Appendix for SAM2Long: Enhancing SAM 2 for Long Video Segmentation with a Training-Free Memory Tree

Shuangrui Ding[1] Rui Qian[1] Xiaoyi Dong[1,2*] Pan Zhang[2]
Yuhang Zang[2] Yuhang Cao[2] Yuwei Guo[1] Dahua Lin[1,2,3] Jiaqi Wang[2*]
[1] The Chinese University of Hong Kong [2] Shanghai AI Laboratory [3] CPII under InnoHK
https://mark12ding.github.io/project/SAM2Long/

A. Dataset Information

SA-V [9] is a large-scale video segmentation dataset designed for promptable visual segmentation across diverse scenarios. It encompasses 50.9K video clips, aggregating to 642.6K masklets with 35.5M meticulously annotated masks. The dataset presents a challenge with its inclusion of small, occluded, and reappearing objects throughout the videos. The dataset is divided into training, validation, and testing sets, with most videos allocated to the training set for robust model training. The validation set has 293 masklets across 155 videos for model tuning, while the testing set includes 278 masklets across 150 videos for comprehensive evaluation.

LVOS v1 [5] is a VOS benchmark for long-term video object segmentation in realistic scenarios. It comprises 720 video clips with 296,401 frames and 407,945 annotations, with an average video duration of over 60 seconds. LVOS introduces challenging elements such as long-term object reappearance and cross-temporal similar objects. In LVOS v1, the dataset includes 120 videos for training, 50 for validation, and 50 for testing.

LVOS v2 [6] expends LVOS v1 and provides 420 videos for training, 140 for validation, and 160 for testing. This paper primarily utilizes v2, as it already includes the sequences present in v1. The dataset spans 44 categories, capturing typical everyday scenarios, with 12 of these categories deliberately left unseen to evaluate and better assess the generalization capabilities of VOS models.

MOSE [2] is a challenging VOS dataset targeted on complex, real-world scenarios, featuring 2,149 video clips with 431,725 high-quality segmentation masks. These videos are split into 1,507 training videos, 311 validation videos, and 331 testing videos.

VOST [10] is a semi-supervised video object segmentation benchmark that emphasizes complex object transformations. Unlike other datasets, VOST includes objects that are broken, torn, or reshaped, significantly altering their appearance. It comprises more than 700 high-resolution videos, captured in diverse settings, with an average duration of 21 seconds, all densely labeled with instance masks.

PUMaVOS [1] is a novel video dataset designed for benchmarking challenging segmentation tasks. It includes 24 video clips, each ranging from 13.5 to 60 seconds (28.7 seconds on average) at 480p resolution with varying aspect ratios. PUMaVOS focuses on difficult scenarios where annotation boundaries do not align with clear visual cues, such as half faces, necks, tattoos, and pimples, commonly encountered in video production.

YouTubeVOS-2019 [11] is a large-scale video object segmentation dataset featuring 3,252 sequences with detailed annotations at 6 FPS across 78 diverse categories, including humans, animals, vehicles, and accessories. Each video clip is between 3 to 6 seconds long and frequently contains multiple objects, which have been manually segmented by professional annotators.

DAVIS2017 [8] is a well-known benchmark dataset comprising 60 training videos and 30 validation videos, with a total of 6,298 frames. It offers high-quality, pixel-level annotations for every frame, making it a standard resource for evaluating different VOS methods.

LaSOT [3] is a large-scale tracking dataset designed for long-term visual object tracking. It contains 1,400 videos spanning 70 object categories, with an average sequence length exceeding 2,500 frames, making it a challenging benchmark for evaluating tracking algorithms. **LaSOText**[4] is an extension of LaSOT, featuring a subset of 15 categories with 150 videos.

GOT-10k [7] is a large-scale generic object tracking dataset that covers over 10,000 video sequences, spanning more than 560 object classes. It provides strict one-shot evaluation settings and diverse real-world tracking scenarios, making it a widely used benchmark for developing and assessing tracking models.

[*] Corresponding Author.

δ_{iou}	SAM 2	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
SA-V	76.3	80.0	80.7	80.6	80.8	80.7	81.0	80.6	80.8	80.0	77.8
LVOS	83.0	84.0	85.1	85.6	85.4	85.4	85.1	85.6	85.2	84.1	83.5

Table 1. Ablation study on IoU threshold δ_{iou}.

δ_{conf}	SAM 2	0	1	2	3	4	5
SA-V	76.3	80.4	80.4	80.8	80.8	79.7	78.8
LVOS	83.0	84.4	84.6	85.4	85.2	84.1	83.9

Table 2. Ablation study on uncertainty threshold δ_{conf}.

$[w_{\text{low}}, w_{\text{high}}]$	SAM 2	$[0.6, 1.4]$	$[0.7, 1.3]$	$[0.8, 1.2]$	$[0.9, 1.1]$	$[0.95, 1.05]$	$[1, 1]$
SA-V	76.3	77.1	79.3	79.8	80.4	80.8	80.5
LVOS	83.0	77.0	78.8	82.9	85.0	85.4	84.9

Table 3. Ablation study on modulation weight $[w_{\text{low}}, w_{\text{high}}]$.

SA-V val	Single-object	Multi-object	Overall
# of seq	73	82	155
SAM 2.1	79.0	78.4	78.6
SAM2.1Long	80.6 (1.6% ↑)	81.3 (2.9% ↑)	81.1 (2.5% ↑)

Table 4. Performance comparison on single-object sequence and multi-object sequence. SAM2Long performs exceptionally well in both single-object and multi-object cases.

B. More Ablation Study

Iou Threshold δ_{iou}. The choice of the IoU threshold δ_{iou} is crucial for selecting frames with reliable object cues. As shown in Table 1, setting $0.1 \geq \delta_{\text{iou}} \geq 0.7$ yields the competitive $\mathcal{J}\&\mathcal{F}$, indicating an effective trade-off between filtering out poor-quality frames and retaining valuable segmentation information. In contrast, setting no quality requirement for masks ($\delta_{\text{iou}} = 0$) lowers the score to 80.0, as unreliable frames with poor segmentation harm SAM 2. Conversely, an overly strict selection ($\delta_{\text{iou}} \geq 0.8$) further degrades performance by excluding important neighboring frames, forcing the model to rely on distant frames as memory.

Uncertainty Threshold δ_{conf}. The uncertainty threshold δ_{conf} controls the selection of hypotheses under uncertain conditions. Our results in Table 2 indicate that setting δ_{conf} to 2 provides the highest $\mathcal{J}\&\mathcal{F}$ score, indicating an optimal level for uncertainty handling. Lower values (e.g., 0) lead to suboptimal performance by committing to incorrect segmentations, causing error propagation. Higher values (e.g., 4) do not improve performance, indicating that beyond a certain threshold, the model efficiently relies on top-scoring masks without needing additional diversity.

Memory Attention Modulation $[w_{\text{low}}, w_{\text{high}}]$. We explore the effect of modulating the attention weights for memory entries using different ranges in Table 3. The configuration $[1, 1]$ means no modulation is applied. We find that the configuration of $[0.95, 1.05]$ achieves the best performance while increasing the modulation range decreases performance. This result indicates that slight modulation sufficiently emphasizes reliable memory entries.

C. More Quantitive Comparison

In this section, we present a detailed comparison of SAM2.1 and SAM2Long across single-object and multi-object sequences, as shown in Table 4. Our experiments demonstrate that SAM2Long outperforms SAM2.1 in both single-object and multi-object scenarios. Specifically, SAM2Long achieves a 1.6% improvement in single-object sequences, a 2.9% improvement in multi-object sequences, and an overall 2.5% enhancement in performance. These results highlight SAM2Long's robustness and effectiveness in various video segmentation tasks.

D. More Visualization

We present additional comparisons between SAM2 and SAM2Long in Figure 1. SAM2Long significantly reduces segmentation errors, showing improved accuracy and consistency in object tracking across frames. Notably, in the Fast & Furious movie scene, SAM2Long successfully tracks the green car, even under challenging dynamic camera movements. Overall, SAM2Long offers substantial improvements over SAM2, especially in handling object occlusion and reappearance, leading to better performance in long-term video segmentation tasks.

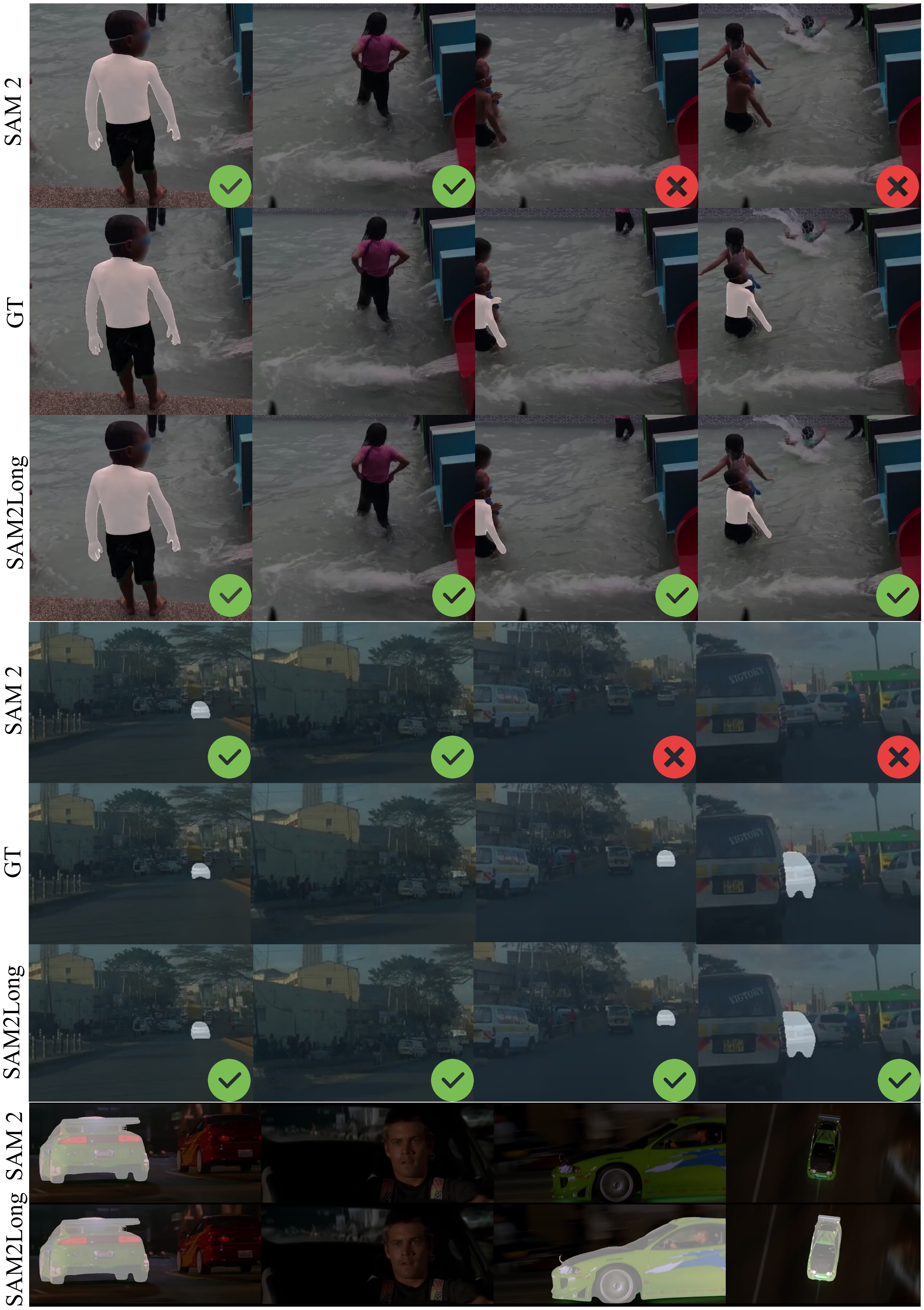

Figure 1. Qualitative comparison between SAM 2 and SAM2Long, with GT (Ground Truth) provided for reference. The last row shows an in-wild case. Best viewed when zoomed in.

References

[1] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 635–644, 2023. 1

[2] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 1

[3] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 1

[4] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129:439–461, 2021. 1

[5] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13480–13492, 2023. 1

[6] Lingyi Hong, Zhongying Liu, Wenchao Chen, Chenzhi Tan, Yuang Feng, Xinyu Zhou, Pinxue Guo, Jinglun Li, Zhaoyu Chen, Shuyong Gao, et al. Lvos: A benchmark for large-scale long-term video object segmentation. *arXiv preprint arXiv:2404.19326*, 2024. 1

[7] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 1

[8] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1

[9] 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

[10] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "object" in video object segmentation. In *CVPR*, 2023. 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. 1