SAM2 for EPIC-KITCHENS VISOR Semi-supervised VOS

Yiqing Wang, Jing He, Lingling Li, Yuting Yang, Long Sun
Intelligent Perception and Image Understanding Lab, Xidian University
Xi'an, Shannxi Province, 710071, China
{24171213882, 24171213874@stu.xidain.edu.cn

Abstract

The Video Object Segmentation (VOS) task on the EPIC-Kitchen VISOR dataset requires continuous segmentation of the M annotated objects in the first frame of video subsequences across subsequent frames, allowing for occlusion, disappearance, and reappearance of objects while excluding those not present in the initial frame. Evaluation follows the DAVIS standard protocol, using the Jaccard Index (J) and Boundary F-Measure (F) to assess generalization capabilities in unseen kitchen scenarios. Our team employed the SAM2 model and achieved a J&F-Mean score of 87.5% on the test set through a series of fine-tuning steps on the EPIC-Kitchen VISOR dataset, demonstrating the effectiveness and superiority of our approach.

1. Introduction

1.1. Datasets

EPIC-KITCHENS VISOR is a large-scale dataset built upon EPIC-KITCHENS-100, providing 272,000 manual semantic masks across 50,700 images of 36 hours, 9.9 million interpolated dense masks, and 67,000 hand-object relation annotations, covering 257 object classes and 36 hours of videos. The dataset employs an AI-assisted annotation pipeline to ensure quality, supporting modeling of long-term object transformations and generalization evaluation in unseen scenarios. Baseline models demonstrate performance variations on the validation set. Released under the CC BY-NC 4.0 license, EPIC-KITCHENS VISOR serves as a critical research resource for fields such as video understanding and embodied intelligence.

The VOS task on this dataset mainly has the following two difficulties:

- Ensuring short-term and long-term consistency of pixel-level annotations when objects undergo transformative interactions. For example, when an onion is peeled, chopped, and cooked, we need to obtain accurate pixel-level annotations of onion skins, onion pieces, chopping boards, knives, pans, and acting hands.
- Any of the M objects may be occluded or invisible and reappear in subsequences.

1.2. Related Works

In recent years, the field of semi-supervised video object segmentation (VOS) has seen profound explorations centered on spatio-temporal information modeling and object semantic association, with a series of innovative approaches emerging continuously. Spatio-temporal memory networks and their improved paradigms (such as STM[5] and STCN[1]) store historical frame features by constructing external memory and achieve long-term dependency modeling through dynamic update mechanisms, effectively addressing object occlusion and appearance changes. Parallel co-attention networks and edge attention gated graph convolutional networks based on attention mechanisms strengthen target boundary and region consistency reasoning by mining inter-frame feature interactions and superpixel spatio-temporal correlations. Feature association-oriented frameworks like STMA[3] and AOT[7] Transformer models achieve precise cross-frame matching and efficient decoding of multiple objects through multi-level feature interaction and unified embedding spaces. The Recurrent Dynamic Embedding (RDE[4]) model significantly reduces error accumulation in long-video segmentation and improves robustness to mask quality through adaptive memory bank updates and self-correction strategies.

Recently, the Cutie model[2] has stood out with its innovative object-level memory reading mechanism. This mechanism introduces object queries and an object transformer to achieve top-down semantic guidance and bottom-up pixel feature interaction. The foreground-background masked attention divides object queries into foreground and background groups, forcibly separating semantic information to effectively avoid cross-region interference. However, in scenarios with numerous objects and severe mutual occlusion, Cutie may suffer from inaccurate target fea-

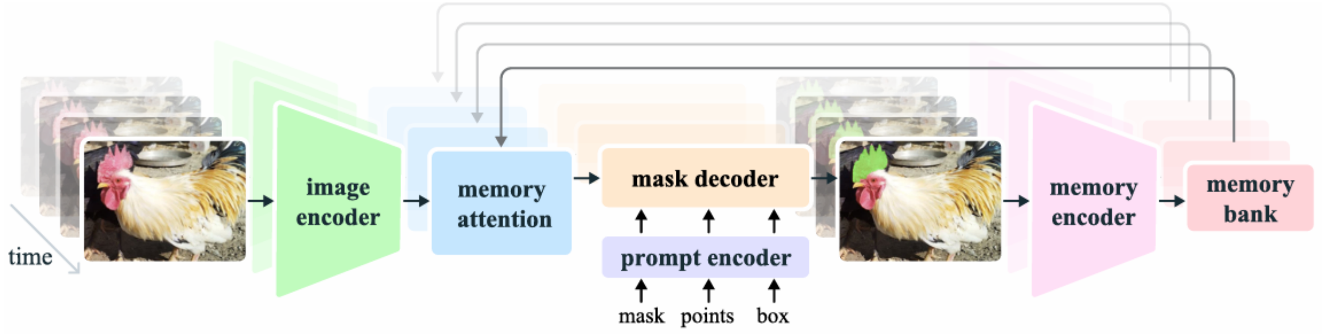

Figure 1. The SAM2 architecture. For a given frame, the segmentation prediction is conditioned on the current prompt and/or on previously observed memories. Videos are processed in a streaming fashion with frames being consumed one at a time by the image encoder, and cross-attended to memories of the 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.

-

ture extraction due to excessive interference of semantic information between objects. Additionally, when objects move extremely fast and exhibit abrupt appearance changes between adjacent frames, its feature matching mechanism based on object queries may fail to keep up with such rapid changes, leading to temporary target tracking loss or segmentation mask drift.

Meanwhile, SAM2 (Segment Anything Model 2)[6] expands segmentation capabilities from images to videos through a streaming memory architecture and ultra-large-scale datasets. Based on a Transformer structure, it dynamically fuses historical frame features via a memory attention module, supporting interactive segmentation for both single images and long videos. It has achieved state-of-the-art (SOTA) performance on multiple datasets. Therefore, our team leverages SAM2 to address the semi-supervised VOS task on the EPIC-Kitchen VISOR dataset.

2. Method

SAM2 (Segment Anything Model 2) is the first foundation model supporting unified segmentation for images and videos, addressing the challenges of interactive visual segmentation in spatio-temporal dimensions through a streaming memory architecture and large-scale dataset-driven approach. Built upon the Transformer structure, the model comprises core modules including an MAE-pretrained Hiera image encoder, a memory attention module that stores historical frame features and interaction information, and a mask decoder that inherits SAM's prompt processing logic while adding an occlusion prediction head. These components enable multi-scale feature decoding, cross-frame semantic alignment, and mask detail optimization. The overall architecture of SAM2 is shown in Figure 1.

In terms of memory mechanisms, SAM2 employs a sequential processing strategy, managing memories of up to N unprompted frames and M prompted frames through a FIFO queue. Combined with a dynamic memory update mechanism, this approach avoids explicit storage of all historical frames, reducing memory usage. The model can generate full-video mask sequences from a single-frame prompt (e.g., a click on the first frame) and requires only one-third of the interactions (compared to traditional methods) for interactive refinement.

Experimental results demonstrate that SAM2 achieves an average J&F metric of 79.3 across 17 video segmentation benchmarks, improves image segmentation accuracy by 1.4 mIoU with a 6× speedup compared to SAM, and exhibits cross-scenario robustness in geographic diversity tests, with performance discrepancies among gender and age groups of less than 3%. This establishes a new paradigm for video segmentation that balances efficiency and generalization capability.

3. Experiments

Our team first converted the EPIC-kitchen VISOR dataset into the standard DAVIS format, where each long video sequence was split into multiple sub-video sequences containing no more than 6 frames. We first performed zero-shot inference, and then conducted a series of fine-tuning on SAM2 using the training set and the combined training-validation set. Finally, it achieved a performance of 87.5% on the test set.

3.1. Zero-Shot Inference

Our team first conducted zero-shot inference on the validation set of the EPIC-Kitchen VISOR dataset using pretrained SAM2's four scale configurations (tiny, small, base_plus, and large). The inference results are shown in the Table 1. The experimental results demonstrate that SAM2 exhibits good performance without any prior exposure to this dataset. The J&F-Mean scores of all four SAM2 scales on the validation set exceeded 75%.Among them, SAM2_large achieved the best performance, with a J&F-Mean score of 77%.

Model	Score
SAM2_t	0.750578
SAM2_s	0.760589
SAM2_b+	0.766896
SAM2_l	0.770411

Table 1. The results of zero-shot inference with SAM2 on the EPIC-Kitchen VISOR dataset.

3.2. Finetuning SAM2 on EPIC-Kitechen VISOR

Our team conducted all training on two 24G 4090 GPUs. First, we fine-tuned SAM2 on the training set. During training, the length of each video sequence was set to 3, the number of objects tracked per frame was set to 3, and the batch size was 2. We first set the base learning rate to 5×10^{-6} and the vision learning rate to 3×10^{-6}, training for 80 epochs. On the validation set, the J&F-Mean reached 87.5%. Next, based on this, we adjusted the base learning rate to 1×10^{-6} and the vision learning rate to 1×10^{-6}, training for an additional 20 epochs. The J&F-Mean on the validation set reached 87.6%, and on the test set, it reached 87.1%.

Subsequently, we merged the training set and validation set. On the basis of the above training, we adjusted the fusion strategy in the FPN neck to bicubic interpolation, set the base learning rate to 5×10^{-6} and the vision learning rate to 3×10^{-6}, and trained for 40 epochs. Finally, the J&F-Mean on the test set reached 87.5%.

4. Conclusion

References

[1] Emre Aksan and Otmar Hilliges. Stcn: Stochastic temporal convolutional networks, 2019. 1

[2] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation, 2024. 1

[3] Mingcong Lei, Yiming Zhao, Ge Wang, Zhixin Mai, Shuguang Cui, Yatong Han, and Jinke Ren. Stma: A spatio-temporal memory agent for long-horizon embodied task planning, 2025. 1

[4] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1322–1331, 2022. 1

[5] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks, 2019. 1

[6] 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, 2024. 2

[7] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation, 2021. 1