SVAC: Scaling Is All You Need For Referring Video Object Segmentation

Li Zhang[1], Haoxiang Gao[2], Zhihao Zhang[1], Luoxiao Huang[3], Tao Zhang[4]
[1]Columbia University, [2]Carnegie Mellon University, [3]New York University, [4]Wuhan University

Problem

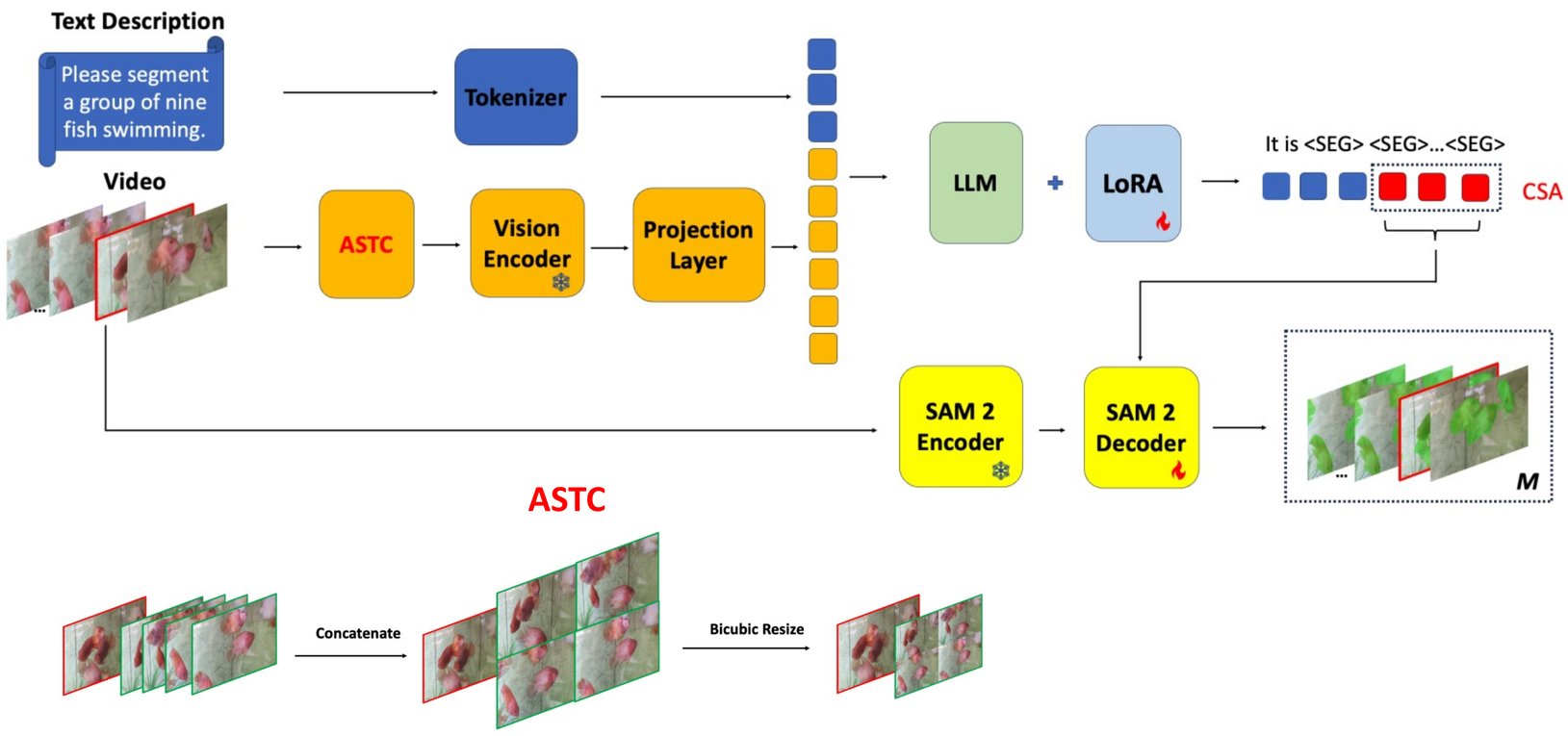
Referring Video Object Segmentation (RVOS) faces lots of challenges:

- Insufficient exploitation of MLLMs' prior knowledge
- Prohibitive computational cost and memory usage for high-resolution, long-duration videos
- Inadequate handling of complex temporal dynamics

Solution

- Scale up video frames input to MLLM with Anchor-Based Spatio-Temporal Compression (**ASTC**)
- Scale up segmentation tokens with Clip-Specific Allocation(**CSA**)

Overall Structure



Experiments

Method	Year	MeVIS$_{val_u}$ [40]			Ref-DAVIS17 [17]			ReVOS [42]		
		J	*F*	*J&F*	*J*	*F*	*J&F*	*J*	*F*	*J&F*
URVOS [32]	ECCV 2020	-	-	-	47.3	56.0	51.6	-	-	-
MTTR [6]	CVPR 2022	-	-	-	-	-	-	25.1	25.9	25.5
ReferFormer [41]	CVPR 2022	-	-	-	58.1	64.1	61.1	26.2	29.9	28.1
LISA-13B [13]	CVPR 2024	-	-	-	63.2	68.8	66.0	39.8	43.5	41.6
TrackGPT-13B [33]	ArXiv2024	-	-	-	62.7	70.4	66.5	43.2	46.8	45.0
VISA-13B [44]	ECCV 2024	-	-	-	67.0	73.8	70.4	48.8	52.9	50.9
VideoLISA [3]	NeurIPS 2024	50.9	58.1	54.5	72.7	64.9	68.8	-	-	-
VideoGLaMM [27]	CVPR 2025	-	-	-	73.3	65.6	69.5	-	-	-
GLUS [24]	CVPR 2025	-	-	61.6	-	-	-	47.5	52.4	49.9
ViLLa [53]	ArXiV 2025	-	-	-	70.6	78.0	74.3	54.9	59.1	57.0
Sa2VA-1B [49]	ArXiV 2025	-	-	53.4	-	-	69.5	-	-	47.6
Sa2VA-4B [49]	ArXiV 2025	-	-	55.9	-	-	73.7	-	-	53.2
Sa2VA-8B [49]	ArXiV 2025	-	-	58.9	-	-	75.9	-	-	57.6
Sa2VA-26B [49]	ArXiV 2025	-	-	61.8	-	-	78.6	-	-	58.4
SVAC-1B (Ours)	BMVC 2025	55.8	64.6	60.2	67.5	76.1	71.8	48.1	50.7	49.4
SVAC-4B (Ours)	BMVC 2025	56.9	65.6	61.3	71.4	80.0	75.7	55.1	56.1	55.6
SVAC-8B (Ours)	BMVC 2025	**58.2**	**67.3**	**62.8**	**74.9**	**83.8**	**79.4**	**58.2**	**60.2**	**59.2**

With only 8B parameters, SVAC achieves SOTA J&F scores of 62.8, 79.4, and 59.2 on MeVIS(val_u), Ref-DAVIS17, and ReVOS datasets respectively, surpassing the previous SOTA model Sa2VA-26B by margins of 1.0, 0.8, and 0.8 points.

Conclusion

In this work, we propose SVAC, a novel framework for Referring Video Object Segmentation (RVOS) that effectively scales up both the number of processed video frames and segmentation tokens. We introduce Anchor-Based Spatio-Temporal Compression (ASTC), which processes more frames to better leverage MLLM's potential by maintaining an anchor frame and a composite compressed representation of other frames for each clip, and Clip-Specific Allocation (CSA), which increases the number of segmentation tokens through clip-wise allocation to better capture motion patterns. Extensive experiments across multiple VOS benchmarks demonstrate that SVAC achieves state-of-the-art performance while maintaining computational efficiency. Our work provides a promising direction for advancing text-guided video understanding and segmentation tasks.