Training-Free Spatio-temporal Decoupled Reasoning Video Segmentation with Adaptive Object Memory

Zhengtong Zhu, Jiaqing Fan*, Zhixuan Liu, Fanzhang Li

School of Computer Science & Technology, Soochow University, Suzhou, China
20245227002@stu.suda.edu.cn, jqfan@suda.edu.cn, 2427405039@stu.suda.edu.cn, lfzh@suda.edu.cn

Abstract

Reasoning Video Object Segmentation (ReasonVOS) is a challenging task that requires stable object segmentation across video sequences using implicit and complex textual inputs. Previous methods fine-tune Multimodal Large Language Models (MLLMs) to produce segmentation outputs, which demand substantial resources. Additionally, some existing methods are coupled in the processing of spatio-temporal information, which affects the temporal stability of the model to some extent. To address these issues, we propose Training-Free **S**patio-temporal **D**ecoupled Reasoning Video Segmentation with **A**daptive Object **M**emory (SDAM). We aim to design a training-free reasoning video segmentation framework that outperforms existing methods requiring fine-tuning, using only pre-trained models. Meanwhile, we propose an Adaptive Object Memory module that selects and memorizes key objects based on motion cues in different video sequences. Finally, we propose Spatio-temporal Decoupling for stable temporal propagation. In the spatial domain, we achieve precise localization and segmentation of target objects, while in the temporal domain, we leverage key object temporal information to drive stable cross-frame propagation. Our method achieves excellent results on five benchmark datasets, including Ref-YouTubeVOS, Ref-DAVIS17, MeViS, ReasonVOS, and ReVOS.

Code — https://github.com/machine928/SDAM

Introduction

The ReasonVOS task was first proposed by VISA(Yan et al. 2025), which differs from traditional Referring Video Object Segmentation (RefVOS) tasks (Yan et al. 2024; He and Ding 2024; Cuttano et al. 2025; Liang et al. 2025). It requires the stable segmentation of target objects in video sequences based on implicit and complex textual inputs, demanding advanced reasoning abilities. Due to the dynamic nature of video data, this task becomes even more challenging, where time-sensitive queries, occlusion, or rapid object movement may complicate the segmentation process.

The ReasonVOS task is an evolution of the RefVOS task, presenting a greater challenge. Previous methods based on MTTR (Botach, Zheltonozhskii, and Baskin 2022; Wu et al. 2022; Zhu et al. 2025) primarily focus on image-level text-image correspondence, mainly segmenting objects in images by analyzing the given text descriptions. However, these methods suffer from coupling in the handling of spatio-temporal information and struggle to fully consider the temporal dependencies between frames, leading to difficulties in consistently producing accurate segmentation results in dynamic and complex scenes. For instance, we observe that in some videos on the Ref-YouTubeVOS dataset, the $\mathcal{J}\&\mathcal{F}$ metric experiences sharp declines in certain rapidly changing frames, which is caused by segmentation ambiguities arising from scene changes. Additionally, some methods at-

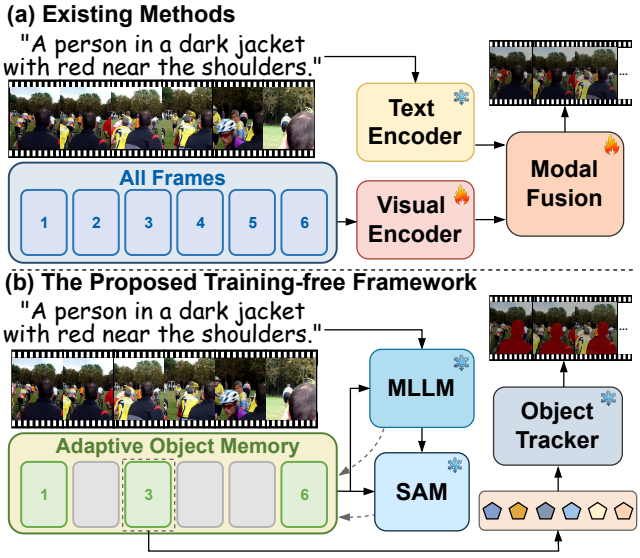

Figure 1: Overview of SDAM. (a) Some existing methods rely on image-level video understanding, establishing correspondence between text descriptions and video frames, while neglecting the inherent temporal information in video tasks. (b) Our **training-free** framework adaptively memorizes key object information based on motion cues in the video and the frame-level confidence jointly obtained from MLLM and SAM. Additionally, we decouple the spatio-temporal information in the video to enhance the temporal stability of the architecture during the segmentation process.

*Corresponding author.

tempt to propagate keyframe masks to achieve stable outputs across the entire video sequence (Cho et al. 2025). However, these methods rely on a fixed global sampling strategy to obtain keyframe candidates, which is challenging to account for the motion cues present in different video sequences. As a result, the spatio-temporal information contained in the candidate frame sequence is limited, making it difficult to obtain accurate keyframe. Moreover, to ensure good text-image correspondence, existing methods typically require significant time and computational resources for model training, this undoubtedly limits the development of the task.

To address these challenges, we propose SDAM, a novel framework based on MLLM with training-free characteristics (Figure 1). Unlike previous methods, SDAM does not solely rely on traditional frame-level image understanding for segmentation. Instead, it combines motion cues and frame-level confidence in a spatio-temporal decoupling approach to consistently output masks. Specifically, our framework integrates existing pre-trained MLLMs (Bai et al. 2025), image segmentation models (Kirillov et al. 2023), and efficient object trackers (Cheng et al. 2024) to create a training-free reasoning video segmentation architecture. First, we analyze the motion information in the video sequence and adaptively select keyframe candidates, thereby obtaining a keyframe candidate set with richer spatio-temporal information (as shown in the Figure 1). Then, using the outputs of MLLM and SAM, we compute confidence values for these candidate frames. The object with the highest confidence will be stored in the object memory bank for later use. Finally, with the help of the object tracker, we effectively track key objects throughout the video sequence, ensuring that the segmentation process remains stable and continuous across the entire video.

To evaluate the effectiveness of our method, we conducted experiments on three RefVOS benchmark datasets (Ref-DAVIS 2017 (Khoreva, Rohrbach, and Schiele 2019), Ref-YouTube-VOS(Seo, Lee, and Han 2020) and MeViS (Ding et al. 2023a)) and two ReasonVOS benchmark datasets (ReasonVOS (Bai et al. 2024) and ReVOS (Yan et al. 2025)). The quantitative and qualitative experimental results demonstrate the superiority of our approach. In summary, we highlight the following main contributions:

- We propose a novel training-free reasoning video segmentation framework that surpasses existing fine-tuned methods by leveraging only pre-trained models.
- We adaptively select and memorize key objects by leveraging motion cues and frame-level confidence across the sequence, thereby capturing richer spatial information.
- We design this architecture from the perspective of spatio-temporal decoupling, first recognizing in the spatial domain and then propagating in the temporal domain, enhancing the temporal stability of the output.
- Our method achieves state-of-the-art results on five benchmark datasets. Specifically, we obtain $\mathcal{J}\&\mathcal{F}$ scores of **65.3**%, **76.0**%, **48.6**%, **55.1**%, and **58.0**% on the Ref-YouTube-VOS, Ref-DAVIS17, MeViS, ReasonVOS, and ReVOS datasets, respectively.

Related Work

Referring Video Object Segmentation (RefVOS). RefVOS aims to segment the target objects in a video based on a given explicit language expression. A recent benchmark, MeViS (Ding et al. 2023a) introduces complex multi-object scenes with extensive motion dynamics, further increasing the challenge of this task. Common approaches (Luo et al. 2023; Tang, Zheng, and Yang 2023; Yan et al. 2024; Liang et al. 2025; Cuttano et al. 2025) focus on attention mechanisms that use language queries to highlight the objects of interest. Additionally, some works (Ding et al. 2023a; He and Ding 2024; Pan et al. 2025; Fang et al. 2025) have proposed motion aggregation techniques to capture motion information and methods like stable diffusion models to achieve modality fusion (Zhu et al. 2025).

Reasoning Segmentation. Reasoning Segmentation (Lai et al. 2024; Rasheed et al. 2024; Xia et al. 2024) has further advanced referring segmentation by generating masks from complex images and implicit text. LISA (Lai et al. 2024) pioneers the field of inference segmentation. It introduces a new token to extend the vocabulary and proposes embedding as a masking paradigm that enhances segmentation to address scenarios requiring complex reasoning and world knowledge. PixelLM (Ren et al. 2024) integrates an innovative pixel decoder and a segmentation codebook with trainable tokens, enabling the efficient generation of high-quality masks without relying on external models. VISA (Yan et al. 2025) brings Reasoning Segmentation into the video domain. It fine-tunes pre-trained MLLM to select keyframes, segments them based on inference, and uses a pret-rained object tracker to propagate masks to other frames. However, it encounters challenges such as the limited representational capacity of a single specialized token and inaccurate keyframe selection, which hinder its segmentation and tracking performance. VRS-HQ (Gong et al. 2025) designs two special tokens and uses autoregressive learning of MLLM to effectively capture both local and global information.

Method

Overall Training-Free Pipeline

The overall pipeline of SDAM is shown in Figure 2. We integrate MLLM (Bai et al. 2025), SAM (Kirillov et al. 2023), and Cutie (Cheng et al. 2024) to form this framework, which mainly consists of two modules: AOM and SD. The AOM module is composed of the Motion Driven Sampler (MDS), Joint Keyframe Selection (JKS), and the Object Memory Bank (OMB). This module adaptively selects and memorizes key objects throughout the video sequence based on motion cues in the video. The SD module integrates MLLM, SAM and Cutie. In the spatial domain, we first input keyframe candidates into both MLLM and SAM. MLLM determines the location of the candidate objects in each frame, and then this location information is passed to SAM to obtain the object masks. In the temporal domain, we feed the outputs of MLLM and SAM into the AOM module to obtain the key object memory, then utilize the object tracking and segmentation capabilities of Cutie to achieve object segmentation across the entire video.

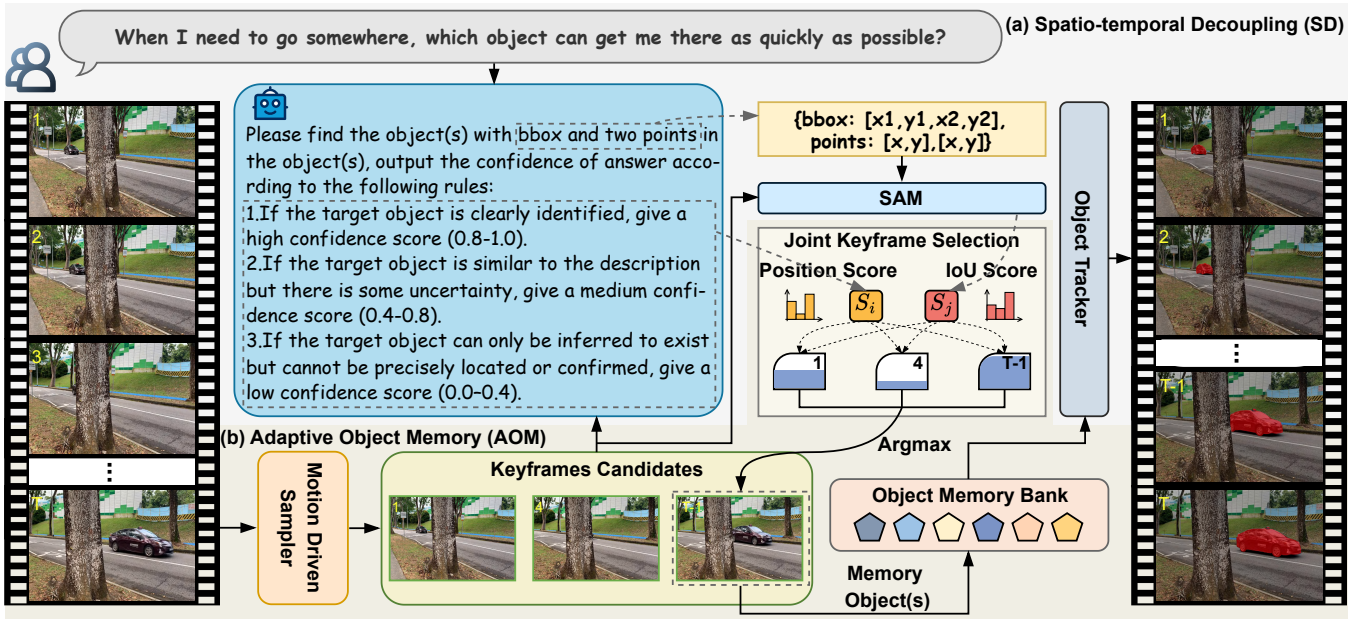

Figure 2: The overall pipeline of SDAM. Our method consists of two parts: (a) Spatio-temporal Decoupling (SD). We pass the keyframe candidates into MLLM and SAM to obtain objects information in the spatial domain, and then use the Object Tracker to propagate the key object across the temporal domain. (b) Adaptive Object Memory (AOM). We first use Motion Driven Sampler to adaptively sample keyframe candidates based on motion cues, then use Joint Keyframe Selection to select the frame with the highest confidence as the keyframe, and store the key object memory in the Object Memory Bank.

Input and Output Settings. Given a video sequence $I = \{I_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{T}$ and a text query q, where T is the length of the original video sequence, we input I into MDS to obtain keyframe candidates $I' = \{I_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{T'}$, with T' representing the length of the keyframe candidate sequence. Then, we input I' and q into MLLM $\mathcal{F}_M(\cdot)$ and input I' into SAM $\mathcal{F}_S(\cdot)$. Subsequently, we obtain the position information of the candidate objects:

$$P = \mathcal{F}_M(I', q) = \{P_i \in \mathbb{R}^N\}_{i=1}^{T'}, \quad (1)$$

where N represents the number of candidate objects in each frame, and P_i is output in JSON format. At the same time, $\mathcal{F}_M(\cdot)$ also outputs the confidence scores $S_{mllm} = \{S_i \in \mathbb{R}^1\}_{i=1}^{T'}$ for each frame. After processing P, it is input into $\mathcal{F}_S(\cdot)$ to obtain the predicted masks for the candidate frames:

$$M' = \mathcal{F}_S(I', P) = \{M_i \in \mathbb{R}^{H\times W\times N}\}_{i=1}^{T'}. \quad (2)$$

Additionally, the IOU scores for each frame $S_{sam} = \{S_j \in \mathbb{R}^1\}_{j=1}^{T'}$ are output. We then input S_{mllm} and S_{sam} into JKS to determine the keyframe $I_{key} \in \mathbb{R}^{H\times W\times 3}$ and its mask $M_{key} \in \mathbb{R}^{H\times W\times N}$, the key object memory O_{key} is obtained as follows:

$$O_{key} = Memory(I_{key}, M_{key}), \quad (3)$$

where $O_{key} \in \mathbb{R}^{N\times C}$, with C being the number of channels for object memory. We input O_{key} into the object memory bank $O = \{O_i \in \mathbb{R}^{N\times C}\}_{i=1}^{T}$, then use the Object Tracker $\mathcal{F}_{Obj}(\cdot)$ to propagate O_{key} across the entire video sequence, resulting in the final masks output:

$$M = \mathcal{F}_{Obj}(I, O) = \{M_i \in \mathbb{R}^{H\times W\times N}\}_{i=1}^{T}. \quad (4)$$

Adaptive Object Memory

In this section, we will provide a detailed description of the design and implementation of the AOM module. Existing sampling-based methods (Cho et al. 2025; Kao, Tai, and Tang 2025) typically adopt a global sampling strategy with fixed intervals. While this approach is simple to implement, it has significant limitations, especially when handling complex text queries. Global sampling may miss critical nodes in the video, particularly when specific actions mentioned in the text description occur at particular moments. For example, given the text query "bird stand on hand, then fly away," the action of "fly away" occurs only in a few frames of the video. If global sampling with fixed intervals is used, it may fail to accurately capture these important time points.

Motion Driven Sampler. To address the above issues, we propose an adaptive keyframe sampling strategy based on motion cues from the video sequence. By analyzing motion information and scene changes within the video, the MDS module can dynamically adjust the sampling strategy to select the frames that best reflect key actions. This adaptive sampling strategy not only improves the hit rate of keyframes but also enhances the spatio-temporal information of the sampling results, ensuring that the core events and actions described in the text are effectively captured.

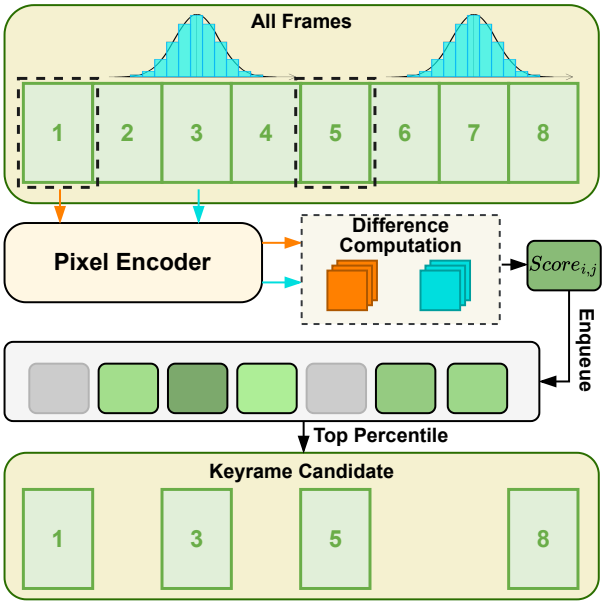

Figure 3: Motion Driven Sampler. To obtain a keyframe candidate set with richer spatio-temporal information, we propose the MDS module, which can adaptively select frames with more significant scene changes as candidates based on the motion cues in the video sequence.

The implementation details of MDS are shown in Figure 3. Given an existing video sequence I, we first select m anchor points $I^a = \left\{I_i^a \in \mathbb{R}^{H\times W\times 3}\right\}_{i=1}^{m}$ at fixed intervals within the video sequence. These anchor points divide the entire video sequence into m segments, each of length n. Within these m segments, we need to calculate the degree of difference between the anchor frame and other frames in the sequence $I^o = \left\{I_j^o \in \mathbb{R}^{H\times W\times 3}\right\}_{j=2}^{n}$, which serves as motion cues. The greater the difference, the more significant the motion change between I_i^a and I_j^o. Specifically, we first input I_i^a and I_j^o into a pixel encoder $\mathcal{F}_{pixelEnc}(\cdot)$ to obtain their respective feature maps $F_i^a \in \mathbb{R}^{h\times w\times c}$ and $F_j^o \in \mathbb{R}^{h\times w\times c}$. We then use F_i^a and F_j^o to calculate the difference metric $D_{i,j}$ between I_i^a and I_j^o, which effectively reflects the motion changes between the anchor frame and the sequence frame. This process is expressed as follows:

$$F_i^a, F_j^o = \mathcal{F}_{pixelEnc}(I_i^a), \mathcal{F}_{pixelEnc}(I_j^o), \tag{5}$$

$$D_{i,j} = \mathcal{D}(F_i^a, F_j^o), \tag{6}$$

where $\mathcal{D}(\cdot)$ represents the difference metric function between feature maps. For the selection of candidate frames, we tend to choose frames with greater differences from the anchor frames, which means that our candidate sequence contains richer spatio-temporal information. However, due to the locality principle, the frames farther from the anchor point tend to have larger differences with the anchor, which inevitably leads to selecting frames that are farther away from the anchor point during sampling, i.e., frames closer to the next anchor point. This negatively impacts the diversity of the sampling. To address this issue, we use a normal distribution to balance this, and the score for each frame in the final subsequence is:

$$Score_{i,j} = \exp\left(-\frac{\left(j-\frac{n}{2}\right)^2}{2\sigma^2}\right)\cdot D_{i,j}, \tag{7}$$

where σ represents the standard deviation. We input the scores of all frames, except for the anchor frames, into a score queue and then select the frames corresponding to the top K-th percentile scores, along with the anchor frames, are output as the keyframe candidates I'.

Joint Keyframe Selection. To determine the frame in the keyframe candidates that best matches the text description, we combine the confidence outputs from MLLM and SAM as the basis for locating the keyframe. Suppose we have already obtained the confidence outputs S_{mllm} and S_{sam} from MLLM and SAM, respectively, using I'. We introduce a harmonizing parameter a to balance the weight of the two confidence values. The final score for the keyframe candidates $S_{key} = \left\{S_k \in \mathbb{R}^1\right\}_{k=1}^{T'}$ is calculated as follows:

$$S_k = a\cdot S_i + (1-a)\cdot S_j, i = j = k, \tag{8}$$

where $S_i \in S_{mllm}, S_j \in S_{sam}$, and $a \in [0,1]$ is used to adjust the importance of S_i and S_j. We select the frame corresponding to the highest score in S_{key} as the keyframe I_{key}.

Object Memory Bank. We introduce OMB to store the high-dimensional semantic information of key objects for more precise object segmentation. Through this memory mechanism, we can continuously identify and track each object throughout the entire video sequence, thereby enhancing the stability of the segmentation output. Now that we have obtained the keyframe I_{key} and its mask M_{key}, we utilize Cutie's (Cheng et al. 2024) object memory capability to implement memory storage for key objects (as in Equation 3). It converts the keyframe and the corresponding segmentation mask into high-dimensional semantic features for storage, which are then used for the bidirectional propagation of object masks across the entire sequence, reducing segmentation errors caused by temporal variations.

Spatio-temporal Decoupling

Previous methods based on MTTR (Wu et al. 2022; Yan et al. 2024; Zhu et al. 2025) have achieved good results, but these architectures are limited to frame-level image understanding, thus ignoring the inherent temporal information in video sequences. This can lead to ambiguity in segmentation when objects in the video sequence resemble the appearance or behavior of the referenced object, causing the model to struggle in consistently outputting the target object's mask. For example, given the same video, if the text queries q_1: "a giraffe is walking on the grass field from the left" and q_2: "a giraffe walking rightwards towards another on the right of the view" are provided, and only frame-level image understanding is used, when the two giraffes switch positions, the model may mistakenly output masks for mismatched instances based on the positional information in the

Method	Publication	Ref-YouTube-VOS			Ref-DAVIS17			MeViS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&F$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&F$	$\mathcal{J}$	$\mathcal{F}$
LBDT (Ding et al. 2022)	CVPR 2022	49.4	48.2	50.6	54.5	-	-	29.3	27.8	30.8
ReferFormer (Wu et al. 2022)	CVPR 2022	62.9	61.3	64.6	61.1	58.1	64.1	31.0	29.8	32.2
VLT+TC (Ding et al. 2023b)	TPAMI 2023	62.7	-	-	60.3	-	-	35.6	33.6	37.3
HTML(Han et al. 2023)	ICCV 2023	63.4	61.5	65.2	62.1	59.2	65.1	-	-	-
OnlineRefer (Wu et al. 2023)	ICCV 2023	63.5	61.6	65.5	64.8	61.6	67.7	-	-	-
LISA (Lai et al. 2024)	CVPR 2024	54.4	54.0	54.8	66.0	63.2	68.8	37.9	35.8	40.0
VISA (Yan et al. 2025)	ECCV 2024	63.0	61.4	64.7	<u>70.4</u>	<u>67.0</u>	<u>73.8</u>	44.5	41.8	47.1
VideoLISA (Bai et al. 2024)	NeurIPS 2024	63.7	61.7	65.7	68.8	64.9	72.7	44.4	41.3	47.6
DMVS (Fang et al. 2025)	CVPR 2025	<u>64.3</u>	<u>62.4</u>	66.2	65.2	62.2	68.2	**48.6**	<u>44.2</u>	<u>52.9</u>
SSA (Pan et al. 2025)	CVPR 2025	<u>64.3</u>	62.2	<u>66.4</u>	67.3	64.0	70.7	**48.6**	44.0	**53.2**
SDAM	Ours	**65.3**	**63.4**	**67.1**	**76.0**	**73.2**	**78.8**	**48.6**	**45.5**	51.7

Table 1: Performance comparison with previous methods on the validation sets of RefVOS datasets. The highest result in each column will be marked in **bold**, and the second-highest result will be <u>underlined</u>.

text description. Therefore, we propose the SD mechanism, which uses MLLM (Bai et al. 2025) and SAM (Kirillov et al. 2023) to locate and segment the target objects, and then employs Cutie (Cheng et al. 2024) to propagate the target object masks across the temporal domain.

Confidence-based Key Object Segmentation. In the spatial domain, we first utilize the real-world reasoning ability of $\mathcal{F}_M(\cdot)$ to output candidate key object location information in JSON format. This is then passed into $\mathcal{F}_S(\cdot)$ to obtain the candidate key object mask M' (as in Equations 1 and 2). The accuracy with which we can locate the object mask most closely matching the text description from the keyframe candidates directly determines the segmentation result of the entire video sequence, which requires the confidence output to be as precise as possible. We use the IOU score of the mask output by $\mathcal{F}_S(\cdot)$ as the segmentation confidence S_{sam}, and the output of $\mathcal{F}_M(\cdot)$ as the object confidence S_{mllm}.

During the experiments, we found that the accuracy of S_{mllm} has a significant impact on the final segmentation result. However, we observed that MLLM, under a broad prompt setting for "output the confidence of answer," has difficulty providing accurate scores. Therefore, we combined its reasoning capability to design a confidence-level output prompt (as shown in Figure 2). Additionally, we also require MLLM to output its reasoning process to guide the model in selecting the appropriate confidence intervals.

Temporal Propagation. Now that we have obtained the key object mask M_{key} in the spatial domain, our next goal is to stabilize the tracking and segmentation of the target object throughout the entire video sequence. We need to leverage the temporal coherence of the object and the visual context in adjacent frames to ensure accurate segmentation across the entire sequence. To achieve this, we introduced the efficient semi-supervised VOS method, Cutie, as our object tracker. We divide the entire video sequence into two parts based on the position of the keyframe: $I^{fw} = \{I_{key}, I_{key-1}, \ldots, I_1\}$ and $I^{bw} = \{I_{key}, I_{key+1}, \ldots, I_T\}$. Then, $\mathcal{F}_{Obj}$ uses the dynamically maintained object memory bank O to bidirectionally propagate the key object mask, starting from the keyframe, and ultimately obtain the mask output for each frame:

$$M_i = \begin{cases} F_{Obj}\left(I^{fw}, O\right), 1 \leqslant i < key \\ F_{Obj}\left(I^{bw}, O\right), \text{key} \leqslant i \leqslant T. \end{cases} \tag{9}$$

Since the presented solution is training-free, it does not rely on any loss function for model optimization.

Experiments

Quantitative Results

Ref-YouTube-VOS. As shown in Table 1, SDAM, as a training-free architecture, surpasses existing state-of-the-art methods that require training on the Ref-YouTube-VOS (Seo, Lee, and Han 2020) dataset. Specifically, our method achieves a $\mathcal{J}\&\mathcal{F}$ score of 65.3% on Ref-YouTube-VOS, leading the previous state-of-the-art methods DMVS(Fang et al. 2025) and SSA(Pan et al. 2025) by 1% on this metric. SDAM demonstrates a certain advantage on this dataset.

Ref-DAVIS17. SDAM significantly outperforms existing state-of-the-art methods on the Ref-DAVIS17 (Khoreva, Rohrbach, and Schiele 2019) dataset, achieving a $\mathcal{J}\&\mathcal{F}$ score of 76.0%, leading the second-ranked method VISA (Yan et al. 2025) by 5.6%, and the third-ranked method VideoLISA (Bai et al. 2024) by 7.2%. SDAM demonstrates exceptional competitiveness on the Ref-DAVIS17 dataset. Due to the presence of similar objects belonging to the same category in multiple videos of the Ref-DAVIS17 dataset, some previous methods struggle to produce stable results. However, SDAM cleverly utilizes the SD mechanism to enhance the temporal stability of the output, significantly improving the model's performance on this dataset.

MeViS. MeViS (Ding et al. 2023a), as a recently emerging benchmark dataset in the RefVOS field, is highly challenging due to its introduction of complex multi-object scenes with extensive motion. Our SDAM achieves a score of 48.6% in $\mathcal{J}\&\mathcal{F}$ on this dataset, tying for first place with the existing state-of-the-art methods DMVS and SSA, while

Method	Publication	Referring			Reasoning			Overall		
		$\mathcal{J}\&F$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&F$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&F$	$\mathcal{J}$	$\mathcal{F}$
Traditional methods without reasoning ability										
ReferFormer (Wu et al. 2022)	CVPR 2022	32.7	31.2	34.3	23.4	21.3	25.6	28.1	26.2	29.9
LMPM (Ding et al. 2023a)	ICCV2023	34.1	29.0	39.1	18.8	13.3	24.3	26.4	21.2	31.7
LLM-based methods with reasoning ability										
TrackGPT-7B (Zhu et al. 2023)	arXiv 2023	48.2	46.7	49.7	39.0	36.8	41.2	43.6	41.8	45.5
TrackGPT-13B (Zhu et al. 2023)	arXiv 2023	49.5	48.3	50.6	40.5	38.1	42.9	45.0	43.2	46.8
LISA-7B (Lai et al. 2024)	CVPR 2024	45.7	44.3	47.1	36.1	33.8	38.4	40.9	39.1	42.7
LISA-13B (Lai et al. 2024)	CVPR 2024	46.6	45.2	47.9	36.7	34.3	39.1	41.6	39.8	43.5
VISA-7B (Yan et al. 2025)	ECCV 2024	50.9	49.2	52.6	43.0	40.6	45.4	46.9	44.9	49.0
VISA-13B (Yan et al. 2025)	ECCV 2024	57.4	55.6	59.1	44.3	42.0	46.7	50.9	48.8	52.9
GLUS (Lin et al. 2025)	CVPR 2025	58.3	56.0	60.7	51.4	48.8	53.9	54.9	52.4	57.3
InstructSeg (Wei et al. 2025)	ICCV 2025	57.0	54.8	59.2	51.9	49.2	54.7	54.5	52.0	56.9
SDAM	Ours	**61.5**	**59.0**	**63.9**	**54.6**	**51.8**	**57.5**	**58.0**	**55.4**	**60.7**

Table 2: Performance comparison with previous methods on the validation set of ReVOS dataset.

Method	Publication	ReasonVOS		
		$\mathcal{J}\&F$	$\mathcal{J}$	$\mathcal{F}$
Traditional methods without reasoning ability				
SOC (Luo et al. 2023)	NeurIPS'23	35.9	33.3	38.5
SgMg (Miao et al. 2023)	ICCV'23	36.2	33.7	38.7
OnlineRefer (Wu et al. 2023)	ICCV'23	38.7	34.6	42.9
LLM-based methods with reasoning ability				
LISA (Lai et al. 2024)	CVPR'24	31.1	29.1	33.1
VideoLISA (Bai et al. 2024)	NeurIPS'24	47.5	45.1	49.9
GLUS (Lin et al. 2025)	CVPR'25	49.9	47.5	52.4
SDAM	Ours	**55.1**	**51.3**	**58.8**

Table 3: Performance comparison with previous methods on the ResonVOS dataset.

leading these methods by 1.3% and 1.5% in the $\mathcal{J}$ metric, respectively. Notably, DMVS and SSA are models specifically designed for MeViS. This demonstrates the robustness of our method when facing complex dynamic scenes.

ReVOS. As shown in Table 2, we significantly outperform existing state-of-the-art methods on all metrics in the large-scale Reasoning-based ReVOS (Yan et al. 2025) dataset. Our method achieves $\mathcal{J}\&\mathcal{F}$ scores of 61.5%, 54.6%, and 58.0% in the "Referring", "Reasoning" and "Overall" categories, respectively, surpassing the second-place methods by 3.2%, 2.7%, and 3.1%.The experimental result shows that our method demonstrates outstanding performance in implicit text and complex multi-object scenarios.

Sampling Strategy	Ref-DAVIS17			Val_u
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
First Frame	69.2	66.3	72.1	50.6
Global	73.4	70.8	75.9	54.4
Motion Driven Sampler	**76.0**	**73.2**	**78.8**	**55.8**

Table 4: Ablation study on sampling strategies.

Confidence Weight		Ref-DAVIS17			ReasonVOS
a	$1-a$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
0.4	0.6	73.9	71.4	76.4	54.1
0.5	0.5	74.7	72.3	77.1	54.8
0.6	0.4	75.8	73.1	78.5	**55.1**
0.75	0.25	**76.0**	**73.2**	**78.8**	54.7

Table 5: Ablation study on confidence weight.

ReasonVOS. As shown in Table 3, we evaluate our method on the ReasonVOS (Bai et al. 2024) dataset. SDAM achieves significant results on this challenging dataset, with a $\mathcal{J}\&\mathcal{F}$ score of 55.1%, surpassing the existing state-of-the-art method GLUS (Lin et al. 2025) by 5.2%.

Ablation Studies

Sampling Strategy. As shown in Table 4, we conduct ablation experiments on three sampling strategies on the Ref-DAVIS17 and MeVis (Val_u)dataset. The anchor frame settings in MDS are the same as the global sampling interval, both set to $\lfloor \frac{T}{4} \rfloor$, and this setting is used for the anchor frames in subsequent experiments. Building upon global sampling,

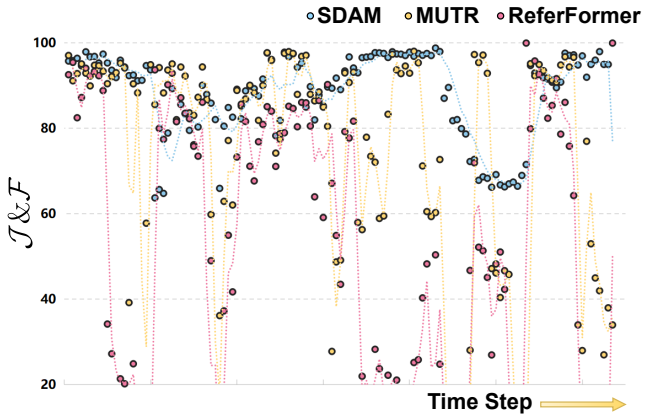

Figure 4: Visualization of temporal stability analysis. We analyze the temporal stability of SDAM, MUTR, and ReferFormer on the Ref-YouTube-VOS dataset.

Number of Keyframes	Ref-DAVIS17		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
$n = 3$	68.6	66.3	70.9
$n = 2$	70.7	69.4	72.0
$n = 1$	**73.4**	**70.8**	**75.9**

Table 6: Ablation study on the number of keyframes.

we use MDS to improve the $\mathcal{J}\&\mathcal{F}$ metric from 73.4% to 76.0% on Ref-DAVIS17, and from 54.4% to 55.8% on MeViS (Val_u). Experimental result shows that our MDS can obtain keyframe candidates with richer spatio-temporal information, thereby improving the keyframe hit rate.

Confidence Weight. As shown in Table 5, we conduct experiments on the confidence weight a based on MDS. We find that when a is set to 0.75, the best performance of 76.0% $\mathcal{J}\&\mathcal{F}$ is achieved on the Ref-DAVIS17 dataset, and when a is set to 0.6, the best performance of $55.1\%\mathcal{J}\&\mathcal{F}$ is achieved on the ReasonVOS dataset. The experimental results suggest that the final performance of the model is positively correlated with the weight of S_{mllm}.

Number of Keyframes. We hypothesize that selecting a single keyframe might lead to incorrect masks for the entire video sequence due to errors in keyframe selection. To enhance the robustness of the method, we try selecting the top n keyframes with the highest confidence and perform bidirectional propagation from multiple keyframe nodes, thus avoiding error propagation caused by mismatching a single keyframe. We conduct experiments on the Ref-DAVIS17 dataset with the global sampling mode, and the experimental results are shown in Table 6. When $n = 1$, we achieve the optimal result of $73.4\%\mathcal{J}\&\mathcal{F}$, which proves that a single keyframe remains the best choice.

Temporal Stability Analysis

As shown in Figure 4, we analyze the temporal stability of SDAM, MUTR (Yan et al. 2024), and ReferFormer (Wu et al. 2022) on the Ref-YouTube-VOS dataset. Specifically, we randomly sample six identical text-video pairs from the Ref-YouTube-VOS dataset, plot the $\mathcal{J}\&\mathcal{F}$ metric of each frame in the video as a scatter plot, and use a Moving Average (MA) to draw the temporal stability trend lines for each method. The experimental result shows that our SD mechanism effectively improves temporal stability in sequence.

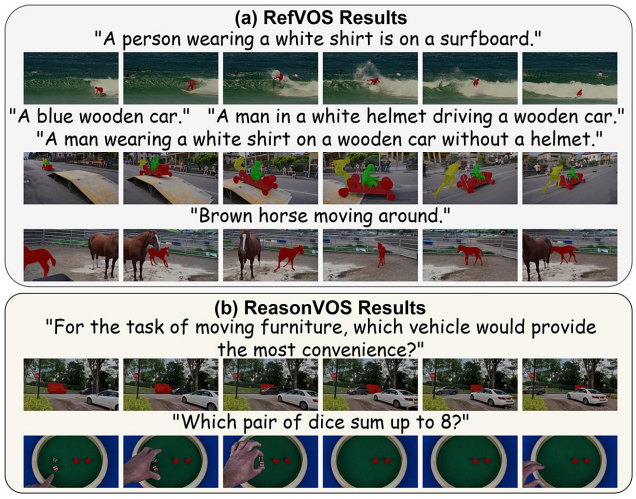

Figure 5: Qualitative results on the RefVOS (a) and ReasonVOS (b) datasets. The time steps are directed from left to right. Zoom in for better viewing.

Qualitative Results

As shown in Figure 5, our method demonstrates excellent performance when facing challenging scenes in the RefVOS and ReasonVOS datasets. Our method performs excellently in challenging scenarios in RefVOS, including cases with small target objects, closely spaced similar objects, and complex prompts containing motion information (as shown in Figure 5 (a)). Notably, the examples on the ReasonVOS dataset are impressive, where our method accurately infers the object described by the complex text query in scenarios with multiple similar objects and occlusions (as shown in Figure 5 for the ReasonVOS). Our method possesses remarkable real-world reasoning abilities, enabling simple mathematical derivations based on information from the video frames (as shown in Figure 5 for the ReVOS).

Conclusion

In this paper, we have proposed novel SDAM, a Spatio-temporal Decoupling Training-Free Reasoning Video Segmentation with Adaptive Object Memory. We introduce an Adaptive Object Memory module to memorize key objects based on motion cues in the video sequence, and leverage a Spatio-temporal Decoupling mechanism to enhance the temporal stability in reasoning video segmentation. We achieve state-of-the-art performance on three RefVOS datasets and two ReasonVOS datasets. In the future, we aim to improve the model's robustness in keyframe selection and the stability in temporal propagation to adapt to diverse scenarios.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62176172, 61672364); partially by the National Key Research and Development Program of China (Grant No. 2018YFA0701701); partially by the Natural Science Foundation of Jiangsu Province (Grant for Young Scholars, Grant No. BK20250789); and partially by Undergraduate Training Program for Innovation and Entrepreneurship, Soochow University (No. 2025C104).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, Z.; He, T.; Mei, H.; Wang, P.; Gao, Z.; Chen, J.; Liu, L.; Zhang, Z.; and Shou, M. Z. 2024. One Token to Seg Them All: Language Instructed Reasoning Segmentation in Videos. In *Advances in Neural Information Processing Systems*, 6833–6859.

Botach, A.; Zheltonozhskii, E.; and Baskin, C. 2022. End-to-End Referring Video Object Segmentation With Multimodal Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4985–4995.

Cheng, H. K.; Oh, S. W.; Price, B.; Lee, J.-Y.; and Schwing, A. 2024. Putting the Object Back into Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3151–3161.

Cho, S.; Lee, S.; Lee, M.; Lee, J.; and Lee, S. 2025. Find First, Track Next: Decoupling Identification and Propagation in Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 3814–3824.

Cuttano, C.; Trivigno, G.; Rosi, G.; Masone, C.; and Averta, G. 2025. SAMWISE: Infusing Wisdom in SAM2 for Text-Driven Video Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3395–3405.

Ding, H.; Liu, C.; He, S.; Jiang, X.; and Loy, C. C. 2023a. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2694–2703.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2023b. VLT: Vision-Language Transformer and Query Generation for Referring Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7900–7916.

Ding, Z.; Hui, T.; Huang, J.; Wei, X.; Han, J.; and Liu, S. 2022. Language-Bridged Spatial-Temporal Interaction for Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4964–4973.

Fang, H.; Cong, R.; Lu, X.; Zhou, X.; Kwong, S.; and Zhang, W. 2025. Decoupled Motion Expression Video Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13821–13831.

Gong, S.; Zhuge, Y.; Zhang, L.; Yang, Z.; Zhang, P.; and Lu, H. 2025. The Devil is in Temporal Token: High Quality Video Reasoning Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 29183–29192.

Han, M.; Wang, Y.; Li, Z.; Yao, L.; Chang, X.; and Qiao, Y. 2023. HTML: Hybrid Temporal-scale Multimodal Learning Framework for Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 13414–13423.

He, S.; and Ding, H. 2024. Decoupling Static and Hierarchical Motion Perception for Referring Video Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13332–13341.

Kao, S.-h.; Tai, Y.-W.; and Tang, C.-K. 2025. ThinkVideo: High-Quality Reasoning Video Segmentation with Chain of Thoughts. *arXiv preprint arXiv:2505.18561*.

Khoreva, A.; Rohrbach, A.; and Schiele, B. 2019. Video Object Segmentation with Language Referring Expressions. In *Computer Vision – ACCV 2018*, 123–141.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9579–9589.

Liang, T.; Lin, K.-Y.; Tan, C.; Zhang, J.; Zheng, W.-S.; and Hu, J.-F. 2025. ReferDINO: Referring Video Object Segmentation with Visual Grounding Foundations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Lin, L.; Yu, X.; Pang, Z.; and Wang, Y.-X. 2025. GLUS: Global-Local Reasoning Unified into A Single Large Language Model for Video Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8658–8667.

Luo, Z.; Xiao, Y.; Liu, Y.; Li, S.; Wang, Y.; Tang, Y.; Li, X.; and Yang, Y. 2023. SOC: Semantic-Assisted Object Cluster for Referring Video Object Segmentation. In *Advances in Neural Information Processing Systems*, 26425–26437.

Miao, B.; Bennamoun, M.; Gao, Y.; and Mian, A. 2023. Spectrum-guided Multi-granularity Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 920–930.

Pan, F.; Fang, H.; Li, F.; Xu, Y.; Li, Y.; Benini, L.; and Lu, X. 2025. Semantic and Sequential Alignment for Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19067–19076.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. GLaMM: Pixel Grounding Large Multimodal Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13009–13018.

Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. PixelLM: Pixel Reasoning with Large Multimodal Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26374–26383.

Seo, S.; Lee, J.-Y.; and Han, B. 2020. URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In *Computer Vision – ECCV 2020*, 208–223.

Tang, J.; Zheng, G.; and Yang, S. 2023. Temporal Collection and Distribution for Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15466–15476.

Wei, C.; Zhong, Y.; Tan, H.; Zeng, Y.; Liu, Y.; Wang, H.; and Yang, Y. 2025. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20193–20203.

Wu, D.; Wang, T.; Zhang, Y.; Zhang, X.; and Shen, J. 2023. OnlineRefer: A Simple Online Baseline for Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2761–2770.

Wu, J.; Jiang, Y.; Sun, P.; Yuan, Z.; and Luo, P. 2022. Language As Queries for Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4974–4984.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. GSVA: Generalized Segmentation via Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3858–3869.

Yan, C.; Wang, H.; Yan, S.; Jiang, X.; Hu, Y.; Kang, G.; Xie, W.; and Gavves, E. 2025. VISA: Reasoning Video Object Segmentation via Large Language Models. In *Computer Vision – ECCV 2024*, 98–115.

Yan, S.; Zhang, R.; Guo, Z.; Chen, W.; Zhang, W.; Li, H.; Qiao, Y.; Dong, H.; He, Z.; and Gao, P. 2024. Referred by Multi-Modality: A Unified Temporal Transformer for Video Object Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6): 6449–6457.

Zhu, J.; Cheng, Z.-Q.; He, J.-Y.; Li, C.; Luo, B.; Lu, H.; Geng, Y.; and Xie, X. 2023. Tracking with Human-Intent Reasoning. arXiv:2312.17448.

Zhu, Z.; Feng, X.; Chen, D.; Yuan, J.; Qiao, C.; and Hua, G. 2025. Exploring Pre-trained Text-to-Video Diffusion Models for Referring Video Object Segmentation. In *Computer Vision – ECCV 2024*, 452–469.