MAVER: MULTI-ANCHOR VERIFICATION AND EVIDENCE RECOVERY FOR TRAINING-FREE VIDEO REASONING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning video object segmentation identifies and segments objects from implicit queries about actions, relations, or temporal events. Although training-free methods use multimodal large language models (MLLMs) to interpret such queries, a plausible interpretation can still yield a coherent track of the wrong object. In this paper, we propose **MAVER** (Multi-Anchor Verification and Evidence Recovery), a training-free framework that treats inferred object descriptions as hypotheses and checks candidate regions against the original query before using them as tracking prompts. Multi-anchor verification searches candidate frames under a fixed hypothesis, providing alternative grounding opportunities. Early stopping ends the search at the first confirmed anchor and uses its frame and box as the single initial prompt. When an entire grounding round confirms no anchor, evidence recovery revises the hypothesis and repeats grounding within a preset round limit, keeping the original query unchanged. If verification remains inconclusive, memory-based fallback can supply an explicitly unconfirmed prompt from the retained candidates. MAVER coordinates frozen reasoning, detection, and segmentation models without task-specific fine-tuning. Extensive experiments on four referring and reasoning video segmentation benchmarks demonstrate that MAVER achieves the best overall performance among the compared methods. Moreover, MAVER improves performance while maintaining inference efficiency.

1 INTRODUCTION

Reasoning video object segmentation identifies and segments objects throughout a video from queries involving actions, relations, temporal events, or implicit intent (Yan et al., 2024; Bai et al., 2024). Such queries often specify a target by its role in an event or its interaction with other objects, requiring reasoning beyond appearance matching. For example, *"Which car ahead of me is obstructing my path and forcing me to switch lanes to the right?"* identifies a vehicle by its effect on the observer's motion (Fig. 1). The relevant interaction may be visible only briefly, while the target must be followed despite occlusion and similar-looking distractors. Reliable reasoning segmentation thus requires deciding which visual evidence suffices to identify the target and initialize mask propagation.

Recent approaches employ multimodal large language models (MLLMs) to connect query interpretation with visual grounding. Supervised methods fine-tune MLLMs to emit segmentation tokens that drive a mask decoder (Yan et al., 2024; Bai et al., 2024; Lin et al., 2025; Yuan et al., 2025); they depend on annotated segmentation data and must be retrained when the underlying models are replaced. Training-free alternatives instead coordinate pretrained models at inference time (Huang et al., 2025). CoT-RVS uses zero-shot chain-of-thought reasoning to generate object descriptions and corresponding keyframes (Kao et al., 2026), but its offline pipeline does not explicitly check the grounded regions against the original query before propagation, so a plausible description can initialize a coherent track of the wrong object. Refer-Agent introduces reflection to check target visibility and query consistency, using feedback to refine frame selection and intent analysis (Jiang et al., 2026). Such reflection answers each failed check with renewed reasoning, yet the failure does not reveal its cause: insufficient evidence in the checked frame, an incorrect interpretation, or an

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.

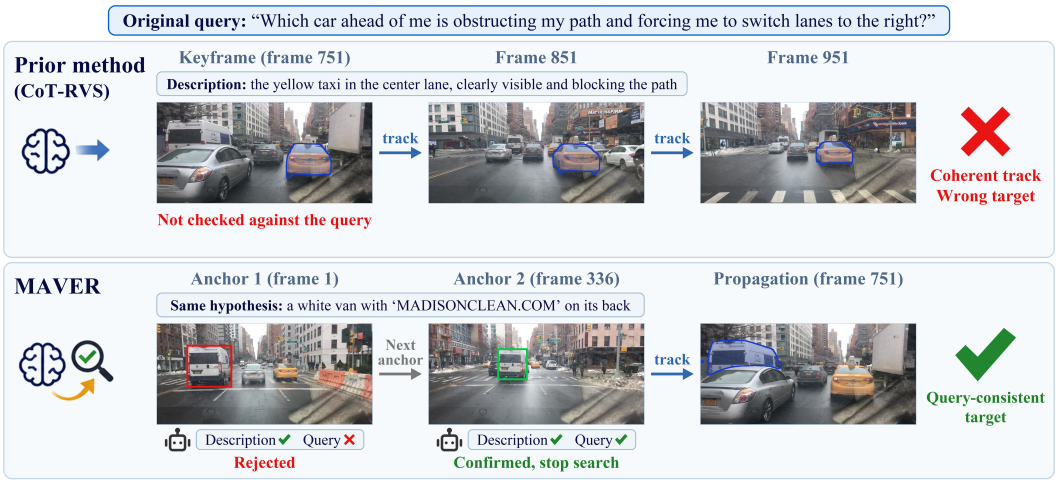

Figure 1: **A coherent track can follow the wrong object.** Top: CoT-RVS tracks a taxi described as blocking the path, unchecked against the query. Bottom: under the same hypothesis, the van fails the query-match check at the first anchor but passes at the second, which alone initializes propagation.

error in verification itself. Handling these cases calls for separate decisions about further search, hypothesis revision, and recovery when no candidate is confirmed.

Two difficulties arise in making these decisions. First, a candidate may match an inferred description without satisfying the original query: in the traffic example, the white van matches its description before it obstructs the observer's path. Checking candidates against the original query and searching alternative frames can help resolve this ambiguity. However, prompting the tracker with every confirmed observation also admits distractors that pass verification by mistake; more search opportunities need not imply more prompts. Second, a valid target may remain unconfirmed after every check, for instance when verification wrongly rejects it or proposals miss it. Discarding all retained candidates in this case can turn uncertainty in verification into a missed segmentation. These difficulties motivate jointly addressing *how to select evidence for propagation and how to recover when verification remains inconclusive*.

To address these challenges, we introduce **MAVER** (Multi-Anchor Verification and Evidence Recovery), a training-free framework for offline reasoning video object segmentation. Its central idea is to treat an inferred object description as a *revisable hypothesis* to be tested against visual evidence. The description guides candidate retrieval, while verification against the *original query* determines whether a candidate is confirmed. This separation keeps verification tied to the requested action, relation, or temporal condition, even when the inferred description is incomplete.

MAVER implements this principle through two complementary mechanisms. *Multi-anchor verification* examines candidate regions across alternative anchor frames under a fixed hypothesis. The search follows detector ranking, advances to another anchor after an unsuccessful check, and stops at the first confirmed anchor, whose frame and box form the single initial prompt (Fig. 2(a)). *Evidence recovery* revises the hypothesis only after an entire grounding round fails, using unsuccessful hypotheses as feedback to restart the search within a preset round limit under the same query and verification rule (Fig. 2(b)). If no round succeeds, *memory-based fallback* supplies an explicitly *unconfirmed* prompt from retained candidates, preserving potentially useful alternatives. Together, these mechanisms distinguish a failed observation from a failed grounding round and provide a recovery path when neither search nor revision yields confirmation. They also keep inference economical, since confirmation ends the search and revision follows only a failed round.

We evaluate MAVER on Ref-YouTube-VOS, MeViS, ReasonVOS, and ReVOS, four challenging benchmarks spanning explicit references, motion descriptions, and implicit reasoning queries. In addition to supervised and training-free baselines, we compare MAVER with Refer-Agent under matched reasoning models, showing that MAVER outperforms it with each tested reasoner. We also conduct in-depth ablation studies to analyze the effectiveness of each component, offering insights into the design of reliable and efficient training-free visual agents.

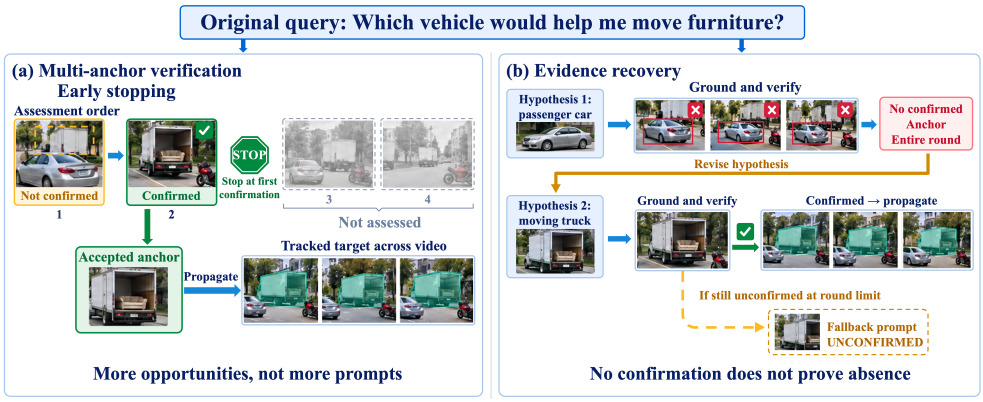

Figure 2: **Multi-anchor verification and evidence recovery.** (a) Anchors are assessed in order until one is confirmed; this anchor alone initializes propagation. (b) A failed round triggers hypothesis revision under the original query; a fallback prompt after the last round remains unconfirmed.

Our contributions are summarized as follows:

- We propose MAVER, a training-free framework whose multi-anchor verification tests a fixed hypothesis on alternative anchors before any reinterpretation. Early stopping selects the first anchor confirmed against the original query as the single initial prompt.
- We introduce evidence recovery through bounded hypothesis revision and memory-based fallback, allowing the framework to revisit unsuccessful grounding without equating a lack of confirmation with target absence.
- Experiments on four benchmarks demonstrate the highest overall performance among the compared methods. Controlled ablations support the roles of verification, anchor search, and evidence recovery.

2 Related Work

Language-guided video segmentation. Referring video object segmentation (RVOS) identifies and segments objects across frames from natural-language descriptions (Seo et al., 2020). Learned approaches develop cross-modal representations (Botach et al., 2022; Wu et al., 2022), temporal association (Wu et al., 2023; Tang et al., 2023), and joint modeling of full and subject-centric expressions (Yuan et al., 2024). MeViS introduces a benchmark centered on motion-based references (Ding et al., 2023), while DsHmp decouples static cues from hierarchical motion perception (He & Ding, 2024). ReferDINO extends visual grounding models to video segmentation (Liang et al., 2025); SAMWISE augments frozen SAM 2 with learned text-conditioning and tracking-correction modules (Cuttano et al., 2025). Reasoning video segmentation emphasizes implicit queries requiring contextual inference and world knowledge. Such queries may identify objects through their actions, relations, or roles in temporally localized events. LISA connects language reasoning to mask prediction (Lai et al., 2024), while VISA and VideoLISA extend this direction to videos and introduce evaluation benchmarks (Yan et al., 2024; Bai et al., 2024). Subsequent models unify global and local context (Lin et al., 2025), learn hierarchical temporal tokens (Gong et al., 2025), or couple MLLMs with promptable segmentation (Yuan et al., 2025). VideoSeg-R1 uses reinforcement learning for reasoning and spatial grounding (Xu et al., 2026). These approaches rely on task-specific training; MAVER instead coordinates pretrained components without additional fine-tuning.

Video object segmentation. Given visual prompts identifying a target, video segmentation methods maintain its identity across frames. XMem addresses long-term correspondence through memory consolidation (Cheng & Schwing, 2022), while Cutie uses object-level memory reading to resist distractors (Cheng et al., 2024). SAM enables promptable image segmentation (Kirillov et al., 2023); SAM 2 adds streaming video memory (Ravi et al., 2025), and SAM 3 supports concept prompts specified by short noun phrases or image exemplars (Carion et al., 2026). SAM2Long, SAMURAI, and DAM4SAM improve propagation through alternative memory pathways, motion cues,

and distractor-aware memory, respectively (Ding et al., 2025; Yang et al., 2024; Videnovic et al., 2025). MAVER addresses the complementary problem of deciding which query-grounded candidates should initialize propagation.

Training-free visual reasoning. Chain-of-thought prompting elicits intermediate reasoning (Wei et al., 2022; Kojima et al., 2022), and ViperGPT composes perception tools (Surís et al., 2023). VideoAgent uses structured memory for tool-based video understanding (Fan et al., 2024), while LVAgent enables multi-round collaboration among MLLM agents (Chen et al., 2025); just-in-time digital twins construct query-dependent scene representations for online segmentation (Shen et al., 2025). AL-Ref-SAM2 uses GPT to select pivot frames and candidate boxes (Huang et al., 2025), CoT-RVS infers object descriptions and corresponding keyframes (Kao et al., 2026), and DecAF prompts SAM 2 from refined MLLM attention maps (Han et al., 2026). Refer-Agent uses Questioner–Responder reflection to check target existence and consistency with the original query, answering each failed check by refining frame selection and intent analysis (Jiang et al., 2026). In contrast, MAVER first searches alternative anchors under a fixed hypothesis and initializes propagation from the first confirmed anchor alone; it revises the hypothesis only after a whole round fails, and if all rounds fail, retained candidates provide an explicitly unconfirmed fallback.

3 MAVER

MAVER uses an inferred description to retrieve candidates and the original query to verify them. It searches multiple anchors under a fixed hypothesis, then invokes evidence recovery if none is confirmed (Fig. 3). Three frozen components support these decisions: a multimodal reasoner $\mathcal{F}_{\mathrm{mllm}}$, which serves the role-specific calls f_{dec}, f_{sel}, f_{ver}, and f_{glob} (Appx. C); an open-vocabulary detector $\mathcal{F}_{\mathrm{det}}$ for proposals and relevance scores; and a video segmenter $\mathcal{F}_{\mathrm{vid}}$ for mask propagation. Model choices and hyper-parameters appear in Sec. 4.1; no component is fine-tuned for the task.

Input and output. Given a video $V = \{I_t \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{T}$ and a text query q, MAVER predicts a binary mask sequence $\hat{y} = \{m_t \in \{0,1\}^{H\times W}\}_{t=1}^{T}$,

$$\hat{y} = \psi(V, q), \tag{1}$$

where T is the video length, $H \times W$ is the frame resolution, and ψ denotes the complete pipeline. Each m_t covers the requested object or the union of requested objects, and is empty when none is visible. We describe the single-target path for explicit and implicit queries; multi-target completion and absent-target handling appear in Appx. B and C.

3.1 Framework components

Hypothesis decomposer. From q and n_d uniformly sampled frames, the reasoner produces a hypothesis $h = (c, s, \mathcal{Q})$: an object category noun c, a short literal description s, and L verification questions $\mathcal{Q} = \{Q_j\}_{j=1}^{L}$. The description expresses the inferred target in visual terms; the questions comprise a main verification question and complementary attribute checks. Repeated decompositions are aggregated by category voting (Wang et al., 2023).

Detector. At each anchor frame I_a (Sec. 3.2), $\mathcal{F}_{\mathrm{det}}$ ranks generated boxes against s, retaining up to k description candidates after score filtering, non-maximum suppression, and containment de-duplication (Appx. B). Category retrieval using c then adds up to k_c same-category instances that the description may miss; these proposals are appended after IoU de-duplication, so the description candidates keep their order. Each box b_a^i in the ordered pool $B_a = \{b_a^i\}_{i=1}^{|B_a|}$ receives identifier i, and numbered overlays (Yang et al., 2023) let the reasoner select detector-provided geometry without generating coordinates.

Selector and verifier. The selector f_{sel} compares candidates and returns a proposal identifier or `null`. The verifier f_{ver} checks the chosen region against q, returning Boolean answers and a box-quality label that determine prompt acceptance. Both calls produce structured decisions without free-form rationales.

Video segmenter. The segmenter $\mathcal{F}_{\mathrm{vid}}$ propagates masks from an anchor box in both temporal directions. Long empty intervals in its predictions provide feedback for subsequent repair.

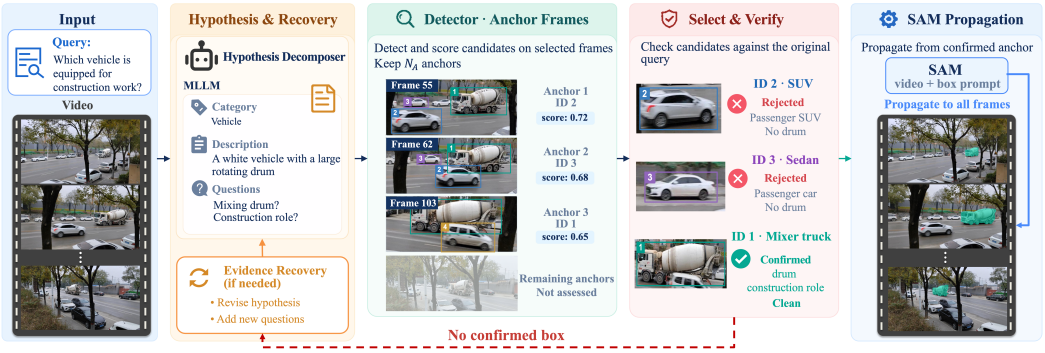

Figure 3: **Overview of MAVER.** A hypothesis is inferred from the query. Anchors are assessed in descending detector score; at each, a numbered proposal is selected and verified against the original query, and the first confirmed box prompts SAM 3. A failed round triggers evidence recovery.

3.2 Multi-anchor verification

Anchor selection. A grounding round fixes one hypothesis and searches alternative observations under it. The first round constructs the hypothesis and anchor set as

$$h = f_{\text{dec}}\big(q, \text{Sample}(V, n_d)\big), \qquad \mathcal{A} = \text{Anchors}\big(V, s; n_s, K, N_A\big), \tag{2}$$

where $\text{Sample}(V, n)$ uniformly samples n frames. The detector's referring scorer ranks n_s uniformly sampled frames by their maximum proposal score against s. Temporal non-maximum suppression retains K separated keyframes, from which Anchors selects N_A evenly spaced entries in temporal order (Appx. A). These anchors are assessed in descending order of their first-ranked proposal scores. Temporal coverage offers alternative views when the target or its defining interaction is obscured; detector scores prioritize the search.

Proposal selection. At anchor a, the selector receives q, h, and the numbered overlay of B_a on I_a. It returns identifier i_a, and a separate call verifies the selected proposal:

$$i_a = f_{\text{sel}}(q, h, I_a, B_a), \qquad v_{ai} = f_{\text{ver}}(q, h, I_a, b_a^i). \tag{3}$$

Here v_{ai} denotes the verification output for a valid proposal i, initially $i = i_a$; a `null` selection leaves the anchor unconfirmed. A singleton pool bypasses selection and proceeds directly to verification. The keyframe sequence provides temporal context where enabled (Appx. A).

Structured verification. The verifier sees the highlighted box and a zoomed crop, and first checks the original query while setting aside s. Confirmation requires passing the hard gates, namely query match and any temporal check posed by the decomposer, a strict majority of positive answers to $\mathcal{Q}$, and a clean or partial box. The output v_{ai} comprises the answers $u_{aij} \in \{0, 1\}$ to Q_j, the box-quality label z_{ai}, and the conjunction $g_{ai} \in \{0, 1\}$ of the hard gates. The acceptance rule is

$$\text{accept}(a, i) = \mathbf{1}\left[\sum_{j=1}^{L} u_{aij} > L/2\right] g_{ai}\, \mathbf{1}[z_{ai} \in \{\texttt{clean}, \texttt{partial}\}], \tag{4}$$

where $\mathbf{1}[\cdot]$ is the indicator function. The gates are evaluated separately from the majority vote, so matching attributes cannot override a failed query or temporal check; temporal checks use video context and equal 1 when inapplicable. The box-quality check rejects regions that merge the target with another instance. If the selected box fails, the first-ranked proposal, when different, is verified once. If no proposal is confirmed, search advances to the next anchor with h unchanged.

Anchor acceptance. The first confirmed proposal ends the anchor search (early stopping). Its frame and box form the single initial prompt for mask propagation:

$$\mathcal{P}_0 = \{(a^\star, b_{a^\star}^{i^\star})\}, \qquad \hat{y}^{(0)} = \mathcal{F}_{\text{vid}}(V, \mathcal{P}_0), \tag{5}$$

where $(a^\star, i^\star)$ is the first pair in assessment order, counting first-ranked alternatives, that satisfies $\text{accept}(a^\star, i^\star) = 1$, and $\hat{y}^{(0)}$ is the initial prediction. Later anchors are skipped: multiple anchors provide search opportunities, while only one confirmed observation initializes propagation.

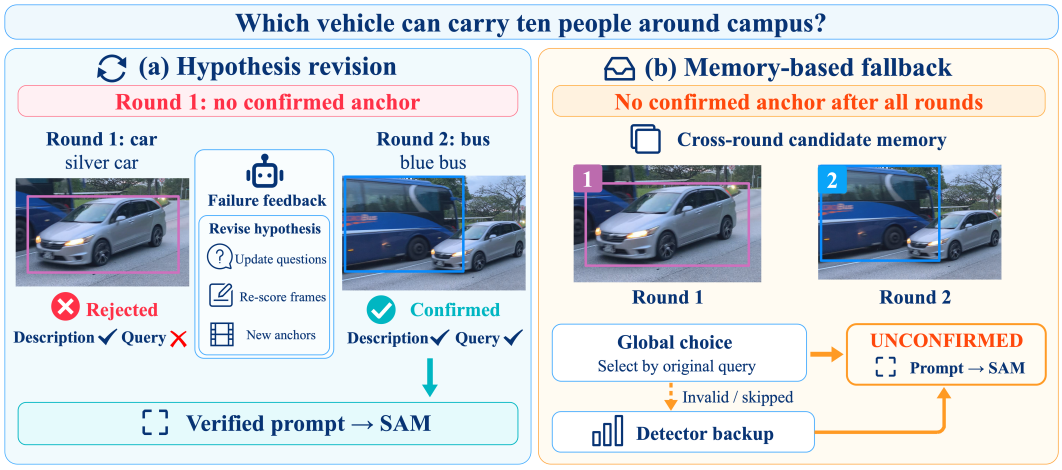

Figure 4: **Evidence recovery in MAVER.** (a) Hypothesis revision: the silver car matches the description but fails the query-match check at every anchor, so the hypothesis is revised to a bus, which is confirmed. (b) Memory-based fallback: if all rounds fail, a global choice over cross-round memory supplies an unconfirmed prompt, with a detector backup for invalid or skipped choices.

Post-propagation repair. After initial propagation, long empty intervals can trigger new grounding under the same hypothesis and original query. Only confirmed probes add prompts for one re-propagation; otherwise, the initial prediction is retained. Appx. B specifies the trigger conditions and maximum number of probes.

3.3 Evidence Recovery

Failure to confirm any anchor in a round may reflect an incorrect hypothesis, missing proposals, or verifier errors. Revision is triggered only after an entire round fails, so that a failed observation at one anchor cannot discard a hypothesis that another anchor may still confirm. We allow up to R grounding rounds, including the initial round; if none yields confirmation, memory-based fallback supplies an explicitly unconfirmed prompt from retained candidates (Fig. 4). Lack of confirmation alone does not establish target absence.

Hypothesis revision. Let $h^{(r)} = (c^{(r)}, s^{(r)}, \mathcal{Q}^{(r)})$ be the round-r hypothesis, with $h^{(1)} = h$. If no anchor is confirmed and $r < R$, the decomposer receives the unsuccessful category–description pairs:

$$h^{(r+1)} = f_{\text{dec}}\big(q, \text{Sample}(V, n_d), \mathcal{H}_r^-\big), \qquad \mathcal{H}_r^- = \{(c^{(\rho)}, s^{(\rho)})\}_{\rho=1}^{r}. \tag{6}$$

The feedback requests an alternative interpretation, which may retain the category while changing the target description. Frame scoring, anchor selection, and grounding restart under $h^{(r+1)}$. Verification questions $\mathcal{Q}$ and their number L may change across rounds; the original query q and acceptance rule remain fixed. Confirmation initializes propagation via Eq. 5.

Memory-based fallback. During search, MAVER retains all anchors' proposals across rounds, including rejected ones, with their source frames in a cross-round memory $\mathcal{M}$. The memory is deduplicated and capped while protecting each round's *leading proposal*, the highest-scored among the anchors' first-ranked proposals (Appx. B), so earlier alternatives remain available after revision. If all R rounds fail and $\mathcal{M}$ is non-empty, the reasoner makes one global choice:

$$\ell^\star = f_{\text{glob}}\big(q, \mathcal{M}, (c^{(R)}, s^{(R)})\big), \tag{7}$$

where $\ell^\star$ is a global candidate identifier. The reasoner compares retained candidates across frames and rounds against q and must choose one. It receives the final category and description as an unreliable prior, without verification questions. If the global choice is unavailable or invalid, a detector backup selects the highest-scored leading proposal across rounds. The chosen frame and box form $\mathcal{P}_0$ for propagation, allowing recovery of a target that verification may have mistakenly rejected. This prompt remains *unconfirmed*; an empty memory yields an empty prediction.

Table 1: Comparison with state-of-the-art methods on Ref-YouTube-VOS, MeViS, and ReasonVOS.

Method	MLLM	Ref-YouTube-VOS			MeViS			ReasonVOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Supervised fine-tuning										
VISA (Yan et al., 2024)	Chat-UniVi-13B	63.0	61.4	64.7	44.5	41.8	47.1	–	–	–
VideoLISA (Bai et al., 2024)	LLaVA-3.8B	63.7	61.7	65.7	44.4	41.3	47.6	47.5	45.1	49.9
GLUS (Lin et al., 2025)	LISA-7B	67.3	65.5	69.0	51.3	48.5	54.2	49.9	47.5	52.4
Sa2VA (Yuan et al., 2025)	InternVL2-8B	70.7	–	–	46.9	–	–	–	–	–
RGA3 (Wang et al., 2025)	Qwen2.5-VL-7B	68.5	66.8	70.1	50.1	47.4	52.8	53.6	51.3	56.0
VRS-HQ (Gong et al., 2025)	Chat-UniVi-13B	71.0	69.0	73.1	50.9	48.0	53.7	–	–	–
Training-free inference										
AL-Ref-SAM2 (Huang et al., 2025)	GPT-4	67.9	65.9	69.9	42.8	39.5	46.2	–	–	–
CoT-RVS (Kao et al., 2026)	Gemma3-12B	–	–	–	44.2	40.3	48.1	50.7	47.5	54.0
CoT-RVS (Kao et al., 2026)	GPT-4o	–	–	–	52.2	48.7	55.7	65.5	62.4	68.7
Refer-Agent (Jiang et al., 2026)	Ovis2.5-9B	71.3	69.0	73.6	54.7	51.6	57.7	69.8	67.0	72.7
MAVER (ours)	Qwen3-VL-8B	**76.6**	**74.5**	**78.6**	**56.7**	**53.3**	**60.1**	**74.1**	**71.2**	**76.9**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Implementation details. We use Qwen3-VL-8B-Instruct (Bai et al., 2025) for reasoning, WeDetect-Uni and WeDetect-Ref (Fu et al., 2026) for proposal generation and relevance scoring, and SAM 3 (Carion et al., 2026) for mask prediction and propagation. All pretrained components remain frozen. We use the same hyper-parameters across datasets by default. Decomposition uses $n_d = 16$ uniformly sampled frames with category voting. From $n_s = 36$ scored frames, we select $K = 10$ keyframes and $N_A = 5$ anchors, retaining up to $k = 5$ description and $k_c = 4$ category proposals per anchor. The $R = 2$ rounds allow one revision of an unconfirmed hypothesis. Appx. A details settings on different datasets; Appx. B–C provide implementation details and prompts.

Datasets. We evaluate referring segmentation on Ref-YouTube-VOS (Seo et al., 2020) and MeViS (Ding et al., 2023), and reasoning segmentation on ReasonVOS (Bai et al., 2024) and ReVOS (Yan et al., 2024). These benchmarks cover explicit references, motion descriptions, and implicit reasoning queries. See Appx. B for dataset descriptions and evaluation protocols.

Evaluation metrics. Following previous works, we evaluate segmentation with region similarity ($\mathcal{J}$), contour accuracy ($\mathcal{F}$), and their average ($\mathcal{J}\&\mathcal{F}$), all reported as percentages. Specifically, $\mathcal{J}$ measures the intersection over union between predicted and ground-truth masks, while $\mathcal{F}$ combines boundary precision and recall to assess contour alignment.

4.2 MAIN RESULTS

Referring video segmentation. As shown in Tab. 1, without any fine-tuning, MAVER outperforms supervised and training-free methods across all reported metrics on Ref-YouTube-VOS and MeViS. Compared with Refer-Agent, the strongest training-free baseline, it improves performance by 5.3% and 2.0%, respectively. It also exceeds the strongest supervised method on each benchmark by more than 5%, even on MeViS, where queries separate similar instances by motion. These results show that the framework handles both explicit references and motion-based descriptions.

Reasoning video segmentation. On ReasonVOS, MAVER surpasses Refer-Agent by 4.3% and, with an 8B open-source reasoner, outperforms CoT-RVS with GPT-4o by 8.6% (Tab. 1). On ReVOS, it leads across all reported metrics, improving overall performance over Refer-Agent by 1.2% (Tab. 2). The gain is larger for reasoning queries (1.6%) than for referring queries (0.7%), suggesting that checking candidates against the original query matters most for implicit queries.

Performance and inference cost. On ReasonVOS, MAVER outperforms Refer-Agent by 15.6% with the same Qwen3-VL-8B reasoner while reducing MLLM calls from 12.1 to 6.3 per query (Tab. 3). Runtime increases from 44 to 57 seconds per query, mainly because category voting repeats multi-frame decomposition. Against Refer-Agent with Ovis2.5-9B, MAVER achieves higher performance with roughly half the calls and one fifth the runtime. GLUS and CoT-RVS offer lower latency at lower performance.

Table 2: Comparison with state-of-the-art methods on ReVOS.

Method	MLLM	Referring			Reasoning			Overall		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Supervised fine-tuning										
VISA (Yan et al., 2024)	Chat-UniVi-13B	57.4	55.6	59.1	44.3	42.0	46.7	50.9	48.8	52.9
GLUS (Lin et al., 2025)	LISA-7B	58.3	56.0	60.7	51.4	48.8	53.9	54.9	52.4	57.3
HyperSeg (Wei et al., 2025a)	Mipha-3B	58.5	56.0	60.9	53.0	50.2	55.8	55.7	53.1	58.4
InstructSeg (Wei et al., 2025b)	Mipha-3B	57.0	54.8	59.2	51.9	49.2	54.7	54.5	52.0	56.9
ViLLa (Zheng et al., 2025)	InternVideo2-6B	–	–	–	–	–	–	57.0	54.9	59.1
RGA3 (Wang et al., 2025)	Qwen2.5-VL-7B	60.5	58.7	62.3	55.4	53.1	57.7	58.0	55.9	60.0
VRS-HQ (Gong et al., 2025)	Chat-UniVi-13B	63.3	61.1	65.5	56.8	54.1	59.4	60.0	57.6	62.5
Training-free inference										
CoT-RVS (Kao et al., 2026)	Gemma3-12B	–	–	–	–	–	–	47.1	43.4	50.9
CoT-RVS (Kao et al., 2026)	GPT-4o	–	–	–	–	–	–	55.9	52.8	59.0
Refer-Agent (Jiang et al., 2026)	Ovis2.5-9B	65.0	62.3	67.7	57.7	54.3	61.0	61.3	58.3	64.3
MAVER (ours)	Qwen3-VL-8B	**65.7**	**62.8**	**68.7**	**59.3**	**56.1**	**62.5**	**62.5**	**59.4**	**65.6**



Figure 5: **Qualitative comparison on ReasonVOS.** Unlike the baselines, MAVER follows the tiger that moves between drinking spots. Predicted masks are blue and ground-truth masks are green.

Qualitative comparison. In Fig. 5, MAVER follows the tiger moving between drinking spots, whereas CoT-RVS and GLUS segment neighboring tigers and Refer-Agent selects another tiger or produces empty masks. At the first assessed anchor, the detector ranks a different tiger highest for the inferred description; MAVER instead selects a tiger added by category retrieval and confirms it against the original query. Further comparisons appear in Appx. F.

4.3 Ablation Study

Multi-anchor verification. Removing verification reduces performance on ReasonVOS by 3.8% (Tab. 4), and omitting only the query-match check costs 2.4%, suggesting that matching the inferred description does not guarantee satisfying the query. Using one anchor while retaining ten keyframes lowers performance by 3.7%, whereas five anchors add only 1.3 MLLM calls per query since the search stops at the first confirmed anchor. Searching alternative observations under a fixed hypothesis thus appears worthwhile. Other anchor and keyframe counts are studied in Appx. E.1.

Evidence recovery. Removing both hypothesis revision and memory-based fallback reduces performance by 5.8% (Tab. 5). Keeping only fallback costs 0.6%, whereas keeping only revision costs 4.4%, so fallback provides most of the recovery gain while adding almost no MLLM calls. Strict verification thus needs a recovery path: fallback turns candidates that verification leaves unconfirmed into a prompt instead of an empty prediction, while revision helps when the hypothesis itself is wrong. Further recovery rounds and fallback memory sizes are studied in Appx. E.2.

Table 3: Performance and inference cost on ReasonVOS.

Method	MLLM	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Calls	s/q
GLUS (Lin et al., 2025)	LISA-7B	47.5	52.4	49.9	–	**25**
CoT-RVS (Kao et al., 2026)	GPT-4o	62.4	68.7	65.5	**2.5**	32
Refer-Agent (Jiang et al., 2026)	Ovis2.5-9B	67.0	72.7	69.8	11.7	269
Refer-Agent (Jiang et al., 2026)	Qwen3-VL-8B	55.6	61.3	58.5	12.1	44
MAVER (ours)	Qwen3-VL-8B	**71.2**	**76.9**	**74.1**	6.3	57

Table 4: Multi-anchor verification.

Setting	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Calls
MAVER (full)	**71.2**	**76.9**	**74.1**	6.3
w/o verification	67.4	73.2	70.3	**2.7**
w/o query-match check	68.8	74.6	71.7	4.1
1 anchor	67.5	73.3	70.4	5.0

Table 5: Evidence recovery.

Setting	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Calls
MAVER (full)	**71.2**	**76.9**	**74.1**	6.3
w/o recovery	65.7	70.8	68.3	**4.9**
Fallback only	70.7	76.4	73.5	5.1
Revision only	67.1	72.3	69.7	6.2

Table 6: Reasoning models; pairwise best in bold.

MLLM	Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Qwen3-VL-8B-Instruct	Refer-Agent	55.6	61.3	58.5
	MAVER	**71.2**	**76.9**	**74.1**
Qwen3-VL-8B-Thinking	Refer-Agent	66.2	72.1	69.1
	MAVER	**69.4**	**75.6**	**72.5**
Ovis2.5-9B	Refer-Agent	67.0	72.7	69.8
	MAVER	**67.1**	**73.4**	**70.2**
Qwen2.5-VL-7B	Refer-Agent	52.4	58.6	55.5
	MAVER	**58.2**	**64.5**	**61.4**

Table 7: Grounding interface and tracker.

Setting	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Calls
MAVER (full)	**71.2**	**76.9**	**74.1**	6.3
Detector top-1 + verification	70.3	76.3	73.3	5.2
Detector top-1 only	67.5	73.6	70.5	**2.1**
Generated box + verification	68.3	74.5	71.4	10.9
Without category proposals	69.3	75.4	72.4	5.2
Raw-query detector scoring	68.1	74.1	71.1	6.6
SAM 3 text grounding	55.2	59.9	57.5	4.8
SAM 2 tracker	69.7	75.5	72.6	6.4

Reasoning models. Tab. 6 pairs MAVER with Refer-Agent under four reasoners, retaining method-specific components. MAVER leads in every pair, most clearly with non-thinking reasoners (15.6% and 5.9%); the margin narrows for Qwen3-VL-8B-Thinking and Ovis2.5-9B, both in thinking mode. Within MAVER, Qwen3-VL-8B-Instruct exceeds its Thinking counterpart by 1.6%, supporting our default choice and suggesting that verification reduces the need for long deliberation.

Grounding interface and tracker. Tab. 7 varies how the initial prompt is obtained and which model propagates it. Verifying the detector's top-ranked proposal closes most of the gap between the detector alone and the full system. Having the reasoner generate box coordinates instead of selecting detector proposals lowers performance by 2.7% with about 1.7 times as many calls, and scoring frames and proposals with the raw query rather than the inferred description, or removing category proposals, also lowers performance. These results favor selecting detector-provided geometry retrieved with the visual hypothesis. Prompting SAM 3 directly with the description is far less effective. With a SAM 2 tracker, as in Refer-Agent and CoT-RVS, MAVER still exceeds Refer-Agent in Tab. 1 by 2.8%, so its advantage does not rely on SAM 3. Further analyses of detector threshold, frame sampling, proposals per anchor, and decomposition appear in Appx. E.3–E.5.

5 Conclusion

We presented MAVER, a training-free framework that treats inferred object descriptions as revisable hypotheses for retrieval and verifies candidates against the original query. Multi-anchor verification searches alternative anchors under a fixed hypothesis and initializes propagation from the first confirmed one; evidence recovery revises failed hypotheses and, if every round fails, falls back to retained candidates. MAVER achieves the highest performance on all four benchmarks and requires fewer MLLM calls than Refer-Agent on ReasonVOS, where it also leads under four matched reasoners. Ablations support the roles of verification, anchor search, and evidence recovery, with fallback providing most of the recovery gain.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript and to provide auxiliary coding support. All AI-assisted manuscript edits were reviewed by the authors, and AI-assisted code was verified and tested by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly released video segmentation benchmarks and pretrained models under their respective licences and involves no new data collection or human-subject studies. Potential concerns include biases inherited from pretrained components, incorrect object associations, and misuse for privacy-invasive surveillance. Responsible use should respect privacy and consent, evaluate performance in the intended setting, and include human oversight in applications where errors could cause harm.

REPRODUCIBILITY STATEMENT

Sec. 3 describes the method, and Sec. 4.1 specifies the pretrained components, datasets, and evaluation metrics. Appx. A and B provide hyper-parameters, inference settings, and benchmark protocols; Appx. C documents prompts and output schemas. Appx. E presents additional ablation results by experimental factor, and Appx. F describes the qualitative comparisons.

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. *Advances in Neural Information Processing Systems*, 37:6833–6859, 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, 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 *The Fourteenth International Conference on Learning Representations*, 2026.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. LVAgent: Long video understanding by multi-round dynamical collaboration of MLLM agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20237–20246, 2025.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In *European Conference on Computer Vision*, pp. 640–658. Springer, 2022.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3151–3161, 2024.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAM-WISE: Infusing wisdom in SAM2 for text-driven video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3395–3405, 2025.

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, 2023.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. SAM2Long: Enhancing SAM 2 for long video segmentation with a training-free memory tree. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13614–13624, 2025.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Shenghao Fu, Yukun Su, Fengyun Rao, Jing Lyu, Xiaohua Xie, and Wei-Shi Zheng. WeDetect: Fast open-vocabulary object detection as retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20377–20387, 2026.

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 Computer Vision and Pattern Recognition Conference*, pp. 29183–29192, 2025.

Su Ho Han, Jeongseok Hyun, Pilhyeon Lee, Minho Shim, Dongyoon Wee, and Seon Joo Kim. Decomposed attention fusion in MLLMs for training-free video reasoning segmentation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13332–13341, 2024.

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*, volume 39, pp. 3715–3723, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39508–39517, 2026.

Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. CoT-RVS: Zero-shot chain-of-thought reasoning segmentation for videos. In *The Fourteenth International Conference on Learning Representations*, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20009–20019, 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 Computer Vision and Pattern Recognition Conference*, pp. 8658–8667, 2025.

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 *The Thirteenth International Conference on Learning Representations*, 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.

Yiqing Shen, Bohan Liu, Chenjia Li, Lalithkumar Seenivasan, and Mathias Unberath. Online reasoning video segmentation with just-in-time digital twins. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24698–24706, 2025.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023.

Jiajin Tang, Ge Zheng, and Sibei Yang. Temporal collection and distribution for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15466–15476, 2023.

Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with SAM2. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24255–24264, 2025.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

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 Computer Vision and Pattern Recognition Conference*, pp. 8931–8941, 2025a.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. OnlineRefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2761–2770, 2023.

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.

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. Springer, 2024.

Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. SAMURAI: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023.

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:2501.04001v1*, 2025.

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.