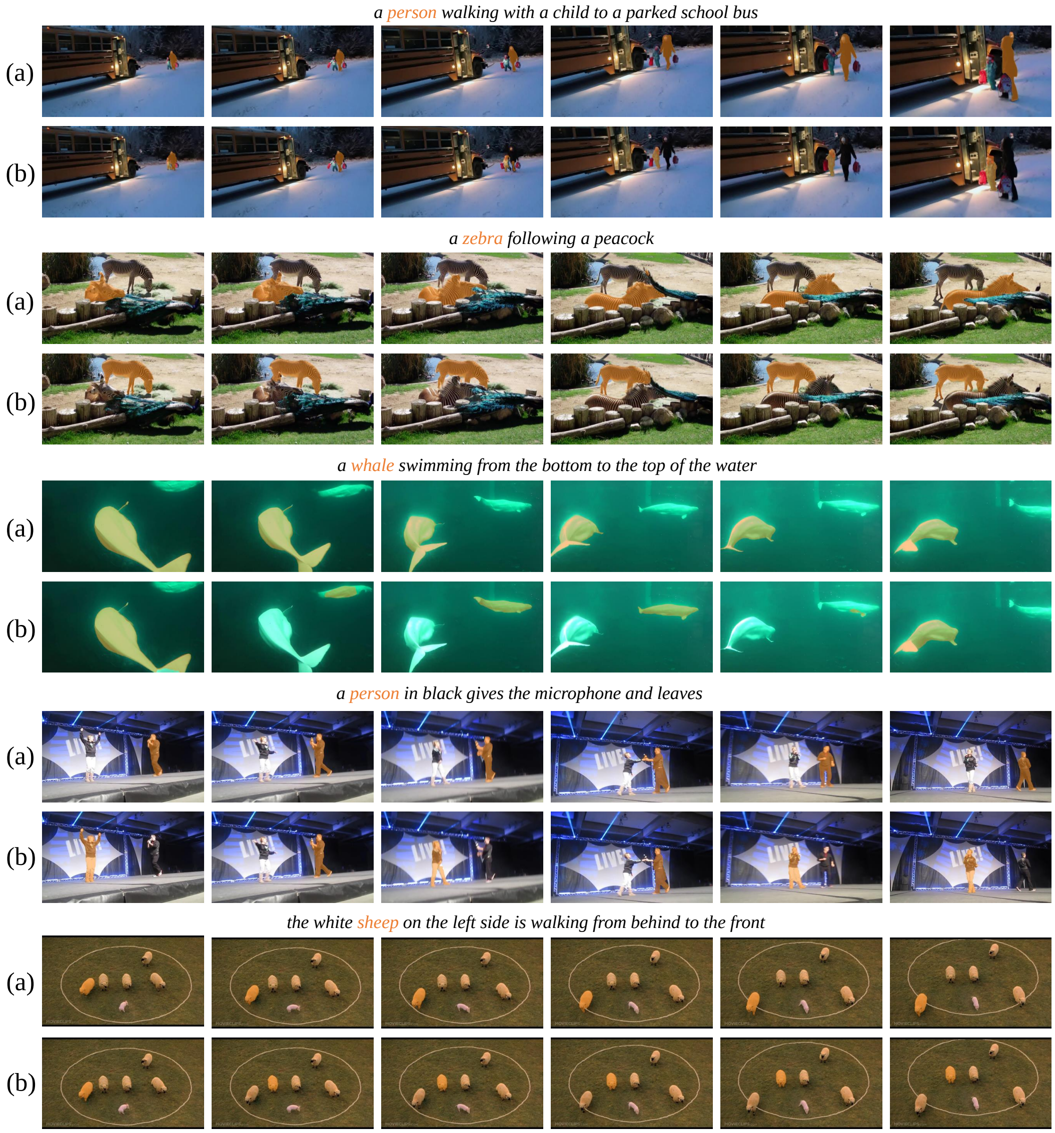

a person walking with a child to a parked school bus
(a)
(b)
a zebra following a peacock
(a)
(b)
a whale swimming from the bottom to the top of the water
(a)
(b)
a person in black gives the microphone and leaves
(a)
(b)
the white sheep on the left side is walking from behind to the front
(a)
(b)

Figure 4: Visualization comparison using text expressions about temporal variation. (a) and (b) are segmentation results of our SOC and ReferFormer [8], respectively.

6 Error Bar

We have retrained our model several times on Ref-YouTube-VOS [6] dataset. The results demonstrate that the randomness of the model has little effect on the performance, *i.e.*, the max deviation is about 0.5% $\mathcal{J}\&\mathcal{F}$.

7 Broader Impact

Malicious use of the RVOS model may lead to potential negative societal impacts, including but not limited to unauthorized surveillance or privacy-infringing tracking. However, we firmly believe that the task itself is neutral with positive implications, such as video editing and human-robot interaction.

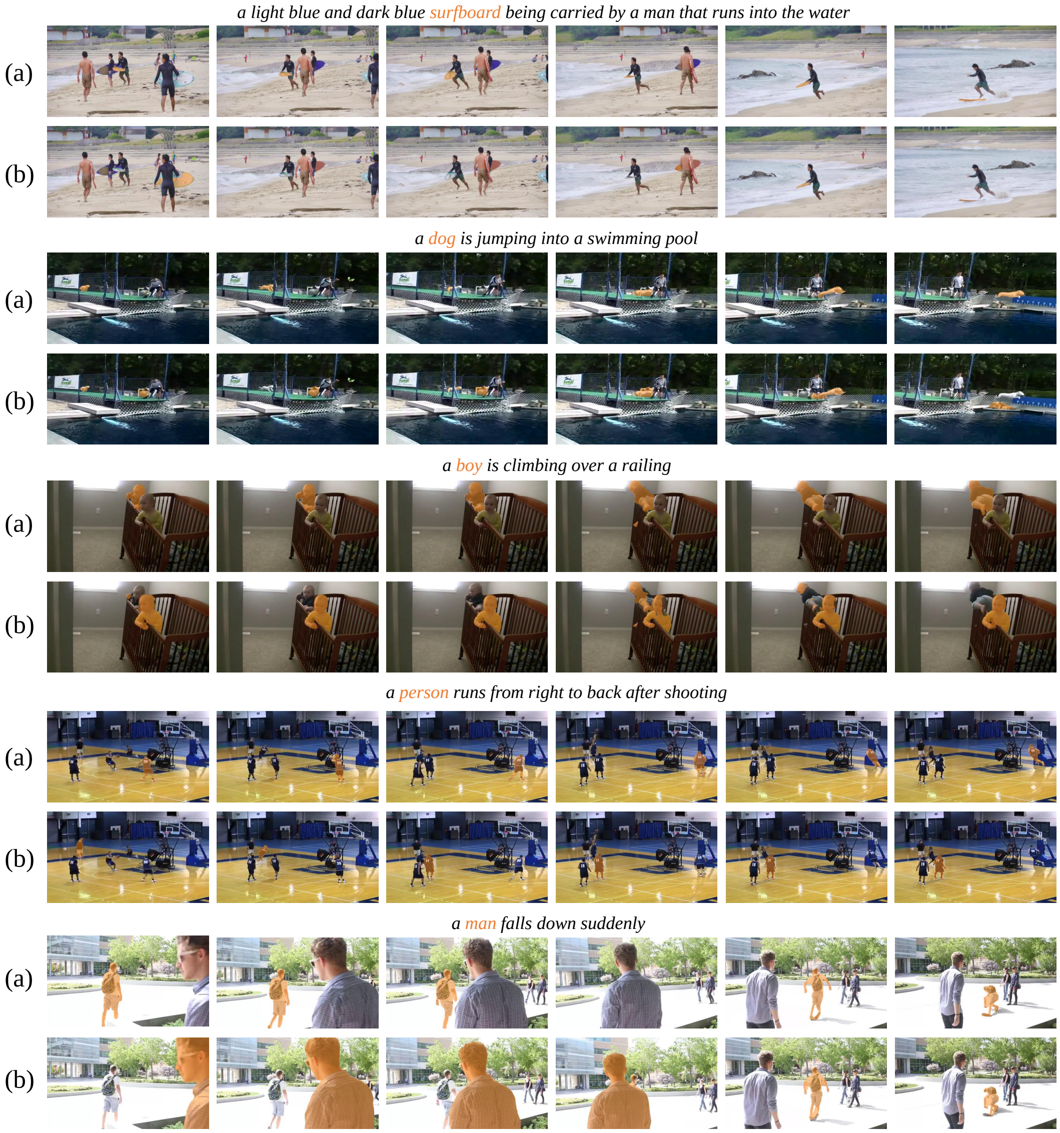

a light blue and dark blue surfboard being carried by a man that runs into the water
(a)
(b)
a dog is jumping into a swimming pool
(a)
(b)
a boy is climbing over a railing
(a)
(b)
a person runs from right to back after shooting
(a)
(b)
a man falls down suddenly
(a)
(b)

Figure 5: Visualization comparison using text expressions about temporal variation. (a) and (b) are segmentation results of our SOC and ReferFormer [8], respectively.

References

[1] Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: CVPR. pp. 4975–4985 (2022) 1, 3

[2] Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.M.: Actor and action video segmentation from a sentence. In: CVPR. pp. 5958–5966 (2018) 1

[3] Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV. pp. 108–124 (2016) 1

[4] Liang, C., Wu, Y., Luo, Y., Yang, Y.: Clawcranenet: Leveraging object-level relation for text-based video segmentation. arXiv preprint arXiv:2103.10702 (2021) 1

[5] Liu, S., Hui, T., Huang, S., Wei, Y., Li, B., Li, G.: Cross-modal progressive comprehension for referring segmentation. TPAMI pp. 4761–4775 (2022) 1

[6] Seo, S., Lee, J., Han, B.: URVOS: unified referring video object segmentation network with a large-scale benchmark. In: ECCV. pp. 208–223 (2020) 4

[7] Wang, H., Deng, C., Yan, J., Tao, D.: Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In: ICCV. pp. 3938–3947 (2019) 1

[8] Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR. pp. 4964–4974 (2022) 1, 2, 3, 4, 5

[9] Ye, L., Rochan, M., Liu, Z., Zhang, X., Wang, Y.: Referring segmentation in images and videos with cross-modal self-attention network. TPAMI pp. 3719–3732 (2022) 1