CER-RVS: CONTRASTIVE EVIDENCE REFLECTION FOR AGENTIC REFERRING VIDEO OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring Video Object Segmentation (RVOS) requires identifying and segmenting a language-referred object throughout a video, often through reasoning about actions, temporal relations, and interactions. Agentic frameworks offer a flexible, training-free route to RVOS by coordinating pretrained Multimodal Large Language Models (MLLMs) with grounding and segmentation tools, avoiding repeated task-specific adaptation as foundation models evolve. However, unreliable candidate judgments and incomplete object tracks can cause premature target rejection and obscure crucial temporal evidence. We propose CER-RVS, a training-free agentic framework that jointly addresses these reasoning and evidence bottlenecks. Contrastive Evidence Reflection mitigates confirmation bias and premature pruning, enabling agents to distinguish genuine query violations from insufficient evidence and keep unresolved candidates revisable. Its role-specialized collaboration combines supporting and opposing perspectives with independent arbitration, turning evidence conflicts into targeted re-examination rather than repetitive self-judgment or simple voting. Complementarily, Anchor-Guided Track Refinement recovers missing object observations, providing agents with more accurate and temporally complete evidence for reasoning. Together, these contributions strengthen both the visual basis of agentic reasoning and the reliability of target selection. Experiments on MeViS, ReVOS (overall), and ReasonVOS demonstrate state-of-the-art training-free performance, with respective J&F gains of 3.5, 1.9, and 5.0 points over AgentRVOS under the same MLLM backbone, while also surpassing representative training-based methods.

1 INTRODUCTION

Referring Video Object Segmentation (RVOS) requires grounding a language-specified target consistently across an entire video, where the decisive evidence may only emerge through its behavior over time. To meet the demands of real-world video understanding, RVOS systems need to interpret referring expressions involving actions, state transitions, temporal order, and inter-object relations. Such expressions often require reasoning beyond static appearance matching to identify the referred instance (Ding et al., 2023; Yan et al., 2024; Bai et al., 2024). Reliable RVOS therefore depends on two complementary capabilities: reasoning over query-specific temporal and relational cues to distinguish visually or semantically similar candidates, and maintaining sufficiently complete spatio-temporal observations to capture targets that are small, occluded, inconspicuous, or only briefly visible. Effectively coupling these two capabilities remains a central challenge in RVOS.

To address this challenge, recent studies have increasingly incorporated Multimodal Large Language Models (MLLMs) (Li et al., 2024; Liu et al., 2023; Team et al., 2023) into RVOS systems, leveraging their strong cross-modal understanding and reasoning capabilities to interpret increasingly complex referring expressions. Along with the rapid development of MLLMs and visual foundation models, the research paradigm has continuously evolved: from early MLLM-based instruction supervised fine-tuning (Bai et al., 2024; Yan et al., 2024; Yuan et al., 2025), to task-specific adaptation of vision and segmentation foundation models (Liang et al., 2025), and further to reinforcement learning for enhancing explicit reasoning in Video Reasoning Segmentation (Li et al., 2025; Xu et al., 2026). Despite these advances, most existing MLLM-based RVOS methods still rely on task-specific super-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

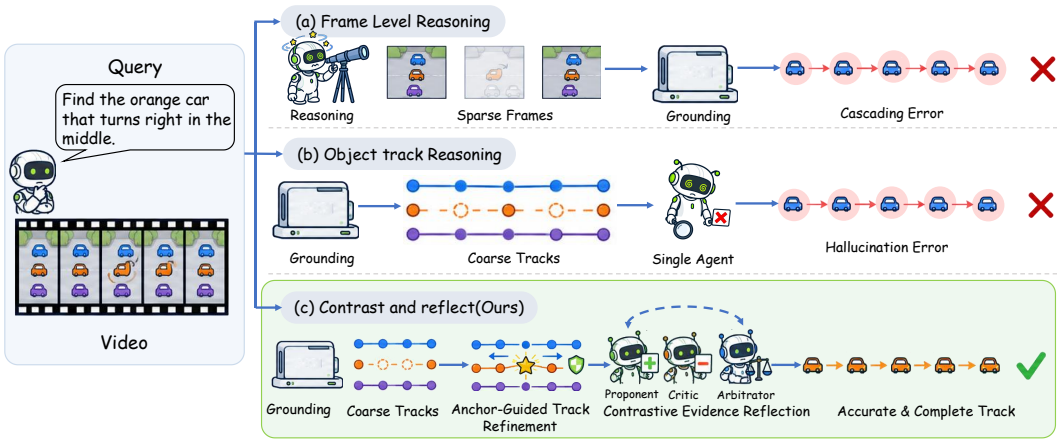


Figure 1: **Comparison of training-free RVOS paradigms.** Frame-level reasoning is vulnerable to sparse visual evidence and cascading grounding errors, while existing object-track reasoning relies on coarse tracks and single-agent decisions. CER-RVS combines Anchor-Guided Track Refinement with Contrastive Evidence Reflection to provide more complete object-level evidence and more reliable reasoning.

vised fine-tuning or reinforcement learning. Such training paradigms typically require large-scale annotated data, carefully designed reasoning supervision, and substantial computational resources. As MLLM backbones continue to evolve rapidly, repeatedly performing costly task adaptation and post-training for every newly released model is expensive and difficult to scale efficiently.

To overcome the reliance on task-specific adaptation, recent studies have explored training-free agentic RVOS by combining the zero-shot reasoning capability of pretrained MLLMs with visual grounding and segmentation models (Ravi et al., 2025; Carion et al., 2026). Existing approaches can be broadly divided into frame-level and object-track-level reasoning, as illustrated in Fig. 1. Frame-level methods (Huang et al., 2025; Kao et al., 2026; Jiang et al., 2026) select a few informative frames for target grounding and propagate the prediction through the video, but sparse observations and unreliable MLLM grounding can lead to missing temporal evidence and cascading errors. Object-track-level methods (Jin et al., 2026) instead reason over candidate trajectories, providing denser visual evidence and better decoupling object discovery from semantic reasoning. However, they typically rely on coarse candidate tracks and single-agent filtering, making them vulnerable to hallucinated judgments, irreversible early pruning, and incomplete trajectories caused by occlusion, scale variation, or weak localization. Thus, although object tracks offer a stronger reasoning substrate than sparse frames, both the reliability of reasoning and the quality of the underlying visual evidence remain key bottlenecks.

To address these limitations, we propose **CER-RVS**, a training-free agentic framework for RVOS that jointly tackles unreliable candidate judgments and incomplete object-track evidence. At the reasoning level, **Contrastive Evidence Reflection (CER)** mitigates confirmation bias and premature target rejection, enabling agents to distinguish genuine query violations from insufficient evidence and to keep unresolved candidates revisable. Rather than relying on repetitive self-judgment or simple voting, its role-specialized multi-agent collaboration directs reflection toward conflicting evidence and unresolved query constraints, improving the reliability of candidate discrimination. At the perception level, **Anchor-Guided Track Refinement (AGTR)** recovers object observations missed by coarse candidate tracks, alleviating gaps in temporal evidence associated with occlusion, scale variation, or weak localization. The resulting improvements in track accuracy and temporal completeness provide agents with a stronger visual basis for reasoning about actions, state transitions, and object interactions. Together, these contributions reduce failures arising from both incomplete observations and unreliable decisions, enabling more robust target identification and segmentation without task-specific training.

We evaluate our method on three major language-guided video object segmentation benchmarks involving complex reasoning, world knowledge, and temporal understanding. Extensive experiments

demonstrate that CER-RVS achieves state-of-the-art performance, substantially outperforming previous training-free approaches and also surpassing representative SFT-based methods. We further conduct comprehensive ablation studies to analyze the effectiveness of each component in training-free RVOS systems. Overall, our contributions are summarized as follows:

- We systematically analyze the key limitations of existing training-free agentic RVOS paradigms in terms of reasoning reliability and object-track quality, and propose CER-RVS, a new evidence-verified agentic framework that jointly addresses both issues without task-specific training.
- We introduce two complementary components: Contrastive Evidence Reflection (CER), which improves reasoning robustness through explicit comparison of supporting and counter-evidence with arbitrator-guided feedback, and Anchor-Guided Track Refinement (AGTR), which reconstructs candidate trajectories through reliability- and scale-aware anchor selection followed by instance-conditioned bidirectional re-tracking.
- Without any additional training, CER-RVS establishes new state-of-the-art performance across multiple RVOS benchmarks, significantly outperforming existing training-free approaches and surpassing representative training-based methods.

2 Related Work

Referring Video Object Segmentation. Referring Video Object Segmentation (RVOS) aims to segment the target object in a video according to a natural-language expression. Early RVOS benchmarks (Gavrilyuk et al., 2018; Khoreva et al., 2018; Seo et al., 2020) mainly focused on appearance-based descriptions, where target objects could typically be identified by static attributes such as category, color, or spatial location. Accordingly, early methods (Botach et al., 2022; Han et al., 2023; Luo et al., 2023; Miao et al., 2023; Wu et al., 2022; Yuan et al., 2024) predominantly adopted query-based architectures built upon DETR (Carion et al., 2020) to associate linguistic expressions with object-level visual representations.

As RVOS benchmarks have evolved (Ding et al., 2023; Yan et al., 2024), referring expressions have progressed beyond appearance-based descriptions toward more challenging queries involving motion, temporal relations, complex reasoning, and world knowledge. To address these challenges, recent approaches have increasingly integrated Multimodal Large Language Models (MLLMs) with foundation segmentation models, typically through large-scale supervised fine-tuning (Bai et al., 2024; Yan et al., 2024; Yuan et al., 2025) or reinforcement-learning-based optimization (Li et al., 2025; Xu et al., 2026). These methods substantially enhance semantic reasoning and generalization, but often require large-scale annotated data and considerable computational resources.

More recently, training-free agentic approaches have emerged to better leverage the zero-shot reasoning capabilities of rapidly evolving foundation models. AL-Ref-SAM2 (Huang et al., 2025) and CoT-RVS (Kao et al., 2026) follow a frame-level reasoning paradigm, where an MLLM first selects informative keyframes and localizes the referred object, after which the predicted masks are propagated throughout the video. Refer-Agent (Jiang et al., 2026) further introduces iterative agentic reasoning and reflection. Recent work (Jin et al., 2026) has also begun to shift from frame-level reasoning toward object-track-level reasoning, where candidate object tracks generated over the entire video provide denser object-level visual evidence for MLLMs.

Multi-Agent Verification and Reflection. Multi-agent debate can improve reasoning and factuality through iterative exchanges among model instances (Du et al., 2024), while adversarial variants introduce opposing roles and a judge (Liang et al., 2024). Self-Refine (Madaan et al., 2023), Reflexion (Shinn et al., 2023), and CRITIC (Gou et al., 2024) further explore feedback-driven refinement and verification. In RVOS, Refer-Agent (Jiang et al., 2026) employs questioner–responder reflection, AgentRVOS (Jin et al., 2026) reasons over candidate tracks using the complete expression and revisits uncertain candidates, and SSUPER (Lee et al., 2026) combines same-prompt multi-model reviews with target-existence verification. Unlike prior approaches, CER couples candidate-wise arbitration of supporting and counter-evidence with conflict-directed re-examination of unresolved tracks under the complete referring expression.

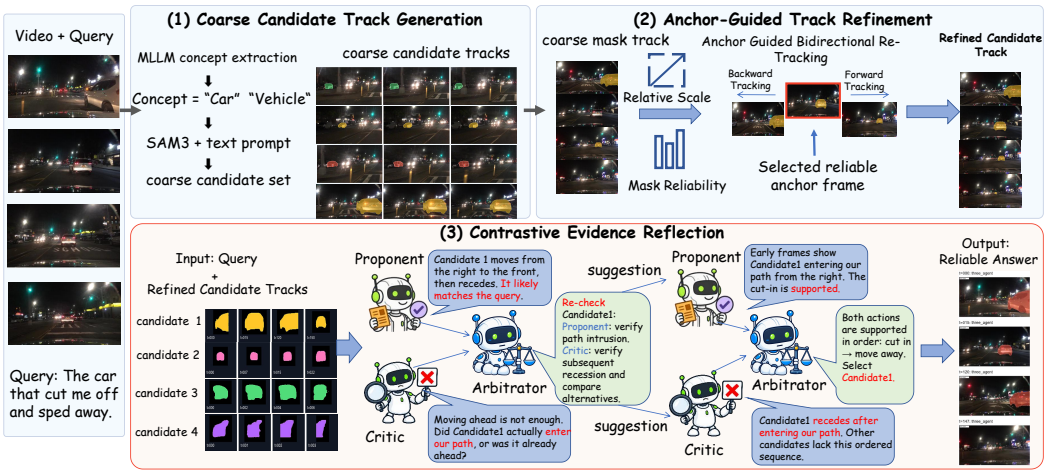


Figure 2: **Overview of the CER-RVS framework.** Given a video-query pair, coarse candidate tracks are first generated and refined by Anchor-Guided Track Refinement (AGTR). Contrastive Evidence Reflection (CER) then verifies each refined candidate through supporting and counter-evidence, with arbitrator-guided reflection for reliable target selection.

3 Methodology

3.1 Problem Formulation and Overview

Given a video $V = \{I_t\}_{t=1}^{T}$ and a natural-language query Q, Referring Video Object Segmentation (RVOS) aims to predict a binary mask sequence $M = \{M_t\}_{t=1}^{T}$, where $M_t \in \{0,1\}^{H\times W}$ denotes the segmentation mask of the referred object at frame t. As illustrated in Fig. 2, CER-RVS first refines coarse candidate tracks with Anchor-Guided Track Refinement (AGTR), and then applies Contrastive Evidence Reflection (CER) to verify the refined candidates and select the referred target.

3.2 Anchor-Guided Track Refinement

Concept-Grounded Candidate Generation. Given a video $V = \{I_t\}_{t=1}^{T}$ and a referring expression Q, an MLLM extracts a specific visual concept and a broader alternative, e.g., "bus" and "vehicle". Each concept is used as a text prompt for SAM3 to generate candidate instance tracks, and we retain the concept that produces more distinct tracks together with its candidate set. Let $c^\star$ denote the retained concept and K the number of candidate tracks.

The coarse mask sequence of candidate k is represented as

$$\widetilde{\mathbf{M}}_k = \big(\widetilde{M}_{k,t}\big)_{t=1}^{T}, \qquad k = 1, \ldots, K,$$

where $\widetilde{M}_{k,t}$ denotes its predicted mask at frame t, with missing predictions represented by all-zero masks. We further define the set of observed frames as

$$\mathcal{I}_k = \Big\{ t \in \{1, \ldots, T\} : |\widetilde{M}_{k,t}| > 0 \Big\},$$

where $|\cdot|$ denotes the number of foreground pixels. These coarse tracks provide the candidate instances for subsequent refinement, but their mask quality may vary across frames, motivating candidate-wise anchor selection.

Candidate-wise Anchor Selection. Coarse candidate tracks may contain missing or inaccurate masks. To repair them through bidirectional re-tracking, we select an *anchor*, consisting of a frame index and its corresponding instance mask, which specifies both the initialization point and the candidate instance to propagate. Because the quality of the refined track depends on this initialization, the anchor should provide both a reliable mask and sufficient target visibility. We therefore first filter

observations by mask reliability and then jointly consider reliability and relative mask area within the reliable set.

For candidate k at an observed frame $t \in \mathcal{I}_k$, we define the mask reliability score as

$$E_{k,t} = s_{k,t}^{\text{obj}} s_{k,t}^{\text{iou}}, \tag{1}$$

where $s_{k,t}^{\text{obj}}$ denotes the segmentation model's confidence that candidate k is present in frame t, and $s_{k,t}^{\text{iou}}$ is its predicted mask IoU, which estimates the quality of $\widetilde{M}_{k,t}$. The latter is a model-predicted quality score rather than an IoU computed against ground truth during inference.

In the first stage, we rank all available observations of candidate k according to their reliability scores and retain only the N most reliable ones:

$$\mathcal{C}_k = \text{TopN}_{t \in \mathcal{I}_k} \left(E_{k,t} \right), \tag{2}$$

where $\mathcal{C}_k \subseteq \mathcal{I}_k$ contains the frame indices associated with the $N_k = \min(N, |\mathcal{I}_k|)$ largest reliability scores. This reliability pre-filtering restricts the subsequent anchor selection to observations that are already reasonably trustworthy, reducing the influence of low-quality masks.

In the second stage, we additionally consider the spatial support of the target within these reliable observations. For each $t \in \mathcal{C}_k$, we define the relative mask area as

$$S_{k,t} = \frac{|\widetilde{M}_{k,t}|}{\max_{\tau \in \mathcal{C}_k} |\widetilde{M}_{k,\tau}|}, \tag{3}$$

where τ indexes the reliable observations in $\mathcal{C}_k$. Thus, $S_{k,t} \in (0, 1]$ measures the mask area in frame t relative to the largest reliable mask of the same candidate. A larger value indicates that the target occupies a larger visible region among the reliable observations and therefore provides stronger spatial support for initializing instance-level re-tracking. We use this quantity only as an auxiliary cue of target visibility and spatial support, rather than as an estimate of segmentation correctness.

We then combine mask reliability and relative spatial support to compute the anchor score:

$$A_{k,t} = \alpha E_{k,t} + (1 - \alpha) S_{k,t}, \qquad t \in \mathcal{C}_k, \tag{4}$$

where $\alpha \in [0, 1]$ controls the trade-off between segmentation reliability and target spatial support. The anchor frame of candidate k is selected as $t_k^\star = \arg\max_{t \in \mathcal{C}_k} A_{k,t}$. Here, $t_k^\star$ denotes the selected anchor-frame index. Its corresponding instance mask is $M_k^{\text{anc}} = \widetilde{M}_{k,t_k^\star}$. The anchor for candidate k is therefore the pair $(t_k^\star, M_k^{\text{anc}})$. Overall, this two-stage strategy first removes unreliable observations and then selects, among the remaining reliable masks, an initialization that jointly provides high mask quality and sufficient target visibility.

Anchor-Guided Bidirectional Re-Tracking. For each candidate k, we initialize an independent single-object segmentation session at frame $t_k^\star$ with mask prompt M_k^{anc}. The specified instance is tracked backward to frame 1 and forward to frame T:

$$\widehat{\mathbf{M}}_k = G_{\text{PVS}} \left(V, t_k^\star, M_k^{\text{anc}} \right) = \left(\widehat{M}_{k,t} \right)_{t=1}^{T}. \tag{5}$$

Here G_{PVS} denotes mask-prompted bidirectional video segmentation, and $\widehat{M}_{k,t}$ is the resulting binary mask of candidate k at frame t. Unlike concept-grounded candidate generation, this pass conditions on an instance mask rather than category-level text. It can revisit missing or unreliable coarse observations, even outside the original observed set $\mathcal{I}_k$, providing temporal evidence for actions and interactions.

The coarse track only selects the anchor; the re-tracked sequence directly replaces it without frame-wise mask fusion. Applying this independently to all K candidates gives

$$\widehat{\mathcal{P}} = \left\{ \widehat{\mathbf{M}}_k \right\}_{k=1}^{K}, \tag{6}$$

These tracks and the complete query Q enter CER. AGTR supplies observations for assessing action and temporal or relational constraints; CER decides which candidate satisfies the full expression.

3.3 CONTRASTIVE EVIDENCE REFLECTION

Although object tracks provide denser visual evidence than sparsely sampled frames, reliably identifying the referred instance remains challenging for complex expressions involving actions, temporal relations, state changes, and object interactions. Existing track-level methods typically rely on a single agent to progressively accept or reject candidates, making the reasoning susceptible to *confirmation bias* and irreversible early errors. In particular, a category-plausible candidate may be favored before the discriminative action, relational, or temporal constraints in the complete query are sufficiently examined. To improve reasoning reliability, we introduce **Contrastive Evidence Reflection (CER)**, which explicitly contrasts supporting and counter-evidence and uses an independent Arbitrator to resolve conflicts and guide iterative reflection.

Contrastive Evidence Collection. For notational simplicity, let $\mathcal{T} = \widehat{\mathcal{P}}$ denote the refined candidate set produced by AGTR, and write its k-th trajectory as $\mathcal{T}_k = \widehat{\mathbf{M}}_k$. For each candidate $\mathcal{T}_k$, CER instantiates two complementary reasoning agents: a **Proponent** $\mathcal{A}^+$ that searches for evidence supporting the hypothesis that $\mathcal{T}_k$ is the referred object, and a **Critic** $\mathcal{A}^-$ that searches for evidence contradicting the same hypothesis. At reasoning round r, the two agents independently produce

$$\mathcal{E}_k^{+,(r)} = \mathcal{A}^+\left(Q, \mathcal{T}_k; \mathcal{S}_k^{(r-1)}\right), \qquad \mathcal{E}_k^{-,(r)} = \mathcal{A}^-\left(Q, \mathcal{T}_k; \mathcal{S}_k^{(r-1)}\right), \tag{7}$$

where $\mathcal{E}_k^{+,(r)}$ and $\mathcal{E}_k^{-,(r)}$ denote the supporting and counter-evidence, respectively, and $\mathcal{S}_k^{(r-1)}$ is the reflection suggestion from the previous round, with $\mathcal{S}_k^{(0)} = \varnothing$.

Both agents are instructed to reason over the *complete referring expression* rather than relying primarily on category-level compatibility. Accordingly, the collected evidence may involve appearance, motion, temporal order, spatial relations, interactions, or state transitions. The asymmetric roles of the two agents turn candidate reasoning from unilateral hypothesis confirmation into explicit evidence competition: the Proponent explains why a candidate could match the query, whereas the Critic actively identifies observations that make it inconsistent with the query.

Contrastive Evidence Arbitration. Independent evidence collection alone does not guarantee a reliable decision, since either agent may hallucinate or over-interpret ambiguous observations. We therefore introduce an independent **Arbitrator** $\mathcal{A}^a$ to explicitly adjudicate between the two sides of evidence:

$$\left(d_k^{(r)}, \mathcal{C}_k^{(r)}, \mathcal{S}_k^{(r)}\right) = \mathcal{A}^a\left(Q, \mathcal{E}_k^{+,(r)}, \mathcal{E}_k^{-,(r)}\right), \tag{8}$$

where $d_k^{(r)} \in \{\texttt{Accept}, \texttt{Reject}, \texttt{Uncertain}\}$ is the candidate-level decision, $\mathcal{C}_k^{(r)}$ summarizes the evidence conflict, and $\mathcal{S}_k^{(r)}$ is a targeted suggestion for subsequent reflection.

Rather than simply voting between the two agents, the Arbitrator evaluates whether their observations are relevant to the discriminative conditions in Q, whether the supporting and counter-evidence are mutually consistent, and which query condition remains unresolved. Importantly, counter-evidence is not treated as an automatic rejection signal. The Arbitrator distinguishes genuine contradiction from insufficient visual evidence, preventing ambiguous or incomplete observations from causing premature candidate elimination.

Arbitrator-Guided Reflection. When the evidence is conflicting or insufficient, CER retains the candidate and uses $\mathcal{S}_k^{(r)}$ to guide the next reasoning round. The suggestion explicitly identifies what should be re-examined, such as an overlooked temporal interval, a specific action, an object interaction, or a query condition interpreted inconsistently by the two agents. The evidence is then revised as:

$$\mathcal{E}_k^{\pm,(r+1)} = \mathcal{A}^{\pm}\left(Q, \mathcal{E}_k^{\pm,(r)}, \mathcal{S}_k^{(r)}\right). \tag{9}$$

Table 1: Comparison with state-of-the-art methods on Referring and Reasoning VOS benchmarks: MeViS (Ding et al., 2023), ReVOS (Yan et al., 2024), and ReasonVOS (Bai et al., 2024). Qwen3-VL-8B-T indicates Qwen3-VL-8B-Thinking. The best-performing results are presented in **bold**, while the second-best results are underlined. † denotes results reproduced and reported by the AgentRVOS authors (Jin et al., 2026), which we quote rather than rerun.

Method	MLLM	MeViS			ReVOS									ReasonVOS		
					Referring			Reasoning			Overall					
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Training-based Methods																
VideoLISA	LLaVA-3.8B	41.3	47.6	44.4	-	-	-	-	-	-	-	-	-	45.1	49.9	47.5
VISA	Chat-UniVi-13B	41.8	47.1	44.5	55.6	59.1	57.4	42.0	46.7	44.3	48.8	52.9	50.9	-	-	-
HyperSeg	Mipha-3B	-	-	-	56.0	60.9	58.5	50.2	55.8	53.0	53.1	58.4	55.7	-	-	-
InstructSeg	Mipha-3B	-	-	-	54.8	59.2	57.0	49.2	54.7	51.9	52.0	56.9	54.5	-	-	-
GLUS	LISA-7B	48.5	54.2	51.3	56.0	60.7	58.3	48.8	53.9	51.4	52.4	57.3	54.9	47.5	52.4	49.9
ViLLa	InternVideo2-6B	46.5	52.3	49.4	-	-	-	-	-	-	54.9	59.1	57.0	-	-	-
Sa2VA	InternVL2-8B	-	-	46.9	-	-	-	-	-	-	-	-	57.6	-	-	-
Sa2VA	InternVL3-14B	-	-	-	-	-	-	-	-	-	-	-	60.7	-	-	-
RGA3	Qwen2.5-VL-7B	47.4	52.8	50.1	58.7	62.3	60.5	53.1	57.7	55.4	55.9	60.0	58.0	51.3	56.0	53.6
VRS-HQ	Chat-UniVi-13B	48.0	53.7	50.9	**61.1**	**65.5**	**63.3**	54.1	59.4	56.8	57.6	62.5	60.0	-	-	-
VideoSeg-R1	Qwen2.5-VL-7B	52.7	57.8	55.3	-	-	-	-	-	-	58.2	**64.0**	61.1	-	-	-
Training-free Methods																
AL-Ref-SAM2	GPT-4	39.5	46.2	42.8	-	-	-	-	-	-	-	-	-	-	-	-
CoT-RVS	Gemma3-12B	40.3	48.1	44.2	-	-	-	-	-	-	43.4	50.9	47.1	47.5	54.0	50.7
CoT-RVS†	Qwen3-VL-8B-T	37.7	43.9	40.8	56.1	61.5	58.8	46.6	53.0	49.8	51.4	57.3	54.3	52.5	58.9	55.7
CoT-RVS	GPT-4o	48.7	55.7	52.2	-	-	-	-	-	-	52.8	59.0	55.9	62.4	68.7	65.5
AgentRVOS	Qwen3-VL-8B-T	59.2	64.5	61.9	58.7	62.5	60.6	56.8	61.3	59.0	57.7	61.9	59.8	65.5	71.8	68.6
CER-RVS(Ours)	Ovis2.5-9B	61.7	67.3	64.5	60.1	64.6	62.4	**58.3**	**62.7**	**60.5**	59.2	63.7	61.5	**71.4**	**77.1**	**74.3**
CER-RVS(Ours)	Qwen3-VL-8B-T	**62.4**	**68.4**	**65.4**	60.8	65.1	63.0	58.1	62.4	60.3	**59.5**	63.8	**61.7**	70.8	76.4	73.6

This makes the reflection process *targeted rather than self-repetitive*: each new round focuses on unresolved conflicts identified by the Arbitrator instead of simply asking the same model to reconsider its previous conclusion. The process terminates when the Arbitrator returns a resolved `Accept` or `Reject` decision, or when a predefined maximum number of rounds is reached. Candidates marked as `Uncertain` remain revisable rather than being irreversibly pruned. Consequently, CER reduces error accumulation from premature decisions and mitigates category-induced reasoning hallucination while remaining entirely training-free.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate our method on three major benchmarks for language-guided video object segmentation: MeViS (Ding et al., 2023), ReVOS (Yan et al., 2024), and ReasonVOS (Bai et al., 2024). These datasets pose distinct challenges. MeViS features complex scenes involving multiple visually similar objects and demands strong motion understanding. ReVOS and ReasonVOS, on the other hand, emphasize reasoning-centric scenarios that require deeper semantic reasoning and world knowledge. Following prior works (Bai et al., 2024; Kao et al., 2026), we report region similarity $\mathcal{J}$ (average IoU), contour accuracy $\mathcal{F}$ (mean boundary similarity), and their average $\mathcal{J}\&\mathcal{F}$.

Comparison Methods. We compare our method with state-of-the-art approaches in referring and reasoning video object segmentation. For training-based methods, we include VideoLISA (Bai et al., 2024), VISA (Yan et al., 2024), HyperSeg (Wei et al., 2025a), InstructSeg (Wei et al., 2025b), GLUS (Lin et al., 2025), ViLLa (Zheng et al., 2025), Sa2VA (Yuan et al., 2026), RGA3 (Wang et al., 2025), VRS-HQ (Gong et al., 2025), and VideoSeg-R1 (Xu et al., 2026). For training-free methods, we compare against AL-Ref-SAM2 (Huang et al., 2025), CoT-RVS (Kao et al., 2026), and AgentRVOS (Jin et al., 2026).

Implementation Details. We adopt various models as our baseline Multimodal Large Language Models (MLLMs), including the open-source models Qwen3-VL-8B-Thinking (Bai et al., 2025) and Ovis2.5-9B (Lu et al., 2025). We choose SAM3 (Carion et al., 2026) as the segmentation model to produce candidate mask tracks. For MLLM-based reasoning, 16 frames are uniformly sampled from each video by default, while segmentation and track refinement operate on the full video. The maximum number of iterations is set to 3 for Contrastive Evidence Reflection. For AGTR,

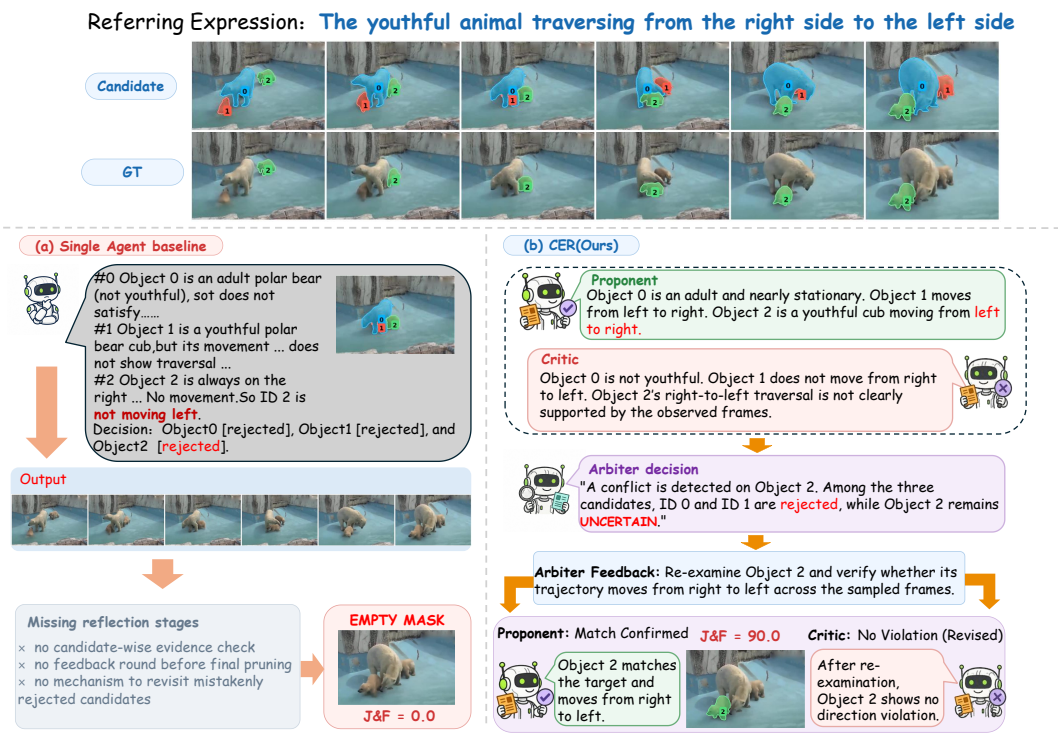


Figure 3: **Qualitative analysis of CER.** The single-agent baseline prematurely rejects the target due to an incorrect motion judgment. CER exposes the conflicting evidence, retains the candidate as *Uncertain*, and triggers targeted re-examination of its cross-frame motion, ultimately recovering the correct target.

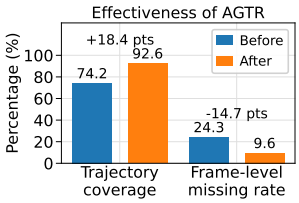
the reliability pre-filter size is set to $N = 5$, and the reliability–scale trade-off parameter is set to $\alpha = 0.6$. All experiments are conducted on 4 NVIDIA RTX 4090 GPUs. Additional implementation details, including the prompts used in each module, are provided in the appendix.

4.2 Main Results

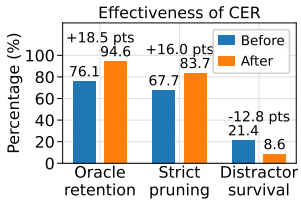
Quantitative Results. As shown in Table 1, CER-RVS establishes a new state of the art among training-free methods across all three benchmarks. With Qwen3-VL-8B-T, our method achieves 65.4, 61.7, and 73.6 $\mathcal{J}\&\mathcal{F}$ on MeViS, ReVOS, and ReasonVOS, respectively. Compared with AgentRVOS under the same MLLM backbone, CER-RVS improves $\mathcal{J}\&\mathcal{F}$ by 3.5 points on MeViS, 1.9 points on ReVOS overall, and 5.0 points on ReasonVOS, demonstrating consistent gains beyond the choice of backbone. Notably, our training-free method also surpasses the best reported training-based results on MeViS, ReVOS overall, and ReasonVOS, while remaining competitive on the ReVOS referring split. Moreover, CER-RVS with Ovis2.5-9B consistently delivers strong performance across all benchmarks, suggesting that the effectiveness of our framework generalizes across different MLLM backbones.

Qualitative Analysis. Figure 3 illustrates how CER corrects a premature rejection caused by an erroneous motion judgment. For the query describing the youthful animal traversing from right to left, the single-agent baseline misinterprets Object 2's trajectory and consequently rejects all candidates, producing an empty mask. In contrast, CER explicitly contrasts supporting and counter-evidence: conflicting assessments of Object 2 cause the Arbitrator to retain it as *Uncertain* rather than pruning it, while providing targeted feedback to re-examine its cross-frame motion. Guided by this reflection, the agents revise the motion judgment, identify Object 2 as the referred target, and improve $\mathcal{J}\&\mathcal{F}$ from 0.0 to 90.0.

Table 2: Component analysis on ReasonVOS: **(a)** AGTR track completeness; **(b)** CER candidate discrimination; **(c)** component ablation.



(a) AGTR



(b) CER

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Baseline	64.7	70.1	67.4
+ CER	68.1	74.7	71.4
+ AGTR	67.8	73.4	70.6
+ AGTR + CER	**71.4**	**77.1**	**74.3**

(c) Ablation

4.3 Analysis

Ablation Study. Table 2(c) evaluates the components on ReasonVOS against our reproduction of AgentRVOS (Jin et al., 2026) using the same Ovis2.5-9B backbone and segmentation backend. CER alone yields $\mathcal{J}/\mathcal{F}/\mathcal{J}\&\mathcal{F}$ of 68.1/74.7/71.4, a 4.0-point $\mathcal{J}\&\mathcal{F}$ gain over the baseline. AGTR alone yields 67.8/73.4/70.6, compared with the baseline's 64.7/70.1/67.4. Combining both reaches 71.4/77.1/74.3, gaining 3.7 points over AGTR alone and 6.9 over the baseline. Notably, the combined improvement is close to the sum of the gains from the two individual modules, suggesting that they address largely distinct rather than redundant failure modes. AGTR strengthens the visual evidence available for reasoning, while CER improves how this evidence is evaluated and resolved during candidate selection. Further CER, anchor-selection, and inference-time analyses appear in Appendix A (Tables A1, A2, and A3).

Effectiveness of AGTR. We evaluate the effectiveness of AGTR in improving object-level visual evidence using GT-aligned trajectory coverage and frame-level missing rate, as shown in Table 2(a). AGTR increases trajectory coverage from 74.2% to 92.6% (+18.4 points) while reducing the frame-level missing rate from 24.3% to 9.6% (-14.7 points). These metrics capture complementary aspects of track quality, measuring trajectory-level completeness and frame-level target availability, respectively. Their simultaneous improvement shows that anchor-guided bidirectional re-tracking effectively repairs fragmented trajectories and restores missing temporal evidence, which is crucial for reasoning about motion, temporal order, and object interactions. Thus, AGTR provides a more complete and reliable visual basis for subsequent track-level reasoning.

Effectiveness of CER. We further isolate the effect of CER on reasoning reliability through a candidate-controlled evaluation, where the compared methods operate on identical candidate pools, as shown in Table 2(b). CER improves oracle-target retention from 76.1% to 94.6% and strict pruning correctness from 67.7% to 83.7%, while reducing the survival rate of low-overlap distractors from 21.4% to 8.6%. Importantly, the improvement in target retention is accompanied by substantially stronger distractor suppression. This rules out the possibility that CER gains performance merely by adopting a more conservative pruning policy that retains more candidates indiscriminately. Instead, CER is better able to distinguish genuinely contradictory evidence from ambiguous or insufficient observations: plausible targets are less likely to be prematurely discarded, while clearly inconsistent distractors are removed more reliably. These results support the role of contrastive evidence collection and arbitrator-guided reflection in improving candidate discrimination and making the final decision process more robust to uncertain temporal and relational cues.

5 Conclusion

We presented CER-RVS, a training-free framework for referring video object segmentation that couples evidence construction with evidence verification. By refining candidate trajectories and contrasting supporting with counter-evidence, CER-RVS addresses incomplete visual observations and unreliable intermediate decisions. Across MeViS, ReVOS, and ReasonVOS, it consistently outperforms existing training-free approaches and achieves competitive or superior performance compared with representative training-based methods. These results show that stronger object-level evidence and verifiable reasoning can improve agentic RVOS without task-specific training.

AI USE STATEMENT

The authors used generative AI tools to assist with language polishing and grammar checking. All AI-assisted content was reviewed and revised by the authors. The authors take full responsibility for the final content of the paper, including all text, claims, experiments, and code.

ETHICS STATEMENT

We evaluate CER-RVS on the publicly available MeViS, ReVOS, and ReasonVOS benchmarks; no new videos or human-subject data were collected. Because the method can identify and track a language-specified target across video frames, including a person, use on footage without appropriate authorization could raise privacy and surveillance concerns. Its pretrained MLLM and segmentation components can also produce incorrect target judgments or incomplete masks when visual evidence is ambiguous or occluded. The outputs should therefore be treated as research predictions rather than evidence for consequential decisions, and any deployment should respect privacy and dataset-use requirements.

REPRODUCIBILITY STATEMENT

CER-RVS requires no task-specific training. We specify the MeViS, ReVOS, and ReasonVOS benchmarks and evaluation metrics, the Qwen3-VL-8B-Thinking and Ovis2.5-9B MLLM backbones, and the SAM3 segmentation model. The main paper reports the 16-frame sampling policy, the maximum of three CER rounds, and the AGTR settings $N = 5$ and $\alpha = 0.6$. The appendix provides the inference algorithm, agent prompts, and evaluation protocols used in the ablations. Code and configuration files will be released upon publication.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. In *Advances in Neural Information Processing Systems*, volume 37, pp. 6833–6859, 2024.

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4985–4995, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229, 2020. doi: 10.1007/978-3-030-58452-8_13.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2694–2703. IEEE, 2023.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 11733–11763. PMLR, 2024.

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5958–5966. IEEE, 2018.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29183–29192, 2025. doi: 10.1109/CVPR52734.2025.02717.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024.

Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, and Yu Qiao. HTML: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13414–13423, 2023.

Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. Unleashing the temporal-spatial reasoning capacity of GPT for training-free audio and language referenced video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3715–3723, 2025. doi: 10.1609/aaai.v39i4.32387.

Haichao Jiang, Tianming Liang, Wei-Shi Zheng, and Jian-Fang Hu. Refer-agent: A collaborative multi-agent system with reasoning and reflection for referring video object segmentation. *arXiv preprint arXiv:2602.03595*, 2026.

Woojeong Jin, Jaeho Lee, Heeseong Shin, Seungho Jang, Junhwan Heo, and Seungryong Kim. Agentrvos: Reasoning over object tracks for zero-shot referring video object segmentation. *arXiv preprint arXiv:2603.23489*, 2026.

Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Cot-rvs: Zero-shot chain-of-thought reasoning segmentation for videos. In *International Conference on Learning Representations*, volume 2026, pp. 58807–58830, 2026.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with referring expressions. In *European Conference on Computer Vision*, pp. 7–12. Springer, 2018.

Jungyoon Lee, Gyuil Lim, Doeon Kim, and Seong-heum Kim. Multi-agent target-existence verification and learned mask geometry refinement: Winning report of the MeViS-Text track at the 8th LSVOS challenge 2026. *arXiv preprint arXiv:2608.11458*, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Yifan Li, Yingda Yin, Lingting Zhu, Weikai Chen, Shengju Qian, Xin Wang, and Yanwei Fu. ReVSeg: Incentivizing the reasoning chain for video segmentation with reinforcement learning. *arXiv preprint arXiv:2512.02835*, 2025.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904. Association for Computational Linguistics, 2024.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. Referdino: Referring video object segmentation with visual grounding foundations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20009–20019. IEEE, 2025.

Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. GLUS: Global-local reasoning unified into a single large language model for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8658–8667, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. SOC: Semantic-assisted object cluster for referring video object segmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26425–26437, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 920–930, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision*, pp. 208–223. Springer, 2020.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22274–22284, 2025.

Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Jie Hu, Dengjie Li, Zheng Zhao, and Yujiu Yang. HyperSeg: Hybrid segmentation assistant with fine-grained visual perceiver. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8931–8941, 2025a. doi: 10.1109/CVPR52734.2025.00835.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. InstructSeg: Unifying instructed visual segmentation with multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20193–20203, 2025b.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4974–4984, 2022.

Zishan Xu, Yifu Guo, Yuquan Lu, Fengyu Yang, Junxin Li, and Lihua Cai. VideoSeg-R1: Reasoning video object segmentation via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11496–11504, 2026. doi: 10.1609/aaai.v40i14.38132.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pp. 98–115, 2024.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with LLaVA for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying sam2 with mllm for dense grounded understanding of images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3720379.

Linfeng Yuan, Miaojing Shi, Zijie Yue, and Qijun Chen. LoSh: Long-short text joint prediction network for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14001–14010, 2024.

Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, and Hengshuang Zhao. ViLLa: Video reasoning segmentation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23667–23677, 2025.

A Additional Results and Analyses

Analysis of CER Components. Table A1 compares CER reasoning strategies with the same Ovis2.5-9B MLLM and number of MLLM calls. Relative to single-agent selection, Proponent–Critic evidence raises $\mathcal{J}\&\mathcal{F}$ from 70.6 to 71.4; independent arbitration raises it to 72.3, and targeted feedback further to 74.3 ($+2.0$). Providing video frames to the Arbitrator instead reduces the full CER score to 72.1. Evidence-only arbitration lets it focus on conflicts and unresolved query conditions while the Proponent and Critic handle visual observations. Overall, full CER improves $\mathcal{J}\&\mathcal{F}$ by 3.7 points over single-agent selection.

Table A1: Ablation of different reasoning strategies in CER. All variants use the same MLLM (Ovis2.5-9B) and the same number of MLLM calls. *Arb. Video* indicates whether the sampled video frames are additionally provided to the Arbitrator; the Proponent and Critic receive visual inputs in all CER variants.

Reasoning Strategy	**Contrastive**	**Arbitrator**	**Feedback**	**Arb. Video**	$\mathcal{J}\&\mathcal{F}$
Single-agent selection	×	×	×	–	70.6
Proponent + Critic (simple aggregation)	✓	×	×	–	71.4
Proponent + Critic + Arbitrator (no feedback)	✓	✓	×	×	72.3
Full CER + visual Arbitrator	✓	✓	✓	✓	72.1
Full CER	✓	✓	✓	×	**74.3**

Ablation of Anchor Selection. Table A2 isolates AGTR anchor selection on ReasonVOS with CER fixed. Reliability-only anchors followed by bidirectional re-tracking improve $\mathcal{J}\&\mathcal{F}$ from 71.4 to 73.2; adding relative mask scale raises it to 74.3. The latter improves $\mathcal{J}$ and $\mathcal{F}$ by 0.4 and 1.7 points over reliability-only selection, respectively, for a 1.1-point gain in $\mathcal{J}\&\mathcal{F}$.

Table A2: **Ablation of anchor selection in AGTR on ReasonVOS.** CER is retained in all variants. The latter two variants use the same bidirectional re-tracking procedure and differ only in the anchor-selection rule. Best results are highlighted in bold.

Variant	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
w/o AGTR (with CER)	68.1	74.7	71.4
Mask reliability only	71.0	75.4	73.2
Mask reliability + relative mask scale	**71.4**	**77.1**	**74.3**

Inference Time Comparison. We measure the average inference time per ReasonVOS sample for the component configurations in Table 2(c). The baseline uses neither AGTR nor CER. All timing runs use Ovis2.5-9B on four customized NVIDIA RTX 5090 GPUs with 50 GB of memory each; this timing setup is separate from the hardware used for the main experiments.

Table A3: **Average inference time on ReasonVOS.** The added time is measured relative to the preceding row.

Variant	Time (s/sample)	Added time (s)
Baseline (w/o AGTR and CER)	106	–
+ AGTR	124	+18
+ AGTR + CER	174	+50

As shown in Table A3, AGTR adds 18 seconds per sample over the baseline. Adding CER on top of AGTR adds a further 50 seconds, for a total increase of 68 seconds per sample.

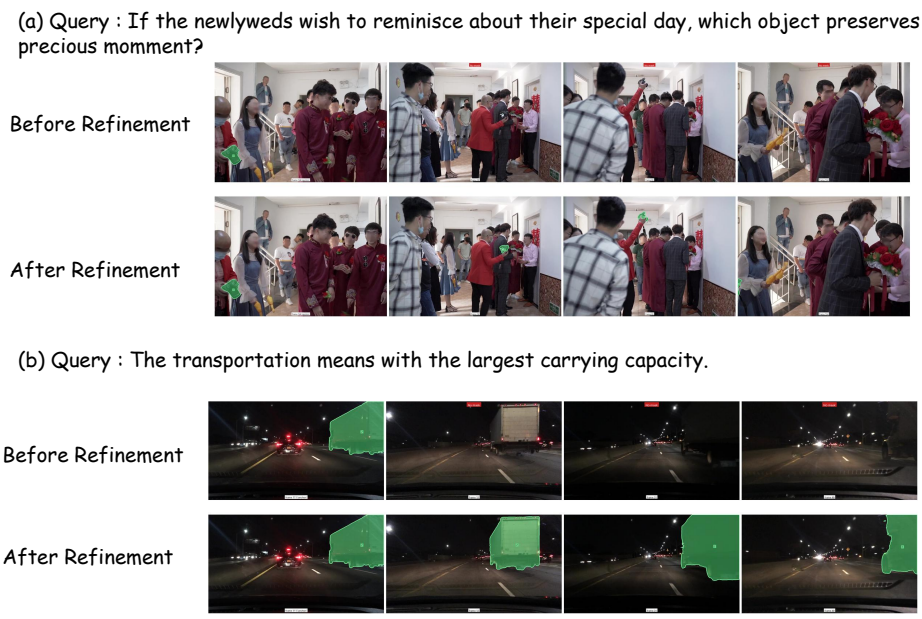


Figure A1: **Qualitative analysis of AGTR.** Green masks show the candidate track before and after refinement in **(a)** a crowded wedding scene and **(b)** a night-driving scene.

Qualitative Analysis of AGTR. Figure A1 visualizes how AGTR refines fragmented coarse candidate tracks. In Fig. A1(a), the queried recording device is small and intermittently visible in a crowded scene, so the coarse track contains only sparse observations. After refinement, AGTR recovers the same candidate in additional frames despite changes in its position and visibility, yielding a more temporally continuous trajectory.

In Fig. A1(b), the target vehicle undergoes substantial spatial and scale variation. The coarse track captures it only intermittently, leaving several frames without valid observations. Starting from a reliable anchor, bidirectional instance-conditioned re-tracking reconstructs the missing portions of the trajectory and maintains the candidate across the sampled frames. These examples show that AGTR improves temporal evidence completeness, providing a denser and more reliable visual basis for subsequent query-conditioned reasoning.

B Complete Inference Procedure

Algorithm 1 Complete CER-RVS inference procedure

Require: Video $V = \{I_t\}_{t=1}^{T}$, query Q, anchor shortlist size N, maximum concept trials L_{con}, CER rounds R, and JSON attempts J
Ensure: Predicted mask sequence $M = \{M_t\}_{t=1}^{T}$

$\widetilde{\mathcal{T}} \leftarrow \varnothing, \mathcal{F} \leftarrow \varnothing$ ▷ failed concepts
for $\ell = 1, \dots, L_{\text{con}}$ **do**
  $\mathcal{C} \leftarrow \textsc{ExtractConcepts}(V, Q, \mathcal{F})$
  for all $c \in \mathcal{C}$ **do**
    $\widetilde{\mathcal{T}}_c \leftarrow \textsc{UniqueTracks}(G_{\text{text}}(V, c))$
  end for
  $c^\star \leftarrow \arg\max_{c \in \mathcal{C}} |\widetilde{\mathcal{T}}_c|; \widetilde{\mathcal{T}} \leftarrow \widetilde{\mathcal{T}}_{c^\star}$
  if $\widetilde{\mathcal{T}} \neq \varnothing$ **then**
    break
  end if
  $\mathcal{F} \leftarrow \mathcal{F} \cup \mathcal{C}$
end for
if $\widetilde{\mathcal{T}} = \varnothing$ **then**
  return $\{\mathbf{0}\}_{t=1}^{T}$
end if
for $k = 1, \dots, K$ **do** ▷ AGTR
  $\mathcal{C}_k \leftarrow \text{TopN}_{t \in \mathcal{I}_k}(E_{k,t})$
  $t_k^\star \leftarrow \arg\max_{t \in \mathcal{C}_k} A_{k,t}$
  $M_k^{\text{anc}} \leftarrow \widetilde{M}_{k,t_k^\star}$
  $\widehat{\mathbf{M}}_k \leftarrow G_{\text{PVS}}(V, t_k^\star, M_k^{\text{anc}})$
end for
$\mathcal{K}_{\text{acc}} \leftarrow \varnothing, \mathcal{K}_{\text{act}} \leftarrow \{1, \dots, K\}$, and $\mathcal{S}_k^{(0)} \leftarrow \varnothing$ for all k ▷ accepted and active IDs
for $r = 1, \dots, R$ **do** ▷ CER
  if $\mathcal{K}_{\text{act}} = \varnothing$ **then**
    break
  end if
  if $r = 1$ **then** ▷ visual evidence collection, Eq. (7)
    $X^{(1)} \leftarrow \textsc{VisualizeAndSample}(V, \{\widehat{\mathbf{M}}_k : k \in \mathcal{K}_{\text{act}}\})$
    $\mathcal{U}^{\pm,(1)} \leftarrow (Q, X^{(1)}; \{\mathcal{S}_k^{(0)}\}_{k \in \mathcal{K}_{\text{act}}})$
  else ▷ text-only reflection, Eq. (9)
    $\mathcal{U}^{\pm,(r)} \leftarrow (Q, \{\mathcal{E}_k^{\pm,(r-1)}\}_{k \in \mathcal{K}_{\text{act}}}, \{\mathcal{S}_k^{(r-1)}\}_{k \in \mathcal{K}_{\text{act}}})$
  end if
  $(\mathcal{E}^{+,(r)}, b^+) \leftarrow \textsc{ValidatedBatchCall}(\mathcal{A}^+, \mathcal{U}^{+,(r)}, \mathcal{K}_{\text{act}}, J)$
  $(\mathcal{E}^{-,(r)}, b^-) \leftarrow \textsc{ValidatedBatchCall}(\mathcal{A}^-, \mathcal{U}^{-,(r)}, \mathcal{K}_{\text{act}}, J)$
  if $\neg(b^+ \wedge b^-)$ **then**
    break
  end if
  $(\mathcal{D}^{(r)}, \mathcal{S}^{(r)}, b^a) \leftarrow \textsc{ValidatedBatchCall}(\mathcal{A}^a, (Q, \mathcal{E}^{+,(r)}, \mathcal{E}^{-,(r)}), \mathcal{K}_{\text{act}}, J)$
  if $\neg b^a$ **then**
    break
  end if
  $\mathcal{K}_{\text{acc}} \leftarrow \mathcal{K}_{\text{acc}} \cup \{k \in \mathcal{K}_{\text{act}} : \mathcal{D}_k^{(r)} = \texttt{Accept}\}$
  $\mathcal{K}_{\text{act}} \leftarrow \{k \in \mathcal{K}_{\text{act}} : \mathcal{D}_k^{(r)} = \texttt{Uncertain}\}$ ▷ Reject IDs are discarded
end for
for $t = 1, \dots, T$ **do**
  $M_t \leftarrow \bigvee_{k \in \mathcal{K}_{\text{acc}}} \widehat{M}_t^{(k)}$ ▷ $\bigvee \varnothing = \mathbf{0}$
end for
return M

Following AgentRVOS (Jin et al., 2026), concept extraction is retried when no candidate track is found. Core and broader concepts are grounded separately, and only the concept with the most unique instance tracks is retained. This avoids pooling tracks from semantically overlapping prompts and counts each tracker instance ID at most once.

Mask Reliability Scores. For each non-empty coarse mask $\widetilde{M}_{k,t}$, the segmentation model provides an object-presence confidence $s^{\mathrm{obj}}_{k,t}$ and a predicted mask-IoU score $s^{\mathrm{iou}}_{k,t}$. The latter estimates mask quality at inference time; it is not computed against a ground-truth mask. Their product $E_{k,t}$ ranks the non-empty observations before the relative-area cue is applied to the shortlist.

The two visual agents are each called once per round on the complete active candidate set $\mathcal{K}_{\mathrm{act}}$, rather than once per candidate, and may be executed in parallel. The Arbitrator is likewise called once, but receives only Q and the two evidence sets; it has no direct access to video frames. A VALIDATEDBATCHCALL succeeds only if the response can be parsed as the required JSON structure, contains every active ID exactly once, contains no extra or duplicate ID, and uses only the allowed decisions. A malformed JSON response, missing ID, or repeated ID triggers a format-correction prompt with the same inputs, for at most J total attempts. If validation still fails, the loop terminates and only candidates accepted in completed rounds are returned.

The output policy contains no ranking or recovery heuristic. Every candidate marked `Accept` in any completed round is retained, and their masks are combined by a frame-wise pixel union. Candidates still `Uncertain` after round R are excluded. Therefore, if no candidate is accepted—including the case in which every candidate is rejected—the method returns an all-zero mask sequence and performs no fallback.

C CER PROMPT AND DECISION DETAILS

Prompts of the CER agents. Figures A2 to A4 show the prompts used by the three CER agents. The *Proponent* judges whether each candidate track matches the referring expression, while the *Critic* judges the inverted question of whether the candidate violates it, so the two agents argue from opposite sides and a `true` judgment from the Critic counts as evidence *against* the corresponding candidate. The *Arbitrator* compares the two sets of judgments, resolves a conflict only when one side provides clearly stronger evidence (otherwise the candidate stays *unknown*), and returns a suggestion of which evidence the next round should re-examine. The Proponent and Critic take as input the video frames with masked object IDs, the referring expression, and the candidate IDs. In contrast, the Arbitrator receives only the referring expression and the supporting and counter-evidence produced by the two agents; it does not access the video frames directly. All three agents respond with JSON according to their respective output specifications. The braced placeholders (e.g., {`expression_string`}) are filled at inference time, and the previous round's suggestion $\mathcal{S}_k^{(r-1)}$ is supplied to the Proponent and Critic as additional context in Eq. (7). The `true`/`false`/`unknown` decision that the prompts return for each candidate corresponds to the candidate-level `Accept`/`Reject`/`Uncertain` decision of Eq. (8).

D EVALUATION PROTOCOL

D.1 EVALUATION PROTOCOL FOR AGTR

We evaluate AGTR from two complementary perspectives: **trajectory-level coverage** and **frame-level target availability**. For the i-th video-expression pair, let $\mathcal{V}_i$ denote the set of GT-valid frames and $\{P_{i,k}\}_{k=1}^{K_i}$ the candidate mask trajectories, before or after refinement. The per-frame quality of candidate k with respect to the ground-truth mask $G_{i,t}$ is

$$q_{i,k,t} = \frac{\mathcal{J}(G_{i,t}, P_{i,k,t}) + \mathcal{F}(G_{i,t}, P_{i,k,t})}{2}, \tag{10}$$

where $\mathcal{J}$ and $\mathcal{F}$ denote region similarity and contour accuracy, and $q_{i,k,t} = 0$ if candidate k has no valid mask at frame t.

Trajectory Coverage. Averaging $q_{i,k,t}$ over GT-valid frames yields the trajectory-level score $Q_{i,k} = |\mathcal{V}_i|^{-1} \sum_{t \in \mathcal{V}_i} q_{i,k,t}$, and a sample is *covered* once its best candidate reaches 0.8:

$$\text{Trajectory Coverage} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\max_k Q_{i,k} \geq 0.8 \right], \tag{11}$$

where N is the number of video-expression samples. Since missing masks contribute zero to $Q_{i,k}$, this metric jointly reflects trajectory-level accuracy and temporal completeness: it asks a *single* candidate trajectory to stay reliable over the whole video.

Frame-Level Missing Rate. Per frame we take the best candidate score $S_{i,t} = \max_k q_{i,k,t}$ and count a frame as *missing* when $S_{i,t} < 0.75$, i.e., no candidate provides a sufficiently accurate target mask:

$$\text{Frame-Level Missing Rate} = \frac{\sum_{i=1}^{N} \sum_{t \in \mathcal{V}_i} \mathbb{I}\left[\max_k q_{i,k,t} < 0.75\right]}{\sum_{i=1}^{N} |\mathcal{V}_i|}. \tag{12}$$

Unlike coverage, the best candidate may differ across frames, so this metric only requires that *some* candidate supplies a reliable mask at each GT-valid frame.

Higher coverage and lower missing rate indicate more accurate and temporally complete candidate evidence. For the before/after comparison, the same protocol is applied independently to the coarse and refined candidate sets.

D.2 Evaluation Protocol for CER

We conduct a candidate-controlled evaluation to isolate the effect of Contrastive Evidence Reflection (CER) on reasoning reliability, where all compared methods operate on identical candidate pools. We reuse the notation of Appendix D.1: for the i-th case, $\mathcal{V}_i$ is the set of GT-valid frames, $P_{i,k}$ the k-th candidate mask trajectory with mask $P_{i,k,t}$ at frame t, $G_{i,t}$ the ground-truth mask, and $Q_{i,k}$ the trajectory-level mean $\mathcal{J}\&\mathcal{F}$.

Evaluation subset. For each case i, we define the *oracle target* $k_i^\star$ as the candidate track with the highest overlap with the ground-truth target, i.e., the one with the largest trajectory-level region similarity

$$J_{i,k} = \frac{1}{|\mathcal{V}_i|} \sum_{t \in \mathcal{V}_i} \mathcal{J}\left(G_{i,t}, P_{i,k,t}\right), \qquad k_i^\star = \arg\max_k J_{i,k}. \tag{13}$$

We consider cases satisfying

$$Q_{i,k_i^\star} \geq 0.75, \tag{14}$$

so that a sufficiently accurate target candidate is available for evaluating the reasoning stage. Among 458 evaluation cases, 406 satisfy this criterion (88.6%); we denote this subset by $\mathcal{E}$.

For each $i \in \mathcal{E}$, the set of low-overlap distractors is defined as

$$\mathcal{D}_i = \{k \neq k_i^\star \mid J_{i,k} \leq 0.1\}. \tag{15}$$

Evaluation metrics. Let $R_{i,k} \in \{0, 1\}$ indicate whether candidate k remains after pruning. We evaluate the following three metrics.

Oracle-target retention is the fraction of eligible cases in which the oracle target is retained:

$$\text{Retention} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \mathbb{I}\left[R_{i,k_i^\star} = 1\right]. \tag{16}$$

Strict pruning correctness is the fraction of eligible cases in which the oracle target is retained and all low-overlap distractors are removed:

$$\text{StrictPruning} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \mathbb{I}\left[R_{i,k_i^\star} = 1 \ \wedge \ \forall k \in \mathcal{D}_i, \ R_{i,k} = 0\right]. \tag{17}$$

Low-overlap distractor survival rate is computed at the candidate level as

$$\text{DistractorSurvival} = \frac{\sum_{i \in \mathcal{E}} \sum_{k \in \mathcal{D}_i} \mathbb{I}\left[R_{i,k} = 1\right]}{\sum_{i \in \mathcal{E}} |\mathcal{D}_i|}. \tag{18}$$

E FULL CER AGENT PROMPTS

For completeness, we provide the full prompts of the Proponent, Critic, and Arbitrator agents below.

```
SELECT_PROMPT = r"""
# Task:
Analyze uniformly sampled video frames and determine whether each prompted object ID matches the referring
expression.

# Input:
- Video frames: visually prompted with object masks and object IDs.
- Referring expression: {expression_string}
- Object IDs: {unknown_obj_ids}
- Candidate Object Category: All object IDs belong to {selected_concept}.

# Step-by-step processing:
For each step, keep only the key conclusions needed for the decision. Do not repeat conclusions from previous
steps.
## Step 1: Global temporal analysis
Examine ALL frames to understand the full context and motion patterns. Identify and track each candidate object
across the sequence, noting evidence relevant to the referring expression.
## Step 2: Constraint matching
For each object ID, compare its visual and temporal evidence against the constraints in the referring expression.
Determine which constraints are satisfied, contradicted, or cannot be reliably verified.
## Step 3: Final classification
Based on Steps 1 and 2, assign each object ID exactly one decision and provide concise reasoning evidence
supporting that decision.

Decision criteria:
- true: The object definitely DOES match the referring expression with high confidence.
- false: The object definitely DOES NOT match the referring expression with high confidence.
- unknown: It cannot be clearly determined whether the object matches the referring expression.

IMPORTANT:
Your response MUST include ALL object IDs from: {unknown_obj_ids}
Each object ID must appear exactly ONCE. Missing or duplicated IDs are errors.
The reasoning_logic should state only the key visual, temporal, or constraint-matching evidence for the decision.

# Output:
Return ONLY the following JSON format enclosed within <answer> tags:

<answer>
[
  {{
    "obj_id": object ID,
    "decision": "true" | "false" | "unknown",
    "reasoning_logic": "concise evidence supporting the decision"
  }}
]
</answer>

""".strip()
```

Figure A2: **Prompt of the Proponent agent in CER.** The agent collects supporting evidence for each candidate track and returns a per-object judgment.

```
SELECT_PROMPT_NEGATIVE = r"""
# Task:
Analyze uniformly sampled video frames and determine whether each prompted object ID violates or does not satisfy
the referring expression*.*

# Input:
- Video frames: visually prompted with object masks and object IDs*.*
- Referring expression: {expression_string}
- Object IDs: {unknown_obj_ids}
- Candidate Object Category: All object IDs belong to {selected_concept}*.*

# Step-by-step processing:
For each step, keep only the key conclusions needed for the decision*.* Do not repeat conclusions from previous
steps*.*

## Step 1: Global temporal analysis
Examine ALL frames to understand the full context and motion patterns*.* Identify and track each candidate object
across the sequence, noting evidence relevant to the referring expression*.*

## Step 2: Constraint matching
For each object ID, compare its visual and temporal evidence against the constraints in the referring expression*.*
Determine which constraints are violated, satisfied, or cannot be reliably verified*.*

## Step 3: Final classification
Based on Steps 1 and 2, assign each object ID exactly one decision and provide concise reasoning evidence supporting
that decision*.*

Decision criteria:
- true: The object definitely violates or DOES NOT satisfy the referring expression with high confidence*.*
- false: The object DOES satisfy the referring expression and does not violate its constraints with high confidence*.*
- unknown: It cannot be clearly determined whether the object violates the referring expression*.*

IMPORTANT:
Your response MUST include ALL object IDs from: {unknown_obj_ids}
Each object ID must appear exactly ONCE*.* Missing or duplicated IDs are errors*.*
The reasoning_logic should state only the key visual, temporal, or constraint-matching evidence for the decision*.*

# Output:
Return ONLY the following JSON format enclosed within <answer> tags:

<answer>
[
  {{
    "obj_id": object ID,
    "decision": "true" | "false" | "unknown",
    "reasoning_logic": "concise evidence supporting the decision"
  }}
]
</answer>

""".strip()
```

Figure A3: **Prompt of the Critic agent in CER.** The agent answers the inverted question—whether each candidate *violates* the referring expression—so its judgments supply the counter-evidence.

```
ARBITRATE_PROMPT = r"""
# Task:
Arbitrate the positive and negative judgments for each object ID and produce a fair final decision based on their reasoning evidence.

# Input:
- Referring expression: {expression_string}
- Object IDs: {unknown_obj_ids}
- Candidate Object Category: All object IDs belong to {selected_concept}.
- Positive result: {positive_result}
- Negative result: {negative_result}

# Step-by-step processing:
For each step, keep only the key conclusions needed for arbitration. Do not repeat previous conclusions.

## Step 1: Cross-check judgments
For each object ID, compare the positive and negative decisions and their reasoning_logic.
Recall that:
- Positive true means the object matches the referring expression.
- Negative true means the object violates or does NOT satisfy the referring expression.

## Step 2: Evidence arbitration
Evaluate the visual, temporal, and constraint-matching evidence from both results. Resolve conflicts only when one side has clearly stronger and more specific evidence; otherwise keep the object unknown.

## Step 3: Final decision and review suggestion
Assign each object ID one final decision:
- true: The object matches the referring expression with high confidence.
- false: The object does NOT match the referring expression with high confidence.
- unknown: The evidence is conflicting or insufficient.

For each unknown object ID, provide concise suggestions for the next positive and negative review, specifying which visual or temporal evidence in the video frames should be checked to resolve the uncertainty.

# Output:
Return ONLY the following JSON format enclosed within <answer> tags:

<answer>
[
  {{
    "obj_id": object ID,
    "decision": "true" | "false" | "unknown",
    "reasoning_logic": "concise evidence supporting the final arbitration",
    "suggestion": {{
      "positive": "next evidence to check for the positive review, or null",
      "negative": "next evidence to check for the negative review, or null"
    }}
  }}
]
</answer>

""".strip()
```

Figure A4: **Prompt of the Arbitrator agent in CER.** The agent adjudicates the supporting and counter-evidence, resolves conflicts, and returns the final decision together with a suggestion that guides the next reflection round.