MPG-SAM 2: Adapting SAM 2 with Mask Priors and Global Context for Referring Video Object Segmentation

Supplementary Material

A. Additional Experimental Studies

Mask-text Similarity Loss. In this part, We validate the generalizability of the mask-text similarity loss function by conducting enhancement experiments with this function on several previous RVOS methods, including ReferFormer [7] and SgMg [2]. Meanwhile, we also validate the impact of this function on the performance of MPG-SAM 2. The experimental results, presented in Tab. 1, indicate performance improvements across all methods, confirming the effectiveness of the proposed similarity function in RVOS tasks.

Model Parameters. In this section, we analyze the parameter count of our model. We supplement our study with a set of low-configuration experiments on the Ref-YouTube-VOS [4] dataset, using SAM 2-Hiera-Large [3] and BEiT-3-Base [5] as initialization parameters, called MPG-SAM 2-Tiny. The experimental results and parameter counts are presented in Tab. 2. Compared to previous methods with relatively small parameter sizes, such as ReferFormer [7] and SgMg [2], our low-configuration model MPG-SAM 2-Tiny exhibits a slightly larger parameter count but achieves a substantial performance gain. Furthermore, although methods like VISA [8] and HyperSeg [6] also employ vision-language models and are trained on additional datasets, in contrast, our full-configuration model MPG-SAM 2, which is not trained on any additional datasets, demonstrates superior performance with a smaller model size. This highlights the effectiveness and efficiency of our approach.

Hierarchical Global-Historical Aggregator. In this division, we perform a more detailed component analysis to evaluate the effectiveness of each part within the hierarchical global-historical aggregator (HGA) on the Ref-YouTube-VOS [4] dataset. The experimental results are presented in Tab. 3. The initial setup involves the MPG-SAM 2 only using the mask prior generator, which achieves 71.9% $\mathcal{J}\&\mathcal{F}$. When incorporating the global enhancement and local enhancement of HGA's object-level fusion part separately, the model reaches $\mathcal{J}\&\mathcal{F}$ scores of 72.2% and 72.4%, respectively, resulting in gains of 0.3% and 0.5% $\mathcal{J}\&\mathcal{F}$. When applying both components of the object-level fusion part to the initial model simultaneously, the model attains 72.8% $\mathcal{J}\&\mathcal{F}$, indicating a 0.9% $\mathcal{J}\&\mathcal{F}$ score improvement compared to the initial setup. Moreover, the inclusion of HGA's pixel-level global fusion part on the initial setup obtains a $\mathcal{J}\&\mathcal{F}$ score of 73.1%. Finally, when all parts of HGA are employed, the full model gains the highest performance of 73.9% $\mathcal{J}\&\mathcal{F}$.

Table 1. Generalizability of the mask-text similarity loss.

Method	Backbone	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [7]	ResNet-50	55.6	54.8	56.5
ReferFormer [7] + $\mathcal{L}_{sim}$	ResNet-50	**56.4**	**55.4**	**57.5**
SgMg [2]	Video-Swin-T	62.0	60.4	63.5
SgMg [2] + $\mathcal{L}_{sim}$	Video-Swin-T	**62.8**	**61.3**	**64.3**
MPG-SAM 2 - $\mathcal{L}_{sim}$	Hiera-L	73.2	71.2	75.2
MPG-SAM 2	Hiera-L	**73.9**	**71.7**	**76.1**

Table 2. Model parameter analysis on Ref-YouTube-VOS dataset. Our model strikes a balance between the number of parameters and performance, demonstrating clear advantages. The best results are high-lighted in bold, and the second best results are underlined.

Method	Reference	All Params	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [7]	CVPR'22	0.24B	62.9	61.3	64.6
SgMg [2]	ICCV'23	0.24B	65.7	63.9	67.4
VISA [8]	ECCV'24	13B	63.0	61.4	64.7
HyperSeg [6]	Arxiv'24	3B	68.5	-	-
MPG-SAM 2-Tiny	-	0.46B	<u>69.9</u>	<u>68.0</u>	<u>71.8</u>
MPG-SAM 2	-	0.92B	**73.9**	**71.7**	**76.1**

We also analyze the effect of the number of the pixel-level fusion layer N_p and object-level fusion layer N_o on the model's performance. For N_p and N_o, we design experiments with 1, 2, and 3 layers, and the results are shown in Tab. 4. The results show that single-layer fusion modules are sufficient to effectively enhance global and historical information at both the pixel level and object level. However, increasing the number of layers introduces redundant information, which may impair the segmentation process for the current frame. As a result, we set both the N_p and N_o to 1 to ensure optimal model performance.

B. Additional Visualization Results

In this section, we present additional visualization results of our MPG-SAM 2 on Ref-YouTube-VOS [4] dataset and MeViS [1] dataset. The visualizations, shown in Fig. 1, highlight target objects covered by blue masks.

For the MeViS [1] dataset, we evaluate the model's segmentation performance in challenging scenarios involving object occlusion and multiple referential targets. As shown in Fig. 1 (a), the target object, a tiger, is initially occluded by an enclosure in the first three frames and becomes vis-

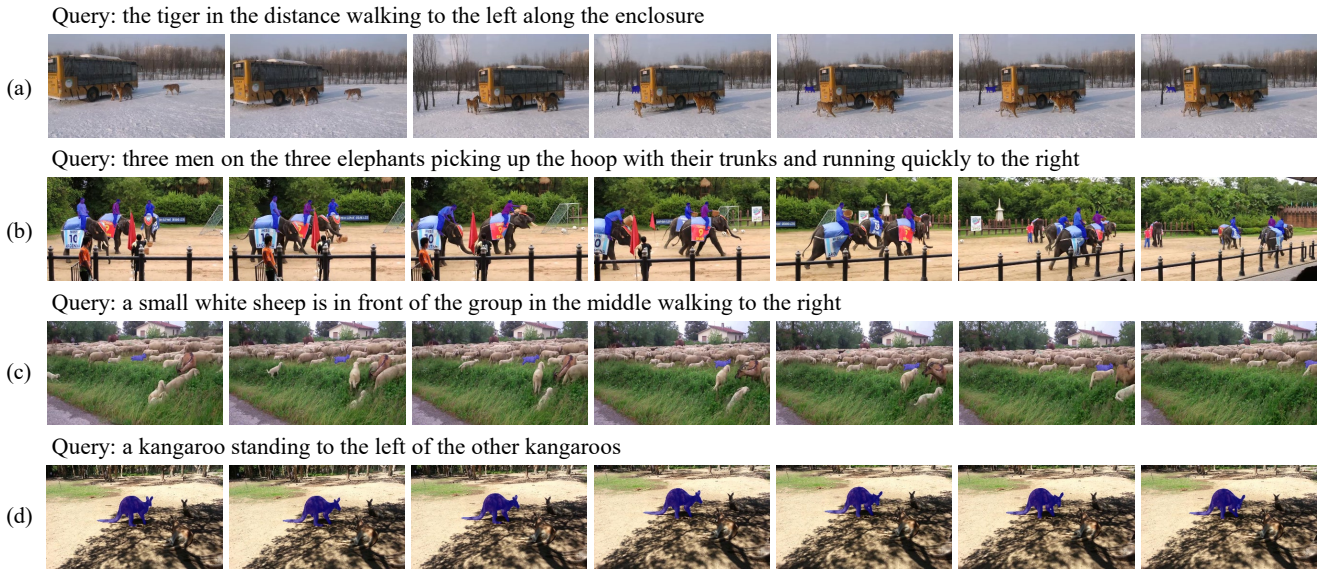

Figure 1. Additional visualization results on several datasets. (a), (b) MeViS, (c), (d) Ref-YouTube-VOS.

Table 3. The ablation experiments of HGA components, where OGF represents the object-level global fusion part, OLF denotes the object-level local fusion part and PGF refers to the pixel-level global fusion part of HGA.

Method	OGF	OLF	PGF	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
MPG-SAM 2				71.9	69.8	73.9
MPG-SAM 2	✓			72.2	70.1	74.3
MPG-SAM 2		✓		72.4	70.1	74.6
MPG-SAM 2	✓	✓		72.8	70.7	75.0
MPG-SAM 2			✓	73.1	71.0	75.3
MPG-SAM 2	✓	✓	✓	**73.9**	**71.7**	**76.1**

Table 4. Performance analysis of the hierarchical global-historical aggregator with varying numbers of layer.

Method	Settings	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
MPG-SAM 2	$N_p = 1$	**73.9**	**71.7**	**76.1**
MPG-SAM 2	$N_p = 2$	73.6	71.6	75.7
MPG-SAM 2	$N_p = 3$	73.0	70.9	75.2
MPG-SAM 2	$N_o = 1$	**73.9**	**71.7**	**76.1**
MPG-SAM 2	$N_o = 2$	73.4	71.2	75.6
MPG-SAM 2	$N_o = 3$	72.9	70.7	75.1

ible in the subsequent four frames. The model effectively detects the absence of the target in the occluded frames and accurately segments it once it appears. Additionally, Fig. 1 (b) presents a scene with three distinct referential targets, all of which are precisely segmented by the model without any omissions across frames.

For the Ref-YouTube-VOS [4] dataset, we select several complex scenarios featuring multiple similar objects to evaluate the model's ability to distinguish between such targets. As shown in Fig. 1 (c), the model accurately segments the specific sheep matching the language description from a group of visually similar sheep. In Fig. 1 (d), the model successfully disregards the interference caused by shadows and accurately segments the kangaroo positioned on the left. These findings highlight the robustness and effectiveness of our model in tackling diverse and challenging scenarios.

References

[1] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703, 2023. 1

[2] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 1

[3] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[4] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Proceedings of the European Con-*

ference on Computer Vision (ECCV), pages 208–223, 2020. 1, 2

[5] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 1

[6] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model. *arXiv preprint arXiv:2411.17606*, 2024. 1

[7] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4974–4984, 2022. 1

[8] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115. Springer, 2024. 1