LiVOS: Light Video Object Segmentation with Gated Linear Matching

Supplementary Material

1. Baselines

We introduced four primary baselines in the main paper. In this section, we present the remaining seven baselines.

AOT [24] and **DeAOT** [22] are two consecutive approaches to improve the efficiency of VOS with multiple objects. Following Cutie [6], we use the model variants with a ResNet-50 backbone as baselines.

CFBI [23] and **CFBI+** [25] propose a collaborate VOS approach that integrates both foreground and background information into embedding learning. As they only use two memory frames, we classify them as non-STM methods with less strict criteria. Both models use RestNet-101 as the backbone, and we adopt them as our baselines.

DEVA [5] decouples task-specific image-level segmentation and mask propagation for universal video segmentation. We use as the model trained solely on YouTube-VOS [21] and DAVIS 2017 [16] as the baseline.

SwiftNet [19] balances accuracy and speed by compressing spatiotemporal redundancy in matching-based VOS with a pixel-adaptive memory. We use the model variant with a ResNet-50 backbone as the baseline.

MobileVOS [15] distills knowledge from a teacher model utilizing large backbone and infinite memory. We use the best-performing model variant with a ResNet-18 backbone as the baseline.

2. Related Work

Interactive VOS. Semi-supervised video object segmentation is highly related to interactive video object segmentation (iVOS), which focuses on segmenting objects in video sequences through user interactions. Traditional methods [1, 4, 9, 14] often rely on user-provided annotations, such as scribbles or clicks, to guide segmentation algorithms. Recent advancements have introduced modular approaches that separate user interaction from mask propagation [4]. Additionally, reinforcement learning techniques have been applied to recommend the most informative frames for annotation, thereby reducing user effort [26]. Another notable development is the use of reliability-based attention maps to assess the trustworthiness of annotated frames, leading to more accurate segmentation with fewer interactions [9]. Furthermore, the integration of global and local transfer modules has been explored to enhance the propagation of segmentation information across frames [8]. Our method can be integrated existing interactive image segmentation approaches [2, 11–13, 18, 20] for interactive video object segmentation.

Segment Anything Model. Recent approaches have integrated the Segment Anything Model (SAM) [10] on images with video trackers based on masks [5]. However, there is no mechanism to interactively refine the tracker's errors. The recent proposed SAM 2 [17] addresses this by proposing a unified model that directly takes prompts for interactive video object segmentation, along with a large and diverse video segmentation dataset. Our work focuses on semi-supervised VOS, which can be viewed as a specific instance of the Promptable Visual Segmentation (PVS) task proposed by SAM 2, utilizing only a mask prompt in the first video frame.

3. Experiments

3.1. Comparisons

CPU latency comparison. Fig. A compares CPU latency (in milliseconds) of softmax matching (blue) and linear matching (red) as the number of objects (N) increases. Note that we only report CPU latency rather than GPU latency because GPU latency is close for both matching methods, despite softmax matching significantly increasing GPU memory usage.

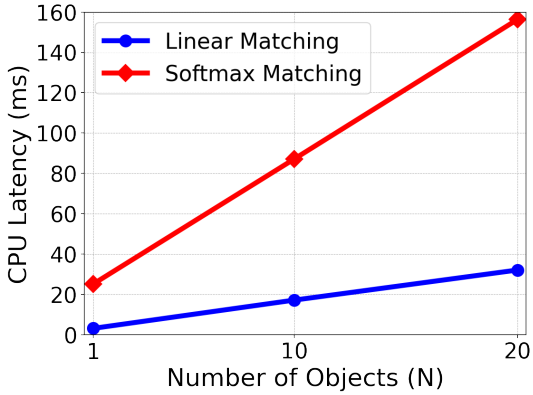

Figure A. *CPU latency comparison between softmax matching and linear matching*. Both linear and softmax attention scale linearly with the number of input objects. However, softmax attention has a significantly steeper slop. Latency is measured on an Intel Core-i7 (2.80GHz) CPU with PyTorch 2.0, batch size 1, and fp32.

Comparisons with downgraded Cutie models. In Table 1 of the main paper, a valid concern arises regarding unfair comparisons with the downgraded Cutie–it uses only one memory frame in inference but multiple during training. To ensure fairness, we train Cutie-small and Cutie-base models

with a single memory frame. As shown in Tab. A, retrained models slightly improve performance but still fall short of LiVOS.

Method	with STM	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
♩ Cutie-small†	✗	76.4	73.0	79.8
♩ Cutie-small‡	✗	77.4 (+1.0)	74.0	80.9
♩ Cutie-base†	✗	79.3	75.8	82.7
♩ Cutie-base‡	✗	80.1 (+0.8)	76.9	83.3
♩ LiVOS	✗	**84.4**	**81.2**	**87.6**

Table A. Comparisons with Cutie models on DAVIS-17 val. † denotes a model that used one memory frame in inference but multiple during training (unfair). ‡ denotes a model that used one memory frame for both training and inference (fair). ♩ represents models trained on YouTube VOS and DAVIS.

Comparisons with downsampling baselines. In Table 3 of the main paper, one can simply downsample the high-resolution input video, process it with a VOS algorithm, and then upsample it back to the original resolution. Tab. B shows the comparisons with downsampling baselines, which downsample input videos to 240p/480p, segment them and upsample the outputs back to 4096p. LiVOS is trained and evaluated on 4096p, achieving notable gains over downsampling baselines. Our method outperforms downsampled non-STM baselines but still falls short of downsampled STM variants. Note that LiVOS has not utilized any modules dedicated for high-res segmentation, leaving room for future improvements.

Method	with STM	Res.	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
♩ Cutie-small	✗	240p	69.9	66.2	73.7
♩ Cutie-small	✗	480p	77.4	74.0	80.9
♩ Cutie-base	✗	240p	70.9	67.1	74.6
♩ Cutie-base	✗	480p	80.1	76.9	83.3
♩ LiVOS	✗	4096p	**82.4**	**79.6**	**85.3**

Table B. Downsampling baseline comparisons on DAVIS-17 val.

3.2. Ablations

Recurrent state. By default, we update the recurrent state for each frame. In this ablation, we investigate how does LiVOS perform with only the first and the last frames in memory for inference. In Tab. C, we evaluate LiVOS in a degraded setting where only the first and the last frames are used to update the state. The degraded LiVOS model has a substantial performance drop, while speed and memory remain stable. This observation indicates that the state update process is efficient and that per-frame updates are highly beneficial.

Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	Mem	FPS
♩ LiVOS	**84.4**	**81.2**	**87.6**	**574M**	40.3
♩ LiVOS (degraded)	78.4	75.3	81.5	**574M**	**42.2**

Table C. Degraded evaluation on DAVIS-17 val.

4. Qualitative Results

We show some qualitative results in this section, including the worse case (Fig. B) and a normal case (Fig. C). We observed that the worst cases arise from thin and scattered structures. Nonetheless, our method is capable of effectively segmenting and tracking some of the objects. In typical cases, such as objects shwon in Fig. C, our method demonstrates exceptional robustness.

5. Implementation Details

Projector. The projector converts a feature map into a gate matrix for element-wise multiplication with the state matrix. We implement the projector with a light-weight convolutional neural network.

Sensory Memory. We adopt sensory memory [3] to maintain low-level information such as object location. A sensory memory stores a hidden state $\mathbf{H}_{t+1} \in \mathbb{R}^{HW \times C_h}$, initialized as a zero vector, and propagated by a Gated Recurrent Unit (GRU) [7]. The hidden state $\mathbf{H}_{t+1}$ is updated every frame using multi-scale features of the mask encoder and decoder, and is added to the value readout $\mathbf{V}_{t+1}$. We set the sensory feature dimension C_h to 256. More details please refer to XMem [3].

Object Memory. We enrich the value readout V_{t+1} for the query frame with object-level semantics using an object transformer [6]. The object transformer takes the initial value readout $\mathbf{V}_{t+1} \in \mathbb{R}^{HW \times C_v \times N}$, a set of M end-to-end trained object queries $\mathbf{Q} \in \mathbb{R}^{M \times C}$ and object memory $\mathbf{O} \in \mathbb{R}^{N \times C}$, and integrates them with L transformer blocks. We set the number of object transformer blocks L to 3 and the number of object queries M to 16. More details, please refer to Cutie [6].

6. Discussions

Limitations. While memory-efficient, our method still lags behind state-of-the-art STM-based approaches across all benchmarks, including short, long, and high-resolution videos. We attribute this gap to excessive input compression, which limits the model's capacity to retain detailed information.

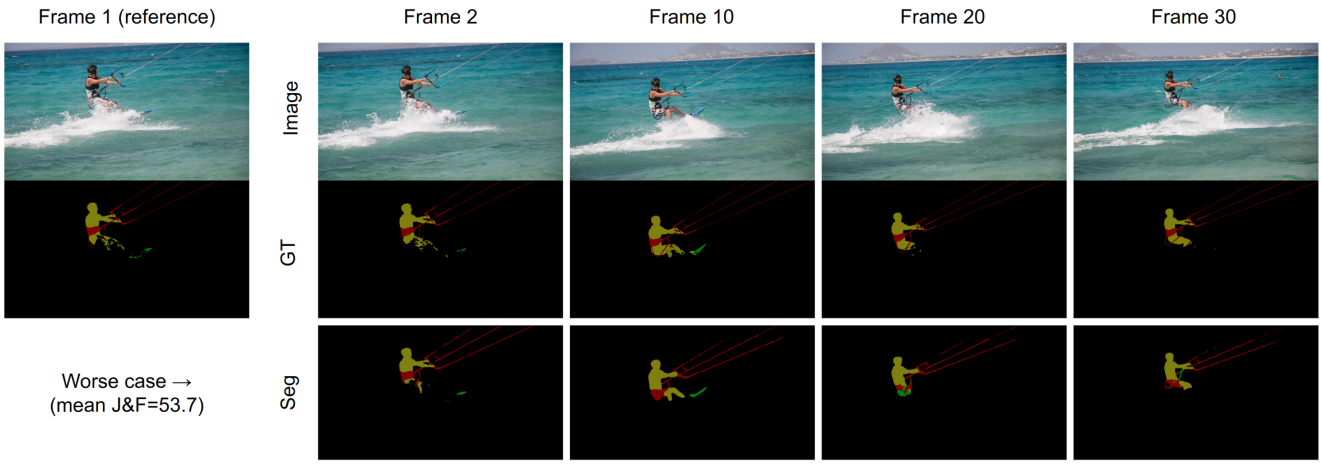

Figure B. *Qualitative results of the worse case in DAVIS 2017 validation set.* Although the objects have thin and scattered structures, our method is capable of effectively segmenting and tracking some of them. Best viewed when zoomed in.

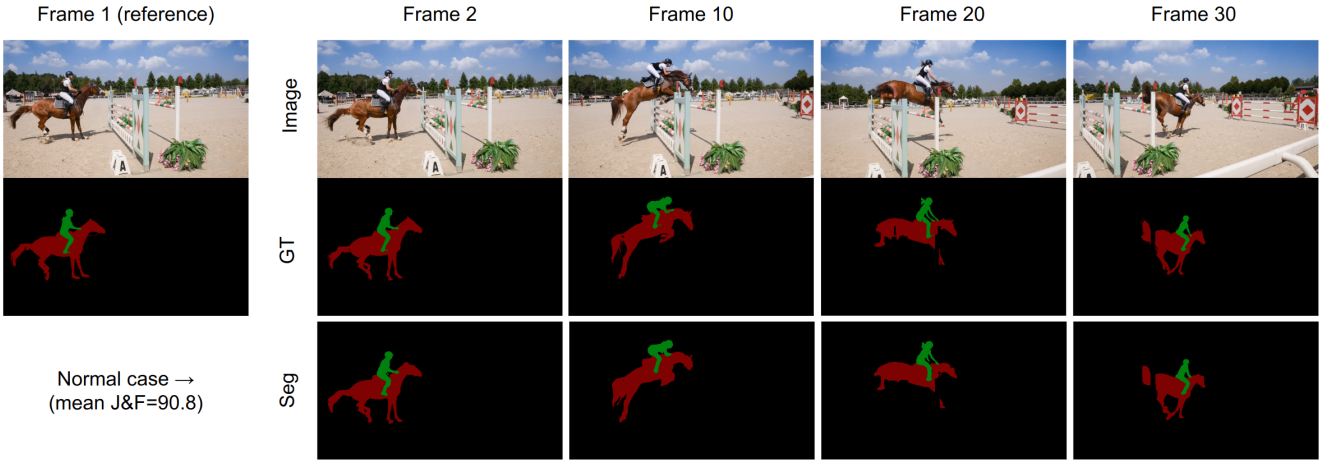

Figure C. *Qualitative results of a normal case in DAVIS 2017 validation set.* Our method works well for these normal objects. Best viewed when zoomed in.

Future Work. A promising future direction is to enhance the single recurrent state with a more advanced representation, better suited for high-resolution video object segmentation and fine-structure delineation. Additionally, due to limited time and space, we do not perform out-of-domain evaluation (e.g., on medical videos). Given the class-agnostic nature of semi-supervised VOS, we remain cautiously optimistic for future exploration.

7. Acknowledgement of AI-Assisted Writing

We acknowledge the use of AI-assisted writing tools, such as large language models, to help improve the clarity, conciseness, and readability of this manuscript. All conceptual content, experimental design, and analysis were solely conducted by the authors.

References

[1] Arnaud Benard and Michael Gygli. Interactive video object segmentation in the wild. *arXiv preprint arXiv:1801.00269*, 2017. 1

[2] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 1

[3] Ho Kei Cheng and Alexander G Schwing. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In *ECCV*, pages 640–658, 2022. 2

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, pages 5559–5568, 2021. 1

[5] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexan-

der Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1316–1326, 2023. 1

[6] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3151–3161, 2024. 1, 2

[7] Kyunghyun Cho. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 2

[8] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Interactive video object segmentation using global and local transfer modules. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 297–313. Springer, 2020. 1

[9] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Guided interactive video object segmentation using reliability-based attention maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7322–7330, 2021. 1

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[11] Qin Liu, Meng Zheng, Benjamin Planche, Srikrishna Karanam, Terrence Chen, Marc Niethammer, and Ziyan Wu. Pseudoclick: Interactive image segmentation with click imitation. In *European Conference on Computer Vision*, pages 728–745. Springer, 2022. 1

[12] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023.

[13] Qin Liu, Jaemin Cho, Mohit Bansal, and Marc Niethammer. Rethinking interactive image segmentation with low latency high quality and diverse prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3782, 2024. 1

[14] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10366–10375, 2020. 1

[15] Roy Miles, Mehmet Kerim Yucel, Bruno Manganelli, and Albert Saà-Garriga. Mobilevos: Real-time video object segmentation contrastive learning meets knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10480–10490, 2023. 1

[16] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 1

[17] 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

[18] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 1

[19] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1296–1305, 2021. 1

[20] Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In *European Conference on Computer Vision*, pages 207–229. Springer, 2024. 1

[21] 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

[22] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 35:36324–36336, 2022. 1

[23] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pages 332–348. Springer, 2020. 1

[24] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021. 1

[25] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4701–4712, 2021. 1

[26] Zhaoyuan Yin, Jia Zheng, Weixin Luo, Shenhan Qian, Hanling Zhang, and Shenghua Gao. Learning to recommend frame for interactive video object segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15445–15454, 2021. 1