

SAMSON: 3rd Place Solution of LSVOS 2025 VOS Challenge

Yujie Xie¹
xyj@pixcakeai.com

Hongyang Zhang^{1,2}
hongyangzhang1@link.cuhk.edu.cn

Zhihui Liu¹
lzh@pixcakeai.com

Shihai Ruan¹
rsh@pixcakeai.com

¹Truesight Research

²School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen

Abstract

*Large-scale Video Object Segmentation (LSVOS) addresses the challenge of accurately tracking and segmenting objects in long video sequences, where difficulties stem from object reappearance, small-scale targets, heavy occlusions, and crowded scenes. Existing approaches predominantly adopt SAM2-based frameworks with various memory mechanisms for complex video mask generation. In this report, we proposed Segment Anything with Memory Strengthened Object Navigation (SAMSON), the **3rd place solution** in the MOSE track of ICCV 2025, which integrates the strengths of state-of-the-art VOS models into an effective paradigm. To handle visually similar instances and long-term object disappearance in MOSE, we incorporate a **long-term memory** module for reliable object re-identification. Additionally, we adopt **SAM2Long** as a post-processing strategy to reduce error accumulation and enhance segmentation stability in long video sequences. Our method achieved a final performance of 0.8427 in terms of $J\&F$ in the test-set leaderboard.*

1. Introduction

Video object segmentation (VOS) is a fundamental problem in computer vision, which aims to segment an arbitrary target throughout a video sequence, given a single annotated mask in the first frame [8, 12, 13]. While conventional VOS benchmarks are typically short video clips, long-term video object segmentation (LVOS) [2, 4, 6] extends this setting to much longer sequences, where substantial appearance changes, occlusions, and scene variations pose additional challenges. The 7th LSVOS challenge tackles these difficulties through three tracks: Complex VOS, standard VOS, and Referring VOS, supported by the MOSEv2 [4], MOSEv1 [2], and MeViS[3], datasets, respectively. Our method is primarily developed for Track 2, which is based

on the MOSEv1 dataset.

The MOSEv1 dataset was introduced to advance video object segmentation (VOS) in complex scenes. Its successor, MOSEv2, presents a more challenging benchmark by intensifying scene complexity and introducing underrepresented factors, including adverse weather (rain, snow, fog), low-light conditions (nighttime, underwater), and multi-shot sequences. These enhancements aim to better approximate real-world scenarios and narrow the gap between existing VOS benchmarks and unconstrained environments. MeViS aims to segment and track target objects in videos based on natural language descriptions of their motions. This dataset involves handling diverse motion expressions that specify target objects within complex environments. It provides a benchmark for advancing language-guided video segmentation, where motion expressions serve as a primary cue to enhance object segmentation in challenging video scenes.

Recent progress in video object segmentation (VOS) has increasingly emphasized memory-based approaches due to their clear advantages over alternatives. Building on image-based SAM, SAM2 [9] introduces a memory module that extends its capability to VOS tasks and delivers notable improvements in segmentation performance. Nevertheless, its greedy segmentation strategy remains vulnerable to challenging scenarios involving frequent occlusions and object reappearances, while the fixed 8-frame memory restricts its effectiveness in long-term video analysis. To address these limitations, Segment Concept (SeC) [15] introduces a concept-driven paradigm that shifts from feature matching to constructing and leveraging high-level object representations. By equipping SAM2 with an enhanced long-term memory module, SeC achieves significant gains in VOS performance.

In this work, we propose SAMSON, our 3rd place solution for the MOSEv1 challenge. Witnessing the enhanced memory module of SeC, we further fine-tune its grounding

encoder through a two-stage training strategy on the MOSEv2 dataset. To mitigate the error propagation inherent in the original SAM2 framework, we incorporate SAM2Long at inference, thereby improving segmentation robustness. Our approach achieves a $\mathcal{J}\&\mathcal{F}$ of 0.8427, with $\mathcal{J} = 0.8182$ and $\mathcal{F} = 0.8671$ on the MOSEv1 track of ICCV 2025. Beyond this primary solution designed for the competition, we also conducted a series of exploratory experiments to address the fundamental trade-off between memory length and computational cost from a different perspective. These explorations focused on designing a new memory paradigm, including enlarging the temporal perception field and refining memory update mechanisms. Although this experimental approach is still in its preliminary stages, it offers valuable insights for tackling extreme challenges like long-term occlusions, which we discuss in a later section.

2. Related Work

2.1. Video Object Segmentation

Video Object Segmentation (VOS) tasks, including semi-supervised and unsupervised segmentation, have been primarily evaluated on datasets like DAVIS [8] and YouTube-VOS [12]. DAVIS offers high-quality, short-term video sequences with dense annotations, focusing on precise segmentation under occlusions and appearance changes. YouTube-VOS provides larger-scale, diverse videos, introducing challenges such as dynamic backgrounds and moderate-length tracking. However, their short-to-medium durations limit their ability to assess long-term temporal consistency required for real-world applications.

Recent datasets such as LVOS [6] and MOSE [2] address long-term VOS with extended sequences featuring challenges like prolonged occlusions, object reappearances, and crowded scenes. LVOS emphasizes diverse categories and long-term tracking across thousands of frames, testing consistency, while MOSE focuses on occlusions and motion blur in cluttered environments, posing significant challenges. The core task is to maintain precision in long-term videos by combining the strengths of LVOS and MOSE. More recently, MOSEv2 [4] advances multi-object video segmentation by scaling data volume and diversity, with over 5,200 videos and 1.2M annotated frames. It covers broader categories, denser interactions, and harder conditions such as severe occlusions and illumination changes, providing a comprehensive benchmark for evaluating robustness and generalization in video segmentation models.

2.2. Memory based methods in VOS

Memory-based frameworks have become an emerging paradigm in video object segmentation, leveraging stored temporal information to ensure consistency across frames. Key methods include XMem [1], which uses an Atkinson-

Shiffrin memory model for long-term VOS, partitioning memory into short-term and long-term components. However, it accumulates errors under heavy occlusions due to pixel-level matching without distractor filtering. Learning Quality-aware Dynamic Memory [7] updates memory by feature quality, enhancing robustness. RMem [16] restricts memory banks for relevant representations, lowering overhead. Still, both face memory overload and distractor interference in long-term, crowded videos.

Recent integrations with foundation models like SAM 2 [9] have advanced the VOS. SAM2Long [5] uses a training-free memory tree for bidirectional propagation, addressing long-term occlusions but increasing complexity. DAM4SAM [11] employs distractor-aware memory for tracking, suppressing similar objects, yet struggles with pure segmentation in extended sequences. SAMURAI [14] leverages motion-aware memory for zero-shot tracking, but may miss static objects, risking context loss. Recently, SeC [15] utilizes Large Vision-Language Models (LVLMs) to progressively construct high-level, object-centric representations by integrating visual cues across frames, enabling robust semantic reasoning. During inference, SeC builds comprehensive semantic representations from processed frames for accurate segmentation of subsequent frames. Additionally, SeC dynamically balances LVLM-based semantic reasoning with enhanced feature matching, adapting computational efforts to scene complexity.

However, they still face the challenges: memory inefficiency and overload in long sequences (LVOS), handling uncertainty from occlusions and deformations (MOSE).

3. Methods

3.1. Overview

Given a video sequence with T frames $\{I_t\}_{t=1}^T$, the ground-truth mask M_1 of the target objects are provided in the first frame. The goal is to predict segmentation masks $\{M_t\}_{t=2}^T$ for the remaining frames through the segmentation model $f_\theta(\cdot)$.

Image Encoder. We adopt Hiera [10], a hierarchical masked autoencoder, as the image encoder. Its multiscale architecture enables effective capture of both local details and long-range dependencies, providing robust representations for video segmentation.

Mask Encoder. The mask encoder in SAM2 encodes segmentation masks by first embedding the input mask through a convolutional module, which projects it into the feature space. This embedding is then element-wise combined with the corresponding frame features from the image encoder, followed by lightweight convolutional layers for feature fusion. During tracking, only initialization masks or predicted masks are used, while interactive inputs such as clicks or bounding boxes are excluded to ensure full automation.

This design refines mask representations in a compact and efficient manner, enabling precise segmentation and seamless integration into the overall SAM2 pipeline.

Memory Bank. The memory bank stores the initialization frame with its ground-truth mask and the six most recent frames with predicted masks. Temporal encodings are applied to recent frames to preserve ordering, while the initialization frame remains unencoded to serve as a target prior.

Mask Decoder. Current-frame features attend to memory frames to obtain memory-conditioned representations, which are decoded into three candidate masks with IoU scores. The mask with the highest score is selected as output, and the memory is updated in a first-in-first-out manner, with the initialization frame permanently retained.

Optimization. The training objective of the proposed method combines complementary losses for pixel-level accuracy, region alignment, overlap quality, and mask-score regression. Concretely, we use a binary cross-entropy (BCE) loss for pixel-wise foreground/background classification, an IoU loss for region-level alignment, a Dice loss to mitigate class imbalance, and a Mask loss to supervise the decoder’s predicted mask quality scores.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{IoU}} + \lambda_3 \mathcal{L}_{\text{Dice}} + \lambda_4 \mathcal{L}_{\text{Mask}}, \quad (1)$$

where the $\mathcal{L}_{\text{Mask}}$ term is defined as:

$$\mathcal{L}_{\text{Mask}} = \frac{1}{K} \sum_{k=1}^K \ell(\hat{s}_k, s_k), \quad s_k = \text{IoU}(\hat{M}_k, M_{\text{gt}}), \quad (2)$$

with \hat{s}_k the decoder’s predicted IoU for candidate mask \hat{M}_k , s_k the ground-truth IoU computed against M_{gt} , K the number of candidates per frame, and $\ell(\cdot, \cdot)$ a regression loss (e.g. Smooth- L_1 or MSE). The weights $\lambda_{1..4}$ balance the terms.

Since the SeC framework adaptively balances LVLM-based semantic reasoning with feature matching and dynamically allocates computation according to scene complexity, and given its superior empirical performance over state-of-the-art methods such as SAM2 and its variants across multiple benchmarks, we adopt it as our baseline. The training framework for the second stage is illustrated in Figure.1.

3.2. Long-Term Memory Update for SAM

We utilize the grounding encoder from SeC model and enhance the memory bank update mechanism by incorporating a distractor-aware memory module, drawing inspiration from DAM4SAM, to improve robustness and accuracy in video object segmentation. In the inference stage, SAM2Long [5] is adopted for robust long-term video object segmentation, using a training-free memory tree to mitigate error accumulation and enable accurate tracking across extended sequences with occlusions.

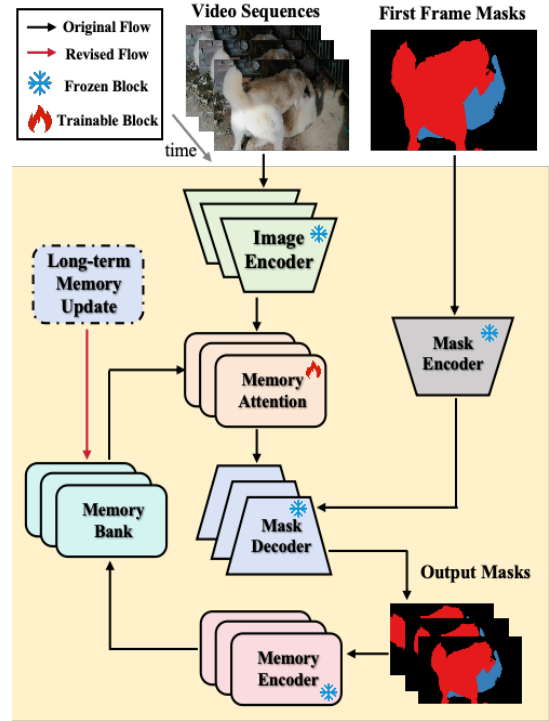


Figure 1. Overview of the proposed second stage of training pipeline, where only the memory attention module is fine-tuned during this process.

3.2.1. Sec Model

Inspired by the Segment Concept (SeC) framework, we adopt its progressive concept-grounding encoder to construct high-level, object-centric representations for video object segmentation. SeC model is trained by the strategy as below:

Concept Guidance with LVLM. To strengthen concept-level reasoning, a sparse keyframe bank is maintained and updated during tracking. It retains the initialization frame and a few representative keyframes to ensure semantic diversity. LVLM encodes this compact set, with a special $\langle \text{SEG} \rangle$ token extracting object-level concept guidance.

Scene-Adaptive Activation. To avoid redundancy, a scene-adaptive strategy applies concept guidance only when notable scene changes occur; otherwise, lightweight pixel-level matching is used. When activated, the LVLM-derived concept vector is fused with current frame features via cross-attention, enriching memory-enhanced representations. This balances semantic priors with fine-grained visual cues, ensuring robust segmentation across challenging scenarios.

3.2.2. Distractor-Aware Memory Strategy

To address long-term dependencies, we scale up the memory module inspired by the design of a distractor-aware

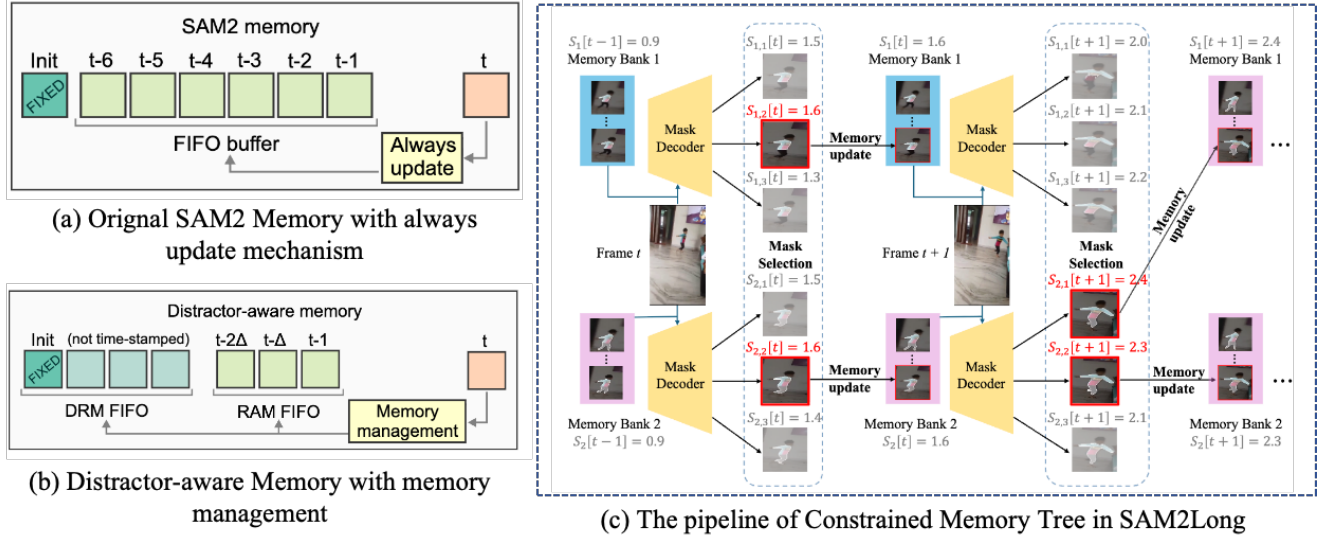


Figure 2. (a) Original design of the SAM2 memory mechanism and (b) proposed distractor-aware memory mechanism, both presented in [11]; (c) At each time step, multiple memory pathways are maintained, with the mask decoder generating candidate masks conditioned on memory banks. The pathway with the highest cumulative score is selected for propagation, adapted from [5].

memory (DAM). Figure 2 illustrates the memory management mechanisms in video object segmentation. Figure 2 (a) depicts the original SAM2 memory system with an always-update mechanism, featuring a FIFO buffer that processes frames from $t = 6$ to $t = 1$, with the initial frame fixed and the most recent frame always updated. Figure 2 (b) shows the distractor-aware memory management, incorporating a DFM FIFO for fixed initial frames (non-time-stamped), an RAM FIFO for dynamic frames from $t = 2$ to $t = 1$, and an integrated memory management module for enhanced robustness against distractors. The target of it mainly focused on tracking design, our memory stores an expanded set of temporal features $\{F_t, C_t\}_{t=1}^T$, with a capacity increased by a factor of k (e.g., $k = 5$) relative to DAM. This larger memory retains detailed object information, critical for handling reappearances after prolonged occlusions. The distractor-aware mechanism computes a similarity score to filter irrelevant objects:

$$S_t = \text{Sim}(C_t, M_t), \quad M_t = \{F_i, C_i \mid i \in [1, t-1]\}, \quad (3)$$

where $\text{Sim}(\cdot, \cdot)$ is a cosine similarity function, and M_t is the memory bank. Low-scoring distractors are suppressed, ensuring focus on the target object. The distractor-aware mechanism, adapted from DAM4SAM, enhances accuracy by mitigating interference from similar objects.

3.3. SAM2Long for inference

During inference, we further introduce SAM2Long to improve robustness without introducing additional training costs. The method adopts a constrained tree memory structure with uncertainty handling.

The detailed information of constrained tree memory is illustrated in Figure 2 (c). Formally, given a set of memory nodes $\{m_i\}_{i=1}^N$, each associated with an uncertainty score σ_i , the aggregated memory feature at time step t is computed as:

$$\hat{M}_t = \sum_{i=1}^N w_i \cdot m_i, \quad w_i = \frac{\exp(-\sigma_i)}{\sum_{j=1}^N \exp(-\sigma_j)}, \quad (4)$$

where the weights w_i are constrained by the tree hierarchy, ensuring that closer parent-child nodes in the memory tree receive consistent weighting. The uncertainty score σ_i is estimated from prediction confidence, allowing unreliable memory nodes to be down-weighted automatically.

The ensemble mechanism then fuses the uncertainty-aware memory \hat{M}_t with the concept representation C_t , yielding the final segmentation prediction:

$$\hat{Y}_t = f_{\text{dec}}(\hat{M}_t, C_t, I_t), \quad (5)$$

where I_t denotes the current frame embedding and f_{dec} is the mask decoder.

This design not only balances adaptability and stability, but also mitigates error accumulation by dynamically suppressing noisy or outdated memory entries. Consequently, SAM2Long achieves consistent tracking under long-term occlusion, re-appearance, and large-scale appearance variations, while maintaining high efficiency at test time.

4. Experiment

4.1. Implementation details

Training Details. In this experiment, we adopt SeC’s ground-encoder as the baseline, given its effectiveness demonstrated across multiple benchmarks, and fine-tune its memory encoder using the MOSEv2 dataset, which contains 3,666 annotated videos. The network architecture is based on DAM4SAM, while the pretrained weights are inherited from SeC. MOSEv2 is chosen as it subsumes MOSEv1 while offering greater diversity and more challenging scenarios, thereby enhancing model generalization. The training process consists of two stages. In the first stage, we fine-tune the entire model with 8-frame inputs only due to the GPU memory constraints. In the second stage, we froze all components except for the memory attention module and extended the input to 24 frames to recover the long-term memory capacity of the designed structure. All experiments are conducted on 8 H800 GPUs with a batch size of 1 per GPU. Input images are resized to 1024×1024 , with data augmentation applied through RandomHorizontalFlip, RandomAffine and ColorJitter. In the training process, the model is optimized by AdamW for 40 epochs. Besides, the optimization hyperparameter of λ_4 is set to 15.0.

Benchmarks. We evaluated our model against two rigorous benchmarks: MOSE v1 and LVOSv1. MOSE v1 offers 2,149 video sequences, totaling over 560K frames, with dense annotations for multi-object video segmentation. This benchmark addresses diverse challenges such as occlusion, rapid motion, and scale variations. LVOSv1 comprises 1,128 long-term videos, exceeding 400K frames, with many sequences over 1,000 frames. This dataset is ideal for assessing robustness to occlusion, target re-appearance, and temporal consistency in long-term video segmentation. The validation sets include 311 video clips from MOSE v1 and 50 from LVOSv1.

Metrics. We adopt standard evaluation metrics: region similarity (\mathcal{J}), contour accuracy (\mathcal{F}) and their average value ($\mathcal{J\&F}$) on the two benchmarks.

4.2. Performance Comparison

We present a series of ablation studies on the validation subsets on the two benchmarks, utilizing SAM2-Large as the default model size following the setting in SeC.

Effectiveness of the proposed module. First, we explore the effectiveness of the DAM module and fine-tuning tricks in the training stage, SAM2Long in the inference stage. As shown in Table 1, introducing DAM brings clear performance improvements over the zero-shot baseline, e.g., from 75.1 to 77.7 in $\mathcal{J\&F}$ on MOSEv1 and from 77.4 to 81.2 on LVOSv1. Further applying fine-tuning tricks provides additional gains, especially on LVOSv1, where \mathcal{F} improves from 84.9 to 86.5. Finally, equipping the model with

SAM2Long during inference achieves the best overall results, reaching 83.6 in $\mathcal{J\&F}$ on LVOSv1 and 77.9 on MOSEv1, which demonstrates the complementary benefits of DAM and SAM2Long.

Effectiveness of Fine-tuning tricks. Furthermore, we compare the performance of the model with full fine-tuning (Full FT), the model only fine-tuned with Memory Attention (Memory FT), and the model with two-stage fine-tuning (two-stage FT). In Table 2, two-stage FT achieves the best performance with a $\mathcal{J\&F}$ score of 77.9, slightly surpassing Full FT (77.7), while Memory FT lags behind at 77.0. This demonstrates that selectively freezing modules and progressively unfreezing them not only reduces computational cost but also recovers long-term memory capability, yielding more robust segmentation.

4.3. Qualitative Results

To evaluate the performance of the proposed method, we present the visualization results between Baseline and our proposed method in Figure 3 and Figure 4, and the comparison clearly demonstrates that our method produces more accurate and consistent segmentation, especially in challenging scenarios with object occlusion, appearance changes, and cluttered backgrounds.

5. Exploratory Study: An Enhanced Memory Paradigm

5.1. Motivation

We recognize a fundamental limitation in existing VOS models: fixed-length, first-in-first-out (FIFO) sliding window memory, as used in SAM2, is inherently constrained when handling long videos. This weakness becomes particularly evident during prolonged occlusions or significant appearance changes, where critical historical information is discarded, leading to tracking failures. To address the trade-off between memory length and computational cost, we conducted an exploratory study to design a more robust and flexible memory paradigm, shown in Figure 5.

5.2. Dynamic Multi-Group Memory Architecture

Instead of simply enlarging memory capacity, our method fundamentally redesigns memory organization and utilization through four key mechanisms.

Effective Frame Selection. We establish an extensive “memory pool” to retain a large number of historical frames. To ensure memory quality, we filter frames using predefined IoU and object score thresholds before adding them to or sampling from the pool, guaranteeing that only high-confidence frames are used for segmentation.

Dynamic Positional Encoding. We critically redesigned SAM2’s positional encoding. Traditional methods use relative batch order, which fails to reflect true temporal separa-

Table 1. Performance Comparison (%) on MOSEv1 and LVOSv1 datasets. All baselines are trained on MOSEv2 dataset, except for zero-shot. * denotes the model with fine-tuning tricks.

Method	MOSEv1 val			LVOSv1 val		
	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
baseline (zero-shot)	75.1	71.1	79.1	77.4	73.2	81.6
baseline + DAM	77.7	73.6	81.8	81.2	77.6	84.9
baseline + DAM*	78.1	73.8	82.3	81.7	76.9	86.5
baseline + DAM* + SAM2Long	77.9	73.7	82.1	83.6	78.8	88.5

Table 2. Ablation Study Performance (%) of different fine-tuning tricks on MOSEv1.

Method	MOSEv1 val		
	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
baseline + Full FT	77.7	73.6	81.8
baseline + Memory FT	77.0	72.8	81.3
baseline + two-stage FT	77.9	73.7	82.1

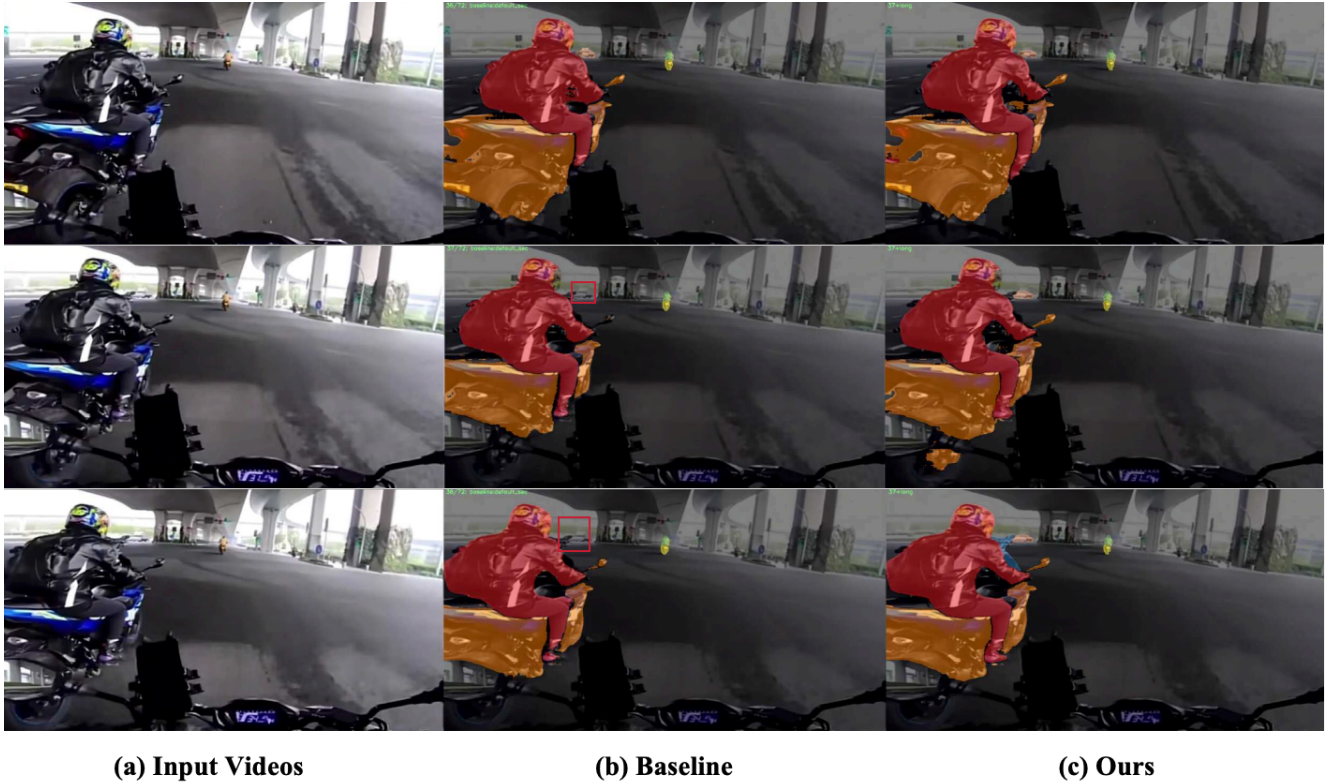


Figure 3. Qualitative comparison of video1 on MOSEv1 dataset. (a) Input video frames. (b) Results of the baseline method. (c) Results of our method. The red bounding box denote the failed examples in baseline.

ration. Our method dynamically calculates sinusoidal positional encoding based on the actual frame index difference between a memory frame and the current frame, normalized by a maximum temporal gap. This provides a more precise, physically intuitive representation of motion and temporal continuity, crucial for robust long-term tracking. Following

is the modified formula of the position encoding:

$$\text{PE}_{\text{orig}}(\text{pos}, d) = \sin\left(\frac{\text{pos}}{10000^{\frac{2d}{d_{\text{model}}}}}\right) \quad (6)$$

$$\text{PE}_{\text{improved}}(t, p, d) = \sin\left(\frac{\min(|t - p|, 128)/128}{10000^{\frac{2d}{d_{\text{model}}}}}\right) \quad (7)$$



Figure 4. Qualitative comparison of video2 on MOSEv1 dataset. (a) Input video frames. (b) Results of the baseline method. (c) Results of our method. The red bounding boxes denote the failed examples in baseline.

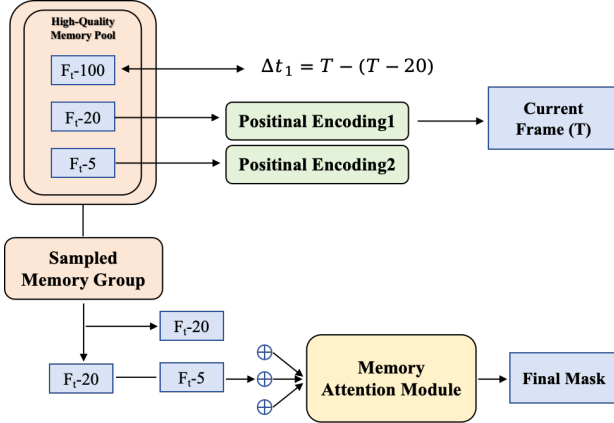


Figure 5. The pipeline of the enhanced Memory Module.

where $PE_{\text{orig}}(\text{pos}, d)$ represents the original position encoding in SAM2, $PE_{\text{improved}}(t, p, d)$ represents our modified version. $|t - p|$ is the frame index difference and is normalized by 128 (a preset maximum timestamp).

Multi-Group Memory Processing. At inference, we sample multiple memory groups from a high-quality pool:

recent frames preserve short-term dynamics, while early frames maintain identity. This flexible strategy, supported by our training design, enables simultaneous exploitation of both local and global temporal context.

Weighted Fusion. Pixel features from different memory groups are aggregated through weighted summation, enabling integration of target information across historical periods. This ensemble-like design improves robustness by reducing reliance on any single, potentially suboptimal memory. In the current implementation, weights are assigned uniformly, leaving room for future exploration of more adaptive strategies.

5.3. Preliminary Results and Analysis

This approach showed strong potential in challenging scenarios. In qualitative tests with long-term object disappearance and reappearance, the model re-identified the target where baselines failed, highlighting the value of retaining longer, higher-quality history.

However, the approach also introduced new issues. Its overall $J\&F$ score on the MOSE dataset was lower than our final SAMSON submission, primarily due to a significant increase in false positives. We hypothesize that while

the expanded memory provides richer historical context, the current filtering and simple fusion mechanisms are not yet sophisticated enough to consistently distinguish the target from visually similar distractors over long periods. This can cause the model to be misled by outdated or irrelevant memory cues, especially in crowded scenes.

6. Conclusion & Future work

In this report, we presented SAMSON, our solution to the 7th LSVOS challenge, which ranked 3rd place in the MOSE track of ICCV 2025. By integrating a long-term memory module with SAM2 and adopting SAM2Long at inference, our method effectively mitigates the challenges of object reappearance, occlusions, and error accumulation in long video sequences. Leveraging a two-stage fine-tuning strategy on MOSEv2, SAMSON achieves a strong performance of $\mathcal{J}\&\mathcal{F} = 0.8427$ on the MOSEv1 leaderboard, demonstrating the importance of memory-enhanced object navigation for large-scale video object segmentation. Furthermore, our exploratory study on an enhanced memory paradigm, while not yet achieving competitive overall accuracy, has shown clear promise for handling extreme long-term occlusions. The preliminary results suggest that a larger memory pool is a viable path forward. Future work will focus on developing more intelligent and adaptive memory sampling and weighting mechanisms. By incorporating attention or quality assessment modules to actively suppress irrelevant historical information, we aim to harness the full potential of an expanded memory pool without sacrificing precision, ultimately developing a more powerful and robust VOS model.

References

- [1] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pages 640–658. Springer, 2022. 2
- [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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20224–20234, 2023. 1, 2
- [3] Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. MeViS: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1
- [4] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. Mosev2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025. 1, 2
- [5] Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. *arXiv preprint arXiv:2410.16268*, 2024. 2, 3, 4
- [6] 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, 2
- [7] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *European Conference on Computer Vision*, pages 468–486. Springer, 2022. 2
- [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, 2
- [9] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2
- [10] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, pages 29441–29454. PMLR, 2023. 2
- [11] Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with sam2. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24255–24264, 2025. 2, 4
- [12] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018. 1, 2
- [13] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. In *ECCV*, 2018. 1
- [14] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024. 2
- [15] Zhixiong Zhang, Shuangrui Ding, Xiaoyi Dong, Songxin He, Jianfan Lin, Junsong Tang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Sec: Advancing complex video object segmentation via progressive concept construction. *arXiv preprint arXiv:2507.15852*, 2025. 1, 2
- [16] Junbao Zhou, Ziqi Pang, and Yu-Xiong Wang. Rmem: Restricted memory banks improve video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18602–18611, 2024. 2