ReasonVOS: Benchmarking and Addressing Spatiotemporal-Semantic Reasoning in Instruction-Guided Video Segmentation

Anonymous ACL submission

Abstract

Existing approaches to Reasoning Video Object Segmentation (ReasonVOS) typically generate mask sequences based on implicit instructions combined with external world knowledge. However, these instructions often focus on static or isolated visual elements (e.g., "Which pants are gray"), neglecting the spatiotemporal dynamics intrinsic to video data. In this work, We introduce DualReasonVOS, a new benchmark for ReasonVOS that combines temporal reasoning over object dynamics with semantic reasoning over implicit language, leveraging both visual context and world knowledge. To this end, we redesign the CReaVOS dataset by incorporating carefully curated implicit instructions that emphasize spatiotemporal reasoning. Furthermore, we propose Complex Video Reasoning Segmentation Framework (CVRS), a novel framework that introduces an adaptive reasoning mechanism to decompose implicit instructions into hierarchical reasoning chains. This enables context-aware identification of query-relevant objects across diverse video scenarios. Experimental results demonstrate that CVRS significantly enhances both temporal and spatial reasoning capabilities, achieving superior mask quality compared to state-of-the-art methods on the CReaVOS and ReVOS benchmarks.

1 Introduction

Reasoning Video Object Segmentation (ReasonVOS) (Yan et al., 2024; Zheng et al., 2024; Bai et al., 2024) has recently emerged as a challenging extension of traditional video object segmentation, which aims to generate a sequence of segmentation masks based on implicit natural language queries. Unlike Referring Video Object Segmentation (RVOS) (Khoreva et al., 2018; Seo et al., 2020; Botach et al., 2022; Wu et al., 2022), which focuses on resolving explicit queries (e.g., "*identify the fish*") through direct visual matching, ReasonVOS addresses unspecified queries such as "*identify the animal that breathes with gills*", requiring reasoning grounded in both world knowledge and video context. By integrating semantic reasoning with precise spatial localization, ReasonVOS unlocks the potential for interactive AI agents to understand natural language instructions and engage with complex environments.

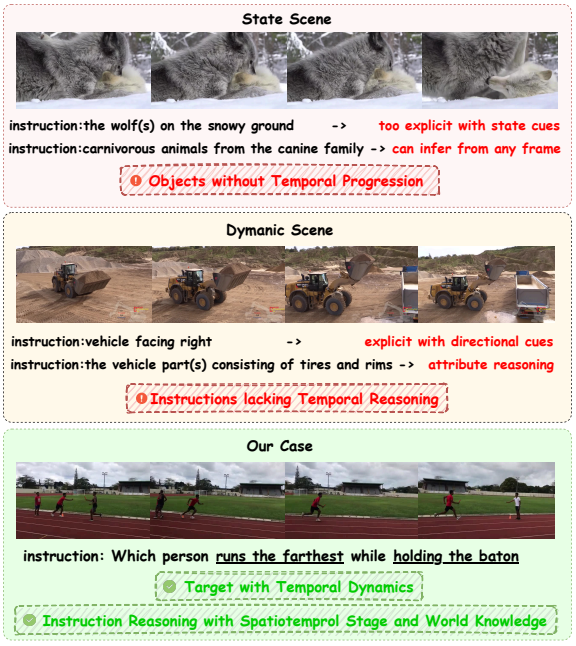

Figure 1: **Comparison between existing instructions in ReasonVOS and our redefined DualReasonVOS.** The first case shows objects in static scene with no temporal progression. The second involves a dynamic video, but the instruction lacks temporal reasoning. In contrast, our case combines temporal object evolution with instructions requiring both spatiotemporal and world knowledge reasoning.

However, a closer look of existing ReasonVOS datasets (Yan et al., 2024) exposes a fundamental limitation: the reasoning required is primarily anchored in the query text and relies heavily on world knowledge, rather than on the temporal dynamics of the video itself. In most cases, the target ob-

ject can be accurately identified within any single frame, rendering temporal modeling unnecessary. This limitation largely stems from two factors, as illustrated in Figure 1: (1) the video scenes tend to be simplistic and lack contextual diversity, and (2) the objects of interest are typically static or exhibit minimal motion. As a result, the task reduces to matching static visual cues with semantic knowledge inferred from the query, rather than engaging in true video-based reasoning across time. While such datasets serve as an important first step, their limited temporal complexity constrains the development and evaluation of models designed to reason over time, thus falling short of realizing the full potential of ReasonVOS to drive progress in dynamic scene understanding.

To address the above shortcomings, we introduce DualReasonVOS, a new benchmark for ReasonVOS that requires dual reasoning—temporal reasoning over object dynamics across frames, and semantic reasoning over implicit natural language. Specifically, Specifically, we first curate a diverse set of videos where either the scene context or the target object exhibits noticeable changes over time. Next, we identify target objects whose states or behaviors evolve throughout the video, ensuring that temporal reasoning is essential for accurate interpretation. Based on these dynamics, we manually craft implicit textual instructions that incorporate both world knowledge and the spatiotemporal behaviors of the target. Finally, we annotate the targets across all frames to produce precise mask sequences, capturing their complete spatiotemporal trajectories. By integrating both temporal and semantic reasoning, DualReasonVOS provides a challenging and realistic testbed for evaluating a model's ability to perform temporally grounded visual reasoning and object identification.

Moreover, we propose Complex Video Reasoning Segmentation (CVRS) framework, a novel framework inspired by the Chain-of-Thought (CoT) paradigm (Wei et al., 2022; Fei et al., 2024). CVRS emulates human reasoning behavior by adaptively decomposing implicit natural language instructions into hierarchical reasoning steps. This decomposition enables context-aware identification of relevant visual cues and target objects across diverse and dynamic video scenarios. The generated reasoning chains not only enhance interpretability but also progressively guide multimodal large language models (MLLMs) toward more precise segmentation by explicitly modeling spatiotemporal dependencies. In contrast to prior methods (Yan et al., 2024; Bai et al., 2024; Zheng et al., 2024), which rely on heavy encoder-decoder architectures and require extensive fine-tuning, CVRS is entirely training-free, offering greater flexibility and ease of deployment. Extensive experiments demonstrate the effectiveness of our method on both ReasonVOS and DualReasonVOS. Overall, our contributions are as follows:

- We introduce DualReasonVOS, a challenging and realistic benchmark for video object segmentation that requires both the temporal reasoning over videos and semantic reasoning over implicit text queries.
- We propose CVRS, a training-free framework with an adaptive reasoning mechanism that decomposes implicit instructions into hierarchical chains, enabling adaptive target identification and guiding fine-grained spatiotemporal reasoning for precise segmentation.
- CVRS achieves state-of-the-art performance on CReaVOS and ReVOS, demonstrating strong spatia-temporal awareness and reasoning capabilities.

2 Related Work

2.1 Referring and Reasoning Video Object Segmentation

RVOS is designed to bridge the vision-language gap by segmenting video objects based on natural language queries (Khoreva et al., 2018; Ding et al., 2023a; Wu et al., 2022; Bellver et al., 2023; Botach et al., 2022). Its effectiveness primarily stems from the use of explicit queries, which enable a direct alignment between textual and visual modalities. To address the limitations of RVOS in handling implicit queries, ReasonVOS extends the task by incorporating world knowledge and video context, enabling more robust and context-aware object segmentation. Early efforts such as VISA (Yan et al., 2024) formulated the ReasonVOS task and established a benchmark for reasoning video segmentation. VideoLISA (Bai et al., 2024) further advanced this direction by introducing a sparse-dense sampling strategy to capture fine-grained spatiotemporal information. ViLLa (Zheng et al., 2024) enhanced temporal context modeling through a temporal-aware encoder and a video-frame decoder. While these approaches

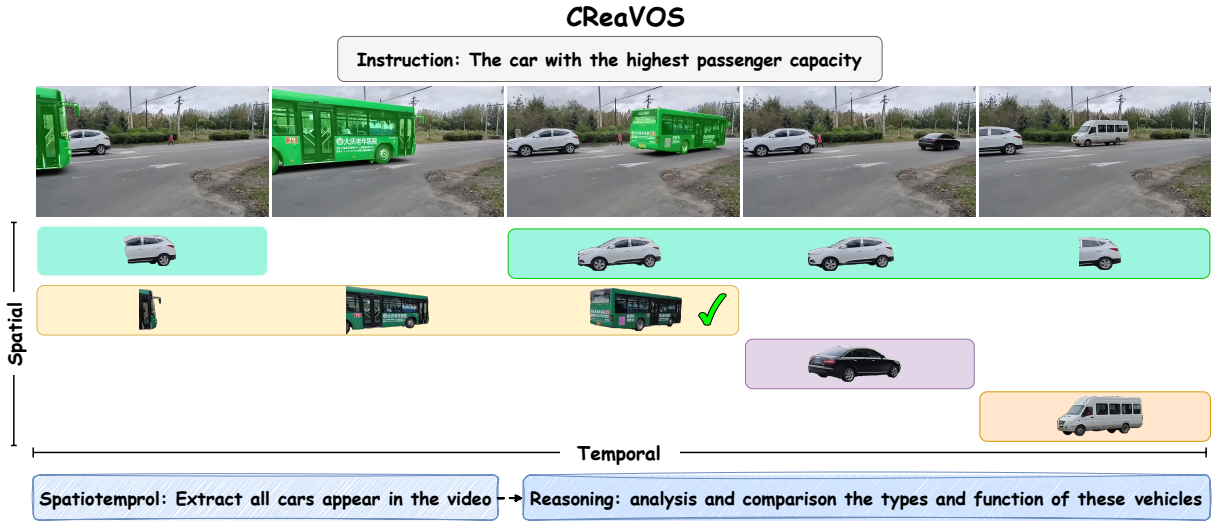

Figure 2: **Illustration of a CReaVOS example.** The presented sample (e.g., "The car with the highest passenger capacity") demonstrates a dual reasoning requirement: spatiotemporal inference over scene and object changes across the video, and semantic understanding involving external world knowledge. This highlights the core design of CReaVOS—grounding complex, dynamic queries within real-world video contexts.

have shown promising results in interpreting implicit queries, they often fall short in fully leveraging the spatiotemporal dynamics of target objects. This work focuses on modeling the spatiotemporal dynamics of target objects to better exploit video-specific features, while incorporating world knowledge to enable more robust and context-aware reasoning.

2.2 Multimodal Chain-of-Thought

Multimodal Chain-of-Thought (MCoT) reasoning has recently emerged as a powerful paradigm in vision-language research, enhancing model interpretability and reasoning ability by generating intermediate, step-by-step inferences. Early works (Zhang et al., 2024b) laid the groundwork for integrating CoT into large multimodal models. Subsequent approaches such as CoCoT (Zhang et al., 2024a) and RelationLMM (Xie et al., 2025) expanded this capability by modeling visual similarities and inter-object relationships, simulating human-like cognitive strategies. MCoT has also shown strong potential in fine-grained tasks. For instance, CoTDet (Tang et al., 2023) and CPSeg (Li, 2024) significantly improved object detection and segmentation by introducing reasoning chains at the instance level. In the video domain, works like CaVIR (Li et al., 2023) and VideoAgent (Wang et al., 2024) addressed long-form video understanding through zero-shot MCoT, while VoT (Fei et al., 2024) proposed a structured five-stage pipeline encompassing task identification, object tracking, and action analysis. Despite these advancements, existing MCoT research has primarily focused on image-level or general video understanding tasks. To the best of our knowledge, no prior work has explored Chain-of-Thought reasoning in the context of ReasonVOS. To fill this gap, we propose an adaptive reasoning mechanism that constructs tailored, hierarchical reasoning steps for each video, enabling fine-grained and context-aware segmentation grounded in world knowledge.

3 A Challenging ReasoningVOS Benchmark

3.1 Task Defination and Formulation

ReasonVOS can be formally defined as follows: given a video $X_V \in \mathbb{R}^{T\times 3\times H\times W}$ comprising T frames of spatial resolution $H \times W$, and an implicit textual instruction X_t, the objective is to learn a model $\mathcal{M}$ that outputs a sequence of binary masks $\{Y_t\}_{t=1}^{T}$, where each $Y_t \in \{0,1\}^{H\times W}$ indicates the mask of target in frame t. Unlike existing datasets ReVOS, which incorporate world knowledge through implicit queries (e.g., "objects for protecting athletes"), these queries primarily emphasize semantic reasoning but fall short in capturing the spatiotemporal dynamics of object throughout the video. ReasonVOS poses significant challenges by requiring integration of world knowledge with robust spatiotemporal and multimodal reasoning capabilities. To support this challenging, we intro-

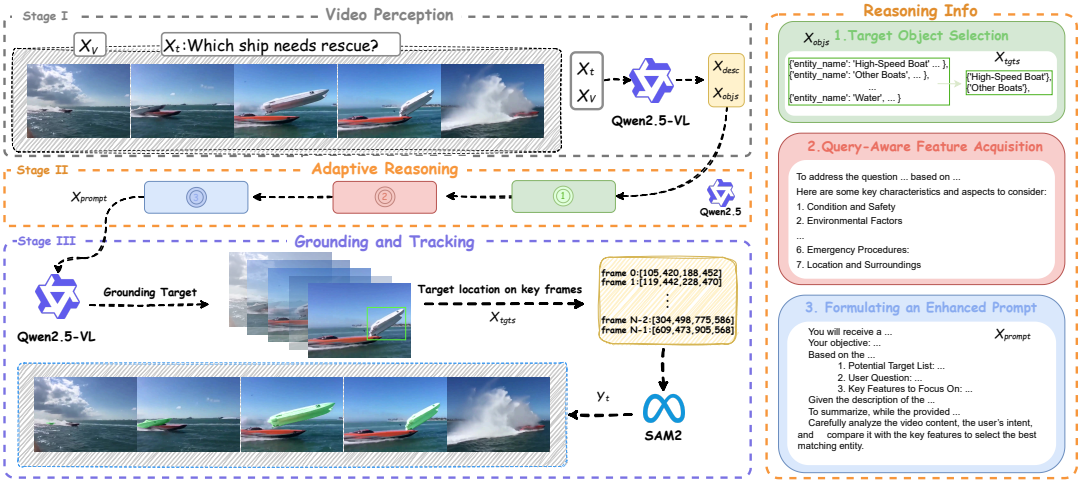

Figure 3: **Overview of the CVRS framework.** The CVRS consists of three stages: (1) Video Perception – the MLLM parses video content and extracts candidate objects; (2) Adaptive Reasoning – the TLM performs text-based reasoning to select the target and generate an enhanced prompt; (3) Target Grounding and Tracking – the MLLM grounds the target using the enhanced prompt, and SAM2 generates the final mask sequence.

duce the CReaVOS dataset, which preserves the implicit nature of instructions (e.g., "the vehicle with the highest passenger capacity") while emphasizing spatiotemporal object evolution. It features dynamic appearances and interactions, requiring joint reasoning over space, time, and external knowledge for accurate segmentation.

3.2 Benchmark

To assess the benchmarking effectiveness of CReaVOS, we construct a standardized evaluation suite comprising implicit reasoning instructions and the corresponding high-quality mask sequences. The benchmark is aligned with the VISA protocol and leverages the ReVOS (Yan et al., 2024) validation set for consistent evaluation. Details of the validation setup are included in Appendix A.

Data Source ReVOS is the first available dataset that collects a diverse set of videos from LV-VIS (Wang et al., 2023), MOSE (Ding et al., 2023b), OVIS (Liu et al., 2022), TAO (Dave et al., 2020), and UVO (Wang et al., 2021). Considering that videos in the TAO dataset are typically several minutes long, significantly increasing the complexity of analysis and computational cost, we exclude TAO during the construction of CReaVOS and sample videos only from the remaining four datasets.

Video Selection and Instruction Design Despite the volume and diversity of ReVOS, most videos with instruction can be resolved from a single frame, thereby simplifying the task to static object recognition rather than spatiotemporal reasoning. To address these limitations, we construct the CReaVOS dataset from two key criterias: video selection and instruction design. First, we ensure that the selected videos contain non-trivial scene dynamics—either through changing environments that lead to evolving object behaviors, or through temporally distinct object states even within static scenes. This ensures that identifying the target requires understanding its temporal trajectory or behavioral evolution. Second, based on the video content, we carefully design implicit instructions that require external world knowledge to identify. These instructions are crafted to refer to objects whose identities can only be inferred by reasoning over their spatial and temporal characteristics across the video. Each video is annotated with corresponding segmentation masks that reflect these temporally grounded targets. Through this approach, CReaVOS not only emphasizes the temporal evolution and contextual variation of target objects, but also integrates rich world knowledge into the instruction design, fostering deeper and more realistic reasoning in video understanding tasks. Figure 2 illustrates examples in CReaVOS.

Data Statistic Following the above construction criteria, we ultimately collected 137 videos from four datasets, each annotated with implicit text instructions and high-quality target mask sequences. Specifically, we collect 76 videos from LV-VIS, 27 from MOSE, 25 from OVIS, and 9 from UVO. Table 1 presents a comparison of the data distribution

between ReVOS and our constructed CReaVOS dataset.

Dataset	ReVOS	CReaVOS	Percentage
LV-VIS	388	76	19.58%
MOSE	208	27	12.98%
OVIS	140	25	17.85%
UVO	255	9	3.50%

Table 1: Distribution of ReVOS and CReaVOS samples across different datasets. Percentage indicates the ratio of CReaVOS to ReVOS samples.

4 The Proposed Framework

4.1 Framework Overall

As shown in Figure 3, Our CVRS framework consists of three main components: a MLLM for video content parsing, object extraction and identification; a text-only language model (TLM) for text-level video reasoning, and Segment Anything 2 (SAM2) as the object tracking model. Specifically, CVRS includes three stages when processing videos. Firstly, **Video Perception**: The MLLM takes the video input and generates a detailed description, from which the objects present in the scene are extracted (4.2). Secondly, **Adaptive Reasoning**: The TLM performs reasoning over the video content based on the video description, extracted objects, and the query. This reasoning process involves three key steps: (1) Target Object Selection, (2) Query-Aware Feature Analysis, and (3) Formulating an Enhanced Prompt for the MLLM (4.3). Finally, **Target Grounding and Tracking**: the video, query, and enhanced prompt are input into the MLLM to identify and localize the target object, which is then tracked across frames using Segment Anything 2 (SAM2) to generate the corresponding mask sequence (4.4).

4.2 Video Perception

In this stage, we employ Qwen2.5-VL as the multimodal large language model to facilitate video perception and the generation of structured textual output. Given an input video X_V comprising T frames, a structured prompt template is designed (see Appendix B.1 for details) to derive a comprehensive textual description, denoted as X_{desc}, along with a list of detected objects, referred to as X_{objs}. Consequently, this process yields rich, high-level video representations that capture scene context, environmental attributes, object presence, and the temporal dynamics associated with each object throughout the video.

4.3 Adaptive Reasoning

The complexity of video understanding by temporal progression, contextual variability, and semantic ambiguity calls for a human-like reasoning strategy that can adaptively interpret visual content in alignment with abstract task objectives. To tackle the challenges posed by evolving object behaviors and complex spatiotemporal interactions in video data, we introduce an Adaptive Reasoning mechanism inspired by human cognitive strategies. This mechanism enables interpretable and efficient reasoning by guiding the model through structured steps aligned with the query semantics. To achieve this in a lightweight and generalizable manner, we incorporate a TLM, which facilitates adaptive reasoning across diverse videos with minimal additional parameters. comprises three key components:

(a) Target Object Selection. From the video perception stage, a diverse set of candidate objects X_{objs} is obtained. However, many of these are irrelevant to the query and may introduce noise. To reduce complexity, we filter this set to obtain X_{tgts}, a subset of objects semantically aligned with the query. For instance, given the query "Which ship needs to be rescued?", we retain relevant objects: {white boat, red boat} and discard unrelated elements like "sky" or "water".

(b) Query-Aware Feature Analysis. Given the inherent complexity and diversity of video targets and scenes, direct alignment between queries and visual content becomes non-trivial. After identifying a set of potential targets, we leverage a lightweight TLM to perform semantic analysis between the candidate objects and the query. Specifically, the TLM reasons about the attributes and temporal behaviors of each object in context, extracting the key discriminative features most relevant to the query intent. This enables adaptive alignment across diverse video scenarios and provides the most informative cues for target identification.

(c) Prompt Construction for MLLM. Since MLLMs are sensitive to prompt design, we construct an enhanced prompt that fuses the visual context and the reasoning output. This adaptive prompt directs the MLLM to focus on temporally and semantically relevant aspects of the video, guiding accurate object localization and segmentation with-

out additional fine-tuning. The formal formulation is given as follows:

$$X_{prompt} = TLM(X_{desc}, X_{objs}, X_t) \quad (1)$$

This stage significantly improves reasoning efficiency and generalization by aligning the model's attention with high-level semantics and temporal cues. Detailed prompt structures are provided in Appendix B.2.

4.4 Target Grounding and Tracking

Methods like VISA, which use LLaMA-Vid to localize targets via keyframes, fall short in scenarios requiring temporal reasoning. Their reliance on isolated frames overlooks the spatiotemporal dependencies essential for complex video understanding. In contrast, our CReaVOS dataset emphasizes targets whose identification hinges on temporal evolution and contextual dynamics, making single-frame resolution insufficient. To tackle this challenge, our adaptive reasoning stage distills the temporal behavior of potential targets and extracts query-relevant discriminative features. These are encoded into an enhanced prompt that explicitly guides the MLLM to focus on critical objects and salient cues, enabling accurate grounding of the target in the video, denotes as X_{tgt}. Leveraging SAM2, we then produce a complete and precise tracklet sequence across the video frames.

$$X_{\text{tgt}} = \mathcal{M}(X_V, X_t, X_{prompt}) \quad (2)$$

$$Y_t = SAM2(X_V, X_{tgt}) \quad (3)$$

5 Experiments

5.1 Dataset and Metrics

To ensure a fair evaluation of our proposed benchmark, we select 800 expression-object pairs from the validation of ReVOS, filtering only those cases where each expression corresponds to a single target object within the video. Additionally, we curate a zero-shot set comprising 137 videos from ReVOS, for which we reconstruct implicit text instructions, ensuring that each expression is aligned with exactly one target object in the video. We evaluate ReasonVOS using region similarity ($\mathcal{J}$), contour accuracy ($\mathcal{F}$), and their mean score ($\mathcal{J}\&F$).

5.2 Implementation Details

Previous end-to-end approaches rely on decoder integration and fine-tuning, leading to high computational costs and limited ability to capture spatiotemporal target dynamics. To overcome these limitations, we introduce CVRS, a fully training-free framework. It leverages the Qwen2.5-VL series for video parsing, target extraction, and grounding, and incorporates the text-only language model Qwen2.5 in the Adaptive Reasoning stage for efficient, high-level semantic reasoning. This stage incrementally generates structured prompts to guide the MLLM toward critical objects and salient cues, enabling accurate and interpretable grounding. Finally, SAM2 is used to produce spatiotemporal tracklets of the identified targets. All experiments run on a single NVIDIA A10 GPU (24 GB VRAM).

5.3 Comparison Results

To showcase the robust video-level, pixel-level perception and generalization of CVRS, we conduct evaluations across ReVOS, CReaVOS datasets. Table 2 illustrates the performance comparison with previous methods on the ReVOS and CReaVOS benchmark. CVRS demonstrates significant improvements over the previous state-of-the-art across three metrics. Since LISA was originally proposed for image reasoning and lacks inherent video reasoning capabilities, we reproduce it by incorporating XMem for video reasoning in this experiment. Remarkably, in terms of $\mathcal{J}\&F$, our CVRS generally achieves over 6.47% $\mathcal{J}\&F$ improvements with LISA in CReaVOS and 16.85% in ReVOS, indicating that LISA can only perform reasoning on static images and is unable to handle continuously changing objects in videos. Moreover, CVRS surpasses VISA by 7.14% in CReaVOS and 13.95% in ReVOS, over VideoLISA by 6.92% in CReaVOS and 16.51% in ReVOS, respectively. These improvements highlight the efficacy of the proposed Adaptive Reasoning CoT module and its dynamic reasoning chains, which substantially enhances the ability of visual perceptino and reasoning of MLLM by directing attention toward semantically relevant aspects of the query, thereby facilitating a more accurate target object.

5.4 Ablation

Ablation of Adaptive Reasoning. The ablation analysis of the Target Object Selection component in the Adaptive Reasoning is presented in Table 3, line 2. Precisely filtering of objects irrelevant to the query improves the prediction of target accuracy (e.g. $\mathcal{J}\&F$ increases from 38.98% to 46.57%) by reducing interference from unrelated objects, thereby effectively narrowing the search space and

Method	CReaVOS(ours)			ReVOS		
	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&F\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&F\uparrow$
LISA+XMem (Lai et al., 2024)	37.07	43.13	40.10	37.87	43.09	40.48
VISA (Yan et al., 2024)	36.87	41.99	39.43	40.56	46.19	43.38
VideoLISA (Bai et al., 2024)	37.32	41.97	39.65	38.94	41.90	40.42
Ours	**45.50**	**47.65**	**46.57**	**53.81**	**60.85**	**57.33**

Table 2: Performance comparison on CReaVOS and ReVOS datasets. Metrics include region similarity ($\mathcal{J}$), contour accuracy ($\mathcal{F}$), and their average ($\mathcal{J}\&F$).

(1)	(2)	(3)	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&F\uparrow$
✓	✓	✓	**45.50**	**47.65**	**46.57**
	✓	✓	37.21	40.74	38.98
✓		✓	35.85	39.12	37.48
✓	✓		32.50	35.37	33.93

Table 3: Ablation study on different components for CReaVOS. (1): Target Object Selection, (2): Query-Aware Feature Analysis, (3): Formulating Enhanced Prompt.

Model Variant	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&F\uparrow$
Qwen2.5-VL-3B	30.00	32.10	31.55
Qwen2.5-VL-7B	45.50	47.65	46.57
Qwen2.5-VL-32B	**73.53**	**71.68**	**72.61**

Table 4: Ablation study on the parameter scale of Qwen2.5-VL for CReaVOS.

Method	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&F\uparrow$
Qwen2.5-0.5B	35.80	38.32	37.06
Qwen2.5-1.5B	42.73	44.72	43.73
Qwen2.5-3B	**45.50**	**47.65**	**46.57**

Table 5: Ablation study on different TLM parameters scale for the CReaVOS dataset

significantly reducing reasoning complexity. To assess the impact ofQuery-Aware Feature Analysis, we compare two variants of our framework, as shown in Table 3, line 3. The results indicate that identifying crucial features related to the query enhances the reasoning chain in human-like thinking, guiding the MLLM to analyze from the most relevant perspective and yielding a 9.09% improvement in $\mathcal{J}\&F$ (from 37.48%). Furthermore, Table 3, line 4 evaluates the effect of customized prompt construction for the MLLM. Compared to directly querying the MLLM, enhanced prompt tailored with potential targets and query-relevant features enable the MLLM to more effectively distinguish the target and enhance its reasoning capabilities based on the provided visual and semantic cues, leading to a 12.64% gain in $\mathcal{J}\&F$ performance.

Ablation of parameters scale of MLLM. Table 4 presents a performance comparison of Qwen2.5-VL models at varying parameter scales: 3B, 7B, and 32B. Results indicate that larger model sizes consistently yield notable improvements in the $\mathcal{J}\&F$ metric. Specifically, Qwen2.5-VL-3B, tailored for edge AI deployment, surpasses the similarly sized VideoLISA-3B by 8.1% (39.65% vs. 31.55%). Compared to VISA-7B, Qwen2.5-VL-7B achieves a 7.14% gain (46.57% vs. 39.43%). Furthermore, Qwen2.5-VL-32B delivers a substantial 26.04% improvement over the 7B variant (72.61% vs. 46.57%). These results underscore the critical role of model capacity in enhancing video perception and understanding—enabling more comprehensive scene interpretation, fine-grained object discrimination, and accurate spatial localization across temporal sequences. While larger models offer superior performance, they come with increased computational and latency costs. To strike a balance between accuracy and efficiency, we select the 7B model as our default configuration for all subsequent experiments.

Ablation of parameters scale of TLM. In the Adaptive Reasoning Phase, we conduct an ablation experience on the parameter scale of the TLM using Qwen2.5 series models with 0.5B, 1.5B, and 3B. Table 5. demonstrate that although the 0.5B model achieves an extremely lightweight configuration, it struggles to perform fine-grained reasoning chain, leading to a performance drop of 9.51% in $\mathcal{J}\&F$ compared to the 3B variant. The 1.5B model achieves a 6.67% improvement over 0.5B and shows only a 2.84% gap from the 3B model, despite having half the parameters. While larger versions of Qwen2.5 (e.g., 7B, 14B) are available, employing such scales contradicts the design philosophy of our Adaptive Reasoning CoT, which aims to achieve optimal reasoning performance with minimal model capacity. Based on this trade-off between performance and efficiency, we adopt

Figure 4: Visualization comparison between the CVRS framework and other benchmark methods on the CReaVOS dataset. Rows from top to bottom represent LISA, VISA, VideoLISA, and Ours.

Figure 5: Visualization of limitations.

the 3B model as the final choice for the TLM.

5.5 Qualitative Comparison

Visualization Analysis. The qualitative comparison between our proposed CVRS framework and existing benchmarks on the CReaVOS dataset across diverse scenarios is shown in Figure 4. the target to be inferred is: the blue truck, which does not appear explicitly throughout the entire video. This necessitates global video-level perception to identify all candidate objects before reasoning. LISA, originally developed for image-level reasoning, lacks the architectural capacity to model temporal dynamics inherent in video data, thereby limiting its applicability in scenarios that require spatiotemporal understanding. Although VideoLISA and VISA are both tailored for the ReasonVOS task and support video-level perception, they still exhibit notable limitations. Specifically, VideoLISA demonstrates ambiguity in spatial localization and overlooks semantically important objects, revealing a shallow understanding of visual semantics. VISA successfully detects a truck appearing later in the video but incorrectly identifies the surrounding sand as the target object. In contrast, our method accurately identifies and localizes the relevant entities in the video. It first filters out potential targets (i.e., the sand), and through the Adaptive Reasoning mechanism, extracts the most crucial features with respect to the input query, ultimately determining the correct target.

6 Conclusion

We redefine the ReasonVOS task with DualReasonVOS, a benchmark requiring both temporal reasoning over object dynamics and semantic reasoning over implicit language. To support this, we reconstruct the CReaVOS dataset, emphasizing queries that reflect spatiotemporal variations and require world knowledge. Building on this, we introduce CVRS, a zero-shot reasoning framework that integrates structured textual reasoning with multimodal grounding to segment target objects from implicit queries. Additionally, we design a multi-step Adaptive Reasoning process guided by a TLM, enabling precise and interpretable reasoning information that support MLLMs in understanding and perceiving complex video content. Extensive experiments on CReaVOS and ReVOS demonstrate the effectiveness and generalization capability of CVRS, achieving state-of-the-art performance on ReasonVOS tasks.

Limitations

Although our model demonstrates strong performance across various benchmarks, it still exhibits certain limitations, which we discuss in this section to inform and motivate future research directions. As illustrated in Figure 5, the inability to locate all chess in the video through text leads to failure in identifying which chess will be moved. The example highlight a core limitation of our method: the inability to localize multiple homogeneous entities in the video solely through textual descriptions. Intuitively, to address this challenge, leveraging a vision encoder pre-trained on video data could significantly enhance the model's spatiotemporal perception for video object understanding. Moreover, integrating multimodal large models with object-aware perception, tracking, and reasoning CoT mechanisms presents a promising direction for future research.

References

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. 2024. One token to seg them all: Language instructed reasoning segmentation in videos. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giró-i-Nieto. 2023. A closer look at referring expressions for video object segmentation. *Multim. Tools Appl.*, 82(3):4419–4438.

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. 2022. End-to-end referring video object segmentation with multimodal transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4975–4985. IEEE.

Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. 2020. TAO: A large-scale benchmark for tracking any object. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 436–454. Springer.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. 2023a. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2694–2703. IEEE.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip H. S. Torr, and Song Bai. 2023b. MOSE: A new dataset for video object segmentation in complex scenes. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20167–20177. IEEE.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. 2018. Video object segmentation with language referring expressions. In *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part IV*, volume 11364 of *Lecture Notes in Computer Science*, pages 123–141. Springer.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. LISA: reasoning segmentation via large language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9579–9589. IEEE.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023. Intentqa: Context-aware video intent reasoning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11929–11940. IEEE.

Lei Li. 2024. Cpseg: Finer-grained image semantic segmentation via chain-of-thought language prompting. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 502–511. IEEE.

Sheng Liu, Kevin Lin, Lijuan Wang, Junsong Yuan, and Zicheng Liu. 2022. OVIS: open-vocabulary visual instance search via visual-semantic aligned representation learning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 1773–1781. AAAI Press.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2020. URVOS: unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV*, volume 12360 of *Lecture Notes in Computer Science*, pages 208–223. Springer.

Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibei Yang. 2023. Cotdet: Affordance knowledge prompting for task driven object detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3045–3055. IEEE.

Haochen Wang, Shuai Wang, Cilin Yan, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. 2023. Towards open-vocabulary video instance segmentation. *CoRR*, abs/2304.01715.

Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. 2021. Unidentified video objects: A benchmark for dense, open-world segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10756–10765. IEEE.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024. Videoagent: Long-form video understanding with large language model as agent. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXX*, volume 15138 of *Lecture Notes in Computer Science*, pages 58–76. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. 2022. Language as queries for referring video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4964–4974. IEEE.

Chi Xie, Shuang Liang, Jie Li, Zhao Zhang, Feng Zhu, Rui Zhao, and Yichen Wei. 2025. Relationlmm: Large multimodal model as open and versatile visual relationship generalist. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3515–3529.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. 2024. VISA: reasoning video object segmentation via large language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XV*, volume 15073 of *Lecture Notes in Computer Science*, pages 98–115. Springer.

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024a. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *CoRR*, abs/2401.02582.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024.

Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. 2024. Villa: Video reasoning segmentation with large language model. *CoRR*, abs/2407.14500.

A Validation of VISA

In the construction of the ReVOS dataset, each video is associated with multiple instructions, some of which correspond to multiple objects within the video. In contrast, our proposed CReaVOS dataset is deliberately designed such that each video is paired with a single implicit textual instruction referring to a specific target object. To ensure fairness in evaluation, we selected 800 instructions from the VISA validation set, each corresponding to a single target object within the video.

B Prompt Design on Each Stage

We provide the template for the prompt used at each stage below.

B.1 Video Perception

Prompt for generating the detailed description and objects with an input video

> Provide a detailed description of the video. List all objects or entities in the video. For each entity in the video, provide a corresponding textual description to distinguish its reference in the video. This description should include appearance and spatial position. Assign a unique identifier to each entity in the following format: [Entity Name: Entity Description]. The output should be in JSON format as follows: { video description: description, entity list:[{ entity name: name, entity description: entity description }, ...] }

B.2 Adaptive Reasoning

Prompt for selecting potential targets from the objects extracted during the video perception stage.

B.2.1 Target Object Selection

To mitigate interference from irrelevant objects, the TLM filters the most semantically relevant candidate target with respect to the query, based on all extracted objects from the video. The corresponding prompt is:

You will receive three pieces of information: Video Description: A description of the overall scene in the video. $\{X_{desc}\}$ Video Question: The question that needs to be answered about the video. $\{X_t\}$ Entity List: A list of extracted entities from the video, each containing an entity name and description. $\{X_{objs}\}$ Objective: Filter the Entity List based on the Video Question, retaining only entities (denotes as $\{X_{tgts}\}$) directly relevant to the question while removing background or environmental details that are unrelated. The selected objects should be returned in JSON format.

B.2.2 Query-Aware Feature Analysis

For each potential target, we examine the key features most relevant to the query semantics. The corresponding prompt is as follows:

After identifying the potential targets: $\{X_{tgts}\}$, which aspects of their characteristics need to be focused on to better address this question: $\{X_t\}$

B.2.3 Enhanced Prompt

After obtaining the potential targets and their query-aware features, we structure this information into an enhanced prompt following the template below and feed it into the MLLM:

Task Description:
You will receive a video, a user question, a list of potential targets related to the video, and the key features of each target.
Your Objective:
Based on the video, user question, and target descriptions, select exactly one most appropriate target from the provided potential target list.
Very Important Constraints:
• You must select one and only one target from the `"Potential Target List"` below.
• You must not generate or infer any entity outside of the given target list.
• Do not return `"None"`, `"Unknown"`, or similar invalid responses.
• Your response must be exactly in this format (no extra text): `<p>entity name</p>` where `entity name` is copied exactly from the list.
1. Potential Target List: $\{X_{tgts}\}$
2. User Question (question): $\{X_t\}$
3. Key Features to Focus On: {target_features}
Carefully analyze the video content, the user's intent, and compare it with the key features to select the best-matching entity.

B.3 Target Grounding and Tracking

After the MLLM identify the target object, we use the following prompt to obtain its grounding information, which is then used by SAM2 to generate the corresponding mask sequence.

Outline the position of {object_name}:{object_description} and output the coordinates in JSON format. Position item should be as form with ["bbox_2d": [x1, y1, x2, y2],"label": {object_name}], these coordinates represent the top-left corner (x1, y1) and the bottom-right corner (x2, y2) of the bounding box.

C Visualizations

Figure 6 presents several visualization examples from CReaVOS.

Figure 6: Visualization of examples from CReaVOS.