# Multi-Granularity Video Object Segmentation

**Sangbeom Lim[1*], Seongchan Kim[1*], Seungjun An[2*], Seokju Cho[3], Paul Hongsuck Seo[1†], Seungryong Kim[3†]**

[1]Korea University
[2]Samsung Electronics
[3]KAIST
{limsbeom, 2020320120, phseo}@korea.ac.kr[1], sjun.an@samsung.com[2], {seokju.cho, seungryong.kim}@kaist.ac.kr[3],

## Abstract

Current benchmarks for video segmentation are limited to annotating only salient objects (i.e., foreground instances). Despite their impressive architectural designs, previous works trained on these benchmarks have struggled to adapt to real-world scenarios. Thus, developing a new video segmentation dataset aimed at tracking multi-granularity segmentation target in the video scene is necessary. In this work, we aim to generate multi-granularity video segmentation dataset that is annotated for both salient and non-salient masks. To achieve this, we propose a large-scale, densely annotated multi-granularity video object segmentation (**MUG-VOS**) dataset that includes various types and granularities of mask annotations. We automatically collected a training set that assists in tracking both salient and non-salient objects, and we also curated a human-annotated test set for reliable evaluation. In addition, we present memory-based mask propagation model (MMPM), trained and evaluated on MUG-VOS dataset, which leads to the best performance among the existing video object segmentation methods and Segment SAM-based video segmentation methods.

**Code** — https://cvlab-kaist.github.io/MUG-VOS

## Introduction

Video segmentation has been one of fundamental tasks in computer vision, aimed at predicting and tracking the masks corresponding to specified targets in video data (Yang, Fan, and Xu 2019; Kim et al. 2020; Wang et al. 2021). Video segmentation showed great performance on identifying object and consistently tracking the object, however has shown poor performance on unknown class that has not trained on supervised stage. To develop a model capable of detecting and segmenting a wide range of categories, it is necessary to construct a larger dataset that includes as many classes as possible. However, this approach is cost-inefficient and challenging, making it difficult to create a model that can handle various types of object masks.

To alleviate this burden, several video segmentation tasks such as video object segmentation (VOS) (Caelles et al. 2017) or interactive video object segmentation (Benard and Gygli 2017) provide additional references for the target by giving segmentation mask at the first frame or giving iterative refinement to clarify it. Unfortunately, these tasks also encounter performance degradation when they aim to segment non-salient objects, as they are only trained on specific targets, such as objects and stuff, according to the training dataset. Their utility in real-world applications often struggles in certain cases, such as interactive video editing and open-world video understanding, which require to return valid segmentation mask tracks for any given segmentation prompt.

Regardless of the model's architecture, segmenting anything in a video scene requires extensive data. SA-1B (Kirillov et al. 2023), an image segmentation dataset for the Segment Anything task, averages 100 masks per image. In contrast, most video segmentation datasets (Athar et al. 2023b) have far fewer mask tracks per video and rely on costly human annotation, which, while reliable, is resource-intensive.

In this work, we introduce **Mu**lti-**G**ranularity **V**ideo **O**bject **S**egmentation (**MUG-VOS**) dataset to extend the success of segment anything to the video domain. Our goal is to provide annotations of varying granularity for target masks, covering a range of objects, parts, and backgrounds. MUG-VOS fundamentally differs from previous video segmentation datasets (Xu et al. 2018; Pont-Tuset et al. 2017; Wang et al. 2021) by addressing limitations of traditional Video Instance Segmentation and Video Panoptic Segmentation, which often relies on a closed vocabulary, such as the 40 categories in YouTube-VIS (Yang, Fan, and Xu 2019) or the 124 classes in VIPSeg (Miao et al. 2022a). While Open-vocab Video Segmentation (Wang et al. 2023) manages unlimited classes via natural language descriptions, it typically focuses on salient objects and often miss background elements or partial objects.

To build a large-scale video segmentation dataset capable of segmenting various targets, we utilized SAM, which can segment anything in images. We propose a SAM-based data collection pipeline that leverages a simple IoU-based tracking method to generate diverse training data. This resulted in a new dataset with 77k video clips and 47M masks, designed to train and evaluate segmentation and tracking at multiple granularities masks. MUG-VOS includes diverse mask tracks—salient, non-salient, and partial objects—enabling

[*]These authors contributed equally.
[†]Corresponding authors.

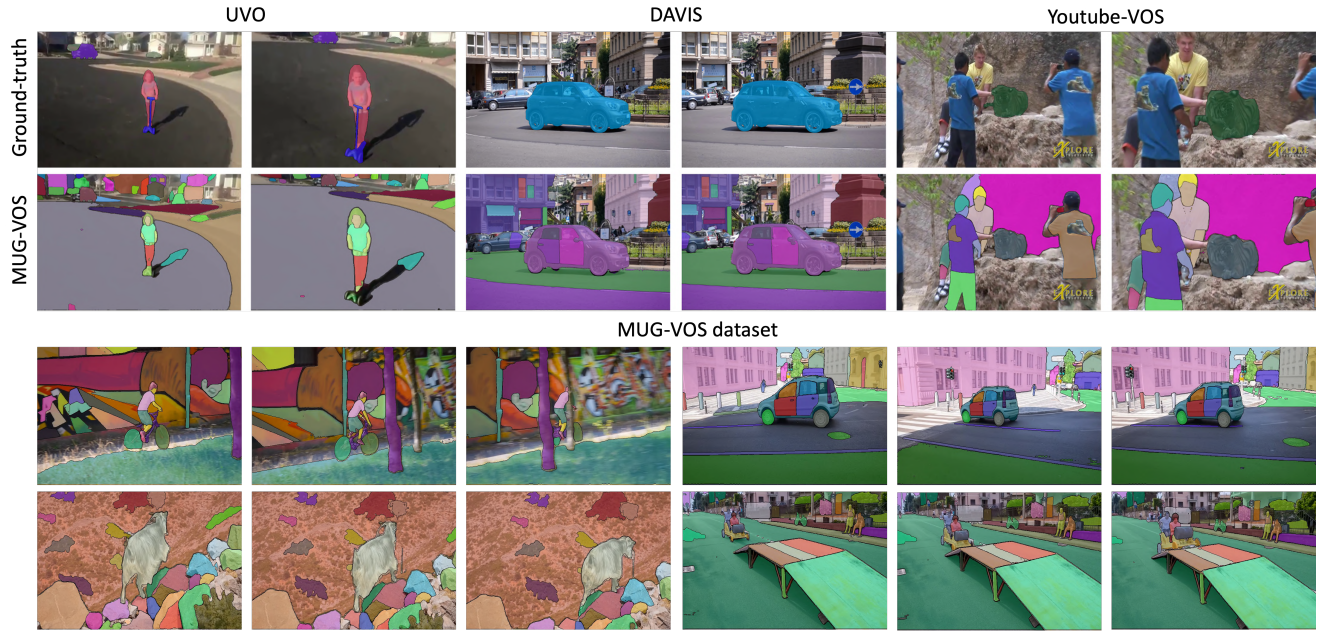


Figure 1: **Comparison of granularities of video segmentation datasets:** Visualization of **(top)** MUG-VOS masks annotated by our data collection pipeline and ground-truth masks of Youtube-VOS (Xu et al. 2018), DAVIS (Pont-Tuset et al. 2017), and UVO (Wang et al. 2021) data and **(bottom)** MUG-VOS dataset. MUG-VOS masks include various types and granularities of objects, parts, stuff, and backgrounds, even those not covered by existing datasets.

training and evaluation of models to predict previously unmanageable masks in video segmentation. Fig. 1 clearly shows the difference between the MUG-VOS dataset and existing video segmentation datasets.

On the other hand, recent attempts to adapt SAM for video segmentation (Zhou et al. 2024) have faced challenges by using SAM primarily as an image segmentation tool. For example, SAM-PT (Rajič et al. 2023) tracks points from the first frame to propagate masks, while DEVA (Cheng et al. 2023) uses SAM to generate candidate masks for each frame, while XMem (Cheng and Schwing 2022) propagates these masks by matching them with IoU.

SAM's strength lies in its ability to generate masks of varying granularity from prompts. If this capabilities extend to the video domain, it could be highly valuable for real-world applications like interactive video editing (Lee et al. 2023; Zhang et al. 2024; Lee, Cho, and Lee 2023) and generation, where traditional methods fall short.

We also propose a Memory-based Mask Propagation Model (MMPM) as a baseline for MUG-VOS. For MMPM, we utilize the pre-trained SAM encoder to leverage the rich knowledge of SAM. Specifically, we introduce a memory module that can store information about target objects from previous outputs. This memory module allows the model to generate consistent segmentation mask tracks throughout the video. MMPM processes video frames sequentially while attending the memories that are relevant to the current frame. This module enables the SAM structure to be directly applied in the video domain. Additionally, we evaluate existing video segmentation methods and MMPM on MUG-VOS test dataset. MMPM shows best performance quantitatively and qualitatively on MUG-VOS dataset.

## Related Work

### Video segmentation

The segmentation task stands out as one of the most extensively researched areas within the field of computer vision, encompassing various sub-tasks such as instance segmentation (Wu et al. 2023; Zhang et al. 2023b; Meinhardt et al. 2023), semantic segmentation (Hu et al. 2020; Wang, Wang, and Liu 2021; Liu et al. 2020), and panoptic segmentation (Li et al. 2022; Qiao et al. 2020; Athar et al. 2023a).

Video semantic segmentation focuses on segmenting the same semantic class referring to segmentation results on different frames. This method mainly focuses on improving the temporal consistency of segmented results. Video instance segmentation further challenges the complex problem of consistently identifying the same instance even on the deformation, blur, and existence of similar object classes. Video panoptic segmentation (Kim et al. 2020) further challenges to segment background stuff as previous methods have not been discovered.

Despite the remarkable performance achieved by these methods, they often struggle when confronted with unseen classes, limiting their applicability in real-world scenarios. To address this limitation, the concept of open-world segmentation (Xu et al. 2023; Liu et al. 2022) has been introduced. In this paradigm, models are designed to handle not only known classes that have been seen on the train but also unseen ones. For example, the BURST (Athar et al. 2023b) dataset contains 482 labels, which can be categorized into 78 seen classes and 404 unseen classes. Similarly, the UVO (Wang et al. 2021) dataset has been developed to facilitate open-world segmentation, aiming to build

class-agnostic segmentation models. These approaches prioritize the ability to segment unseen classes, albeit often at the expense of overall performance. In contrast, our model is specifically engineered to segment any class based on given input conditions, without compromising performance or being restricted by class distinctions.

### Segment anything

SAM (Kirillov et al. 2023) has emerged as a powerful solution for image segmentation tasks, demonstrating impressive performance. What sets SAM apart is its ability to interpret user input in various forms, including points, bounding boxes, and text. This ability to interpret any prompt format enables users to provide segmentation guidance through multiple modalities, enhancing the model's usability and flexibility. Another key strengths of SAM lies in its iterative training approach, facilitated by the utilization of SA-1B (Kirillov et al. 2023), a dataset containing various granularity of masks. Through this iterative process, SAM refines its segmentation capabilities, learning from the data in SA-1B and continuously improving its performance. This iterative training strategy plays a crucial role in enhancing SAM's effectiveness, allowing it to adapt to diverse inputs. As a result, SAM has proven to be a highly effective tool for a wide range of image segmentation applications, offering promising results and paving the way for further advancements in this field.

### Segment anything in video domain

In recent years, a variety of approaches have emerged to extend SAM to the video level. For instance, Tracking Anything (Zhu et al. 2023; Cheng et al. 2023) and DEVA (Cheng et al. 2023) fused SAM at the image level and propagated masks to future frames using a video object segmentation model. SAM-PT (Rajič et al. 2023) adopted a different strategy by sampling points from the mask and leveraging a point tracking model to estimate their positions in future frames. These estimated points were then utilized to generate masks using SAM. Meanwhile, SAM-PD (Zhou et al. 2024) tackled video segmentation as a prompt denoising task. Their method involved spreading jittered and scaled bounding boxes and then selecting them based on semantic similarity. However, these prior works employ independent methods for propagating masks. This approach introduces noise that accumulates errors as the timestep increases. In contrast, our method takes an end-to-end approach to handle video segmentation, eliminating the need for additional steps and thus mitigating error accumulation over time.

## Dataset

Existing video segmentation datasets (Xu et al. 2018; Pont-Tuset et al. 2017; Wang et al. 2021; Miao et al. 2022b; Athar et al. 2023b) typically focused on salient target objects that can be defined as specific classes. However, there currently exists no dataset for training or evaluating a diverse granularity of masks in video segmentation task.

Despite the numerous datasets for image-level segmentation, video-level segmentation datasets are scarce. Additionally, the lack of densely annotated masks in current video datasets hinders the adaptation of SAM on video segmentation. In order to train and evaluate MMPM, a large-scale video segmentation dataset is needed. To satisfy such conditions, Annotating a mask in a handcrafted way is accurate, but costs a lot. Therefore, we employ a semi-automatic data generation process involving minimal human annotators. To create a sufficiently large and diverse dataset, we developed a data collection pipeline to collect video segmentation data with varied granularity masks. Our data collection pipeline produces the segmentation data in a step-wise manner, ensuring both the quality and quantity of the masks.

### Data collection pipeline

We propose a data collection pipeline designed to autonomously derive video segmentation data from pre-existing video sources. Our data collection pipeline exploits SAM to generate dense masks for each frame and leverages an optical flow model to establish temporal connections between masks across consecutive frames, thereby generating pseudo-masks for video segmentation.

More specifically, for the given video $V = \{v_t | t \in [1, T]\}$, where $v_t$ is $t$-th frame image, and $T$ is the number of frames, target masks $M_t^i$ are defined as those obtained from the initial frame using SAM through grid point prompts. Masks from subsequent frames that are temporally consistent to the masks from the initial frame, where $i$ is the index of the track containing the target masks. Following the initial frame, prior to obtaining the target mask in the $t$-th frame, candidate masks potentially serving as the target mask are predicted.

To generate the candidate masks $C_t^i$, the points $P_t^i$ are sampled from the target mask $M_{t-1}^i$. Subsequently, the points $W_t^i$ are derived by warping the sampled points $P_t^i$ into the $t$-th frame utilizing a flow map $F_{t-1\to t}$ predicted by a flow model.

$$P_t^i = \mathrm{sample}(M_{t-1}^i) \tag{1}$$

$$W_t^i = \mathrm{warp}(P_t^i, F_{t-1\to t}) \tag{2}$$

$$F_{t-1\to t} = \mathrm{flow}(v_{t-1}, v_t) \tag{3}$$

The $\mathrm{sample}$ function denotes the process of randomly selecting point coordinates on the mask, while the $\mathrm{warp}$ function represents the warping of a point or a mask according to the flow map. The $\mathrm{flow}$ function refers to the estimation of a flow map between two adjacent frames. Candidate masks $C_t^i$ in the $t$-th frame are derived by applying the point prompts acquired from the warped points $W_t^i$ to SAM. The target mask $M_{t-1}^i$ is further warped to the $t$-th frame via the flow map $F_{t-1\to t}$ to obtain the warped mask $\tilde{M}_{t-1}^i$. Subsequently, the Intersection over Union (IoU) between the warped mask $\tilde{M}_{t-1}^i$ and the candidate masks $C_t^i$ are computed, with the candidate mask exhibiting the highest IoU being designated as the target mask $M_t^i$ in the $t$-th frame.

$$C_t^i = \mathrm{SAM}(v_t, W_t^i) \tag{4}$$

$$\tilde{M}_{t-1}^i = \mathrm{warp}(M_{t-1}^i, F_{t-1\to t}) \tag{5}$$

$$M_t^i = \arg\max_{C_t^i} \mathrm{IoU}(\tilde{M}_{t-1}^i, C_t^i) \tag{6}$$

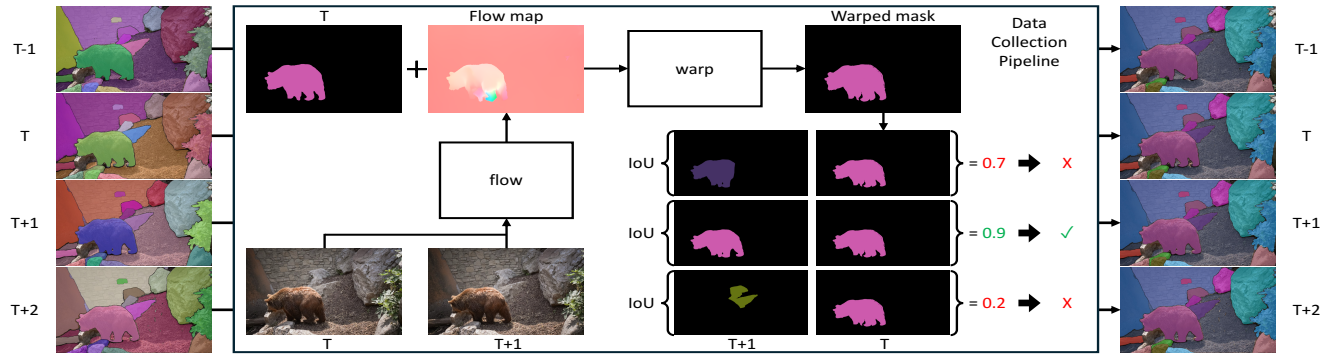


Figure 2: **MUG-VOS data collection pipeline.** We propose a data collection pipeline to generate a dataset to curate multi-granularity mask tracks completely automatically. Using SAM, we generate a large number of masks per frame and find a temporal connection through the IoU between the mask warped from the previous frame and the mask from the current frame.

**Quality assurance.** Fully automated processes can induce significant errors in the evaluation procedure. To prevent errors from accumulating in the automated process, human annotators were tasked with manually tracking and generating masks. For efficiency, annotators were instructed to accept or reject the mask tracks produced by the data collection pipeline. If an error occurred, the annotators refined the mask at the frame level using SAM. More details can be found in the appendix.

To ensure the quality of the mask tracks in the dataset, we adopted a verification process. As human annotators conduct the tracking process, completed tracks are sent to a supervisor for approval. The supervisor reviews the annotated mask tracks alongside the original video to determine whether the tracks are satisfactory or unsatisfactory. Rejected mask tracks are sent back to the annotators for refinement. Quality assurance process was only done on the MUG-VOS test dataset.

## MUG-VOS Dataset

The MUG-VOS dataset was built using our data collection pipeline and, to the best of our knowledge, is larger than any other video segmentation dataset.

**Video collection.** In the pursuit of crafting the video segmentation dataset, a judiciously selected subset of videos from the HD-VILA-100M (Xue et al. 2022) dataset was scrupulously curated for inclusion. The HD-VILA-100M dataset is devised to function as an expansive, high-resolution, and diversified video-language dataset, with the overarching goal of fostering multi-modal representation learning. This dataset encompasses a total of 3.3 M videos, characterized by their high caliber and equitable distribution across 15 categories. A subset of videos from this dataset was processed to get 77,994 video clips.

In addition, We create additional annotations for the 30 videos in the DAVIS-17(Pont-Tuset et al. 2017) validation set, which widely used in VOS task, for MUG-VOS test sets. We use SAM to generate average 29.6 mask tracks of varying types and granularities for each video.

**Data generation.** The MUG-VOS dataset facilitates training and evaluation on masks with various types and granularities, which are not covered by existing video segmentation datasets. Our data collection pipeline simplifies this process while ensuring high quality. To achieve this, we adapted our pipeline to the HD-VILA-100M and DAVIS-17 videos, resulting in a dataset that is both high quality and large in scale.

**Data statistics.** Table 1 presents the evaluation of related datasets and benchmarks for video segmentation. The density $D$ of the segmentation masks is notated in Table 1 can be given by the equation 7:

$$D = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} M_{i,j}}{H \times W} \tag{7}$$

where as $H$ is the height of the image, $W$ is the width of the image, $M_{i,j}$ is the value of the mask at pixel $(i, j)$, which is 1 if the pixel is covered by any mask and 0 otherwise.

For VOS tasks, common benchmarks include YouTube-VOS (Xu et al. 2018) and DAVIS (Pont-Tuset et al. 2017). Both datasets are annotated on in-the-wild videos, each lasting approximately 5-10 seconds. However, these datasets contain only a few mask tracks per video, which limits their ability to evaluate tracking performance across diverse masks.

The UVO (Wang et al. 2021) dataset primarily focuses on evaluating performance on unseen mask classes during training. It provides mask annotations for a variety of salient objects that humans typically recognize as "objects". While UVO covers a wide range of masks from a human perspective, it does not include non-salient objects that are not easily defined by humans, which remains as a limitation.

VIPSeg (Miao et al. 2022b) is part of the video panoptic segmentation benchmark. Video panoptic segmentation aims to predict the semantic class of every pixel in the temporal dimension. Compared to the datasets in Table 1, VIPSeg covers a higher density of pixels than any other dataset in Table 1. However, the masks in VIPSeg are much simpler compared to MUG, as the number of masks per frame is significantly lower than in MUG.

BURST (Athar et al. 2023b) is a universal dataset designed to cover multiple segmentation tasks, including video object segmentation, video instance segmentation,

| | Density | Masks | Mask Tracks | Masks per Frame | Annotated Frames |
|---|---|---|---|---|---|
| OVIS (Qi et al. 2022) | 0.186 | 296k | 5,223 | 5.8 | 51,059 |
| YT-VIS (Yang, Fan, and Xu 2019) | 0.167 | 132K | 4,866 | 1.69 | 79,260 |
| YT-VOS (Xu et al. 2018) | 0.184 | 17K | 8,614 | 1.63 | 123,265 |
| DAVIS (Pont-Tuset et al. 2017) | 0.120 | 27K | 386 | 2.6 | 10,459 |
| UVO (Wang et al. 2021) | 0.425 | 593K | 104,898 | 12.3 | 58,140 |
| BURST (Athar et al. 2023b) | 0.167 | 600K | 16,089 | 3.1 | 195,713 |
| VIPSeg (Miao et al. 2022b) | 0.979 | 926K | 38,592 | 10.9 | 84,750 |
| MOSE (Ding et al. 2023) | 0.083 | 431K | 5,200 | 2.4 | 110,067 |
| **MUG-VOS Train** | 0.714 | 47M | 4.7M | 66.3 | 77,9940 |
| **MUG-VOS Test** | 0.663 | 59K | 887 | 29.6 | 1,999 |

Table 1: **Data comparison with previous video segmentation datasets.** Comparing MUG-VOS with existing video segmentation datasets in terms of statistics.

and point-guided segmentation (Guo et al. 2024; Zulfikar et al. 2024). While the aforementioned datasets primarily focus on annotating salient objects (e.g., humans, cars, animals), MUG-VOS includes a diverse range of class-agnostic masks.

## Model

Figure 3 shows the overview of MMPM. For simplicity, the figure assumes a single object, but the MMPM model is designed to handle multiple objects simultaneously. MMPM consistently tracks objects and generates segmentation masks sequentially, frame-by-frame. Thanks to the memory module, which is designed to handle occlusion, motion blur, and deformation, the model can effectively track objects even in challenging video scenes. To build a robust video segmentation model, we incorporated two types of memory modules that function independently. For each frame, the RGB image is encoded by an image encoder, which serves as a query for the memory module. The encoded query from the image encoder accesses the memory bank to retrieve memory features that provide crucial information for generating the current mask. MMPM uses two types of memory: sequential memory and temporal memory, both of which are updated for selected frame. Frames selected for memory update are selected at regular intervals. Specifically, the temporal memory has a storage limit, $T_{max}$ , to prevent running out of memory. When the temporal memory reaches a store limit, we randomly filter out memory entries, except for those from the first and previous frames. These memory features works as a enhancement for image feature to generate high-quality masks, even in long videos.

### Temporal memory

The temporal memory stores information about previous predictions for the same target object in the video by maintaining high-resolution features for up to $T_{max}$ recent frames. We split the memory into key and value components, which effectively retrieve relevant information. The key is encoded in the same embedding space as the query feature, while the value is encoded from a value encoder that fuses semantic and spatial information by encoding the image and binary mask together. Specifically, the key $\mathbb{R}^{C^k \times THW}$ exists in the same embedding space as the query feature, while the value is encoded independently.

For a selected frame, we copy the query feature from the image encoder as a key for the current frame. The generated key is then processed along with the binary mask from the mask decoder through the value encoder. The new key and value from the current frame are then appended to the memory and provide important information for future decoding processes. When the number of temporal values reaches $T_{max}$, we randomly sample values, excluding those from the first and previous frames. The first frame's value provides initial information about the target object in the VOS task and is considered to best represent the target object. The value from the last frame contains the latest information about the target object.

This dynamic filtering of the temporal memory enables the model to represent both short-term and long-term object motion effectively.

### Sequential memory

While temporal memory focuses on high-resolution features from previous frames, sequential memory focuses on low-resolution features such as geometric, semantic, and location information. Sequential memory complements temporal memory by updating the value on selected frame.

Specifically, sequential memory $S \in \mathbb{R}^{C^s \times HW}$ is updated every select frame using the query from the current frame. To effectively propagate sequential information, we use the GRU (Cho et al. 2014) method, which dynamically forgets previous information and updates with the current frame's information. The output from the value encoder serves as a candidate for updating the sequential memory. The GRU discards outdated information from the previous sequential memory and updates the sequential memory in parallel. This propagation process compensates for the lack of object information in temporal memory and stores new low-resolution information.

## Experiments

### MUG-VOS evaluation

In order to evaluate the capability of multi-granularity segmentation in a video scene, we tested our models on the MUG-VOS dataset, which contains high-quality masks with a variety of granularities. To demonstrate that our MMPM model is best suited to the MUG-VOS dataset, we also tested

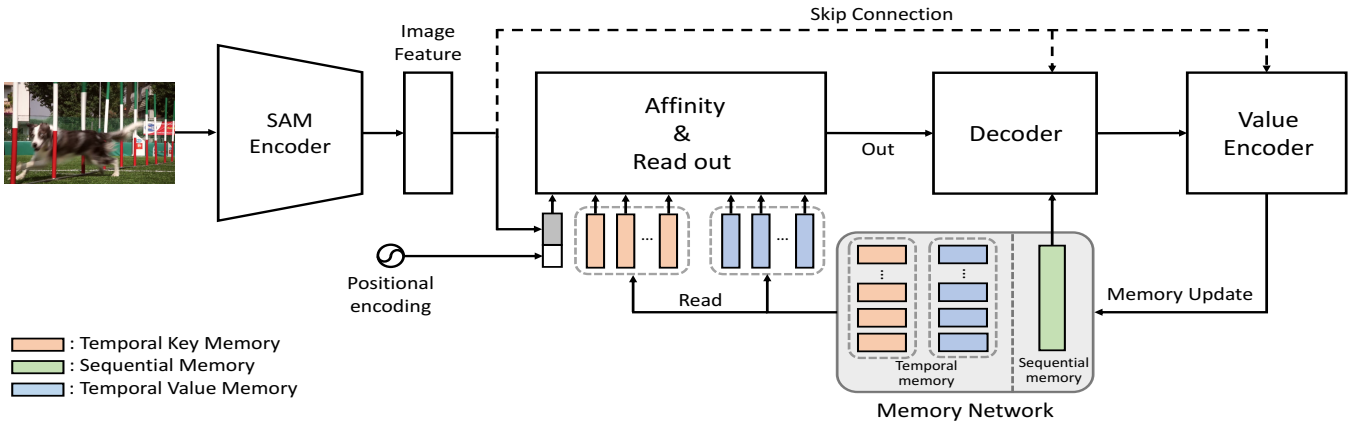


Figure 3: **MMPM overview.** We introduce the MMPM model, which generates masks based on previous results. Starting from an initial mask that indicates the target object, the MMPM model consistently tracks and segments the target throughout the entire video. Sequential memory stores low-resolution features, updated at every selected frames, while temporal memory retains high-resolution features from previous frames, capturing a variety of information gathered from multiple frames.

| MUG-VOS Test | SAM | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|---|
| *zero-shot transfer* | | | | |
| SAM (Grid 32) + IoU | ✓ | 22.2 | 20.2 | 24.2 |
| SAM (Grid 64) + IoU | ✓ | 26.7 | 24.5 | 28.9 |
| PerSAM (Zhang et al. 2023a) | ✓ | 33.1 | 30.8 | 35.3 |
| PerSAM-F (Zhang et al. 2023a) | ✓ | 41.5 | 39.1 | 43.9 |
| SAM-PT (Rajič et al. 2023) | ✓ | 78.3 | 76.2 | 80.4 |
| XMem (Cheng and Schwing 2022) | - | 83.0 | 86.9 | 79.1 |
| DEVA (Cheng et al. 2023) | - | 85.6 | 88.2 | 82.9 |
| MMPM | - | 86.1 | 86.0 | 86.1 |

Table 2: Quantitative evaluation of video object segmentation on DAVIS-2017 (Pont-Tuset et al. 2017) validation set. All models are trained on the MUG-VOS train set. "SAM" corresponds to architectures using SAM (Kirillov et al. 2023) Encoder and Decoder.

the MUG-VOS dataset with other existing models (Zhang et al. 2023a; Rajič et al. 2023; Cheng and Schwing 2022; Cheng et al. 2023). Specifically, since our MMPM leverages SAM's knowledge for video segmentation, we implemented a SAM-based baseline model that retrieves the mask proposal with the highest IoU score relative to the previous frame. The masks were initialized by providing grid points to the SAM on every frame and subsequently the mask tracks were generated by sequentially selecting the mask with the highest IoU score.

Table 2 shows the quantitative results on MUG-VOS dataset. For the SAM-based baseline, "Grid" refers to the number of points per side when sampling points from the grid. The baseline demonstrated poor performance because it failed to generate the current frame's mask without relying on prior information from the previous frame. Additionally, relying solely on IoU for track connection makes it vulnerable to significant motion changes and occlusions. Although XMem and DEVA have shown great performance on the VOS task, they performed poorly on the MUG-VOS test set. Our experiments not only demonstrate that the MMPM model achieved the highest score compared to other models but also highlight that the MUG-VOS dataset is more challenging compared to existing VOS benchmarks. Previous VOS benchmarks only evaluate performance on salient objects, allowing models to be tuned to find the most prominent objects. However, to be effective in real-world settings, models should be able to segment a variety of objects regardless of the spatial position of the target.

| Methods | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| Not filter | 85.0 | 85.7 | 84.2 |
| FIFO | 85.9 | 85.8 | 85.9 |
| Random | 86.0 | 85.9 | 86.0 |
| + P. first & last | 86.1 | 86.0 | 86.1 |

(a) **Memory filtering methods.**

| Methods | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| No memory | 81.8 | 85.3 | 78.2 |
| + Only sequential | 81.9 | 85.3 | 78.5 |
| + Only temporal | 85.7 | 85.9 | 85.8 |
| + Both | 86.1 | 86.0 | 86.1 |

(b) **Memory usage types.**

Table 3: Ablation studies on (a) memory filtering methods and (b) number of temporal memory values.

Figure 4 shows the qualitative results compared with other methods. We will also provide more qualitative results in the appendix. MMPM demonstrated superior performance both quantitatively and qualitatively.

## Ablation study

We conducted ablation studies on the MMPM model using the MUG-VOS test dataset. These studies explore the effects


Figure 4: Qualitative comparison between MMPM, DEVA (Cheng et al. 2023), PerSAM-F (Zhang et al. 2023a), and SAM-PT (Rajič et al. 2023) from MUG-VOS test set.

| $r$ | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| 1 | 66.6 | 50.4 | 82.8 |
| 3 | 68.3 | 52.2 | 84.4 |
| 5 | 86.1 | 86.0 | 86.1 |

(a) **Memory update interval.**

| $N$ | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| 5 | 85.7 | 85.8 | 85.5 |
| 10 | 86.1 | 86.0 | 86.1 |
| 15 | 86.1 | 86.0 | 86.1 |

(b) **Number of temporal memory values.**

Table 4: Ablation studies on (a) memory update interval and (b) memory usage types.

of different memory filtering methods, the number of memory values, memory update intervals, and the characteristics of each memory module.

Table 3a shows the performance of the MMPM model with different memory filtering methods. Firstly, no filtering is applied (i.e. Not filter) when memory reaches the limit $T_{max}$ and the memory retains the first inserted values for the entire video. Secondly, to append the latest values to memory, we implemented the First-In-First-Out (FIFO) method. Performance gain from FIFO method enables MMPM model to generate segmentation masks with the latest information enables the model to find corresponding objects efficiently. Thirdly, randomly filtering values from the memory allows the model to access diverse memories with random interval. Lastly, applying a constraint to the random filtering method to preserve the first and last frame values (i.e. P. first & last), enhanced model performance compared to any other methods. Since, in VOS tasks, the initial mask from the first frame best represents the target object, and providing the latest mask to the MMPM model has a similar effect to the initial mask.

Table 3b shows the performance comparison between different memory usage types. To demonstrate the advantage of memory usage, we experimented with the MMPM model without memory values, where the mask is generated by referring only to the previous mask, indicated as 'No memory'. Adding sequential memory, which is updated from selected frame, improved performance. The low-resolution sequential memory helps the model retrieve the target object. When temporal memory is added, the high-resolution information further enhances the model's ability to segment high-quality masks, as temporal memory stores diverse target information (e.g., shape, color, motion) from multiple frames.

Table 4a shows the performance variation based on the interval period, denoted as *r*, which determines when to update the memory. Increasing the interval between memory updates allows the model to perform faster inference while enabling it to store diverse information in memory. This also demonstrates that our model is robust enough to handle significant appearance changes while effectively mitigating drifting and error accumulation (Oh et al. 2019; Yang, Wei, and Yang 2021). However, selecting an interval value that is too high can result in the loss of important intermediate information.

Table 4b shows the performance variation when implementing different number of memory values in the temporal memory, denoted as *N*. Increasing number of value stored in temporal memory gave MMPM model to look over various memories. Abundant information from past frames enabled model to enhance to image embedding, resulting to MMPM to generate high quality segmentation mask (Duke et al. 2021). Interestingly, MMPM model performance has converged around the number of values 10.

## Conclusion

In this work, we introduce the MUG-VOS dataset for training and evaluating multi-granularity masks in the video domain. We developed a data collection pipeline to produce a large-scale, densely annotated video segmentation dataset. MUG-VOS features diverse segmentation masks, covering both salient and non-salient objects. Additionally, we present MMPM, which outperforms existing models on the MUG-VOS test set. We anticipate that MUG-VOS will be a valuable benchmark for advancing multi-granularity.

## Acknowledgments

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279,RS-2024-00436857, RS-2020-II201819) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2023-00266509, RS-2024-00333068), and National Research Foundation of Korea (RS-2024-00346597) and the Hyundai Motor Chung Mong-Koo Foundation and the National Supercomputing Center with supercomputing resources including technical support (KSC-2023-CRE-0416).

## References

Athar, A.; Hermans, A.; Luiten, J.; Ramanan, D.; and Leibe, B. 2023a. TarViS: A Unified Architecture for Target-based Video Segmentation. In *CVPR*.

Athar, A.; Luiten, J.; Voigtlaender, P.; Khurana, T.; Dave, A.; Leibe, B.; and Ramanan, D. 2023b. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1674–1683.

Benard, A.; and Gygli, M. 2017. Interactive video object segmentation in the wild. *arXiv preprint arXiv:1801.00269*.

Caelles, S.; Maninis, K.-K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; and Van Gool, L. 2017. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 221–230.

Cheng, H. K.; Oh, S. W.; Price, B.; Schwing, A.; and Lee, J.-Y. 2023. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1316–1326.

Cheng, H. K.; and Schwing, A. G. 2022. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, 640–658. Springer.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Ding, H.; Liu, C.; He, S.; Jiang, X.; Torr, P. H.; and Bai, S. 2023. MOSE: A New Dataset for Video Object Segmentation in Complex Scenes. In *ICCV*.

Duke, B.; Ahmed, A.; Wolf, C.; Aarabi, P.; and Taylor, G. W. 2021. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5912–5921.

Guo, D.; Fan, D.-P.; Lu, T.; Sakaridis, C.; and Van Gool, L. 2024. Vanishing-Point-Guided Video Semantic Segmentation of Driving Scenes. *arXiv preprint arXiv:2401.15261*.

Hu, P.; Caba, F.; Wang, O.; Lin, Z.; Sclaroff, S.; and Perazzi, F. 2020. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8818–8827.

Kim, D.; Woo, S.; Lee, J.-Y.; and Kweon, I. S. 2020. Video Panoptic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Lee, S.; Cho, S.; and Lee, S. 2023. One-shot video inpainting. *arXiv preprint arXiv:2302.14362*.

Lee, Y.-C.; Jang, J.-Z. G.; Chen, Y.-T.; Qiu, E.; and Huang, J.-B. 2023. Shape-aware text-driven layered video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14317–14326.

Li, X.; Zhang, W.; Pang, J.; Chen, K.; Cheng, G.; Tong, Y.; and Loy, C. C. 2022. Video k-net: A simple, strong, and unified baseline for video segmentation. In *CVPR*.

Liu, Y.; Shen, C.; Yu, C.; and Wang, J. 2020. Efficient Semantic Video Segmentation with Per-frame Inference. *ECCV*.

Liu, Y.; Zulfikar, I. E.; Luiten, J.; Dave, A.; Ramanan, D.; Leibe, B.; Ošep, A.; and Leal-Taixé, L. 2022. Opening up open world tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19045–19055.

Meinhardt, T.; Feiszli, M.; Fan, Y.; Leal-Taixé, L.; and Ranjan, R. 2023. NOVIS: A Case for End-to-End Near-Online Video Instance Segmentation. *ArXiv*, abs/2308.15266.

Miao, J.; Wang, X.; Wu, Y.; Li, W.; Zhang, X.; Wei, Y.; and Yang, Y. 2022a. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21033–21043.

Miao, J.; Wang, X.; Wu, Y.; Li, W.; Zhang, X.; Wei, Y.; and Yang, Y. 2022b. Large-Scale Video Panoptic Segmentation in the Wild: A Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21033–21043.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9226–9235.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

Qi, J.; Gao, Y.; Hu, Y.; Wang, X.; Liu, X.; Bai, X.; Belongie, S.; Yuille, A.; Torr, P. H.; and Bai, S. 2022. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039.

Qiao, S.; Zhu, Y.; Adam, H.; Yuille, A.; and Chen, L.-C. 2020. ViP-DeepLab: Learning Visual Perception with

Depth-aware Video Panoptic Segmentation. *arXiv preprint arXiv:2012.05258*.

Rajič, F.; Ke, L.; Tai, Y.-W.; Tang, C.-K.; Danelljan, M.; and Yu, F. 2023. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*.

Wang, H.; Wang, W.; and Liu, J. 2021. Temporal memory attention for video semantic segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2254–2258. IEEE.

Wang, H.; Yan, C.; Wang, S.; Jiang, X.; Tang, X.; Hu, Y.; Xie, W.; and Gavves, E. 2023. Towards open-vocabulary video instance segmentation. In *proceedings of the IEEE/CVF international conference on computer vision*, 4057–4066.

Wang, W.; Feiszli, M.; Wang, H.; and Tran, D. 2021. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10776–10785.

Wu, J.; Jiang, Y.; Liu, Q.; Yuan, Z.; Bai, X.; and Bai, S. 2023. General object foundation model for images and videos at scale. *arXiv preprint arXiv:2312.09158*.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023. Open-Vocabulary Panoptic Segmentation With Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2955–2966.

Xu, N.; Yang, L.; Fan, Y.; Yue, D.; Liang, Y.; Yang, J.; and Huang, T. 2018. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*.

Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; and Guo, B. 2022. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5188–5197.

Yang, Z.; Wei, Y.; and Yang, Y. 2021. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34: 2491–2502.

Zhang, R.; Jiang, Z.; Guo, Z.; Yan, S.; Pan, J.; Ma, X.; Dong, H.; Gao, P.; and Li, H. 2023a. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*.

Zhang, T.; Tian, X.; Zhou, Y.; Ji, S.; Wang, X.; Tao, X.; Zhang, Y.; Wan, P.; Wang, Z.; and Wu, Y. 2023b. DVIS++: Improved Decoupled Framework for Universal Video Segmentation. arXiv:2312.13305.

Zhang, Z.; Wu, B.; Wang, X.; Luo, Y.; Zhang, L.; Zhao, Y.; Vajda, P.; Metaxas, D.; and Yu, L. 2024. AVID: Any-Length Video Inpainting with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7162–7172.

Zhou, T.; Luo, W.; Ye, Q.; Shi, Z.; and Chen, J. 2024. SAM-PD: How Far Can SAM Take Us in Tracking and Segmenting Anything in Videos by Prompt Denoising. *arXiv preprint arXiv:2403.04194*.

Zhu, J.; Chen, Z.; Hao, Z.; Chang, S.; Zhang, L.; Wang, D.; Lu, H.; Luo, B.; He, J.-Y.; Lan, J.-P.; et al. 2023. Tracking anything in high quality. *arXiv preprint arXiv:2307.13974*.

Zulfikar, I. E.; Mahadevan, S.; Voigtlaender, P.; and Leibe, B. 2024. Point-VOS: Pointing Up Video Object Segmentation. *arXiv preprint arXiv:2402.05917*.