MomentSeg: Moment-Centric Sampling for Enhanced Referring Video Object Segmentation

Ming Dai[1], Sen Yang[2], Boqiang Duan[2], Wankou Yang[*1], and Jingdong Wang[2]

[1] Southeast University, Nanjing, China
[2] Baidu Inc., Beijing, China

Abstract. Referring Video Object Segmentation (RefVOS) seeks to segment target objects in videos guided by natural language descriptions, demanding both temporal reasoning and fine-grained visual comprehension. Existing sampling strategies for LLM-based approaches typically rely on either handcrafted heuristics or external keyframe models. The former often overlooks essential temporal cues, while the latter increases system complexity. To address this, we propose a unified framework that jointly optimizes Temporal Sentence Grounding (TSG) and RefVOS, naturally incorporating key moment grounding capability. During training, we introduce a novel TSG paradigm that employs a dedicated `[FIND]` token for key moment identification through temporal token similarity matching, thereby avoiding the need for external timestamp encodings. For inference, we design a Moment-Centric Sampling (MCS) strategy that densely samples informative moments while sparsely sampling non-essential frames, preserving both motion details and global context. To further enhance tracking stability, we develop Bidirectional Anchor-updated Propagation (BAP), which leverages the most relevant moment as start point for high-quality mask initialization and dynamically updates at sampled points to mitigate accumulated errors. Code will be released at `https://github.com/Dmmm1997/MomentSeg`.

Keywords: Video Referring Object Segmentation · Moment-Centric Sampling · Temporal Sentence Grounding

1 Introduction

RefVOS has attracted considerable attention in recent years. The task involves localizing and segmenting target objects at the pixel level in videos based on natural language expressions [18, 29, 52]. It requires models to interpret temporal action semantics described in text and to maintain consistent tracking of the referred object across frames. Recent Transformer-based methods [3, 9, 23, 34, 40, 44, 51] mainly focus on improving temporal reasoning and inter-frame consistency. In contrast, large multimodal models (LMMs) [1, 6, 32], pre-trained on

* Corresponding author.

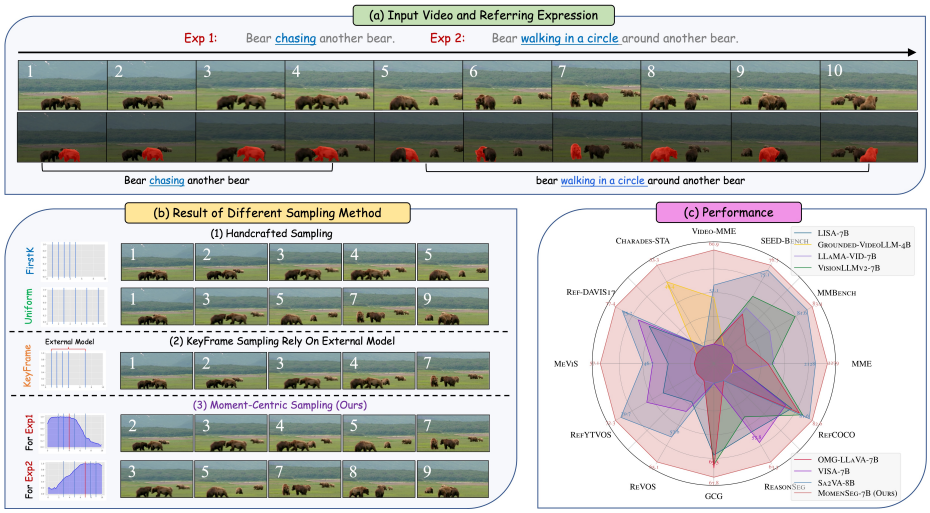

Fig. 1: Illustration of Moment-Centric Sampling (MCS). (a) Input videos and referring expressions. (b) Comparison of sampling strategies: handcrafted, external keyframe–based, and our proposed MCS. MCS performs dense sampling at critical moments and sparse sampling elsewhere, without relying on external keyframe models. (c) MomentSeg achieves superior performance across RefVOS, TSG, and QA tasks.

massive video-text corpora, demonstrate stronger potential for advancing pixel-level video understanding tasks [35, 39, 42, 68].

A central challenge in RefVOS lies in effectively sampling keyframes that are relevant to the linguistic expression. Given the variable and often large number of frames, feeding all frames into the model is both computationally prohibitive and redundant. Consequently, identifying and selecting semantically meaningful keyframes is worth exploring. Existing approaches employ diverse strategies for keyframe selection, as summarized in Table 1. For instance, Sa2VA [68] simply selects the first K frames, disregarding content-specific cues such as object presence or action timing. VideoLISA [2] compresses temporal information and samples frames uniformly to retain coarse timing information but still overlooks referentially critical frames. In contrast, methods such as VISA [65], ViLLA [72], VRS-HQ [20], and GLUS [35] depend on separately trained models to pinpoint the most relevant timestamps, then either use tracking networks [7] to generate mask trajectories or select key sampling frames to guide large-model segmentation. While these methods enable the identification of keyframes, they are dependent on these additional models.

Building on prior explorations, we first pose a fundamental question: ***Is keyframe selection necessary?*** From an ***intuitive*** perspective, failing to capture critical moments in a video risks missing the target object's appearance or the described action, which directly hinders accurate segmentation. Conversely, precisely identifying and leveraging text-relevant key moments mit-

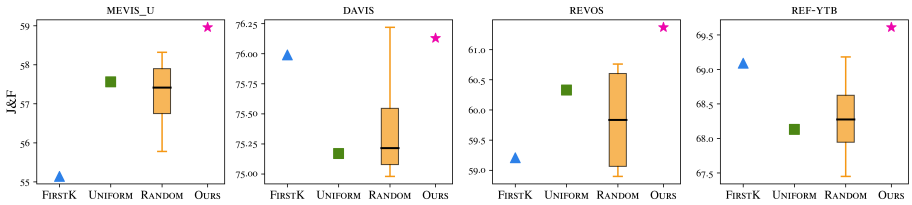

Fig. 2: Impact of Sampling Strategies. We evaluate various sampling strategies for the RefVOS task on four datasets. Five independent runs of random sampling show high variance, highlighting the need for a robust sampling mechanism. Our proposed MCS consistently outperforms alternative methods across all datasets.

igates the misleading effects of irrelevant frames on LMMs while reducing computational redundancy. From a ***quantitative*** perspective, as shown in Fig. 2, we evaluate several sampling strategies across multiple RefVOS datasets. Two consistent patterns emerge: **(1)** random sampling exhibits high variance, highlighting the strong influence of sampling strategy; and **(2)** for scenarios involving motion-driven (e.g., MeVIS [15]) and complex semantic understanding (e.g., ReVOS [65]), uniform sampling often surpasses the firstK approach. In contrast, firstK performs better on datasets [29, 52] where targets appear early in the video and descriptions rely less on temporal cues. Likewise, MeVIS, which includes numerous action-centric descriptions and scenes with multiple visually similar objects, poses a particular challenge: if sampled frames miss the described action while static cues match several candidates, the model struggles to disambiguate the target. This analysis underscores that *keyframe selection is critical for RefVOS*.

Next, we consider the following question: ***How can we effectively identify keyframes, and how can these keyframes enhance RefVOS performance?*** Prior work offers two primary strategies: **(1)** Keyframe for tracking initialization. VISA [65] employs LLAMA-VID [33] to select the most critical frame as the tracking initialization timestamp, and applies XMem [7] for subsequent tracking. **(2)** Keyframes for LMM understanding. ViLLA [72] leverages Grounded-VideoLLM [54] to generate relevant temporal intervals and then samples frames inside and outside these intervals using handcrafted ratios, subsequently feeding these frames to LMM. VRS-HQ [20] further combines CLIP-based [47] vision-language similarity with SAM2 scores to filter keyframes before passing them to the LMM. While effective, all these methods depend on external models for keyframe selection, inevitably increasing system complexity.

In this paper, we propose a unified framework for the synchronized optimization of TSG and RefVOS. ***During training***, as shown in Fig. 3, we introduce a special `[FIND]` token for the TSG task, similar to the `[SEG]` token in RefVOS. This token computes similarity with sampled video frame tokens, offering two key advantages: **First**, for RefVOS, it eliminates the need for external models for keyframe selection, as the model natively identifies key moments. **Second**, it

Table 1: Comparison of *keyframe models*, *sampling*, and *information* utilized in existing RefVOS MLLMs. "Cont." means continuous sampling. Unlike prior methods, our approach inherently supports keyframe selection while simultaneously leveraging both "Global" and "Essential" information during inference.

Method	Keyframe Model	Training		Inference		`[SEG]`
		Sampling	Information	Sampling	Information	
VideoLISA [2]	/	Uniform	Global	Uniform	Global	1
VISA [65]	LLaMA-VID [33]	Random	Random	Uniform + Cont.	Global + Local	1
ViLLa [72]	G-VideoLLM [54]	Uniform	Global	Cont.	Local	N
GLUS [35]	Chat-UniVi [27]	Uniform + Cont.	Global + Local	Uniform + Cont.	Global + Local	N
Sa2VA [68]	/	Random	Random	FirstK	Local	1
VRS-HQ [20]	CLIP [47]	Uniform	Global	Selection	Essential	N
MomentSeg	Inherently	Uniform + Cont.	Global + Local	Selection	Global + Essential	1

obviates the need for explicit timestamp encoding [5] by leveraging a frame-level matching mechanism between `[FIND]` and temporal tokens. ***During inference***, as depicted in Fig. 4, we propose a Moment-Centric Sampling (MCS). MCS leverages the similarity distribution from the `[FIND]` token to densely sample video moments highly relevant to the description, while sparsely sampling less critical frames. This approach effectively preserves crucial motion cues while maintaining a global temporal context. For the RefVOS task, consistent with Sa2VA [68], a single `[SEG]` token is used to represent the referential object throughout the video. We further introduce a novel Bidirectional Anchor-updated Propagation (BAP) strategy, which is composed of two key components: **(1)** *Bidirectional Propagation*: The central keyframes identified by MCS serve as starting points for propagation in both forward and backward temporal directions. **(2)** *Anchor-updated*: At these sampled time points, the target mask is updated and the prior memory is adaptively refreshed, which mitigates cumulative tracking errors.

In summary, our key contributions are as follows: **(1)** We propose **MomentSeg**, a unified framework that integrates temporal sentence grounding (TSG) and referring video object segmentation (RefVOS). For TSG, MomentSeg obviates explicit timestamp encoding by leveraging the similarity between the `[FIND]` token and temporal tokens. For RefVOS, our model inherently identifies text-relevant keyframes, thereby removing the dependency on external keyframe-selection models. **(2)** We introduce **Moment-Centric Sampling (MCS)**, a similarity-driven sampling strategy that densely samples frames around relevant moments and sparsely samples non-essential frames. MCS preserves salient motion cues while retaining global temporal context. **(3)** We present **Bidirectional Anchor-updated Propagation (BAP)**, which initiates mask propagation from query-relevant keyframes and adaptively updates the target mask at sampled temporal points. BAP enhances tracking robustness and mitigates cumulative propagation errors. (4) MomentSeg achieves new SOTA performance, delivering a 5% improvement on the MeVIS and a 6% gain on the ReVOS.

2 Related Work

2.1 Referring Video Object Segmentation

RefVOS aims to segment targets in a video conditioned on a given description. Existing approaches typically employ object queries to fuse expressions with visual features for referent identification [40, 53, 59, 60, 69], while a line of work emphasizes motion aggregation to capture dynamic cues [8, 15, 16, 23]. Concurrently, LMMs [11, 37] have been used to reason over compositional and complex expressions [13, 14, 19, 39, 42, 63, 65, 73, 74]. To address these gaps, we propose **MomentSeg**, which performs keyframe selection natively, removing the need for external modules. We further introduce Bi-directional Anchor-updated Propagation (BAP), which propagates masks forward and backward from key moments and adaptively updates memory at sampled nodes, enhancing segmentation robustness and reducing cumulative errors.

2.2 LLM-based RefVOS Sampling Strategies

Existing LLM-based RefVOS methods adopt diverse frame-sampling and keyframe-selection schemes (Table 1). VideoLISA [2] and Sa2VA [68] adopt handcrafted sampling strategies, using uniform or firstK sampling to generate a global `[SEG]` token. In contrast, VISA [65], ViLLA [72], and GLUS [35] rely on external models (e.g., LLaMA-VID [33], Grounded-VideoLLM [54], Chat-UniVi [27]) for keyframe selection, while VRS-HQ [20] selects keyframes via CLIP-based vision–language similarity [47]. Such dependencies increase pipeline complexity and limit adaptability. By comparison, **MomentSeg** performs keyframe selection natively and introduces Moment-Centric Sampling, which applies dense sampling around text-relevant moments and sparse sampling elsewhere, preserving salient motion cues while maintaining a compact global temporal context.

2.3 Timestamps Encoding in Temporal Sentence Grounding

TSG [4, 17, 31, 38, 54, 70] aims to localize a video segment corresponding to a language query. A key challenge for LMM-based methods lies in developing an effective timestamp encoding strategy. Approaches such as TimeChat [50], VTimeLLM [25], and LITA [26] leverage instruction tuning for plain text temporal grounding. Momentor [45] injects explicit temporal position encodings into frame-level features to improve temporal localization, while VTG-LLM [22] incorporates absolute-time tokens. Vid2Seq [66] and Grounded-VideoLLM [54] adopt discrete or relative temporal tokens to circumvent direct timestamp encoding. NumberIT [62] injects sequential frame indices into images for grounding. In contrast, **MomentSeg** avoids timestamp encoding by employing a specialized `[FIND]` token that performs frame-level temporal matching.

3 Methods

3.1 Preliminaries

Referring Video Object Segmentation: Given an input video with T frames $I_{1:T} \in \mathbb{R}^{T\times H\times W\times 3}$ and a referring language expression R, RefVOS aims to train

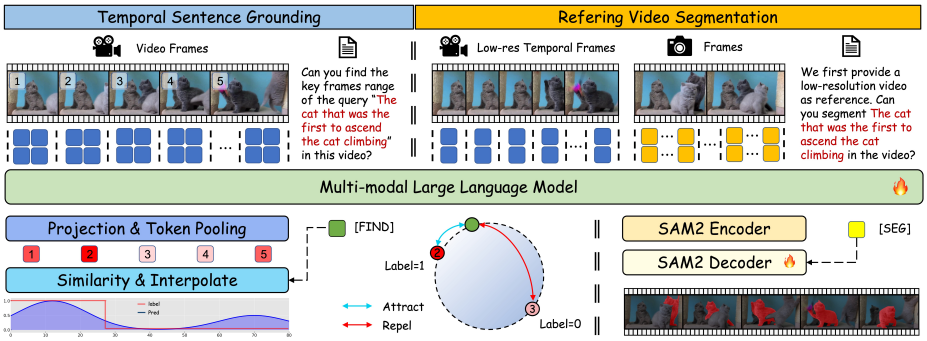

Fig. 3: Training framework of MomentSeg. In the TSG paradigm, we employ the Qwen2.5-VL [1] video encoder with low-resolution inputs and introduce a `[FIND]` token optimized via contrastive learning. For RefVOS, the model processes both low-resolution uniform frames and high-resolution sparsely frames. Supervision is mediated by the `[SEG]` token, with masks generated through the SAM2 decoder.

a model ϕ_θ that predicts binary segmentation masks $M_{1:T}$ for the referred object:

$$M_{1:T} = \phi_\theta(I_{1:T}, R). \tag{1}$$

LMMs-based RefVOS Paradigm: Recent LMM-based RefVOS approaches build upon image-level referring segmentation models such as LISA [30], which employ a dedicated `[SEG]` token to represent the target and a segmentation decoder `Dec` to generate masks:

$$\texttt{[SEG]} = \texttt{LMM}(I, R), \quad M = \texttt{Dec}(I, \texttt{[SEG]}). \tag{2}$$

This paradigm extends to videos by adopting a multi-image formulation, where the LMM processes multiple frames $I_{1:N}$ and segmentation tokens decode masks for each frame:

$$\texttt{[SEG]} = \texttt{LMM}(I_{1:N}, R), \quad M_{1:N} = \texttt{Dec}(I_{1:N}, \texttt{[SEG]}). \tag{3}$$

This equation exemplifies a "One Token Seg All" paradigm [2], in which a single `[SEG]` token represents the referential object throughout the video. Here, N denotes the number of frames provided to the LMM, which is typically smaller than the full sequence length T.

3.2 Model Architecture of MomentSeg

Training Pipeline Fig. 3 presents the training paradigm of MomentSeg for both TSG and RefVOS. For RefVOS, we follow a scheme similar to Sa2VA [68], while additionally incorporating low-resolution, uniformly sampled frames $I^l_{1:L}$ and high-resolution, densely sampled frames $I^h_{1:N}$ from temporal segments where

the referential object is present. The sampling of these low-resolution frames is referred to as *Temporal Token Injecting* in this paper. Notably, the number of tokens introduced into the LMM remains minimal, at no more than 16 tokens per frame in our setting. The low-resolution frames $I^l_{1:L}$ supply global temporal context, whereas the high-resolution frames $I^h_{1:N}$ capture fine-grained spatial details. A SAM2 decoder is then applied to the high-resolution frames and is supervised as follows:

$$\texttt{[SEG]}^h = \texttt{LMM}(I^l_{1:L}, I^h_{1:N}, R), \quad M_{1:N} = \texttt{Dec}(I^h_{1:N}, \texttt{[SEG]}^h). \tag{4}$$

For TSG, we introduce a special token `[FIND]` to identify relevant frames through similarity matching with temporal features. This design removes the need for external timestamp encoding, as temporal correspondence is established directly via frame-level matching. Specifically, we first extract the temporal tokens corresponding to the LMM output. These tokens are projected into a feature space using an MLP layer and subsequently average-pooled to obtain frame representations $T^v \in \mathbb{R}^{L_t \times C}$. The `[FIND]` tokens, $T^f \in \mathbb{R}^{N_f \times C}$ (where N_f denotes the number of `[FIND]` tokens in a sample under the multi-round dialogue paradigm), are mapped through the same MLP layer. We then compute the similarity matrix $\ell \in \mathbb{R}^{N_f \times L_t}$ between frame tokens and `[FIND]` tokens to identify text-relevant frames. The matching loss is defined as:

$$\ell_{ij} = \frac{(T^f_i)^\top T^v_j}{\|T^f_i\| \, \|T^v_j\| \, \tau}, \tag{5}$$

$$\mathcal{L}_{find} = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \Big[-\lambda_p y_{ij} \log \sigma(\ell_{ij}) - (1 - y_{ij}) \log(1 - \sigma(\ell_{ij})) \Big], \tag{6}$$

where σ denotes the sigmoid function, $T^v_i \in \mathbb{R}^C$ is the i-th pooled frame token, $T^f_j \in \mathbb{R}^C$ is the j-th `[FIND]` token, $y_{ij} \in \{0, 1\}$ is the match label, Ω is the set of valid pairs, and τ is the temperature factor, set to 0.07. The positive sample weight λ_p is set to 2.0. Notably, since Qwen2.5-VL compresses the temporal dimension during video encoding ($L_t = 0.5 \times L$), we apply bilinear interpolation to the similarity distribution to ensure alignment with the original length.

Inference Pipeline As shown in Fig. 4, inference begins with densely sampled low-resolution frames $I^l_{1:L}$ to compute the frame-query similarity distribution relative to the description. Based on this distribution, the Moment-Centric Sampling (MCS) strategy selects key frames $I^h_{1:N}$. These key frames, together with RefVOS instructions, are then input to the model for segmentation. Finally, Bidirectional Anchor-updated Propagation (BAP) is applied to enhance segmentation robustness.

Moment-Centric Sampling (MCS) MCS selectively samples text-relevant frames while preserving global temporal context. Specifically, we first quantify

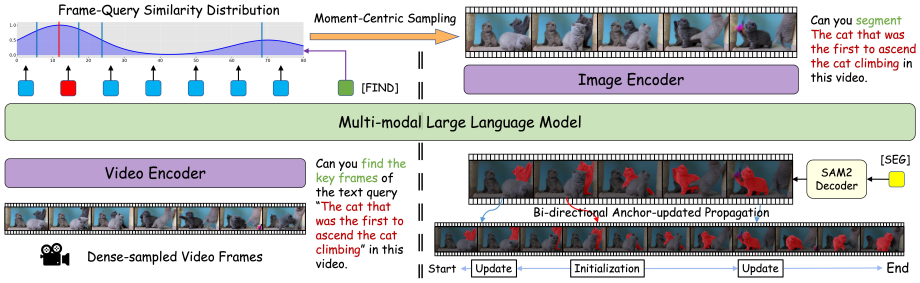

Fig. 4: Inference pipeline of MomentSeg. We first compute the frame-query similarity distribution from densely sampled frames to guide Moment-Centric Sampling (MCS) in selecting keyframes. Subsequently, Bidirectional Anchor-updated Propagation (BAP) is employed to enhance segmentation robustness.

the semantic relevance of each frame by computing the similarity between temporal frame tokens $I^l_{1:L}$ and the `[FIND]` token T^f. Based on the resulting similarity distribution, intuitive strategies such as NearbyK and TopK sampling often lead to densely clustered selections. In contrast, MCS adaptively modulates sampling density according to frame-query similarity, concentrating coverage in high-relevance regions while maintaining a sparse representation elsewhere, as validated in Table 8. Concretely, Gaussian smoothing is applied to mitigate noise, yielding a refined similarity distribution $\mathcal{S} \in \mathbb{R}^T$. The moment center c^* is then identified via a sliding window of size w that maximizes the window-level cumulative similarity:

$$i^* = \underset{i=0 \to T-w}{\arg\max} \sum_{j=i}^{i+w-1} \mathcal{S}_j, \quad c^* = i^* + \lfloor \frac{w}{2} \rfloor, \tag{7}$$

where i^* denotes the starting index of the optimal window. Alg. 1 details the process: *(1) Compute partitioned mass:* cumulative similarities (w_L, w_R) are aggregated for regions flanking c^*. *(2) Allocate budget:* the remaining $K-1$ samples are distributed to the left (k_L) and right (k_R) partitions proportional to their weights. *(3) Perform stratified sampling:* InverseCDF is applied to each partition, and the resulting indices are combined with c^* to form the final set. The InverseCDF procedure takes as input a sub-distribution $\mathcal{S}_{a:b}$ and a target sample size k. It normalizes the weights to form probabilities $p_i = \mathcal{S}_i / \sum_{j=a}^{b} \mathcal{S}_j$ and computes

Algorithm 1 Moment-Centric Sampling

input: similarity scores $\mathcal{S} \in \mathbb{R}^T$, total samples $K \in \mathbb{R}^1$, moment center $c^* \in \mathbb{R}^1$
output: sampled indices $\mathcal{I} \subseteq \{0, \ldots, T-1\}$
(1) *Compute partitioned similarity mass*
$w_L \leftarrow \sum_{i=0}^{c^*-1} \mathcal{S}_i, \quad w_R \leftarrow \sum_{i=c^*+1}^{T-1} \mathcal{S}_i.$
(2) *Allocate partition sampling budget*
$k_L \leftarrow \left\lfloor (K-1) \frac{w_L}{w_L + w_R} \right\rceil,$
$k_R \leftarrow K - 1 - k_L.$
(3) *Perform stratified sampling*
$\mathcal{I}_L \leftarrow \text{InverseCDF}(\mathcal{S}_{0:c^*-1}, k_L),$
$\mathcal{I}_R \leftarrow \text{InverseCDF}(\mathcal{S}_{c^*+1:T-1}, k_R).$
return $\text{sort}(\mathcal{I}_L \cup \{c^*\} \cup \mathcal{I}_R)$

$F(i) = \sum_{j=a}^{i} p_j$. Then, the sampled indices are determined by querying uniformly spaced quantiles $\hat{u}_m \in (0, 1)$:

$$i_m = \min\{\, i \in [a, b] \mid F(i) \geq \hat{u}_m \,\}, \qquad \hat{u}_m = \frac{m - 0.5}{k}, \qquad m = 1, \ldots, k. \quad (8)$$

This ensures that moments with higher similarity receive denser sampling while still maintaining representation across the entire temporal span. The key advantage of MCS lies in its ability to enhance the capacity to capture both fine-grained actions and their broader temporal dependencies.

Bidirectional Anchor-updated Propagation (BAP) To improve segmentation robustness, we propose BAP, which leverages key moments as anchors for bidirectional propagation. The primary goal is to provide SAM2 with a strong initialization while ensuring stable tracking through mask updates during propagation. Its two key features are: (1) SAM2 propagation starts from critical temporal anchors and proceeds bidirectionally (*Initialization* in Fig. 4); and (2) mask updates occur at sampled key moments with selective memory clearing (*Update* in Fig. 4). This design enables dynamic error correction, enhancing long-term tracking. Specifically, the most relevant moment c^* identified by MCS is selected as the starting point for bidirectional propagation. At each sampled key moment $\mathcal{I}$, a memory cleaning mechanism adaptively clears the cache based on the mask prediction score S^p and the tracking score S^t, as determined by:

$$U_i = \begin{cases} 1, & \text{if } \prod_{j=1}^{t} S_j^t < \lambda \cdot S_i^p \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

where U_i is the binary indicator (1 for clearing memory, 0 for retention), S_j^t is the tracking score at step j, $\prod_{j=1}^{t} S_j^t$ denotes the cumulative tracking confidence up to time t, S_i^p is the prediction confidence at node i, and λ is a sensitivity hyperparameter (set to 0.9). An update is triggered when accumulated confidence falls below the threshold determined by the current prediction quality, ensuring that memory is refreshed only when beneficial for tracking accuracy.

4 Experiments

4.1 Implementation Details

We construct our baseline by integrating Qwen2.5-VL [1] with SAM2 [49]. Following prior work [68], segmentation masks are generated by decoding the hidden state of the `[SEG]` token using the SAM2 decoder. The LLM is fine-tuned with LoRA [24], with a maximum sequence length of 8,192 tokens. Training is conducted on 8 NVIDIA H20 GPUs (96GB memory each) and takes approximately 24 hours for the MomentSeg-3B model. For evaluation, we adopt lmms-eval [71] for both image and video chat tasks. The segmentation training data mainly follows Sa2VA, with additional TSG-related data; see Appendix B.1 for details.

Table 2: Performance comparison on temporal sentence grounding benchmarks.

Method	Charades-STA				ActivityNet-Grounding			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Momenter-7B [46]	42.6	26.6	11.6	28.5	42.9	23.0	12.4	29.3
VTimeLLM-7B [25]	51.0	27.5	11.4	31.2	44.0	27.8	14.3	30.4
TimeChat-7B [50]	-	32.2	13.4	-	-	-	-	-
VTG-LLM-7B [21]	-	33.8	15.7	-	-	-	-	-
HawkEye-7B [55]	50.6	31.4	14.5	33.7	49.1	29.3	10.7	32.7
Grounded-VideoLLM-4B [54]	54.2	36.4	19.7	36.8	46.2	30.3	19.0	36.1
NumPro-LongVA-7B [62]	63.8	42.0	20.6	41.4	55.6	37.5	20.6	38.8
Qwen2.5-VL-7B [1]	-	-	-	43.6	-	-	-	-
MomentSeg-3B (Ours)	**76.1**	**58.2**	**25.8**	**50.2**	**67.5**	**44.7**	**23.2**	**45.4**

4.2 Main Results

Temporal Sentence Grounding Task We evaluate MomentSeg on the Charades-STA [17] and ActivityNet [4] datasets. As shown in Table 2, MomentSeg consistently outperforms a wide range of SOTA methods. Notably, despite its compact parameter scale, MomentSeg-3B surpasses some large-scale Video-LLMs. This strong performance validates the effectiveness of the proposed paradigm, which is underpinned by the novel mechanism where the `[FIND]` token interacts with temporal tokens via cosine similarity to achieve frame-level matching.

Referring/Reasoning Video Segmentation Task For **referring** tasks (Table 3), MomentSeg-3B achieves $\mathcal{J}\&\mathcal{F}$ scores of 62.0 on MeViS [15] (val^u) and 54.8 on MeViS (val), surpassing GLUS-7B [35] by 1.1 and 3.5 points, respectively. On Ref-Youtube-VOS [52] and Ref-DAVIS17 [28], it attains scores of 72.0 and 76.4, improving over Sa2VA-4B [68] by 2.0 and 6.4 points. Scaling to 7B further enhances performance across benchmarks. For **reasoning** tasks (Table 4), MomentSeg-3B obtains 62.6 on ReVOS [65], exceeding VRS-HQ-7B [20] by 3.5 points. On ReasonVOS [2], it achieves 61.7, outperforming ViLLa-6B [72] by 6.3 points, with consistent improvements at larger scales. Furthermore, as shown in Table 5, on the Ref-SAV [68] dataset, MomentSeg-3B achieves a $\mathcal{J}\&\mathcal{F}$ of 64.0 without fine-tuning, exceeding the Sa2VA-8B method by 22.7 points.

Table 5: Performance on Ref-SAV [68] benchmark. FT means fine-tuning on Ref-SAV dataset.

Method	Size	FT	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
UniRef++ [61]	–	✗	11.6	9.5	10.5
UNINEXT [64]	–	✗	8.8	6.4	7.6
LMPM [15]	–	✗	12.2	9.8	10.3
VISA [65]	7B	✗	13.2	11.3	11.8
Sa2VA [68]	8B	✗	39.6	43.0	41.3
Sa2VA [68]	8B	✓	48.3	51.7	50.0
MomentSeg	3B	✗	62.9	65.2	64.0

Referring/Reasoning Image Segmentation Task As shown in Table 6, MomentSeg-3B achieves 82.1/76.9/78.8 on the RefCOCO/+/g [41, 43, 67] validation sets, outperforming Sa2VA-4B by 3.2/5.2/4.7 points, respectively. For reasoning tasks, MomentSeg-7B attains 65.5 on the ReasonSeg [30] test set, surpassing the previous method, LISA-7B, by 2.6 points. Furthermore, on the

Table 3: Performance comparison on referring video object segmentation benchmarks.

Method	MeViS (val^u)			MeViS (val)			Ref-Youtube-VOS			Ref-DAVIS17		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Methods without LLMs												
ReferFormer [60]	-	-	-	31.0	29.8	32.2	62.9	61.3	64.6	61.1	58.1	64.1
DsHmp [23]	55.3	51.0	60.4	46.4	43.0	49.8	67.1	65.0	69.1	64.9	61.7	68.1
DeRVOS [8]	60.6	56.1	65.1	51.8	48.1	55.4	70.0	68.0	71.9	70.9	68.3	73.5
Methods with LLMs												
LISA-7B [30]	43.2	39.9	46.5	37.2	35.1	39.4	53.9	53.4	54.3	64.8	62.2	67.3
VideoLISA-3.8B [2]	54.5	50.9	58.1	44.4	41.3	47.6	63.7	61.7	65.7	68.8	64.9	72.7
VISA-7B [65]	-	-	-	43.5	40.7	46.3	61.5	59.8	63.2	69.4	66.3	72.5
ViLLa-6B [72]	-	-	-	49.4	46.5	52.3	67.5	64.6	70.4	74.3	70.6	78.0
GLUS-7B [35]	60.9	-	-	51.3	48.5	54.2	67.3	65.5	69.0	-	-	-
InstructSeg-3B [58]	-	-	-	-	-	-	67.5	65.4	69.5	71.1	67.3	74.9
Sa2VA-4B [68]	-	-	-	46.2	-	-	70.0	-	-	70.0	-	-
Sa2VA-8B [68]	-	-	-	46.9	-	-	70.7	-	-	75.2	-	-
Sa2VA-Qwen2.5-VL-3B [68]	57.5	53.6	61.3	50.0	46.9	53.1	71.0	68.8	73.1	74.4	70.1	78.7
MomentSeg-3B (Ours)	62.0	58.1	65.9	54.8	51.7	58.0	72.0	69.8	74.3	76.4	72.2	80.6
MomentSeg-7B (Ours)	**62.6**	**58.7**	**66.5**	**57.1**	**53.9**	**60.2**	**72.3**	**70.1**	**74.5**	**77.4**	**73.2**	**81.7**

GCG [48] benchmark, MomentSeg-7B achieves 67.8/67.9 mIoU (val/test), exceeding GLaMM-7B by 2.0/3.3 points.

4.3 Ablations Studies

Key Components Design To assess the effectiveness of individual components, we conduct ablation studies on six RefVOS datasets. As shown in Table 7, starting from the baseline, introducing Temporal Token Injecting (TTI) consistently improves performance, with smaller gains on Ref-Youtube-VOS [52] and Ref-DaVIS17 [28] due to shorter videos and static expression. Incorporating MCS provides further improvements, particularly on motion-intensive datasets such as MeViS, where key frame sampling enhances action understanding. Finally, adding BAP strengthens tracking and error correction, with the largest benefits observed on reasoning-oriented benchmarks involving long video sequences, underscoring the importance of robust temporal tracking mechanisms. The detailed analysis of TTI are discussed in Appendix E.1 and E.2.

Effectiveness of Moment-Centric Sampling Sampling strategies impact performance in a dataset-specific manner. MeViS requires action understanding in long videos where targets appear only in specific segments, Ref-DAVIS17 contains targets present throughout most of the video, and ReasonVOS emphasizes relational reasoning with targets often emerging mid-video. As shown in Table 8, *FirstK* underperforms relative to *Uniform* on MeViS and ReasonVOS, as early frames frequently omit target segments. Keyframe-based strategies mitigate this by prioritizing high-similarity moments: *NearbyK* samples adjacent

Table 4: Performance comparison on reasoning video object segmentation tasks.

Method	ReVOS									ReasonVOS		
	Referring			Reasoning			Overall					
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
TrackGPT-13B [74]	49.5	48.3	50.6	40.5	38.1	42.9	45.0	43.2	46.8	-	-	-
VISA-13B [65]	57.4	55.6	59.1	44.3	42.0	46.7	50.9	48.8	52.9	-	-	-
VideoLISA-3.8B [2]	-	-	-	-	-	-	-	-	-	47.5	45.1	49.9
HyperSeg-3B [57]	58.5	56.0	60.9	53.0	50.2	55.8	55.7	53.1	58.4	-	-	-
ViLLa-6B [72]	-	-	-	-	-	-	57.0	54.9	59.1	55.4	-	-
GLUS-7B [35]	58.3	56.0	60.7	51.4	48.8	53.9	54.9	52.4	57.3	49.9	47.5	52.4
Sa2VA-4B [68]	-	-	-	-	-	-	53.2	-	-	-	-	-
VRS-HQ-7B [20]	62.1	59.8	64.5	56.1	53.5	58.7	59.1	56.6	61.6	-	-	-
Sa2VA-Qwen2.5-VL-3B [68]	61.5	59.0	64.1	55.9	53.0	58.8	58.7	56.0	61.4	49.6	46.9	52.4
MomentSeg-3B (Ours)	65.4	63.0	67.9	59.9	57.1	62.6	62.6	60.0	65.2	61.7	58.2	65.3
MomentSeg-7B (Ours)	**67.4**	**64.9**	**69.8**	**62.8**	**59.8**	**65.7**	**65.1**	**62.3**	**67.8**	**62.7**	**59.2**	**66.1**

frames, while *TopK* selects frames with globally highest similarity. Both strategies rely on the similarity distribution from `[FIND]`, yet they tend to concentrate samples in highly correlated contiguous regions, leading to redundancy in local frames and loss of global context. This limitation is particularly pronounced for long videos with complex textual descriptions. Motivated by this, our proposed **MCS** combines dense sampling near key moments with sparse sampling elsewhere, effectively capturing both crucial local cues and global context. This approach achieves the best overall performance across benchmarks, yielding improvements of 3.0 and 4.9 percentage points over *FirstK* on MeViS and ReasonVOS, respectively. Importantly, MCS introduces negligible computational overhead, contributing less than 0.1% to the total latency.

Effectiveness of Bidirectional Anchor-updated Propagation Mask propagation often struggles with poor initialization when the target appears late in the video. BAP overcomes this by initializing at the most probable target moment, propagating bidirectionally, and applying adaptive memory updates to limit error accumulation. Table 9 shows ablations of its components. *Moment-anchored Updating* notably improves performance, particularly on long video scenarios, by correcting accumulated tracking errors, while *Adaptive Memory Cleaning* provides smaller but consistent improvements. Overall, *Bidirectional Propagation* strategy improves robustness by initializing from the most probable target moment without adding inference cost.

Analysis of Inference Latency As shown in Table 10, MomentSeg achieves notable performance gains on the RefVOS task (+10% on MeVIS and +7% on ReVOS) with only a slight increase in inference latency (approximately 3%). To analyze latency sources, we divide the process into three stages: *Preprocessing*, *MLLM*, and *SAM2*. The runtime of Preprocessing and SAM2 varies across

Table 6: Comparison of model results across image segmentation tasks. RefCOCO/+/g and ReasonSeg use the cIoU metric unless otherwise specified, while GCG employs the mIoU metric. † indicates that the RefCOCO/+/g results are reported with the mIoU metric.

Method	RefCOCO			RefCOCO+			RefCOCOg		ReasonSeg		GCG	
	val	testA	testB	val	testA	testB	val(U)	test(U)	val	test	val	test
Methods without LLMs												
ReLA† [36]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	-	-	-	-
C3VG† [12]	81.4	82.9	79.1	77.1	79.6	72.4	76.3	77.1	-	-	-	-
InstanceVG† [10]	81.4	83.1	79.3	76.6	79.5	71.6	75.9	76.6	-	-	-	-
SEEM [75]	-	-	-	-	-	-	65.7	-	25.5	21.2	-	-
Methods with LLMs												
LISA-7B [30]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	61.3	62.9	62.0	61.7
GLaMM-7B [48]	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9			65.8	64.6
LaSagnA-7B [56]	76.8	78.7	73.8	66.4	70.6	60.1	70.6	71.9	48.8	47.2	-	-
VISA-7B [65]	72.4	75.5	68.1	59.8	64.8	53.1	65.5	66.4	52.7	57.8	-	-
VRS-HQ-7B [20]	73.5	77.5	69.5	61.7	67.6	54.3	66.7	67.5	51.8	52.9	-	-
Sa2VA-4B [68]	78.9	-	-	71.7	-	-	74.1	-	-	-	-	-
MomentSeg-3B (Ours)	82.1	83.7	79.2	76.9	81.1	71.8	78.8	79.2	62.0	64.3	67.0	65.9
MomentSeg-7B (Ours)	**82.6**	**85.1**	**80.2**	**78.2**	**81.9**	**72.3**	**80.1**	**80.1**	**63.3**	**65.5**	**67.8**	**67.9**

Table 7: Effectiveness of the key components. We report the $\mathcal{J}\&\mathcal{F}$ metric.

ID	TTI	MCS	BAP	MeViS (val^u)	MeViS (val)	Ref-DAVIS17	Ref-Youtube-VOS	ReasonVOS	ReVOS (Overall)
1				57.0	51.3	75.5	70.1	54.1	58.3
2	✓			58.7(↑1.7)	52.3(↑1.0)	76.2(↑0.7)	70.3(↑0.2)	58.1(↑4.0)	59.2(↑0.9)
3	✓	✓		61.3(↑2.6)	53.9(↑1.6)	76.8(↑0.6)	70.8(↑0.5)	60.8(↑2.7)	60.4(↑1.2)
4	✓	✓	✓	62.0(↑0.7)	54.1(↑0.2)	76.4(↓0.4)	72.0(↑1.2)	61.7(↑0.9)	62.6(↑2.2)

datasets due to differences in video length and resolution, while MLLM remains stable, accounting for about 20% of total latency. The MLLM overhead in MomentSeg is slightly higher because the TSG stage introduces an additional 7% computation cost, which is still minor compared with SAM2 propagation. This overhead can be further reduced by decreasing the number or resolution of temporal frames in the TSG stage.

Analysis of Robustness We evaluate the robustness of BAP on the MeViS (val^u) dataset using four metrics: *Initial Shot Ratio*, *Initial mIoU*, *Anchor mIoU* with/without updating, and *Correction Ratio*. As shown in Table 11, MomentSeg significantly improves both the Initial Shot Ratio and Initial mIoU by leveraging MCS to select more reliable starting frames. Moreover, anchor updating increases the anchor-frame mIoU by 2.6 points compared with the non-updating setting, highlighting its effective correction ability. Overall, BAP corrects about 7% of low-IoU cases, demonstrating its robustness and practical benefit.

Table 8: Impact of moment-centric sampling.

Method	MeViS (val^u)	Ref-DAVIS17	ReasonVOS
FirstK (Baseline)	58.3	76.3	55.9
Uniform	59.6(↑1.3)	76.2(↓0.1)	57.1 (↑1.2)
KeyFrame (TopK)	60.1	76.9	60.2
KeyFrame (NearbyK)	60.9 (↑0.8)	77.1(↑0.2)	59.9(↓0.3)
MCS	61.3(↑3.0)	77.3(↑1.0)	60.8(↑4.9)

Table 9: Effectiveness of bidirectional anchor-updated propagation.

Method	MeViS (val^u)	Ref-DAVIS17	ReasonVOS
Forward Propagation (Baseline)	58.9	76.2	57.4
+Moment-anchored Updating	61.1(↑2.2)	76.9(↑0.7)	60.8(↑3.4)
+Adaptive Memory Cleaning	61.3(↑0.2)	77.0(↑0.1)	61.0(↑0.2)
+Bidirectional Propagation	62.0(↑0.7)	76.8(↓0.2)	61.7(↑0.7)

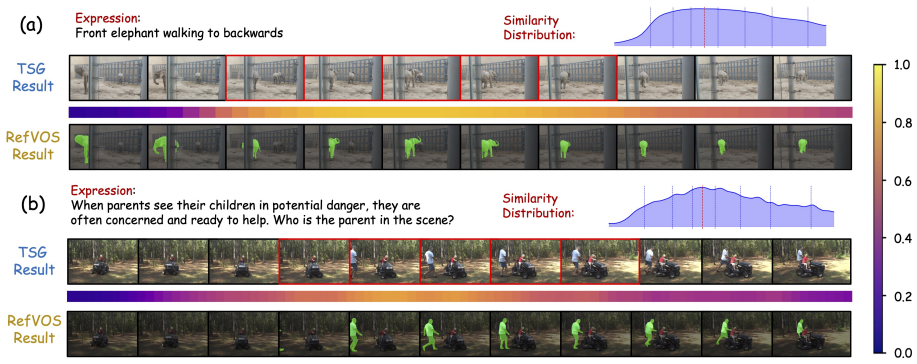

Fig. 5: Qualitative results of MomentSeg on TSG and RefVOS tasks. The figure displays the input expression, frame-query similarity distribution, sampled frames, and the predicted segmentation masks for TSG and RefVOS. (a) illustrates an example for the referring task, while (b) presents an example for the reasoning task.

5 Qualitative Results

As illustrated in Fig. 5, we show qualitative results of MomentSeg on TSG and RefVOS tasks. For the referring task (Fig. 5a), the model accurately localizes the described action and, through dense moment sampling with MCS, delivers precise segmentation. For the reasoning task (Fig. 5b), MomentSeg identifies the temporal segment aligned with complex relational descriptions, highlighting its strong temporal reasoning. More visualizations are provided in Appendix F.

Table 10: Comparison of inference efficiency and performance. The "Latency" denotes the end-to-end inference time per video, measured on a single A100 GPU (40 GB).

Method	Ref-Youtube-VOS		MeVIS (*val*)		ReVOS	
	$\mathcal{J}\&\mathcal{F}$	Latency (s)	$\mathcal{J}\&\mathcal{F}$	Latency (s)	$\mathcal{J}\&\mathcal{F}$	Latency (s)
Sa2VA-8B	70.7	6.8	46.9	7.5	57.6	6.3
MomentSeg-7B (Ours)	72.3(↑1.6)	7.0(↓3%)	57.1(↑10.2)	7.7(↓3%)	65.1(↑7.5)	6.4(↓2%)
↪ Preprocess	-	1.8 (26%)	-	3.3 (43%)	-	2.4 (38%)
↪ MLLM	-	1.3 (19%)	-	1.4 (18%)	-	1.3 (20%)
↪ TSG	-	0.5 (7%)	-	0.6 (8%)	-	0.5 (8%)
↪ RefVOS	-	0.8 (12%)	-	0.8 (10%)	-	0.8 (12%)
↪ SAM2	-	3.9 (55%)	-	3.0 (39%)	-	2.7 (42%)

Table 11: Robustness evaluation of BAP on the MeViS (val^u) dataset. Initial Shot Ratio denotes the percentage of initial frames containing the referred target. Initial mIoU measures the segmentation quality on the initial frame. Anchor mIoU compares performance without and with anchor updating. Correction Ratio indicates the proportion of frames with IoU < 0.3 that are corrected to IoU > 0.7 via updating.

Method	Initial Shot Ratio	Initial mIoU	Anchor mIoU (w/o → w/ Updating)	Correction Ratio
Sa2VA-8B	88.7	54.0	–	–
MomentSeg-7B	96.6 (↑7.9)	63.3 (↑9.3)	61.0 → 63.6 (↑2.6)	7.1

6 Conclusion

In this paper, we introduce MomentSeg, a unified framework that optimizes both TSG and RefVOS tasks simultaneously. Our approach eliminates the reliance on external models for keyframe selection, addressing the inherent complexity of previous methods. By leveraging a `[FIND]` token during training, MomentSeg identifies text-relevant key moments directly, bypassing the need for explicit timestamp encodings and enhancing temporal understanding. For inference, we propose Moment-Centric Sampling (MCS), which intelligently selects relevant frames while preserving essential motion cues and global temporal context. Additionally, our Bidirectional Anchor-updated Propagation (BAP) further improves tracking robustness by mitigating cumulative errors and handling occlusions effectively. Experiments on several RefVOS and TSG benchmarks demonstrate that MomentSeg outperforms existing methods, achieving SOTA results.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Nos. 62276061 and 62436002. This work is also supported by Research Fund for Advanced Ocean Institute of Southeast University (Major Program MP202404). This work is also supported by the SEU Innovation Capability Enhancement Plan for Doctoral Students (CXJH SEU 25125).

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Liu, L., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. NeurIPS (2024)
3. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: CVPR (2022)
4. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR. pp. 961–970 (2015)
5. Chen, S., Lan, X., Yuan, Y., Jie, Z., Ma, L.: Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. arXiv preprint arXiv:2411.18211 (2024)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)
7. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: ECCV (2022)
8. Cheng, W., Dai, M., Lu, H., Yang, W.: Dervos: Decoupling consistent trajectory generation and multimodal understanding for referring video object segmentation. In: CVPR. pp. 24651–24662 (2026)
9. Cuttano, C., Trivigno, G., Rosi, G., Masone, C., Averta, G.: Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In: CVPR. pp. 3395–3405 (2025)
10. Dai, M., Cheng, W., Liu, J.J., Yang, L., Feng, Z., Yang, W., Wang, J.: Improving generalized visual grounding with instance-aware joint learning. TPAMI (2025)
11. Dai, M., Cheng, W., Liu, J.j., Yang, S., Cai, W., Sun, Y., Yang, W.: Deris: Decoupling perception and cognition for enhanced referring image segmentation through loopback synergy (2025)
12. Dai, M., Li, J., Zhuang, J., Zhang, X., Yang, W.: Multi-task visual grounding with coarse-to-fine consistency constraints. In: AAAI. vol. 39, pp. 2618–2626 (2025)
13. Dai, M., Yang, S., Duan, B., Tong, B., Zhuang, J., Yang, W., Wang, J.: Videoseg-o3: A multi-turn reinforcement learning framework for reasoning video object segmentation. arXiv preprint arXiv:2606.06819 (2026)
14. Deng, A., Chen, T., Yu, S., Yang, T., Spencer, L., Tian, Y., Mian, A.S., Bansal, M., Chen, C.: Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In: CVPR. pp. 8625–8636 (2025)
15. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: ICCV (2023)
16. Fang, H., Cong, R., Lu, X., Zhou, X., Kwong, S., Zhang, W.: Decoupled motion expression video segmentation. In: CVPR. pp. 13821–13831 (2025)
17. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: CVPR. pp. 5267–5275 (2017)
18. Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.: Actor and action video segmentation from a sentence. In: CVPR. pp. 5958–5966 (2018)

19. Gong, S., Zhang, L., Zhuge, Y., Jia, X., Zhang, P., Lu, H.: Reinforcing video reasoning segmentation to think before it segments. arXiv preprint arXiv:2508.11538 (2025)
20. Gong, S., Zhuge, Y., Zhang, L., Yang, Z., Zhang, P., Lu, H.: The devil is in temporal token: High quality video reasoning segmentation. In: CVPR. pp. 29183–29192 (2025)
21. Guo, Y., Liu, J., Li, M., Cheng, D., Tang, X., Sui, D., Liu, Q., Chen, X., Zhao, K.: Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3302–3310 (2025)
22. Guo, Y., Liu, J., Li, M., Tang, X., Chen, X., Zhao, B.: Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. arXiv preprint arXiv:2405.13382 (2024)
23. He, S., Ding, H.: Decoupling static and hierarchical motion perception for referring video segmentation. In: CVPR (2024)
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. ICLR (2021)
25. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: CVPR. pp. 14271–14280 (2024)
26. Huang, D.A., Liao, S., Radhakrishnan, S., Yin, H., Molchanov, P., Yu, Z., Kautz, J.: Lita: Language instructed temporal-localization assistant. arXiv preprint arXiv:2403.19046 (2024)
27. Jin, P., Takanobu, R., Zhang, W., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. In: CVPR (2024)
28. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: ACCV. pp. 123–141 (2018)
29. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: ACCV. Springer (2019)
30. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR (2024)
31. Li, H., Chen, J., Wei, Z., Huang, S., Hui, T., Gao, J., Wei, X., Liu, S.: Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In: CVPR. pp. 8592–8603 (2025)
32. Li, J., Lu, W., Fei, H., Luo, M., Dai, M., Xia, M., Jin, Y., Gan, Z., Qi, D., Fu, C., et al.: A survey on benchmarks of multimodal large language models. arXiv preprint arXiv:2408.08632 (2024)
33. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: ECCV (2025)
34. Liang, T., Lin, K.Y., Tan, C., Zhang, J., Zheng, W.S., Hu, J.F.: Referdino: Referring video object segmentation with visual grounding foundations. ICCV (2025)
35. Lin, L., Yu, X., Pang, Z., Wang, Y.X.: Glus: Global-local reasoning unified into a single large language model for video segmentation. In: CVPR. pp. 8658–8667 (2025)
36. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR. pp. 23592–23601 (2023)
37. Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al.: Llava-plus: Learning to use tools for creating multimodal agents. In: ECCV. pp. 126–142. Springer (2024)
38. Liu, Y., Lin, K.Q., Chen, C.W., Shou, M.Z.: Videomind: A chain-of-lora agent for long video reasoning. arXiv preprint arXiv:2503.13444 (2025)

39. Liu, Y., Ma, Z., Pu, J., Qi, Z., Wu, Y., Shan, Y., Chen, C.: Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. NeurIPS **38**, 126078–126108 (2026)
40. Luo, Z., Xiao, Y., Liu, Y., Li, S., Wang, Y., Tang, Y., Li, X., Yang, Y.: SOC: semantic-assisted object cluster for referring video object segmentation. In: NeurIPS (2023)
41. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR. pp. 11–20 (2016)
42. Munasinghe, S., Gani, H., Zhu, W., Cao, J., Xing, E., Khan, F.S., Khan, S.: Videoglamm: A large multimodal model for pixel-level visual grounding in videos. arXiv preprint arXiv:2411.04923 (2024)
43. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: ECCV. pp. 792–807 (2016)
44. Pan, F., Fang, H., Li, F., Xu, Y., Li, Y., Benini, L., Lu, X.: Semantic and sequential alignment for referring video object segmentation. In: CVPR. pp. 19067–19076 (2025)
45. Qian, L., Li, J., Wu, Y., Ye, Y., Fei, H., Chua, T.S., Zhuang, Y., Tang, S.: Momentor: Advancing video large language model with fine-grained temporal reasoning. In: ICML (2024)
46. Qian, L., Li, J., Wu, Y., Ye, Y., Fei, H., Chua, T.S., Zhuang, Y., Tang, S.: Momentor: Advancing video large language model with fine-grained temporal reasoning. arXiv preprint arXiv:2402.11435 (2024)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
48. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: CVPR (2024)
49. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
50. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR. pp. 14313–14323 (2024)
51. Rong, F., Lan, M., Zhang, Q., Zhang, L.: Mpg-sam 2: Adapting sam 2 with mask priors and global context for referring video object segmentation. ICCV (2025)
52. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV (2020)
53. Tang, J., Zheng, G., Yang, S.: Temporal collection and distribution for referring video object segmentation. In: ICCV (2023)
54. Wang, H., Xu, Z., Cheng, Y., Diao, S., Zhou, Y., Cao, Y., Wang, Q., Ge, W., Huang, L.: Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. arXiv preprint arXiv:2410.03290 (2024)
55. Wang, Y., Meng, X., Liang, J., Wang, Y., Liu, Q., Zhao, D.: Hawkeye: Training video-text llms for grounding text in videos. arXiv preprint arXiv:2403.10228 (2024)
56. Wei, C., Tan, H., Zhong, Y., Yang, Y., Ma, L.: Lasagna: Language-based segmentation assistant for complex queries. arXiv preprint arXiv:2404.08506 (2024)

57. Wei, C., Zhong, Y., Tan, H., Liu, Y., Zhao, Z., Hu, J., Yang, Y.: Hyperseg: Towards universal visual segmentation with large language model. CVPR (2025)
58. Wei, C., Zhong, Y., Tan, H., Zeng, Y., Liu, Y., Zhao, Z., Yang, Y.: Instructseg: Unifying instructed visual segmentation with multi-modal large language models. ICCV (2025)
59. Wu, D., Wang, T., Zhang, Y., Zhang, X., Shen, J.: Onlinerefer: A simple online baseline for referring video object segmentation. In: ICCV (2023)
60. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR (2022)
61. Wu, J., Jiang, Y., Yan, B., Lu, H., Yuan, Z., Luo, P.: Segment every reference object in spatial and temporal spaces. In: ICCV. pp. 2538–2550 (2023)
62. Wu, Y., Hu, X., Sun, Y., Zhou, Y., Zhu, W., Rao, F., Schiele, B., Yang, X.: Number it: Temporal grounding videos like flipping manga. In: CVPR. pp. 13754–13765 (2025)
63. Xu, Z., Guo, Y., Lu, Y., Yang, F., Li, J.: Videoseg-r1: Reasoning video object segmentation via reinforcement learning. arXiv preprint arXiv:2511.16077 (2025)
64. Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., Lu, H.: Universal instance perception as object discovery and retrieval. In: CVPR. pp. 15325–15336 (2023)
65. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. ECCV (2024)
66. Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., Schmid, C.: Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: CVPR. pp. 10714–10726 (2023)
67. Yu, L., Poisson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV (2016)
68. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025)
69. Yuan, L., Shi, M., Yue, Z., Chen, Q.: Losh: Long-short text joint prediction network for referring video object segmentation. In: CVPR (2024)
70. Zala, A., Cho, J., Kottur, S., Chen, X., Oguz, B., Mehdad, Y., Bansal, M.: Hierarchical video-moment retrieval and step-captioning. In: CVPR. pp. 23056–23065 (2023)
71. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., Liu, Z.: Lmms-eval: Reality check on the evaluation of large multimodal models (2024)
72. Zheng, R., Qi, L., Chen, X., Wang, Y., Wang, K., Qiao, Y., Zhao, H.: Villa: Video reasoning segmentation with large language model. arXiv preprint arXiv:2407.14500 (2024)
73. Zhong, H., Zhu, M., Du, Z., Huang, Z., Zhao, C., Liu, M., Wang, W., Chen, H., Shen, C.: Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. arXiv preprint arXiv:2505.20256 (2025)
74. Zhu, J., Cheng, Z.Q., He, J.Y., Li, C., Luo, B., Lu, H., Geng, Y., Xie, X.: Tracking with human-intent reasoning. arXiv preprint arXiv:2312.17448 (2023)
75. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. NeurIPS **36**, 19769–19782 (2023)