Segment Every Reference Object in Spatial and Temporal Spaces (Supplementary Material)

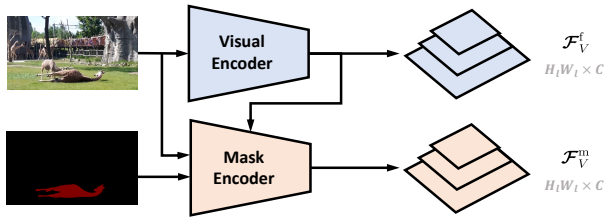
Jiannan Wu[1], Yi Jiang[2], Bin Yan[3], Huchuan Lu[3], Zehuan Yuan[2], Ping Luo[1,4]

[1]The University of Hong Kong [2]ByteDance

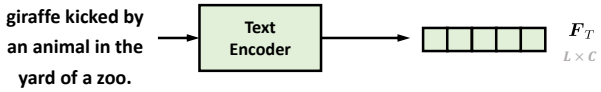
[3]Dalian University of Technology [4]Shanghai AI Laboratory

Appendix A. Architecture

Reference Encoding Figure 1 illustrates the process of reference encoding. (i) For mask references, we employ the same visual encoder $\mathbf{Enc}_V$ for both the current and reference frames to generate multi-scale features (*i.e.*, C3, C4, C5). We denote the encoded features of the reference frame as $\boldsymbol{\mathcal{F}}_{\boldsymbol{V}}^{\mathbf{f}}$, where the ℓ-th feature ($\ell = 2, 3, 4$) has a size of $H_\ell \times W_\ell \times C$, with a spatial stride of $2^{\ell+1}$ relative to the original size. Next, we use a lightweight mask encoder (ResNet-18 in all our experiments) that takes the reference frame and annotated mask as inputs. We concatenate the last three layer features with the corresponding level features in $\boldsymbol{\mathcal{F}}_{\boldsymbol{V}}^{\mathbf{f}}$ and further process them with two ResBlocks [2] and a CBAM block [10] to obtain the final outputs, denoted as $\boldsymbol{\mathcal{F}}_{\boldsymbol{V}}^{\mathbf{m}}$. Finally, we flatten each level feature in $\boldsymbol{\mathcal{F}}_{\boldsymbol{V}}^{\mathbf{f}}$ and $\boldsymbol{\mathcal{F}}_{\boldsymbol{V}}^{\mathbf{m}}$ into 1-dimensional vectors. (ii) For language references, we directly use off-the-shelf text encoder RoBERTa [6] to extract the 1-d linguistic features.



(a) Mask Reference



(b) Language Reference

Figure 1: The process of reference encoding for (a) mask references and (b) language references.

Appendix B. Implementation Details

Training Details. Our training process consists of three sequential stages: VG pretraining, image-level training and video-level training. We train models on NVIDIA A100 GPUs and it takes 2-3 days (depends on the visual backbone) to complete the whole training. The text encoder is unfrozen during the first two stages and then frozen for the final stage. The detailed configurations are summarized in Table 1. We follow the implementation of Detic [13] for the multi-dataset training. The learning rate is reduced by the factor of 10 when the iteration reaches the specified step in the table. Data augmentation includes random horizontal flip and scale jitter for resizing the input images. In the table, short side means the range of values for the shortest side and long side represents the maximum value for the longest side. During video-level training, for COCO [5] and RefCOCO/+/g [12, 7], we apply two different augmentations on the same image to generate the pseudo videos for training. And for OVIS [8], we convert the dataset into a class-agnostic format to make it suitable for VOS training.

Inference Details. For both RVOS and VOS tasks, all the videos are rescaled to 480p for inference. And the score thresholds are set as 0.4 for VOS datasets and 0.3 for RVOS datasets, respectively. For these two tasks, both the masks in the first frame and previous frame are adopted as references.

Appendix C. More Results

How to Use Both Language and Mask Reference for RVOS? As shown in Figure 2, we design three strategies to utilize both language and mask references for RVOS: sequential M->L, sequential L->M and parallel. The ablation results are presented in Table 2, where it is evident that the sequential strategies yield poor performance, while the parallel strategy emerges as the most effective way to integrate both mask and language references. This finding is reasonable since the parallel strategy is a post-fusion process so it does not impact the visual features of the current frame.

Reference Frames for Mask Propagation. In this study, we investigate the effect of reference frames for mask prop-

Table 1: The detailed configurations for the three training stages.

Stage	Task	Dataset	Sampling Weight	Batch Size	Short Side	Long Side	GPU Number	Learning Rate	Weight Decay	Max Iteration	Step
I	RIS	Visual Genome [4]	1	2	480 ∼ 800	1333	32	0.0002	0.0001	90000	80000
II	RIS	RefCOCO/+/g [12, 7]	1	2	480 ∼ 800	1333	16	0.0001	0.0001	90000	80000
III	VOS	COCO [5]	0.40	2	480 ∼ 800	1333	16	0.0001	0.05	90000	80000
		Youtube-VOS2019 [11]	0.30	2	320 ∼ 640	768					
		LVOS [3]	0.20	2	320 ∼ 640	768					
		OVIS [8]	0.10	2	320 ∼ 640	768					
	RVOS	RefCOCO/g/+ [12, 7]	0.45	2	480 ∼ 800	1333					
		Ref-Youtube-VOS [9]	0.55	2	320 ∼ 640	768					

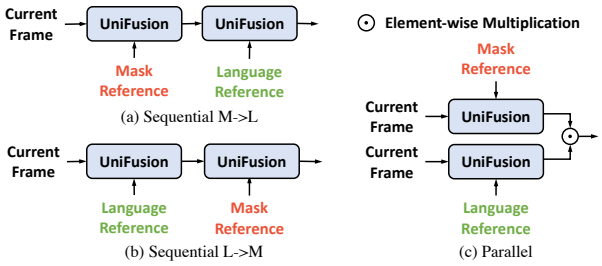


Figure 2: Three strategies for utilizing both language and mask references for RVOS. (a) Sequential M->L: the visual features of current frame are fused with mask and language references sequentially. (b) Sequential L->M: the visual features of current frame are fused with language and mask references sequentially. (c) Parallel (used in the paper): the visual features of current frame are fused with mask reference and language reference, respectively. The two fused features are multiplied in the end.

Table 2: Ablation on the strategies for utilizing both language and mask references for RVOS. Experiments are conducted on Ref-Youtube-VOS. Our default settings are marked in gray.

Variants	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Sequential M->L	13.8	14.6	13.0
Sequential L->M	9.5	9.6	9.5
Parallel	60.1	58.9	61.4

agation, as presented in Table 3. Specifically, we analyze the impact of discarding the first frame and the previous frame on performance for two datasets, namely Youtube-VOS2018 and Ref-Youtube-VOS. These two datasets are evaluated for VOS and RVOS tasks, respectively. On Youtube-VOS2018, the first frame provides a reliable annotated mask, while the previous frame has the highest similarity with the current frame. Therefore, discarding either of these frames would result in a significant drop in performance. on Ref-Youtube-VOS, no provided mask is available, and thus the performance drop is less noticeable. Nevertheless, our findings support the conclusion that utilizing both the first frame and the previous frame as references yields the best results for mask propagation.

Table 3: Ablation on the reference frames used for mask propagation during inference. We use the final model with ResNet-50 visual backbone in this ablation. Our default settings are marked in gray.

First	Previous	Youtube-VOS2018					Ref-Youtube-VOS		
		$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✓		75.2	76.7	80.4	69.0	74.9	60.1	58.6	61.7
	✓	79.2	79.8	83.9	72.8	80.3	59.6	58.1	61.2
✓	✓	81.4	81.6	85.9	75.6	82.4	60.6	59.0	62.3

Efficiency Comparison with Memory-based Methods. We compare the efficiency of our UniRef and the representative memory-based method STCN [1] in Table 4. The results indicate that while our method is slightly slower than STCN on the Youtube-VOS dataset, it is much more efficient than STCN on the long-term video LVOS dataset. This is because the memory-based methods have linear memory complexity with respect to the video duration, while our method has constant memory cost. To better highlight the advantages of our method, we further plot the single-object FPS in Figure 3.

Appendix D. Visualization Results

We provide the visualization results of UniRef-L for RVOS tasks in Figure 4 and Figure 5. It can be seen that our model can segment the referred objects correctly and accurately in various challenging scenes, *e.g.*, partial display, similar objects and fast moving, as illustrated in Figure 4.

Visualization results for the VOS tasks are presented in Figure 6 and Figure 7. Notably, our model reveals strong ability in handling long-term videos that typically last for

Table 4: Efficiency comparison between our method and the representative memory-based method STCN. 'YT-VOS18' represents Youtube-VOS2018 dataset.

Dataset	Mean Frames	Mean Objects	FPS	
			STCN	UniRef
YT-VOS18	27	1.9	13.2	10.5
LVOS	574	1.3	4.8	19.3

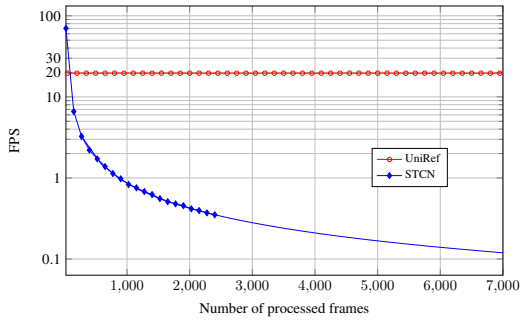


Figure 3: FPS scaling of our method and the representative memory-based method STCN.

over a minute, such as those in LVOS [3]. As shown in Figure 7, our model can accurately segment the target objects throughout the whole video, despite the objects have significant pose variation. We further provide a video demo in the supplementary material.

References

[1] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. *arXiv preprint arXiv:2211.10181*, 2022. 2, 3

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[7] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1, 2

[8] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8):2022–2039, 2022. 1, 2

[9] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 2

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

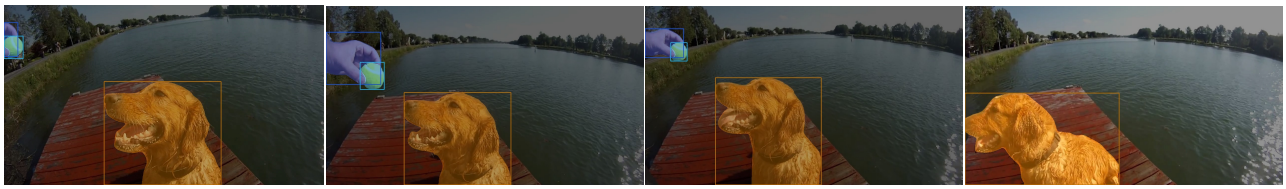
[11] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2

[12] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 1, 2

[13] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022. 1

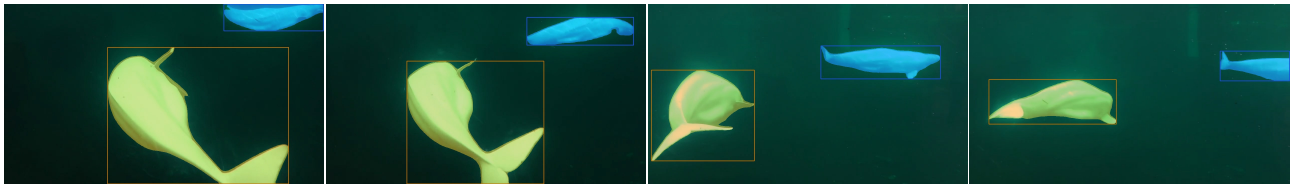
a dog is waiting to catch the ball shown to him.

a hand is showing a ball to a dog. a lawn tennis ball in the hand of a person.



a whale swimming from the bottom to the top of the water.

a whale on the top right swimming underwater.



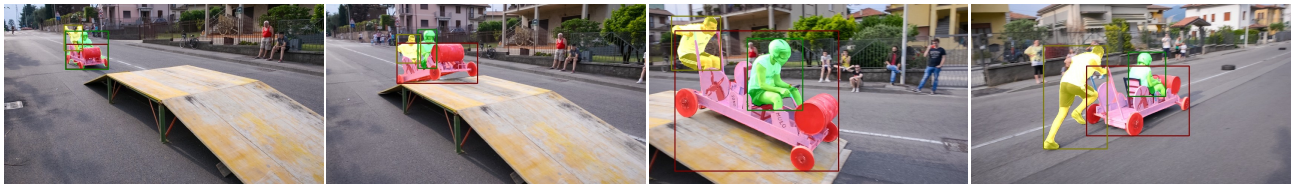
a skateboard being rolled through a road filled with cars and people.

a boy in black shorts and white tee shirt roller skating.

Figure 4: Visualization results on Ref-Youtube-VOS validation set.

a go-cart type car. a person driving the go cart.

person at the back of the go-cart without a helmet.



a man wearing a green helmet. a motor-bike.

a blonde haired girl dancing in a blue dress.

Figure 5: Visualization results on Ref-DAVIS17 validation set.

Figure 6: Visualization results on Youtube-VOS2018 validation set.

Figure 7: Visualization results on LVOS validation set.