Online Reasoning Video Object Segmentation

Jinyuan Liu[1,2], Yang Wang[2], Zeyu Zhao[2], Weixin Li[1], Song Wang[2], and Ruize Han[2] *

[1] School of Computer Science and Engineering, Beihang University, Beijing, China
{jinyuanliu, weixinli}@buaa.edu.cn
[2] Faculty of Computer Science and Artificial Intelligence, Shenzhen University of Advanced Technology, Shenzhen, China
{wangyang, hanruize, wangsong}@suat-sz.edu.cn

https://github.com/KN1GHT9/ORVOSB

Abstract. Reasoning video object segmentation predicts pixel-level masks in videos from natural-language queries that may involve implicit and temporally grounded references. However, existing methods are developed and evaluated in an offline regime, where the entire video is available at inference time and future frames can be exploited for retrospective disambiguation, deviating from real-world deployments that require strictly causal, frame-by-frame decisions. We study Online Reasoning Video Object Segmentation (ORVOS), where models must incrementally interpret queries using only past and current frames without revisiting previous predictions, while handling referent shifts as events unfold. To support evaluation, we introduce ORVOSB, a benchmark with frame-level causal annotations and referent-shift labels, comprising 210 videos, 12,907 annotated frames, and 512 queries across five reasoning categories. We further propose a baseline with continually-updated segmentation prompts and a structured temporal token reservoir for long-horizon reasoning under bounded computation. Experiments show that existing methods struggle under strict causality and referent shifts, while our baseline establishes a strong foundation for future research.

Keywords: online video segmentation · reasoning segmentation · strict causality · referent shifts

1 Introduction

A video is not a sequence of images – it is a temporal narrative. Remove the flow, and you erase its meaning.

—André Bazin

Reasoning video object segmentation [15, 32, 50] aims to temporally predict pixel-level masks in videos from natural-language queries, where the referential relationship is implicit requiring logical reasoning. With the rapid progress of multimodal large language models (MLLMs) [2, 20, 26], recent work [50] has

* Corresponding author.

established a unified paradigm that maps video frames and language prompts to segmentation tokens for mask decoding, improving the ability of open-world grounding and compositional reasoning.

Despite this progress, existing methods [13, 36, 44, 50] are all developed and evaluated in an offline regime, where the entire video is available at inference stage. This setting has two significant limitations. First, **retrospective disambiguation**: the model may use future observations to identify the referent and then interpret earlier frames with hindsight [34]. For instance, a vehicle may begin to move at the middle moment of the video, it will be segmented along the whole video, including the first half of the video when it remains stationary. Second, **online application limitation**: while suitable for post-hoc analysis, existing setting does not explicitly assess the frame-by-frame, causal decision making required in real deployments such as robotics [35], embodied agents [53], and on-device video analytics [12], where frames arrive sequentially and future information is unavailable [6].

From the above observations, in this work, we propose to study a new and practical problem, i.e., online reasoning video object segmentation (**ORVOS**). Specifically, the model must interpret the query incrementally using only past and current frames, without revisiting or revising previous predictions [18]. Although practical, the online setting also brings new challenges for the reasoning video object segmentation: i) **Strict Causality**: the referred target can be determinate and segmented only after relevant actions or interactions occur. For instance, 'the moving vehicle' can not be segmented until it starts moving. ii) **Referent Shifts**: many event-driven expressions induce referent shifts, where the queried entity changes over time. For instance, queries such as 'the moving vehicle' or 'the animal that is attacked' may correspond to different instances at different stages of an unfolding event. Such properties introduce challenges beyond global reasoning and temporal association in classical RVOS, including causal disambiguation, progressive interpretation, and shift-aware target tracking.

Considering the above problem characteristics and challenges, existing works mainly encounter the following problems. On the one hand, existing benchmarks are not designed to evaluate these online inference capabilities. Most of the referring or reasoning video object segmentation datasets [4, 15, 32, 50, 56] assume full-video access and typically treat the referent as fixed for each query. Consequently, they cannot faithfully measure incremental decision making, adaptation to evolving targets, or robustness under strictly causal inference. On the other hand, from the modeling perspective, existing MLLM-based approaches are also mostly optimized for offline processing and are not explicitly tailored to incremental inference. Many pipelines summarize the video into a single representation and reuse it for segmentation: some perform keyframe-level reasoning and then propagate predictions via tracking [15, 50], while others segment all frames with a shared video-level feature [4]. More recent methods [25, 56] incorporate global–local fusion to obtain frame-level tokens, but typically do not maintain a causal mechanism to accumulate and reuse past disambiguation cues as tem-

poral memory. This motivates an online design that continually updates the segmentation prompt over time and explicitly stores historical prompt features for progressive disambiguation.

In this work, to establish the study of ORVOS, we build a new benchmark and propose a new baseline method. Specifically, to **fill the benchmark gap**, we introduce ORVOSB, a new Benchmark for Online Reasoning Video Object Segmentation. In ORVOSB, we annotate the referred targets under the strict causality constraint by accurately determining whether the target satisfies the referring queries frame-by-frame. Moreover, ORVOSB includes various referent shifts, which require models to resolve ambiguity over time and dynamically adapt as evidence accumulates. In total, it contains 210 videos and 12,907 annotated frames, together with 512 reasoning queries spanning five categories: attribute, spatial relation, action or state change, interaction, and external knowledge. To **build the methodology basics**, we further propose a simple and effective baseline for ORVOS. The proposed method maintains a continually-updating segmentation prompt: at each time step, the MLLM generates a frame-specific target-aware token conditioned on the current frame, a short window of context frames, and aggregated historical memory, which are together used to construct a history-aware prompt for mask prediction. To support long-term reasoning under bounded computation, we further maintain a structured temporal token reservoir with dense-to-sparse retention and query-based aggregation. Experimental results show that state-of-the-art RVOS methods degrade substantially under strict causality and referent shifts on ORVOSB, while the proposed method establishes a strong baseline on ORVOS, which provides a foundation for future research on the proposed causally-grounded and referent-alterable reasoning segmentation task. In summary, we make the following contributions:

- We propose a new and practical problem of online reasoning video object segmentation, which gets rid of the retrospective disambiguation in classical RVOS and can be applied in real-world online applications.
- We build ORVOSB, a new benchmark with various referent-shift annotations and causality evaluation protocol, which enables accurate assessment of online causality and target switching.
- We propose a simple and effective baseline with past information enhancing and dynamic embedding updating, which uses a continually-updating segmentation prompt and a structured temporal token reservoir for online reasoning mask prediction.
- We provide comprehensive experiments that reveal the limitations of existing methods under online RVOS setting and validate the usefulness of the proposed benchmark and the effectiveness of the proposed method.

2 Related Work

2.1 Referring Video Object Segmentation

Referring video object segmentation (RVOS) [5, 13, 47, 51] aims to segment target objects in videos according to explicit natural language expressions that

directly describe the target. Early approaches primarily rely on multimodal fusion, aligning visual representations with linguistic cues to associate targets across space and time. For instance, URVOS [44] employs cross-modal attention and memory mechanisms to jointly address RVOS and semi-supervised video object segmentation, while Hui *et al.* [19] dynamically recombine linguistic and visual features to highlight spatiotemporal regions of interest. Inspired by DETR-style architectures [59], query-based end-to-end methods such as MTTR [5] and ReferFormer [47] leverage language queries to directly guide target localization and mask decoding within a unified framework. More recently, MLLM-based approaches [4, 32, 50, 56] further enhance compositional reasoning capability to better handle complex expressions.

2.2 Reasoning Video Object Segmentation

Reasoning object segmentation [23, 43, 54] focuses on implicit and compositional expressions that require higher-level inference, where accurate mask prediction depends on intent understanding and sometimes external knowledge. Recent works establish a unified paradigm that connects an MLLM [11, 20, 26] with a mask decoder [22, 42], enabling video frames and language queries to be translated into a special `<SEG>` token that is subsequently decoded into pixel-level masks. Extending this formulation to videos, VISA [50] performs keyframe-level reasoning using a pretrained MLLM and propagates masks temporally to remaining frames. Subsequent works improve spatiotemporal modeling from different perspectives. VideoLISA [32] adopts sparse–dense frame sampling to balance temporal coverage and spatial precision. VRS-HQ [15] introduces hierarchical tokens for enhanced temporal reasoning, followed by mask propagation. GLUS [25] unifies global referring reasoning and local temporal continuity within a transformer framework. Despite stronger reasoning capability, these methods typically rely on offline video access.

2.3 Online Video Understanding

Recent advances in MLLMs [1,3,26] extend video understanding from offline processing to real-time streaming scenarios, where frames arrive sequentially under causal constraints. VideoLLM-Online [9] introduces a streaming framework for temporally aligned, long-context video conversation. StreamingVLM [49] aligns training and inference for stable real-time understanding of unbounded visual streams. LiveVLM [33] reduces inference cost through key-value cache pruning and frame-wise token merging, while StreamChat [48] incorporates hierarchical memory for multi-round dialogue. VideoStreaming [38] further enables arbitrary-length video processing with a constant number of adaptively selected tokens. Although these approaches enable streaming inference, they mainly operate at the language level and focus on conversational understanding. In contrast, pixel-level tasks such as video object segmentation require strict spatial-temporal consistency under causal constraints, yet remain underexplored in online settings.

3 Online Reasoning Video Object Segmentation Benchmark

3.1 Motivation

Most existing referring and reasoning video object segmentation benchmarks [13, 36, 44, 50] are built around an offline protocol, where the full video is accessible during inference. As a result, evaluation may implicitly allow retrospective disambiguation with future frames [34], rather than measuring strictly causal, frame-by-frame segmentation in online video scenarios [6].

More importantly for dataset design, existing benchmarks [4, 50, 56] typically assume a fixed referent for each query. Masks are annotated with a single target identity across the entire sequence, leaving no supervision for event-driven cases where the semantically correct referent changes over time. Without explicit referent-shift annotations and their temporal boundaries, current datasets cannot evaluate whether a model adapts its target online, nor can they disentangle failures caused by early ambiguity from those caused by evidence revealed later.

Recent efforts on online video understanding emphasize sequential input and causal constraints [24, 45, 48, 55], but they mainly focus on language-level tasks and do not provide pixel-level annotations or protocols for causal mask prediction. Therefore, a dedicated benchmark is still missing that simultaneously offers (i) a strictly causal evaluation protocol and (ii) referent-shift annotations to assess online target disambiguation and temporal consistency.

These gaps motivate ORVOSB, a benchmark for online reasoning video object segmentation under strict causality, featuring referent-shift annotations and a causal evaluation protocol.

3.2 Problem Definition

We formally define the task of online reasoning video object segmentation (ORVOS). Given a streaming video and a textual query, the model is required to predict the segmentation mask of the referred object at each time step, without accessing any future frame.

Streaming Video Input. Let $\mathcal{V} = \{I_1, I_2, \ldots, I_T\}$ denote a video of length T, where $I_t \in \mathbb{R}^{H\times W\times 3}$ is the frame at time step t. In the online setting, frames arrive sequentially. At time step t, the model only observes the prefix $\mathcal{V}_{\leq t} = \{I_1, \ldots, I_t\}$.

Textual Query. Let q denote a referring expression that specifies the target object. The query remains fixed throughout the video stream.

Online Segmentation Objective. At each time step t, the model predicts a binary mask $M_t \in \{0,1\}^{H\times W}$ corresponding to the object referred to by q in frame I_t. Formally, an online model f_θ with parameters θ satisfies

$$M_t = f_\theta(\mathcal{V}_{\leq t}, q), \tag{1}$$

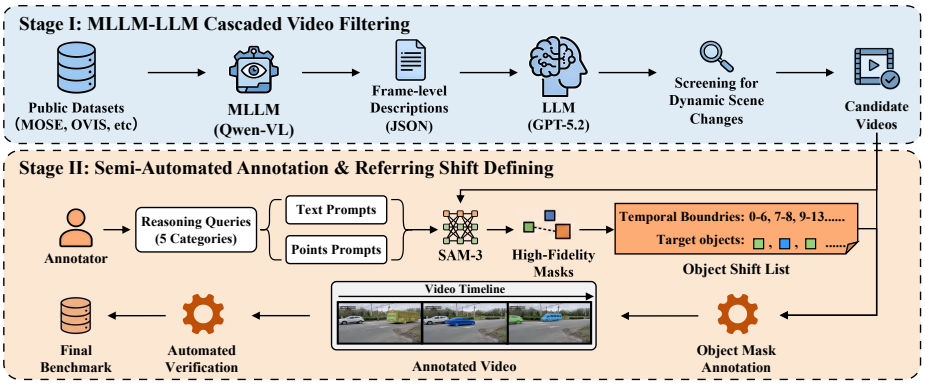

Fig. 1: Overview of ORVOSB data construction and annotation pipeline.

where the prediction at time t must not depend on any future frame $I_{t'}$ with $t' > t$.

Causality Constraints. In ORVOS, the causality constraint is a new yet important concept, which is to say that an object should be segmented iff. It satisfies the condition of the referring expression q. For instance, under the referring expression q indicating 'which watermelon has been cut', in ORVOS, the watermelon should be segmented only after it has been cut. The previous offline benchmarks aim to go through the entire video and then segment the watermelon before it is cut.

Evaluation. The predicted mask sequence $\{M_t\}_{t=1}^{T}$ is evaluated against ground-truth masks $\{M_t^*\}_{t=1}^{T}$ using standard video segmentation metrics under strictly causal inference.

3.3 Data Construction and Annotation

Existing reasoning video object segmentation datasets typically assume a *fixed target to be segmented* throughout the video, which is inadequate for evaluating online reasoning models where *target identities dynamically shift* in tandem with unfolding events. To address this, we construct a new **evaluation** benchmark namely ORVOSB (Online Reasoning Video Object Segmentation Benchmark), containing various *referent shifts*.

Given a referring expression, the referent shifts from one object to another in the video, which is not uncommon. This way, we collect the raw video sequences from established datasets, including MOSE [14], OVIS [37], YouTube-VIS 2022 [52], DAVIS17 [36], and MeViS [13], by selecting videos that exhibit dynamic temporal transitions. Based on the videos in the above datasets, we employ a two-stage MLLM-LLM cascaded automated screening pipeline to efficiently identify candidate sequences. First, a Multimodal Large Language Model (e.g.,

Qwen-VL [2]) processes the video frame sequences to generate detailed, frame-level textual descriptions formatted in JSON. Subsequently, these dense textual representations are fed into a Large Language Model (e.g., GPT-5.2) to identify sequences with potential dynamic scene changes. Human annotators then rapidly review these candidates. For each retained video and its corresponding query, we develop a human-in-the-loop annotation framework. We leverage the SAM-3 [8] model to generate initial high-fidelity segmentation masks based on human-provided prompts, with a small number of manual corrections. Crucially, to capture the *referent shifts*, annotators carefully determine the start and end frames for each specific object identity as the video event unfolds. These temporal boundaries are recorded into a shift list, dynamically associating masks with the evolving query context. Finally, a verification script is employed to cross-check all annotated data, preventing missing frames of the mask annotation to ensure the temporal consistency. The overall pipeline is shown in Fig. 1.

3.4 Dataset Statistics and Analysis

In total, the final constructed benchmark comprises 512 evaluation samples, which come from 210 diverse video sequences, yielding a total of 12,907 annotated frames. The video lengths vary significantly to reflect real-world streaming scenarios, ranging from a minimum of 12 frames to a maximum of 342 frames, with an average length of 61.46 frames per sequence. In this benchmark, we consider five types of reasoning query: attribute, spatial relation, action or state change, interaction, and external knowledge, which are shown in Fig 2.

Unlike static offline referring segmentation benchmarks, our dataset uniquely emphasizes dynamic referring target transitions. On average, each query contains 3.66 referent shifts, meaning the target identity changes multiple times within a single streaming sequence. Furthermore, 55.86% of the queries involve scenarios of discontinuous target grounding, where an object temporarily loses its target status due to unfolding events but later regains it, as shown in Fig 2(a). This requires the model to maintain long-term representations of inactive objects, posing a severe challenge to the memory mechanisms of online reasoning models. This distribution ensures that our benchmark comprehensively evaluates a model's ability to perform causality-grounded, event-aware object segmentation under streaming video input.

4 Method

4.1 Overview

Given a streaming video $\mathcal{V} = \{I_1, \ldots, I_T\}$ and a textual query q, our goal is to perform causal reasoning video object segmentation under strict online constraints. To this end, we propose a memory-enhanced multimodal framework that integrates a multimodal large language model (MLLM), a dynamic token reservoir, and a segmentation head into a unified architecture, as illustrated in Fig. 3.

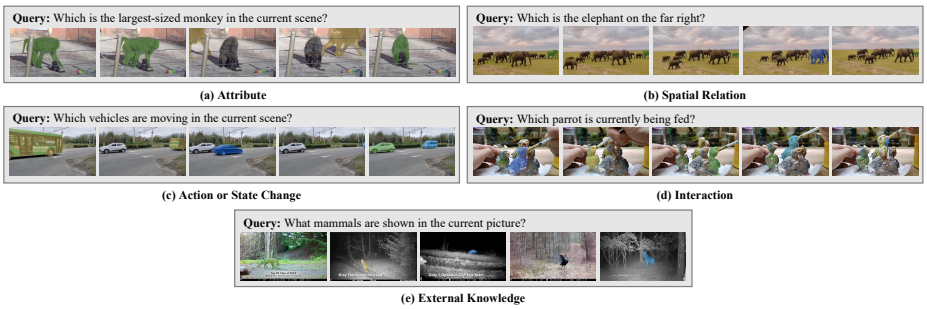

Fig. 2: Illustrative examples of the five reasoning query types in ORVOSB.

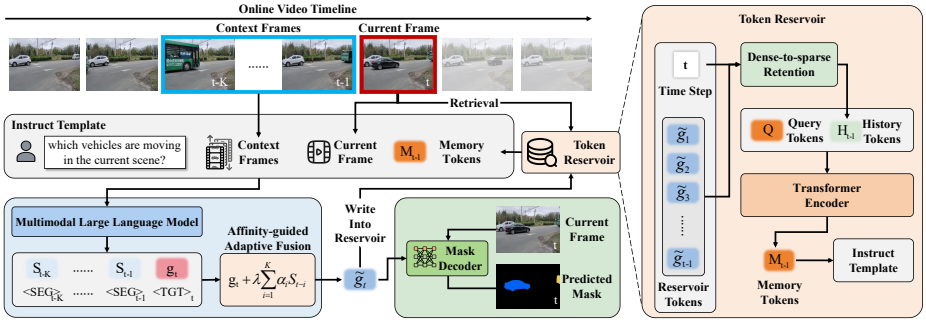

Fig. 3: Overview of the proposed online reasoning video object segmentation framework. At each time step, the model processes context frames and the current frame together with a dynamic token reservoir. The multimodal large language model produces segmentation tokens, which are adaptively fused to form a history-aware prompt for mask prediction. The fused token features are written back to the token reservoir, enabling structured long-term reasoning under strict causal constraints.

At each time step t, the model processes the current frame I_t, a short prefix of recent context frames $I_{t-K:t-1}$ (in terms of a K-frame sliding window), a dynamic token reservoir $\mathcal{M}_{t-1}$, and the query q. The MLLM generates frame-level segmentation tokens for context frames $I_{t-K:t-1}$ and a target-aware token for the current frame t. These tokens are adaptively fused to construct a history-aware query representation, allowing the segmentation prompt to evolve over time as more evidence is accumulated (Section 4.2). The resulting representation is fed into the segmentation head to predict the mask M_t, and the fused token-level features are subsequently stored in the 'token reservoir' for future steps. Here, to ensure stable long-term reasoning under streaming settings, the token reservoir adopts a temporally-aware strategy that retains denser representations for recent observations while progressively sparsifying distant history (Section 4.3). This design enables the model to leverage both short-term temporal context and structured long-term memory while strictly preserving causality during online inference.

4.2 Continually-Updating Segmentation Prompt

MLLM encoding and token extraction. At time step t, we construct the multimodal input sequence according to a predefined prompt template:

$$\mathbf{X}_t = [\mathbf{T}_{\text{inst}};\ \mathbf{T}_q;\ \mathbf{T}_{\text{vis}};\ \mathbf{T}_{\text{mem}}], \tag{2}$$

where $\mathbf{T}_{\text{inst}}$ denotes fixed instruction tokens, $\mathbf{T}_q$ is the tokenized textual query, $\mathbf{T}_{\text{vis}}$ are visual tokens extracted from the context frames $I_{t-K:t-1}$ and current frame I_t, and $\mathbf{T}_{\text{mem}} = \mathbf{M}_{t-1}$ are memory tokens from the token reservoir (Section 4.3). The instruction template further includes predefined placeholder tokens `<TGT>` and `<SEG>` at fixed positions, serving as semantic anchors for the current target frame and context frames.

The sequence is processed autoregressively by the MLLM, yielding contextualized embeddings

$$\mathbf{E}_t = \text{MLLM}(\mathbf{X}_t) \in \mathbb{R}^{L_t \times d}. \tag{3}$$

As shown in Fig. 3, we read the predefined placeholder positions from the L_t contextualized embeddings and obtain: the embedding at position `<TGT>` forms the target-aware feature $\mathbf{g}_t \in \mathbb{R}^d$, while embeddings at positions `<SEG>` form context segmentation features $\{\mathbf{s}_{t-i}\}_{i=1}^{K} \in \mathbb{R}^{K \times d}$. This design allows the MLLM to encode target and contextual information in a unified sequence, while preserving explicit structural roles within the prompt.

Affinity-guided adaptive fusion. To adaptively incorporate historical evidence, we compute the affinity between the current target token and historical segmentation tokens:

$$u_i = \cos(\mathbf{s}_{t-i}, \mathbf{g}_t), \qquad \alpha_i = \frac{\exp(u_i)}{\sum_{j=1}^{K} \exp(u_j)}. \tag{4}$$

The fused prompt representation is obtained by injecting weighted contextual information into the target token:

$$\tilde{\mathbf{g}}_t = \mathbf{g}_t + \lambda \sum_{i=1}^{K} \alpha_i\, \mathbf{s}_{t-i}. \tag{5}$$

where λ is a pre-set parameter. This affinity-guided fusion emphasizes semantically relevant historical cues while suppressing irrelevant context, leading to a history-aware target representation.

4.3 Structured Temporal Token Reservoir

In online video reasoning segmentation, the model should reason over as much temporal context as possible under strict causality while considering a bounded computational budget. Storing all historical features causes unbounded memory growth, while naive truncation may discard critical semantic evidence. Furthermore, recent frames typically provide precise spatial cues, whereas distant

Algorithm 1: Dense-to-Sparse Temporal Sampling

Input: Current step n, maximum memory length $N_{\max}$
Output: Sampled index set $\mathcal{I} \subseteq \{1, \dots, n\}$
if $n \leq N_{\max}$ **then**
  return $\{1, 2, \dots, n\}$;
Initialize $\mathcal{I} \leftarrow \emptyset$;
for $k = 0$ **to** $N_{\max} - 1$ **do**
  $u \leftarrow \frac{k}{N_{\max}-1}$; // Uniform coordinate in $[0, 1]$
  $\phi(u) \leftarrow 1 - (1-u)^2$; // Nonlinear time warping
  $t_k \leftarrow \lfloor \phi(u)(n-1) \rfloor + 1$; // Map to $\{1, \dots, n\}$
  Add t_k to $\mathcal{I}$;
return $\mathcal{I}$;

frames contribute complementary high-level semantics. Motivated by these observations, we design a structured temporal token reservoir that compactly summarizes past information with temporally varying importance.

At time $t-1$, after obtaining the fused prompt feature $\tilde{\mathbf{g}}_{t-1}$ in Section 4.2, we maintain a token reservoir storing semantic features from *all* previous steps. Let $\tilde{\mathbf{g}}_\tau \in \mathbb{R}^d$ denote the feature written at time step τ ($\tau \leq t-1$), and define the token reservoir

$$\mathcal{B}_{t-1} = \{\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_\tau, \dots, \tilde{\mathbf{g}}_{t-1}\}. \tag{6}$$

Dense-to-sparse retention. With the increase of time, storing *all* previous steps is high-cost. To balance recency and long-term coverage, when $|\mathcal{B}_{t-1}| > N_{\max}$, we retain a fixed-length memory by selecting a temporally non-uniform index set $\mathcal{I}_{t-1} \subseteq \{1, \dots, t-1\}$, which samples recent steps more densely and distant steps more sparsely. Concretely, we construct $\mathcal{I}_{t-1}$ via a nonlinear time-warping function and map uniformly spaced coordinates to discrete time indices (Algorithm 1). The resulting sampled memory sequence is

$$\mathbf{H}_{t-1} = [\mathbf{h}_i]_{i \in \mathcal{I}_{t-1}} \in \mathbb{R}^{N \times d}, \tag{7}$$

where $N = N_{\max}$ when $|\mathcal{B}_{t-1}| > N_{\max}$, and $N = |\mathcal{B}_{t-1}|$ otherwise.

Query-based aggregation. Besides history token $\mathbf{H}_{t-1}$, we further introduce L learnable memory query tokens $\mathbf{Q} \in \mathbb{R}^{L \times d}$. We add sinusoidal positional encoding $\mathbf{P}$ to $\mathbf{H}_{t-1}$ and jointly aggregate queries and history by applying a lightweight Transformer encoder $\mathrm{Agg}(\cdot)$ as

$$\mathbf{M}_{t-1} = \mathrm{Agg}\big([\mathbf{Q};\ \mathbf{H}_{t-1} + \mathbf{P}]\big)_{1:L} \in \mathbb{R}^{L \times d}. \tag{8}$$

The resulting **memory tokens** $\mathbf{M}_{t-1}$ serve as an external mnemonic input to the MLLM.

Note that, the initial memory (i.e., $t = 1$) tokens are obtained from $\mathbf{Q}$ only.

Memory update. As shown in Fig. 3, for current time t, we write $\tilde{\mathbf{g}}_t$ into the token reservoir:

$$\mathcal{B}_t = \mathcal{B}_{t-1} \cup \{\tilde{\mathbf{g}}_t\}. \tag{9}$$

4.4 Prompt-Guided Mask Decoding

The fused prompt $\tilde{\mathbf{g}}_t$ is projected into the mask decoder embedding space via a lightweight MLP. Given visual features $\mathbf{F}_t$ extracted from the current frame I_t, the segmentation head predicts the mask:

$$M_t = \mathcal{D}(\mathbf{F}_t, \mathbf{z}_t), \tag{10}$$

where $\mathbf{z}_t$ denotes the projected prompt feature of $\tilde{\mathbf{g}}_t$.

By conditioning mask prediction on the evolving prompt representation, the model integrates visual evidence and accumulated semantic memory while preserving online causality.

4.5 Training and Implementation

Training Objective. We train the proposed framework end-to-end under a streaming protocol. Given a video $\mathcal{V} = \{I_1, \ldots, I_T\}$ and a query q, we unroll the model along time. At each step t, the model takes the current temporal window $I_{t-K:t}$, the aggregated memory tokens $\mathbf{M}_{t-1}$, and the query q, and produces the predicted mask M_t as well as the autoregressive text output Y_t. We optimize a joint objective that combines token-level language modeling and mask supervision. Specifically, the step-wise loss is

$$\mathcal{L}_t = \mathcal{L}_{\text{CE}}(Y_t, \hat{Y}_t) + \lambda_{\text{bce}}\,\mathcal{L}_{\text{BCE}}(M_t, \hat{M}_t) + \lambda_{\text{dice}}\,\mathcal{L}_{\text{DICE}}(M_t, \hat{M}_t), \tag{11}$$

where $\hat{Y}_t$ denotes the ground-truth token sequences, and $\hat{M}_t$ is the ground-truth mask for frame I_t. $\mathcal{L}_{\text{CE}}$ is the standard autoregressive cross-entropy loss [16], $\mathcal{L}_{\text{BCE}}$ is the pixel-wise binary cross-entropy [46], and $\mathcal{L}_{\text{DICE}}$ is the soft Dice loss [31] that encourages mask overlap.

Implementation Details. We build on Chat-UniVi-7B-v1.5 [20] as the MLLM and adapt it via rank-8 LoRA [17]. We freeze the Chat-UniVi backbone and optimize the LoRA adapters and task-specific modules (SAM-2 mask decoder, prompt-projection MLP, and token reservoir). We use AdamW [28] with a learning rate of 3×10^{-4}. We set λ=0.1 in Eq. (5) and λ_{bce}=2, λ_{dice}=0.5 in Eq. (11). Training uses DeepSpeed [41] for 9,000 iterations on 8 RTX 4090 GPUs, with batch size 1 per GPU and 16-step gradient accumulation (effective batch size 128). Unless specified, the context window is K=4 frames. The structured temporal token reservoir performs query-based aggregation with L=32 learnable query tokens via a 2-layer, 8-head Transformer encoder and retrieves up to 32 historical tokens.

5 Experiments

5.1 Setup

Datasets. We train our model on a mixed corpus of image and video data, combining segmentation supervision with multimodal instruction and video question answering data, including LLaVA-Instruct150k [26] and the video QA datasets from Video-ChatGPT [29]. For segmentation, we use standard image segmentation and referring segmentation datasets (e.g., ADE20K [57], COCO-Stuff [7], PACO-LVIS [40], PASCAL-Part [10], refCOCO/refCOCO+/refCOCOg [21, 30], and ReasonSeg [23]), together with video segmentation benchmarks (e.g., Ref-Youtube-VOS [44], Ref-DAVIS17 [36], MeViS [13], and ReVOS [50]). We evaluate our method on ReVOS and our proposed ORVOSB benchmark.

Evaluation Metrics. We evaluate video object segmentation using region similarity $\mathcal{J}$ (mean IoU) and contour accuracy $\mathcal{F}$ (boundary F-measure), and report their average $\mathcal{J}\&\mathcal{F}$ as the primary metric following standard practice in reasoning video object segmentation.

5.2 Results on Online Benchmark

We evaluate on ORVOSB and compare with the video-based methods VISA, VRS-HQ, GLUS, UniPixel and VideoLISA, as well as the image-based baselines LISA and READ, as shown in Table 1. VISA and VRS-HQ achieve only 33.3% and 43.4% $\mathcal{J}\&\mathcal{F}$, as their segment-then-track paradigm fixes the target on key frames and propagates it forward, making it difficult to handle referent shifts where the queried identity changes over time. In contrast, LISA and READ achieve 46.6% and 51.5% $\mathcal{J}\&\mathcal{F}$ by constructing frame-specific segmentation prompts, allowing the grounding representation to adapt to the evolving temporal context. VideoLISA reaches 50.3% $\mathcal{J}\&\mathcal{F}$; however, it does not explicitly incorporate a causality-preserving prompt update mechanism for sequential target re-grounding, which is crucial in the online setting. Our method achieves the best overall performance at 61.3% $\mathcal{J}\&\mathcal{F}$, surpassing READ by +9.8% and VideoLISA by +11.0% with consistent gains across all reasoning categories, validating the benefits of continually-updating segmentation prompt and structured temporal token reservoir for progressive referent grounding under strict causality.

5.3 Ablation Study

We conduct ablation studies on ReVOS and ORVOSB to evaluate the contribution of each module in our framework, as shown in Table 2. The baseline processes only the current frame (and the same context window) into the MLLM, but uses only the current-frame target token as the segmentation prompt. Introducing the continually-updating segmentation prompt (CUSP) yields consistent gains, improving $\mathcal{J}\&\mathcal{F}$ from 52.0% to 53.0% on ReVOS and from 52.4% to 55.1% on

Table 1: Performance comparison with previous methods on the ORVOSB. Type indicates whether a method is image-based (Img) or video-based (Vid).

Methods	Type	Attr.	Spatial	Action	Interact.	Ext. Know.	Overall		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
VISA [50]	Vid	41.6%	31.0%	40.3%	27.5%	39.1%	30.8%	35.7%	33.3%
UniPixel [27]	Vid	35.5%	33.2%	47.7%	27.7%	57.4%	33.5%	38.7%	36.1%
VRS-HQ [15]	Vid	44.4%	44.3%	41.8%	35.8%	43.7%	40.4%	46.4%	43.4%
GLUS [25]	Vid	43.0%	45.8%	47.0%	32.9%	51.1%	42.7%	47.7%	45.2%
LISA [23]	Img	53.6%	46.4%	47.7%	35.1%	49.5%	44.7%	48.4%	46.6%
VideoLISA [32]	Vid	53.2%	52.1%	46.1%	34.0%	55.0%	47.5%	53.1%	50.3%
READ [39]	Img	56.2%	52.4%	50.7%	38.6%	48.9%	49.8%	53.2%	51.5%
Ours	Vid	**63.8%**	**64.9%**	**52.4%**	**40.0%**	**62.0%**	**58.3%**	**64.2%**	**61.3%**

Table 2: Ablation study on ReVOS and ORVOSB. CUSP refers to the continually-updating segmentation prompt, AGAF to affinity-guided adaptive fusion, TR to token reservoir, US to uniform sampling retention, and D2S to dense-to-sparse retention.

Methods	ReVOS			ORVOSB		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Baseline	49.5%	54.5%	52.0%	49.2%	55.6%	52.4%
+ CUSP (w/o AGAF)	50.5%	55.4%	53.0%	51.8%	58.3%	55.1%
+ CUSP	51.3%	56.1%	53.7%	55.0%	61.3%	58.1%
+ CUSP + TR (US)	51.8%	56.4%	54.1%	58.0%	63.9%	60.9%
+ CUSP + TR (D2S)	**52.0%**	**56.6%**	**54.3%**	**58.3%**	**64.2%**	**61.3%**

ORVOSB. Removing affinity-guided adaptive fusion from CUSP slightly reduces the gain, indicating that adaptive fusion better leverages contextual cues. Adding a token reservoir (TR) with uniform retention further improves $\mathcal{J}\&\mathcal{F}$ to 54.1% on ReVOS and 60.9% on ORVOSB, demonstrating the benefit of explicit historical evidence accumulation. Replacing uniform retention with dense-to-sparse retention achieves the best $\mathcal{J}\&\mathcal{F}$ of 54.3% on ReVOS and 61.3% on ORVOSB, balancing recent detail preservation and long-term coverage under a bounded memory budget. Overall, temporal prompt evolution and structured memory are complementary, and their combination performs best.

5.4 Comparison with Offline Methods on General Benchmark

We evaluate on ReVOS and compare with prior methods, as shown in Table 3. Our method achieves 54.3% $\mathcal{J}\&\mathcal{F}$ overall, which is comparable to GLUS at 54.8% $\mathcal{J}\&\mathcal{F}$ and substantially improves over earlier baselines such as LISA and TrackGPT. UniPixel attains the best overall performance at 63.7% $\mathcal{J}\&\mathcal{F}$. Notably, GLUS, VRS-HQ, and UniPixel all incorporate global video understanding, which is particularly effective on ReVOS where the referent remains fixed throughout the sequence and long-range context can be consolidated into a globally consistent representation. In contrast, we evaluate our model under a causal online constraint, producing predictions sequentially without access to

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

Methods	Referring			Reasoning			Overall		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
LISA [23]	44.3%	47.1%	45.7%	33.8%	38.4%	36.1%	39.1%	42.7%	40.9%
TrackGPT [58]	46.7%	49.7%	48.2%	36.8%	41.2%	39.0%	41.8%	45.5%	43.6%
VISA [50]	49.2%	52.6%	50.9%	40.6%	45.4%	43.0%	44.9%	49.0%	46.9%
GLUS [25]	56.3%	60.7%	58.3%	48.8%	53.9%	51.4%	52.6%	57.3%	54.8%
VRS-HQ [15]	59.8%	64.5%	62.1%	53.5%	58.7%	56.1%	56.6%	61.6%	59.1%
UniPixel [27]	63.9%	67.8%	65.8%	59.4%	63.7%	61.5%	61.7%	65.7%	63.7%
Ours	55.7%	59.9%	57.8%	48.4%	53.3%	50.8%	52.0%	56.6%	54.3%

future frames. This setting reduces the ability to leverage full-video cues, yet our method remains highly competitive, indicating that continually-updating segmentation prompt and structured temporal token reservoir can accumulate sufficient evidence over time and transfer robustly to conventional offline benchmarks even under stricter inference constraints.

5.5 Qualitative Results

Fig. 4 presents qualitative comparisons on ORVOSB and ReVOS against strong baselines. On ORVOSB, we compare with VideoLISA under queries involving dynamically evolving referents. As shown in Fig. 4(a), VideoLISA tends to produce inconsistent masks when the queried identity changes over time, since its segmentation representation is largely shared across frames and lacks explicit temporal adaptation. In contrast, our method maintains stable and accurate masks as the referent evolves. This advantage stems from the continually-updating segmentation prompt, where the fused prompt token is updated at each time step, enabling the model to inject newly observed temporal evidence into the segmentation representation and adapt to dynamic changes. On ReVOS, we compare with VRS-HQ as shown in Fig. 4(b). VRS-HQ relies on a segment-then-track paradigm that determines the target on key frames and propagates it forward. Once an early prediction is suboptimal, subsequent masks may drift without correction. In contrast, our framework progressively updates prompt embedding and the token reservoir accumulates disambiguation cues over time. This design allows the model to refine its target representation sequentially and recover from ambiguous early frames, resulting in more temporally consistent segmentation.

6 Conclusion

In this paper, we introduce Online Reasoning Video Object Segmentation (ORVOS), a causally grounded setting that requires frame-by-frame prediction without access to future observations. In contrast to conventional offline RVOS, ORVOS prohibits retrospective disambiguation and instead demands incremental interpretation, particularly under referent shifts. To support this formulation, we

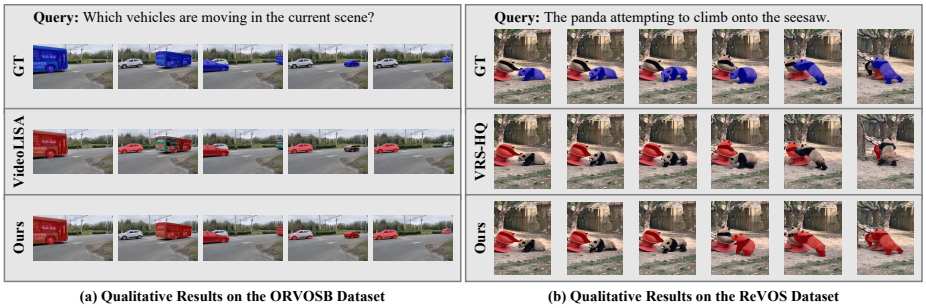

Fig. 4: Qualitative comparison of segmentation results on ORVOSB and ReVOS.

present ORVOSB, a benchmark with frame-level causal annotations and referent-shift labels for systematic evaluation of online reasoning and dynamic target adaptation. We further establish a strong baseline that maintains history-aware segmentation representations via continual prompt updates and structured temporal evidence aggregation under bounded computation. Experiments show that existing RVOS methods degrade substantially in the online regime, underscoring the importance of causality-preserving temporal modeling for future research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62402490, 62576225, 62276018, and the Shenzhen Basic Research Foundation under Grant QNXMC20250701101037019.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. NeurIPS **35**, 23716–23736 (2022)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. NeurIPS **37**, 6833–6859 (2024)
5. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: CVPR. pp. 4985–4995 (2022)
6. Bothra, C., Gao, J., Rao, S., Ribeiro, B.: Veritas: Answering causal queries from video streaming traces. In: SIGCOMM. pp. 738–753 (2023)
7. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR. pp. 1209–1218 (2018)

8. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
9. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: Videollm-online: Online video large language model for streaming video. In: CVPR. pp. 18407–18418 (2024)
10. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: CVPR. pp. 1971–1978 (2014)
11. 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)
12. Dai, P., Chao, Y., Wu, X., Liu, K., Guo, S.: Context-aware offloading for edge-assisted on-device video analytics through online learning approach. IEEE Transactions on Mobile Computing **23**(12), 12761–12777 (2024)
13. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: CVPR. pp. 2694–2703 (2023)
14. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: Mose: A new dataset for video object segmentation in complex scenes. In: CVPR. pp. 20224–20234 (2023)
15. 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)
16. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
18. Huang, Z., Li, X., Li, J., Wang, J., Zeng, X., Liang, C., Wu, T., Chen, X., Li, L., Wang, L.: Online video understanding: Ovbench and videochat-online. In: CVPR. pp. 3328–3338 (2025)
19. Hui, T., Huang, S., Liu, S., Ding, Z., Li, G., Wang, W., Han, J., Wang, F.: Collaborative spatial-temporal modeling for language-queried video actor segmentation. In: CVPR. pp. 4187–4196 (2021)
20. 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. pp. 13700–13710 (2024)
21. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP. pp. 787–798 (2014)
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: CVPR. pp. 4015–4026 (2023)
23. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)
24. Lin, J., Fang, Z., Chen, C., Wan, Z., Luo, F., Li, P., Liu, Y., Sun, M.: Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. arXiv preprint arXiv:2411.03628 (2024)
25. 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)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)

27. 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. Advances in Neural Information Processing Systems **38**, 126078–126108 (2026)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
29. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: ACL. pp. 12585–12602 (2024)
30. 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)
31. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571. Ieee (2016)
32. 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. In: CVPR. pp. 19036–19046 (2025)
33. Ning, Z., Liu, G., Jin, Q., Ding, W., Guo, M., Zhao, J.: Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. arXiv preprint arXiv:2505.15269 (2025)
34. Niu, J., Li, Y., Miao, Z., Ge, C., Zhou, Y., He, Q., Dong, X., Duan, H., Ding, S., Qian, R., et al.: Ovo-bench: How far is your video-llms from real-world online video understanding? In: CVPR. pp. 18902–18913 (2025)
35. Obrenovic, B., Gu, X., Wang, G., Godinic, D., Jakhongirov, I.: Generative ai and human–robot interaction: implications and future agenda for business, society and ethics. AI & society **40**(2), 677–690 (2025)
36. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
37. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H., Bai, S.: Occluded video instance segmentation: A benchmark. IJCV **130**(8), 2022–2039 (2022)
38. Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models. NeurIPS **37**, 119336–119360 (2024)
39. Qian, R., Yin, X., Dou, D.: Reasoning to attend: Try to understand how< seg> token works. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24722–24731 (2025)
40. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., et al.: Paco: Parts and attributes of common objects. In: CVPR. pp. 7141–7151 (2023)
41. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: SIGKDD. pp. 3505–3506 (2020)
42. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
43. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. In: CVPR. pp. 26374–26383 (2024)
44. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV. pp. 208–223. Springer (2020)

45. Wang, H., Feng, B., Lai, Z., Xu, M., Li, S., Ge, W., Dehghan, A., Cao, M., Huang, P.: Streambridge: Turning your offline video large language model into a proactive streaming assistant. arXiv preprint arXiv:2505.05467 (2025)
46. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. In: CVPR. pp. 7303–7313 (2021)
47. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR. pp. 4974–4984 (2022)
48. Xiong, H., Yang, Z., Yu, J., Zhuge, Y., Zhang, L., Zhu, J., Lu, H.: Streaming video understanding and multi-round interaction with memory-enhanced knowledge. arXiv preprint arXiv:2501.13468 (2025)
49. Xu, R., Xiao, G., Chen, Y., He, L., Peng, K., Lu, Y., Han, S.: Streamingvlm: Real-time understanding for infinite video streams. arXiv preprint arXiv:2510.09608 (2025)
50. 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. In: ECCV. pp. 98–115. Springer (2024)
51. Yan, S., Zhang, R., Guo, Z., Chen, W., Zhang, W., Li, H., Qiao, Y., Dong, H., He, Z., Gao, P.: Referred by multi-modality: A unified temporal transformer for video object segmentation. In: AAAI. vol. 38, pp. 6449–6457 (2024)
52. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV. pp. 5188–5197 (2019)
53. Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T.V., Movahedi, M., Li, M., et al.: Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. arXiv preprint arXiv:2502.09560 (2025)
54. Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J.: Lisa++: An improved baseline for reasoning segmentation with large language model. arXiv preprint arXiv:2312.17240 (2023)
55. Yang, Z., Hu, Y., Du, Z., Xue, D., Qian, S., Wu, J., Yang, F., Dong, W., Xu, C.: Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. arXiv preprint arXiv:2502.10810 (2025)
56. Zheng, R., Qi, L., Chen, X., Wang, Y., Wang, K., Qiao, Y., Zhao, H.: Villa: Video reasoning segmentation with large language model. In: ICCV. pp. 23667–23677 (2025)
57. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. pp. 633–641 (2017)
58. 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)
59. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)