Techinical Report for EPIC-KITCHENS VISOR Semi-supervised VOS Challenge 2025

Kehuan Song, Xinglin Xie, Licheng Jiao, Fang Liu, Shuo Li

School of Artificial Intelligence, Xidian Universitly

Abstract

Our team, DJHVNH, achieved a validation score of 0.864 on the VISOR[1] dataset. The dataset presents several challenges, including complex scenes, multiple targets, and issues such as blurring, occlusion, and small target objects, making instance segmentation a non-trivial task. To address these challenges, we adopted the SAM2 model as our base model. SAM2 is known for its real-time, zero-shot tracking capabilities, continuous tracking, efficiency, and robustness, enabling it to quickly and accurately track target objects in dynamic scenes. In terms of video segmentation, SAM2 supports prompt-based segmentation, streaming, multimodal segmentation, and zero-shot segmentation, offering high precision, efficiency, and strong generalization ability, making it suitable for various complex scenarios and applications with limited data. We fine-tuned the SAM2 model and employed specific strategies to improve its performance. Experiments have shown that our approach yields promising results.

1. Introduction

The VISOR dataset is a comprehensive video segmentation and object relations dataset built upon the EPIC-KITCHENS-100 dataset. It features 272K manual semantic masks across 257 object classes, 9.9M interpolated dense masks, and 67K hand-object relations, covering 36 hours of untrimmed videos. The dataset introduces an AI-powered annotation pipeline to ensure scalability and quality, capturing both short-term and long-term consistency of pixel-level annotations as objects undergo transformative interactions in egocentric videos. VISOR also includes three benchmark challenges: Semi-Supervised Video Object Segmentation (VOS), Hand Object Segmentation (HOS), and a Where Did This Come From (WDTCF) challenge for long-term reasoning. The dataset aims to support research in long-term video understanding and complex video segmentation tasks.

At present, mainstream video object segmentation (VOS) techniques generally store the prediction results of previous frames in a memory module and use attention mechanisms to associate the frame representations in memory with the features of the current frame, thereby increasing the contextual information available to the model[2][3][4][5]. For example, XMem[6] achieves a J&F score of 87.7% on DAVIS2017[7] and 86.1% on YouTube-VOS[8]. These methods effectively store and utilize features from past frames in memory modules, allowing models to consider multi-frame information over time when processing current frames, thereby making more accurate judgments about object appearance changes and motion trajectories. However, such pixel-level matching methods are prone to noise interference in complex environments, especially in the presence of occlusion and other factors, leading to significant performance degradation. The re-emergence of objects and the presence of confusing objects (similar neighboring objects) are the main causes of tracking failures.

To promote the application of VOS in Kitchen scenarios, the competition has constructed the VISOR dataset for complex video object segmentation, which is used to study object tracking and segmentation in complex kitchen environments. The VISOR dataset features complex scenes with severe occlusion, disappearance, and re-emergence of objects, as well as small-scale issues[9], which not only challenge object segmentation in images but also greatly increase the difficulty of tracking occluded objects over time. Our methods that perform well on mainstream VOS datasets show significant performance drops on the VISOR dataset. Based on these promblems, We have chosen SAM2 as our base model and trained it on the VISOR dataset.

2. Methods

We optimized the proposed solution in the training finetuning and inference stages. In the training stage,

we finetuned the SAM2 on the VISOR dataset to make them more suitable for the requirements of video object segmentation tasks in complex kitchen scenarios.

SAM2, relying on the memory attention module, can correct predictions based on the context of object memories from previous frames, enabling stable segmentation of objects in videos. Meanwhile, it possesses a powerful zero - shot generalization ability, allowing it to operate on unseen data. The SAM2 model is an advanced AI framework designed for video object segmentation and tracking tasks. It builds upon the strengths of its predecessor, incorporating enhanced features and optimizations to improve performance and efficiency. SAM2 excels in handling complex scenes and dynamic objects, thanks to its robust architecture and sophisticated algorithms. Its ability to process video streams in real-time makes it particularly suitable for applications requiring immediate feedback and decision-making. In our project, we have chosen SAM2 as our base model due to its proven capabilities and potential for high accuracy in video object segmentation tasks. We have further fine-tuned SAM2 on the VISOR dataset to better adapt it to the specific challenges and nuances of our video segmentation tasks, ensuring optimal performance and results.

To enhance the performance on the VISOR dataset while maintaining the strong generalization ability of SAM2, we freeze the encoder part to retain its powerful feature extraction capability and conduct targeted fine-tuning on the decoder. To reduce the training time, we use the large checkpoints of checkpoint 2.1 as the pre-trained weights and train for 40 epochs using 4 3090 GPUs each having 24 GB of memory. We configure the hyperparameters, setting the base_lr to 8×10^{-6}, the btchsize to 2, and define the optimizer as Adam. During the training process, we iteratively process the training data, calculate the Focal loss, Dice loss, and IoU loss, and obtain the total loss through weighted summation. We update the model weights through gradient accumulation and backpropagation, and save the model every 5 epochs. We test the J&F score of each saved model on the validation set. We performed full-parameter fine-tuning of the model based on the base pre-trained weights, with the batch size set to 1 and the learning rate (lr) set to 5×10^{-6}. We trained the model for 80 epochs on the VISOR dataset. As a result, the model achieved a score of 0.864 on the validation set, which is approximately a 10% improvement compared to direct inference.We conducted the training on four NVIDIA GeForce RTX 3090 GPUs, each equipped with 24 GB of memory.We present some quantitative results before and after finetuning in Fig1

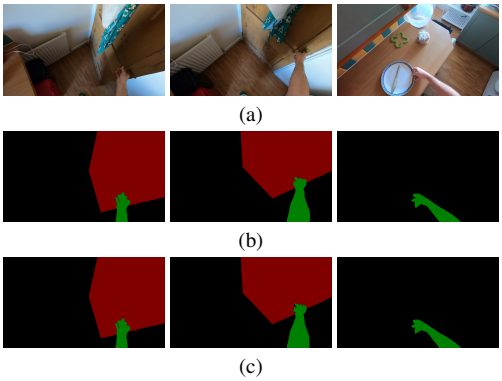

Figure 1. (a) shows the original images of the VISOR val set. (b) presents ground truth. (c) shows that the segmentation results of the fine-tuned SAM2 are almost consistent with the ground truth.

3. Experiments

3.1 Dataset Introduction

The EPIC-KITCHENS VISOR dataset, derived from the EPIC-KITCHENS collection, presents unique challenges not found in existing video segmentation datasets. It demands maintaining both short-term and long-term consistency in pixel-level annotations as objects undergo various transformations, such as peeling, dicing, and cooking an onion. Our goal is to provide precise pixel-level annotations for all relevant elements, including the onion's peel, pieces, chopping board, knife, pan, and the hands performing the actions. To achieve this, VISOR employs a partially AI-driven annotation pipeline designed to ensure scalability and high-quality annotations.

3.2 Evaluation Metrics

We evaluate segmentation accuracy using the average Jaccard (J) index, average boundary F-score, and the average J&F value. After inference by the recommended model, we assess its performance using the J&F score (Joint and Fusion metric). The specific calculation methods are as follows:

Segmentation Accuracy (J) Compare the inferred results with the Ground Truth segmentation masks to calculate the IoU value, and take the average of all targets as the J score. For a predicted segmentation mask P and a ground truth segmentation mask G, the Jaccard value is defined as:

$$J = \frac{|P \cap G|}{|P \cup G|} = \frac{\sum_i P_i \cdot G_i}{\sum_i P_i + \sum_i G_i - \sum_i P_i \cdot G_i}, \quad (1)$$

where P_i and G_i denote the value of the i-th pixel in the predicted and ground truth masks, respectively.

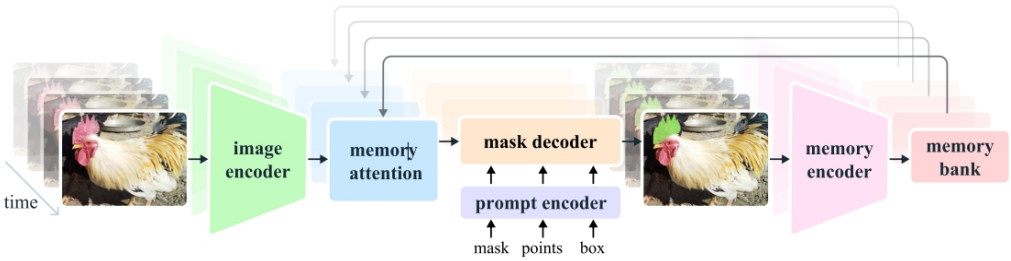

Figure 2. Overview of the SAM2 Framework. For a given frame, the segmentation prediction is conditioned on the current prompt and/or on previously observed memories. Videos are processed in a streaming fashion with frames being consumed one at a time by the image encoder, and cross-attended to memories of the target object from previous frames. The mask decoder, which optionally also takes input prompts, predicts the segmentation mask for that frame. Finally, a memory encoder transforms the prediction and image encoder embeddings (not shown in the figure) for use in future frames.

The Jaccard value ranges from 0 to 1, with higher values indicating better performance.

Tracking Effectiveness (F) Compare the inferred results with the Ground Truth tracking IDs to calculate the MOTA (Multi-Object Tracking Accuracy) value, which measures the accuracy and stability of tracking.It is calculated as follows:

$$F = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{2}$$

where

$$\text{Precision} = \frac{|P \cap G|}{|P|} = \frac{\sum_i P_i \cdot G_i}{\sum_i P_i}, \tag{3}$$

and

$$\text{Recall} = \frac{|P \cap G|}{|G|} = \frac{\sum_i P_i \cdot G_i}{\sum_i G_i}. \tag{4}$$

The F-Measure also ranges from 0 to 1, with higher values indicating better model performance in handling positive and negative samples.

Composite Score Weight J and F (e.g., 0.5:0.5) to derive the final J&F score. A higher J&F score indicates better overall model performance on the video frame.

3.3 Results of Experiments

The experimental results on the validation set demonstrate that finetuning can improve the accuracy of the model.The comparison results of SAM2 before and after finetuning are presented in the Table. 1

Table 1. Comparison before and after finetuning SAM2

Model	epoch	J	F	J&F
base	0	75.3%	83.7%	77.5%
base	60	84.3%	88.4%	86.4%
large	0	76.8%	80.8%	78.8%
large	50	82.9%	87.1%	85.0%

4. Conclusion

In our technical report, we present a comprehensive evaluation of the SAM2 model's performance on the VISOR dataset following fine-tuning. The results, as depicted in Table 1, demonstrate significant improvements across both base and large models after undergoing 60 and 50 training epochs, respectively.

For the base model, the Jaccard Index (J) improved from 75.3

These outcomes underscore the effectiveness of fine-tuning SAM2 on the VISOR dataset, which is crucial for tasks involving video object segmentation. The substantial increase in performance metrics post-training suggests that the model has adeptly learned to handle the complexities of the VISOR dataset, leading to more accurate and reliable segmentation results. This advancement is pivotal for applications requiring precise object tracking and segmentation in egocentric videos, thereby validating the utility of SAM2 in real-world scenarios.

References

[1] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 1

[2] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *European Conference on Computer Vision*, pages 629–645. Springer, 2020. 1

[3] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 1

[4] Jiaming Zhang, Yutao Cui, Gangshan Wu, and Limin Wang. Joint modeling of feature, correspondence, and a compressed memory for video object segmentation. *arXiv preprint arXiv:2308.13505*, 2023. 2023b. 1

[5] Ho Kei Cheng, Yu Wing Tai, and Chi Keung Tang. Rethinking space - time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021. 2021a. 1

[6] 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. 1

[7] 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

[8] 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

[9] Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with sam2. *arXiv preprint arXiv:2411.17576v1*, 2024. License: CC BY 4.0; Submitted on 26 Nov 2024. 1