DIAGNOSING TARGET PRESERVATION FOR RELIABLE MLLM-BASED REFERRING VIDEO OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring video object segmentation requires identifying a language-specified target and preserving its identity throughout a video. However, mean segmentation accuracy alone provides limited insight into system reliability and the sources of failure. In this work, we investigate target preservation in a controlled multimodal large language model-based pipeline by tracing frame-level proposals, reference masks, and propagated predictions. We find that reference formation is a key bottleneck in this pipeline, even when the target is visible in the selected frame, and that propagation failures can persist despite ground-truth initialization. Guided by these findings, we propose a modular framework that separates video-level target resolution, single-frame reference formation, and video propagation. Support-aware proposal control withholds proposals when visual evidence is insufficient, while conditional recovery revisits these queries with additional observations. Experiments on MeViS demonstrate competitive segmentation accuracy, while controlled comparisons reveal trade-offs among output coverage, mask quality, and target-absent recall, highlighting the need to evaluate reliability beyond mean accuracy. The source code will be released.

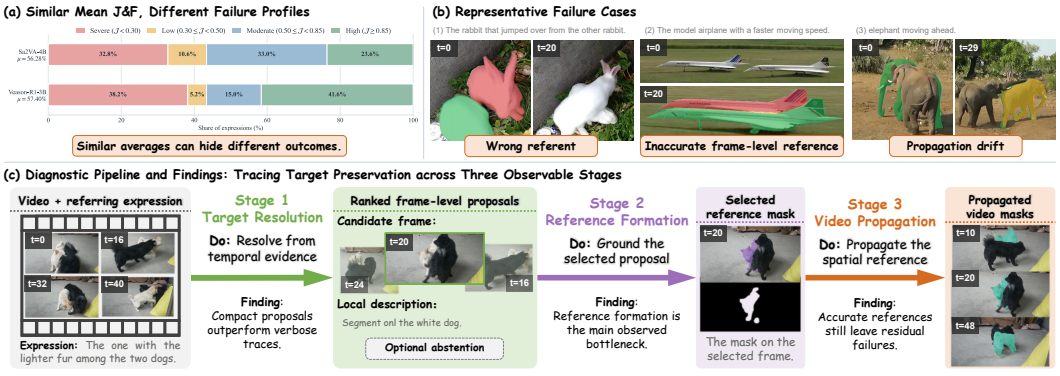


Figure 1: Diagnosing target preservation in MLLM-based RVOS. (a) Upper left: stacked bars show the proportions of queries in different video-level $\mathcal{J}$ ranges, revealing different outcome distributions despite similar mean $\mathcal{J}\&\mathcal{F}$. (b) Upper right: the examples show a wrong referent, an inaccurate reference mask, and propagation drift, from left to right. In these visualizations, green, red, and yellow denote ground-truth-only, prediction-only, and overlapping regions, respectively. (c) Bottom: the diagnostic pipeline traces frame-level proposals, reference masks, and propagated video masks to locate failures. Annotations at each stage summarize its operation and the corresponding diagnostic finding.

1 INTRODUCTION

Referring video object segmentation (RVOS) identifies and segments a target specified by a natural-language expression throughout a video. The expression may describe appearance, motion, interaction, cardinality, temporal order, or identity across disappearance and reappearance (Ding et al.,

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.

2023, 2025, Liang et al., 2026, Yan et al., 2024, Bai et al., 2024). The system must resolve these cues, maintain identity through temporal changes and interactions with similar instances, and recognize target-absent frames.

Existing RVOS methods connect language understanding to segmentation through different modeling and inference strategies. For example, conventional end-to-end models jointly process language and video features to predict target masks across frames (Botach et al., 2022, Wu et al., 2022, Liang et al., 2025). Multimodal large language model (MLLM)-based segmentation models couple multimodal reasoning with mask prediction through learned segmentation interfaces (Bai et al., 2024, Yan et al., 2024, Yuan et al., 2026). Agentic RVOS systems coordinate MLLM reasoning, visual inspection, and segmentation tools to produce and refine predictions (Jiang et al., 2026, Jin et al., 2026). Across these approaches, failures may include incorrect target identification, inaccurate masks, or loss of target identity over time (Fig. 1(b)). However, mean $\mathcal{J}\&\mathcal{F}$ alone provides limited insight into these failure patterns, as illustrated in Fig. 1(a), where systems with similar average scores exhibit different distributions of segmentation quality.

In this work, we assess RVOS reliability beyond mean accuracy by jointly evaluating mask quality, output coverage, and target-absent recall. To examine errors before and after propagation, we trace frame-level proposals, reference masks, and propagated predictions in a controlled pipeline (Fig. 1(c)). An MLLM proposes ranked candidate frames with local descriptions, an image segmenter forms reference masks on these frames, and a video model propagates a selected reference mask. We compare reference and final-mask quality on the same queries and use matched ground-truth (GT) references to examine propagation failures. This analysis reveals that reference formation is a key bottleneck in this pipeline, even when the target is visible in the selected frame. Reference-mask quality is strongly associated with final video quality, while some propagation failures persist even with GT initialization. Controlled comparisons further support compact frame-level proposals and single-frame reference formation.

Guided by these findings, we propose a modular MLLM-based framework with specialized modules for video-level target resolution, single-frame reference formation, and video propagation. Support-aware proposal control withholds proposals when sampled observations provide insufficient target evidence, while conditional recovery revisits these queries with additional observations. Comparisons with and without recovery characterize the trade-offs among output coverage, mask quality, and target-absent recall. On the MeViS `valid_u` split, our framework achieves mean $\mathcal{J}\&\mathcal{F}$ comparable to Refer-Agent (Jiang et al., 2026), with higher target-absent recall and lower estimated GPU time. Our main contributions are as follows:

- We analyze target preservation in a controlled MLLM-based RVOS pipeline and identify reference formation as a key bottleneck, with some propagation failures persisting under matched ground-truth initialization.
- We propose a modular RVOS framework with support-aware proposal control to withhold proposals when visual evidence is insufficient and conditional recovery to revisit these queries with additional observations.
- We jointly evaluate mask quality, output coverage, and target-absent recall on MeViS, and use controlled comparisons to examine how proposal policies and inference budgets affect these outcomes.

2 Related Work

RVOS Models. Conventional end-to-end RVOS models jointly process language and video features to predict target masks across frames (Botach et al., 2022, Wu et al., 2022, Liang et al., 2025). Recent methods connect MLLMs to image or video segmenters through mask tokens, visual prompts, or persistent object states (Lai et al., 2024, Bai et al., 2024, Yan et al., 2024, Munasinghe et al., 2025, Yuan et al., 2026, Gong et al., 2025, Lin et al., 2025b). Evaluation on final video masks combines the effects of target identification and segmentation. We additionally inspect intermediate reference masks to characterize errors before and after propagation.

Modular and Agentic RVOS. Long and ambiguous videos motivate temporal retrieval, event reasoning, and iterative tool use (Lin et al., 2025a, Li et al., 2026, Kao et al., 2026, Xu et al., 2025,

Jin et al., 2026, Tran et al., 2025). Modular methods such as FindTrack separate target identification from mask propagation (Cho et al., 2025). Agentic MLLM-based systems coordinate reasoning and segmentation tools: AgentRVOS reasons over object tracks, while Refer-Agent iteratively verifies and refines intermediate predictions (Jin et al., 2026, Jiang et al., 2026). VideoSEG-O3 uses multi-turn reinforcement learning to gather and verify video evidence (Dai et al., 2026). Our study examines reference formation and proposal control within a modular MLLM-based pipeline. Negative video-text pairs also inform proposal control: Robust R-VOS studies mismatched references (Li et al., 2023), and we use negative cases to supervise whether the available observations support a proposal.

Reference Formation and Video Propagation. Promptable segmenters convert points, boxes, language, or mask tokens into spatial predictions (Kirillov et al., 2023, Ravi et al., 2025, Cuttano et al., 2025, Zhou et al., 2026, Rong et al., 2025). Memory-based and promptable video models then propagate a reference state through occlusion, deformation, and reappearance (Cheng & Schwing, 2022, Ravi et al., 2025, Carion et al., 2026). The reference mask specifies the object regions that the video model will propagate. We examine the relationship between reference-mask quality and final video segmentation, and use matched GT references to assess residual propagation errors.

3 Diagnosing Target Preservation

3.1 Controlled Analysis Setup

Given video $V = (I_1, \ldots, I_T)$ and expression q, we trace the target through three stages:

$$(V, q) \xrightarrow{R} \mathcal{C} \xrightarrow{G} (t^*, \tilde{m}_{t^*}) \xrightarrow{P} \hat{M}_{1:T}. \tag{1}$$

Here, target resolution R identifies the referred object and proposes candidate frames with local descriptions. Reference formation G grounds a proposal as the selected reference mask $\tilde{m}_{t^*}$. Video propagation P propagates this mask into video predictions $\hat{M}_{1:T}$. We define *target preservation* as retaining the identity, relation, cardinality, and scope specified by q. We measure reference-mask IoU and final video $\mathcal{J}\&\mathcal{F}$ as proxies for target preservation. These scores locate mask errors before and after propagation but do not separately verify each semantic requirement. A low reference IoU can result from the proposal description, selected frame, or image segmenter.

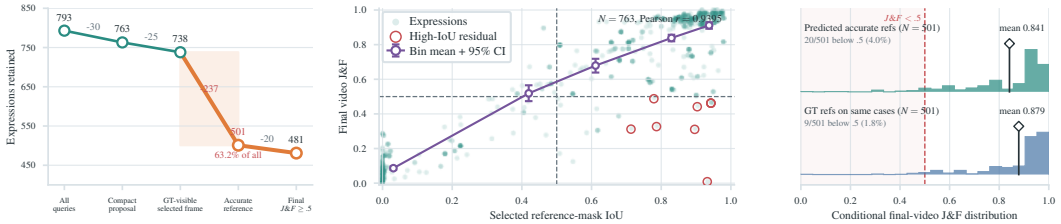
The controlled analysis uses Qwen3-VL-4B (Bai et al., 2025), frozen SAMTok (Zhou et al., 2026), and frozen SAM3 (Carion et al., 2026). Unless specified otherwise, the diagnostic route uses compact proposals with at most three candidates and follows the top-ranked reference through the frozen propagator. An *accurate reference* has frame-level $\mathcal{J} \geq 0.5$, determined offline and never used for inference or routing. Coverage and accurate-reference rates are computed over all queries. Local quality is measured on queries with proposals. The funnel and correlation analyses use the selected reference, following candidate rank.

3.2 Reference Formation and Propagation Failures

Losses at reference formation. Fig. 2(a) follows all 793 MeViS `valid_u` expressions through the compact-proposal diagnostic route with up to three candidates. The route produces proposals for 763 queries. The selected frame contains the target in 738 cases, and the selected reference mask reaches local $\mathcal{J} \geq 0.5$ in 501 cases. Of these 501 cases, 481 retain final video $\mathcal{J}\&\mathcal{F} \geq 0.5$. Among the 262 references below threshold, 25 use a target-absent frame and 237 fail despite GT visibility. The 237 GT-visible failures can reflect proposal, selection, or mask-generation errors. Proposal semantics remain unannotated.

Among the 763 proposal-bearing expressions, 481 pass both the reference and final thresholds, 241 fail both, 21 recover after a below-threshold reference, and 20 fail after an accurate reference. Propagation can both recover weak references and degrade accurate ones. The same ordering of checkpoint decreases holds for matched thresholds from 0.3 to 0.7 (Appendix Table 9).

Reference quality and propagated masks. Reference-mask IoU correlates strongly with final video $\mathcal{J}\&\mathcal{F}$ in the compact route (Fig. 2(b)): Pearson 0.9395 and Spearman 0.8647. Among the



(a) Stage-wise target retention (b) Reference quality is associated with video quality (c) Accurate references leave propagation residuals

Figure 2: Target-preservation diagnostics in the compact route. (a) Of 793 queries, 763 yield proposals, 738 select GT-visible frames, 501 form accurate references, and 481 of those retain final $\mathcal{J}\&\mathcal{F} \geq 0.5$. Retained-query counts quantify these checkpoints. (b) Reference-mask IoU is associated with final video quality. (c) Matched GT references leave residual propagation failures on the same 501 eligible cases.

501 accurate references, 481 retain final $\mathcal{J}\&\mathcal{F} \geq 0.5$, supporting reference quality as an empirical indicator of downstream outcomes.

Residual propagation failures. Twenty of the 501 expressions with accurate predicted references finish below $\mathcal{J}\&\mathcal{F}$ 0.5, including nine whose local IoU is at least 0.7. Replacing predicted references with GT on these 501 expressions raises mean final $\mathcal{J}\&\mathcal{F}$ from 0.841 to 0.879, while nine expressions remain below 0.5 (Fig. 2(c)). This intervention characterizes the eligible subset and excludes the other 292 expressions.

These findings highlight the importance of reference formation for target preservation. Target visibility in a selected frame does not guarantee an accurate reference mask, and propagation errors can persist with GT initialization. They motivate an explicit interface between target resolution and reference formation, together with a decision about whether the available observations support a proposal or additional observations are needed.

4 Reliability-Oriented Framework

To address these challenges, we propose a modular framework that makes the transition from video-level target reasoning to reference-mask generation explicit. Frame-level proposals convey the resolved target to a dedicated image segmenter, whose reference masks initialize video propagation. Support-aware proposal control withholds proposals when visual evidence is insufficient. Since withholding proposals reduces output coverage, conditional recovery revisits these queries with additional observations while preserving predictions already produced.

4.1 Video-Level Target Resolution

Expressions may identify a target through motion or relations that require observations across multiple frames. We therefore resolve the target at the video level and translate the expression into concise descriptions tied to selected frames. Qwen3-VL-4B (Bai et al., 2025) takes the expression and sampled frames as input and returns ordered frame-level proposals

$$\mathcal{C} = \{(t_k, \tilde{q}_k, s_k)\}_{k=1}^{K}, \qquad K \leq 3, \tag{2}$$

where t_k is a candidate frame, $\tilde{q}_k$ is a concise frame-local description, and s_k records rank or confidence. The description must preserve the target identity, relations, cardinality, and scope specified by the original expression.

4.2 Reference Formation and Video Propagation

Each proposal links the target identified from the video to its appearance in a selected frame. We use the selected image and local description to form a reference mask, which provides the target initialization for video propagation.

Frozen SAMTok (Zhou et al., 2026) receives each selected image and local description,

$$\tilde{m}_{t_k} = G(I_{t_k}, \tilde{q}_k). \tag{3}$$

We initialize frozen SAM3 (Carion et al., 2026) with the accepted reference masks. SAM3 propagates each reference mask through the video without reinterpreting the expression. Our framework propagates up to three references independently and combines their masks, allowing propagation to start from several selected frames.

4.3 Support-Aware Control and Conditional Recovery

The proposal interface assumes that sampled observations provide enough evidence to identify the target. This assumption can fail when sampling misses a defining event or the observations do not support the expression. We train a support-aware controller with LoRA (Hu et al., 2022) to decide whether to produce or withhold a proposal. Its 20,000 training examples come exclusively from the MeViS and Long-RVOS training splits, with a 2:1:1 mixture of target-supported, observation-missing, and target-absent cases. Supervision is restricted to the decision to produce or withhold a proposal. No validation or test example or mask is used for optimization. The vision tower and both downstream modules remain frozen.

Inference configurations. We first apply support-aware proposal control to 32 uniformly sampled frames. This *quality-first configuration* withholds proposals when observations are insufficient, at the cost of reduced output coverage. Some of these queries may become answerable with additional observations. The *full configuration* therefore applies conditional recovery only to queries for which proposals were withheld, using 32 relevance-guided and 32 uniform observations while preserving predictions from the initial pass. Both configurations share the 4B MLLM, frozen SAMTok and SAM3, and up to three propagated references. Masks from separate passes remain separate. The tables label the configurations *Ours quality-first* and *Ours*, respectively. The compact single-reference route in Section 3 is a diagnostic control, with proposal coverage reported separately from deployed output coverage.

5 Experiments

5.1 Experimental Setup

Benchmarks and evaluation protocols. We use MeViS (Ding et al., 2023) as the main benchmark for stage-wise analysis. Its `valid_u` split supports controlled intermediate diagnostics, while the larger `valid` split provides an additional end-to-end evaluation. MeViSv2 extends the evaluation to more temporally diverse and target-absent cases (Ding et al., 2025). Long-RVOS evaluates long-term identity preservation through disappearance and reappearance (Liang et al., 2026). ReVOS and ReasonVOS evaluate dynamic relations and reasoning-oriented disambiguation, respectively (Yan et al., 2024, Bai et al., 2024). Our framework is evaluated zero-shot on both benchmarks. The controller is trained only on the MeViS and Long-RVOS training splits. We also use MeViS `valid_u` to develop prompts, routing rules, framework configurations, and reporting thresholds. This split therefore contributes to framework development.

Compared systems and implementation. On MeViS `valid_u`, we compare our framework with VideoLISA-3.8B (Bai et al., 2024), DsHmp (He & Ding, 2024), Sa2VA-1B/4B (Yuan et al., 2026), Veason-R1-3B (Gong et al., 2026), and Refer-Agent-9B (Jiang et al., 2026). Public checkpoints and predictions are evaluated with the same local evaluator. Inference runs on an NVIDIA RTX PRO 6000D. GPU-hours denote aggregate device time.

Outcome measures. We evaluate every video–language query, retain missing outputs as empty predictions, and treat empty predictions on target-absent frames as valid decisions.

Besides mean and median $\mathcal{J}\&\mathcal{F}$, High-$\mathcal{J}@\tau$ and Low-$\mathcal{J}@\tau$ denote the fractions of queries with video $\mathcal{J}$ at least and at most threshold τ, respectively. We use High-$\mathcal{J}$@70/85/90 for increasingly stringent success and Low-$\mathcal{J}$@30/50 for failure tails. We additionally report output coverage, mean quality on queries with outputs, native target-absent recall, and total GPU time. Native target-absent recall macro-averages, over queries containing GT-empty frames, the fraction of those frames assigned empty predictions. We use each system's original empty-output decisions without calibrating an additional rejection threshold.

Table 1: Outcome profiles on MeViS `valid_u` complement mean $\mathcal{J}\&\mathcal{F}$.

System	Reasoner params	Mean		Median	High-$\mathcal{J}$			Low-$\mathcal{J}$		Absent	Inference
		$\mathcal{J}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	@70 ↑	@85 ↑	@90 ↑	@30 ↓	@50 ↓	Recall ↑	GPU-h ↓
VideoLISA	3.8B	0.1576	0.2346	0.1450	0.0530	0.0164	0.0151	0.7945	0.8600	0.0090	4.59
DsHmp	0.2B	0.4787	0.5410	0.5294	0.2711	0.1614	0.1311	0.3077	0.5624	0.1149	1.65
Sa2VA	1B	0.4889	0.5411	0.5978	0.3859	0.2068	0.1311	0.3594	0.4855	0.1283	1.10
Sa2VA	4B	0.5108	0.5628	0.6615	0.4073	0.2358	0.1185	0.3279	0.4338	0.2864	1.15
Veason-R1	3B	0.5313	0.5740	0.7796	0.5120	**0.4161**	0.3153	0.3821	0.4338	0.3489	2.95
Refer-Agent	9B	0.5850	**0.6289**	**0.7929**	0.5511	0.4073	0.3102	0.2900	0.3556	0.4974	40.66
Ours	4B	**0.5891**	0.6272	0.7876	**0.5523**	0.4111	**0.3228**	**0.2774**	**0.3430**	**0.5696**	12.38
Ours†	30B-A3B	0.6209	0.6669	0.7991	0.5485	0.4224	0.3190	0.2371	0.3153	0.4581	13.37

Bold/underline rank the upper block. Green values in the 30B row improve on 4B. GPU-hours is total GPU time. The 4B total includes estimated recovery. † 32 uniform frames with no support-aware LoRA or recovery.

Table 2: The operating modes expose a trade-off between coverage, conditional quality, and native target-absent behavior.

Mode	Output cov.	$\mathcal{J}\&\mathcal{F}$	Cov. $\mathcal{J}\&\mathcal{F}$	Abs. rec.
Ours quality-first	71.37%	0.5331	0.7009	0.6946
Ours	89.79%	0.6272	0.6748	0.5696

Abs. rec. reports native, uncalibrated target-absent behavior.

5.2 End-to-End Accuracy and Reliability

End-to-end outcome profiles. In Table 1, the deployed 4B Ours reaches mean $\mathcal{J}\&\mathcal{F}$ 0.6272, close to Refer-Agent. The 30B-A3B variant is excluded from the ranking because its configuration differs. Refer-Agent and 4B Ours have High-$\mathcal{J}$@90 rates of 0.3102 and 0.3228, respectively, a difference of approximately ten queries. Our framework with the 4B MLLM also has lower Low-$\mathcal{J}$@30/50 rates than Refer-Agent and the highest native target-absent recall in the table. The 30B-A3B variant further reduces low-quality predictions, with lower target-absent recall than 4B Ours. We have not estimated paired confidence intervals that account for queries from the same video.

Coverage and quality trade-off. In Table 2, the additional pass raises output coverage from 71.37% to 89.79% and all-query mean $\mathcal{J}\&\mathcal{F}$ by 0.0941. Covered mean $\mathcal{J}\&\mathcal{F}$ falls by 0.0261, and native target-absent recall falls by 0.1250. The covered-output curve in Fig. 3(a) should therefore be read with its coverage, while the GT-reference curve provides a diagnostic control. Fig. 3(b) reports native target-absent behavior.

Cross-dataset evaluation. Table 3 reports frozen configurations across datasets. Its MeViS `valid_u` row combines deployed outcomes with the compact diagnostic reference rate. We follow the Long-RVOS definitions of temporal IoU (tIoU) and video IoU (vIoU) (Liang et al., 2026). Static Long-RVOS expressions exceed dynamic expressions by 14.67 points in output coverage, 17.07 points in accurate-reference rate, and 0.1155 in full $\mathcal{J}\&\mathcal{F}$. The covered-to-full $\mathcal{J}\&\mathcal{F}$ gap is 0.0476/0.0400 on MeViS/MeViSv2 and rises to 0.1337 on dynamic Long-RVOS, 0.1614 on ReVOS, and 0.1381 on ReasonVOS. These differences occur across benchmarks with the same Qwen3-VL, SAMTok, and SAM3 components.

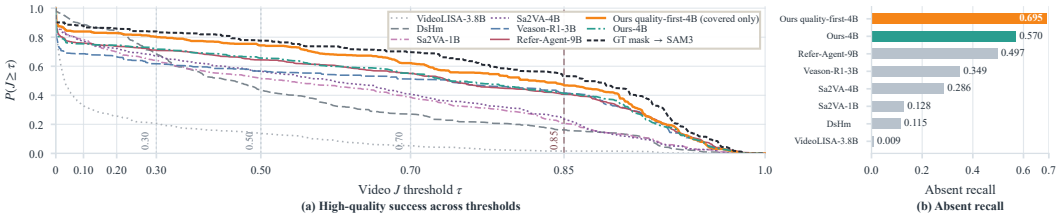


Figure 3: End-to-end outcome profiles on MeViS `valid_u`. In (a), Ours quality-first is evaluated on produced outputs, while the GT-reference curve is a diagnostic control. Panel (b) reports native, uncalibrated target-absent behavior.

Table 3: Output coverage, reference accuracy, and video-mask quality across RVOS benchmarks.

Capability	Benchmark	Split / subset	N	Output cov.	Acc. ref.	Full $\mathcal{J}\&\mathcal{F}$	Cov. $\mathcal{J}\&\mathcal{F}$	tIoU	vIoU	Abs. rec.
Reliability	MeViS	`valid_u`	793	89.79%	63.18%†	0.6272	0.6748	0.8643	0.5696	0.5696
		`valid`	2236	73.17%	–	0.5485	–	–	–	–
Absence	MeViSv2	`valid_u`	907	85.78%	58.88%	0.6179	0.6579	0.8491	0.5602	0.6159
Long horizon	Long-RVOS (test)	static	914	80.20%	67.72%	0.6897	0.7618	0.6953	0.5434	0.7351
		dynamic	847	65.53%	50.65%	0.5742	0.7079	0.5546	0.4113	0.7511
		hybrid	880	68.52%	54.09%	0.6187	0.7325	0.5908	0.4445	0.7651
		overall	2641	71.60%	57.71%	0.6290	0.7366	0.6154	0.4681	0.7502
Dynamic relations	ReVOS	full set	5822	65.84%	45.81%	0.5250	0.6864	0.6585	0.4489	0.7886
Reasoning	ReasonVOS	`valid`	458	75.98%	60.26%	0.5870	0.7251	0.7131	0.5111	0.7384

† Compact diagnostic route. Outcome columns use deployed predictions. Cov. $\mathcal{J}\&\mathcal{F}$ conditions on produced outputs. ReVOS and ReasonVOS use zero-shot evaluation.

Table 4: Compact direct proposals yield stronger reference metrics than evidence-first reasoning under matched components and observations.

Proposal form	Prop. cov.	GT vis.	Local $\mathcal{J}\&\mathcal{F}$	Acc. ref.
Compact proposal	96.22%	96.72%	0.6276	63.18%
Verbose reasoning	87.39%	94.81%	0.6191	56.75%

Acc. ref.: at least one candidate reaches $\mathcal{J} \geq 0.5$, over all queries. Local $\mathcal{J}\&\mathcal{F}$ and GT visibility use the selected reference. GT visibility conditions on proposed queries.

Table 5: Direct multi-frame reference generation does not improve reference masks in the evaluated settings and frequently produces no mask at high visual load.

Reference input	Obs.	Prompt	Mask cov. ↑	Local $\mathcal{J}\&\mathcal{F}$ ↑	Local $\mathcal{J} \geq 0.5$ ↑
Selected image	None	Local description	100.00%	0.6191	64.94%
Multi-image (262k px/image)	25–59	Expression	98.99%	0.5156	51.95%
Multi-image (262k px/image)	25–59	Local description	94.95%	0.5351	54.98%
Multi-image (590k px/image)	25–59	Expression	55.27%	0.3378	34.49%
Multi-image (590k px/image)	25–59	Local description	46.61%	0.3069	30.59%

5.3 Controlled Comparisons of Design Choices

The tables distinguish any-candidate reference accuracy from selected-reference local $\mathcal{J}\&\mathcal{F}$. Mask output coverage denotes the fraction of reference-generation inputs that produce a mask.

Compact proposals versus verbose reasoning. With components and observations fixed, compact frame-level proposals outperform verbose reasoning traces in proposal coverage (96.22% vs. 87.39%), conditional local $\mathcal{J}\&\mathcal{F}$ (0.6276 vs. 0.6191), and accurate-reference rate over all queries (63.18% vs. 56.75%). Table 4 compares the complete proposal policies, whose prompts are given in Appendix B.3.

Reference-formation inputs. We test additional temporal context by fixing the verbose-output proposals and varying only the reference-generator input. Local descriptions improve the moderate multi-image route by 0.0195 $\mathcal{J}\&\mathcal{F}$ (Table 5), yet both multi-image variants remain below single-frame reference formation. At high visual load, mask output coverage falls below 56% and local quality decreases. These results support single-frame reference formation within the tested observation and pixel budgets.

Candidate budget. To assess the value of alternative frame-level proposals, we increase the candidate limit from one to three while fixing the support-aware 4B model, observations, prompt, and SAMTok. The larger budget improves all-query accurate-reference rate and conditional local quality, with little change in proposal coverage (Table 6).

Table 6: A small candidate budget improves reference formation with similar proposal coverage.

Budget	Prop. cov.	Acc. ref.	Local $\mathcal{J}\&\mathcal{F}$
$K = 1$	96.97%	59.27%	0.5968
$K \leq 3$	96.22%	63.18%	0.6276

Acc. ref. uses any qualifying candidate, while local $\mathcal{J}\&\mathcal{F}$ uses the selected reference.

Table 7: Observation count, proposal-model variants, and reference redundancy have different reliability effects.

(a) Observation count

Obs.	Prop. cov.	Acc. ref.	Local $\mathcal{J}\&\mathcal{F}$
32 uniform	84.24%	65.95%	0.6570
64 uniform	86.89%	66.33%	0.6486

Frozen 4B proposal and reference stages.

(b) Proposal-model variants

Model	Out. cov.	Acc. ref.	Median $\mathcal{J}\&\mathcal{F}$	Mean $\mathcal{J}\&\mathcal{F}$	Abs. rec.
4B	89.41%	65.70%	0.7772	0.6100	**0.5887**
8B	94.07%	71.63%	0.7659	0.6114	0.4799
30B-A3B	**98.11%**	**78.69%**	**0.7991**	**0.6669**	0.4581

30B-A3B: 32 uniform observations, no LoRA.

(c) Reference redundancy

Refs.	Mean $\mathcal{J}\&\mathcal{F}$	W/T/L	Abs. rec.
Rank-1	0.6667	–	0.5221
Up to 3	0.7009	286/31/249	0.4743

In (c), both rows are evaluated on the same 566 quality-first outputs. W/T/L denotes wins, ties, and losses relative to Rank-1.

Support-aware proposal control. Under the matched verbose-output control, adaptation raises conditional local $\mathcal{J}\&\mathcal{F}$ from 0.6012 to 0.6191, but proposal coverage falls from 93.95% to 87.39% and the all-query accurate-reference rate falls from 59.39% to 56.75%. The state-migration audit records both recovered failures and lost qualifying candidates (Appendix C.2). The control therefore changes the set of proposed queries as well as their conditional quality.

5.4 Resource Allocation and Computational Cost

Observation count. Moving from 32 to 64 uniform observations recovers 21 additional proposals and raises the all-query accurate-reference rate by only 0.38 points, while conditional local $\mathcal{J}\&\mathcal{F}$ falls by 0.0084, as shown in Table 7(a). This comparison ends before propagation, with limited reference-accuracy gains.

Proposal-model variants. In Table 7(b), with the observation policy, proposal format, candidate budget, and frozen downstream backend fixed, moving from 4B to 8B raises output coverage by 4.66 points and the all-query accurate-reference rate by 5.93 points, but changes mean $\mathcal{J}\&\mathcal{F}$ by only $+0.0014$ (paired-bootstrap 95% CI: $[-0.0191, 0.0219]$). The 30B-A3B variant yields broader gains: relative to 4B, it raises output coverage by 8.70 points, accurate-reference rate by 12.99 points, and mean $\mathcal{J}\&\mathcal{F}$ by 0.0569. Its improvement is concentrated in recovering omissions and severe failures: Low-$\mathcal{J}$@30 falls from 29.63% to 23.71%, whereas High-$\mathcal{J}$@90 changes only from 31.65% to 31.90%. Native target-absent recall nevertheless falls by 13.06 points, while aggregate computation increases from 6.01 to 13.37 GPU-hours. The 30B-A3B variant uses 32 uniform observations and no support-aware LoRA, while the 4B/8B variants use 32 relevance-guided plus 32 uniform observations. Model size, adaptation, and observation policy all differ.

Reference redundancy. On the same 566 quality-first outputs, increasing the number of propagated references from one to three raises mean final $\mathcal{J}\&\mathcal{F}$ from 0.6667 to 0.7009, but yields 286 improvements, 31 ties, and 249 regressions, as shown in Table 7(c). On this paired set, native target-absent recall decreases from 0.5221 to 0.4743. Further details are provided in Appendix C.3.

Computational cost. At comparable mean accuracy (Table 1), the estimated total GPU time for Ours is 12.38 GPU-hours on MeViS `valid_u`, versus 40.66 for Refer-Agent-9B. Compact end-to-end baselines remain faster. In Ours quality-first, target resolution, reference formation, and propagation account for 21.7%, 14.2%, and 64.0% of aggregate device time. The Ours total includes an estimated recovery-pass increment (Appendix D.2).

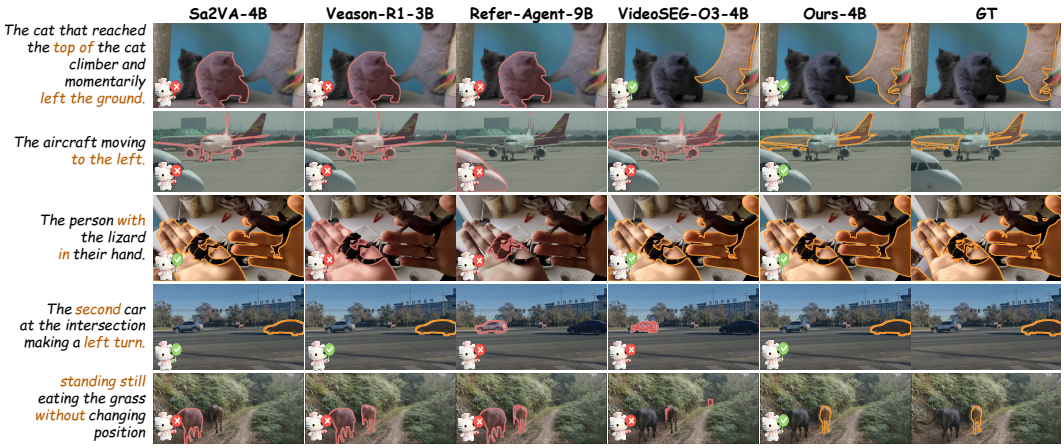


Figure 4: Qualitative comparisons. Ours preserves the referred instance, whereas several baselines switch to a similar distractor.

5.5 Qualitative Comparisons

Fig. 4 shows examples in which several baselines segment a similar distractor while our framework retains the referred instance. These cases illustrate target-identity errors that aggregate mask scores do not localize.

6 Discussion

Frame roles and target preservation. A frame that reveals the action or relation defining the target may still be difficult to segment because of occlusion or a similar distractor. A clearer frame may provide a better reference mask once the video-level reasoning has identified the target. The local description must connect that selected frame to the object specified by the full expression. This distinction explains why target visibility alone does not guarantee an accurate reference.

Model capacity and reliability control. The 30B-A3B base variant produces more outputs, improves mean mask quality, and reduces severe failures (Tables 1 and 7(b)). Our framework with the 4B MLLM has higher target-absent recall, a slightly higher High-$\mathcal{J}$@90 rate, and lower estimated GPU time. Its support-aware control and recovery also provide a trade-off between coverage and quality (Table 2). These results motivate combining stronger reference formation with policies that decide when to abstain and retry. This combination remains untested at 30B-A3B. Differences in sampling and adaptation also prevent attributing the current gains solely to model size.

7 Conclusion

We trace RVOS errors from frame-level proposals through reference masks to final video predictions. Reference formation accounts for the largest measured loss, and GT initialization shows that propagation can still fail with an accurate reference. These findings motivate separate reasoning and segmentation modules with support-aware control and conditional recovery. Evaluating mask quality alongside output coverage and target-absent recall makes the benefits and costs of these choices explicit.

Limitations. Our analysis uses one combination of models, so the observed bottleneck may depend on the chosen components. Mask-overlap metrics measure segmentation quality but do not isolate errors in target interpretation, frame selection, and mask generation. The GT-reference intervention is restricted to 501 cases with accurate predicted references, and the multi-frame study covers 25–59 observations. MeViS `valid_u` also informs framework development, limiting its independence as an evaluation set.

AI USE STATEMENT

AI tools assisted language editing and typesetting checks.

REPRODUCIBILITY STATEMENT

The main paper reports the controlled interfaces, evaluation measures, and framework configurations. The appendix documents frozen components, training and decoding settings, prompt policies, evaluation details, robustness checks, and computational-cost accounting. The appendix reproduces the exact proposal prompts and documents the implementation and evaluation settings.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, et al. 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. doi: 10.52202/079017-0219. NeurIPS 2024, Vancouver, BC, Canada, December 9–15, 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 (CVPR)*, pp. 4985–4995, June 2022. CVPR 2022, New Orleans, LA, USA, June 18–24, 2022.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, et al. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. ICLR 2026, Rio de Janeiro, Brazil, April 23–27, 2026.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In *Computer Vision, ECCV 2022*, volume 13688 of *Lecture Notes in Computer Science*, pp. 640–658. Springer, 2022. doi: 10.1007/978-3-031-19815-1_37. 17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022.

Suhwan Cho, Seunghoon Lee, Minhyeok Lee, Jungho Lee, and Sangyoun Lee. Find first, track next: Decoupling identification and propagation in referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3873–3883, October 2025. ICCV Workshops 2025, Honolulu, HI, USA, October 19–23, 2025.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAMWISE: Infusing wisdom in SAM2 for text-driven video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3395–3405, June 2025. doi: 10.1109/CVPR52734.2025.00322. CVPR 2025, Nashville, TN, USA, June 11–15, 2025.

Ming Dai, Sen Yang, Boqiang Duan, Boyuan Tong, Jiedong Zhuang, et al. VideoSEG-O3: A multi-turn reinforcement learning framework for reasoning video object segmentation. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026. ICML 2026, Seoul, South Korea, July 7–9, 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 (ICCV)*, pp. 2694–2703, October 2023. ICCV 2023, Paris, France, October 2–6, 2023.

Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, et al. MeViS: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11400–11416, December 2025. doi: 10.1109/TPAMI.2025.3600507.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, et al. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29183–29192, June 2025. CVPR 2025, Nashville, TN, USA, June 11–15, 2025.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Jiazuo Yu, Pingping Zhang, et al. Reinforcing video object segmentation to think before it segments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3835–3844, June 2026. CVPR 2026, Denver, CO, USA, June 3–7, 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 (CVPR)*, pp. 13332–13341, June 2024. CVPR 2024, Seattle, WA, USA, June 17–21, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. ICLR 2022, virtual conference, April 25–29, 2022.

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 (CVPR)*, pp. 39508–39517, June 2026. CVPR 2026, Denver, CO, USA, June 3–7, 2026.

Woojeong Jin, Jaeho Lee, Heeseong Shin, Seungho Jang, Junhwan Heo, et al. AgentRVOS: Reasoning over object tracks for zero-shot referring video object segmentation, 2026. arXiv preprint arXiv:2603.23489.

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 (ICLR)*, 2026. ICLR 2026, Rio de Janeiro, Brazil, April 23–27, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, October 2023. doi: 10.1109/ICCV51070.2023.00371. ICCV 2023, Paris, France, October 2–6, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, et al. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, June 2024. doi: 10.1109/CVPR52733.2024.00915. CVPR 2024, Seattle, WA, USA, June 17–21, 2024.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, et al. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. arXiv preprint arXiv:2601.04720.

Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, et al. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22236–22245, October 2023. ICCV 2023, Paris, France, October 2–6, 2023.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, et al. ReferDINO: Referring video object segmentation with visual grounding foundations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20009–20019, October 2025. ICCV 2025, Honolulu, HI, USA, October 19–23, 2025.

Tianming Liang, Haichao Jiang, Yuting Yang, Chaolei Tan, Shuai Li, et al. Long-RVOS: A comprehensive benchmark for long-term referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39497–39507, June 2026. CVPR 2026, Denver, CO, USA, June 3–7, 2026.

Ci-Siang Lin, Min-Hung Chen, I-Jieh Liu, Chien-Yi Wang, Sifei Liu, et al. Temporal prompting matters: Rethinking referring video object segmentation, 2025a. arXiv preprint arXiv:2510.07319.

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 (CVPR)*, pp. 8658–8667, June 2025b. CVPR 2025, Nashville, TN, USA, June 11–15, 2025.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, et al. VideoGLaMM: A large multimodal model for pixel-level visual grounding in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19036–19046, June 2025. CVPR 2025, Nashville, TN, USA, June 11–15, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, et al. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. ICLR 2025, Singapore, April 24–27, 2025.

Fu Rong, Meng Lan, Qian Zhang, and Lefei Zhang. MPG-SAM 2: Adapting SAM 2 with mask priors and global context for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23979–23989, October 2025. ICCV 2025, Honolulu, HI, USA, October 19–23, 2025.

Tuyen Tran, Thao Minh Le, and Truyen Tran. Towards agentic AI for multimodal-guided video object segmentation, 2025. arXiv preprint arXiv:2508.10572. Accepted as a short paper at the ICCV 2025 Instance-Level Recognition and Generation Workshop. Non-archival.

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 (CVPR)*, pp. 4974–4984, June 2022. CVPR 2022, New Orleans, LA, USA, June 18–24, 2022.

Huihui Xu, Jiashi Lin, Haoyu Chen, Junjun He, and Lei Zhu. EventRR: Event referential reasoning for referring video object segmentation, 2025. arXiv preprint arXiv:2508.07171.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, et al. VISA: Reasoning video object segmentation via large language models. In *Computer Vision, ECCV 2024*, volume 15073 of *Lecture Notes in Computer Science*, pp. 98–115. Springer, 2024. doi: 10.1007/978-3-031-72633-0_6. 18th European Conference on Computer Vision, Milan, Italy, September 29–October 4, 2024.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, et al. 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.

Yikang Zhou, Tao Zhang, Dengxian Gong, Yuanzheng Wu, Ye Tian, et al. SAMTok: Representing any mask with two words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37852–37863, June 2026. CVPR 2026, Denver, CO, USA, June 3–7, 2026.

A QUALITATIVE EXAMPLES

Figs. 5–7 show selected MeViS comparisons. Fig. 8 illustrates successful and failed reference-to-video outcomes, including identity, scope, occlusion, and temporal-coverage errors. These examples show the mask errors summarized by the main-paper metrics.

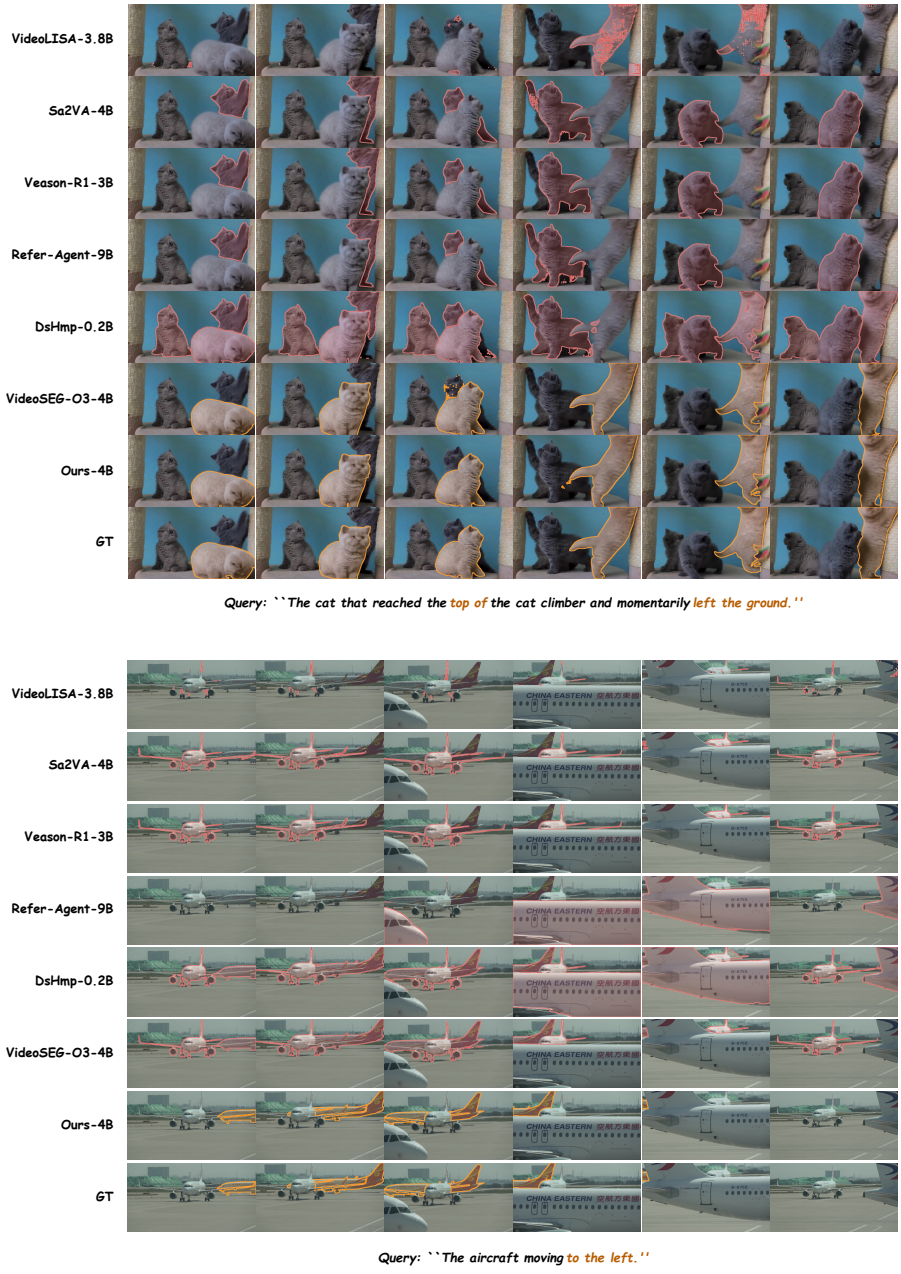


Figure 5: Additional comparisons on MeViS `valid_u`. Each row uses the same six frames. The selected sequences illustrate target identity preservation.



Figure 6: Additional comparisons on MeViS `valid_u`. Each row uses the same six frames. The selected sequences illustrate target identity preservation.



Figure 7: Additional comparisons on MeViS `valid_u`. Each row uses the same six frames. The selected sequences illustrate target identity preservation.

B Implementation and Evaluation Details

B.1 Components, Inference, and Evaluation

Our framework uses Qwen3-VL-4B-Instruct, frozen SAMTok-4B, and frozen SAM3. The 8B model is used in the proposal-model comparison. The 30B-A3B base variant also appears in Table 1. It uses 32 uniform observations, no support-aware LoRA, and no recovery pass. Candidate rank is preserved. SAMTok processes each selected frame independently. Duplicate frame indices are removed before the reference masks initialize SAM3 objects. No GT overlap is used for selection or routing. The quality-first and conditional-recovery policies follow Section 4. The recovery pass preserves accepted quality-first outputs and keeps masks from the two passes separate.

Relevance-guided observations use expression, clip, and overview embeddings e, c_i, and v:

$$r_i = e^\top (c_i - v), \qquad \tilde{r}_i = \begin{cases} r_i, & r_i \geq 0, \\ -\infty, & r_i < 0. \end{cases} \tag{4}$$

A softmax over finite scores defines the sampling distribution. The recovery pass combines 32 relevance draws with 32 uniform observations, removes duplicate indices, and restores temporal order.

The MeViS `valid_u` comparison uses the same 793 expressions and the same evaluator. Predictions are obtained from the official VideoLISA, Sa2VA, and Veason-R1 checkpoints, the released DsHmp checkpoint, and complete Refer-Agent predictions. Conversion scripts change only the prediction-file format. The predicted masks remain unchanged. All-query evaluation retains missing outputs as empty masks, including valid empty decisions on target-absent frames. Query-level output coverage and native target-absent recall follow Section 5.1: quality-first emits for 566 expressions and abstains on 227, while Ours emits for 712 and abstains on 81. Native target-absent recall averages over queries containing GT-empty frames.

The reported bootstrap intervals resample expressions independently and do not account for dependence between expressions from the same video. We also use MeViS `valid_u` to develop prompts, routing rules, framework configurations, and reporting thresholds, as stated in the main paper. Adapter training uses only the two training splits.



Figure 8: Audited reference-mask-to-video outcomes. Each row shows early, middle, and late frames with GT-only regions in green, prediction-only regions in red, and overlap in yellow. Cases show ambiguous target extent or annotations, recovery from an inaccurate reference, occlusion, switches to another instance, missing masks over time, and errors in the number of target instances. Seed IoU denotes reference-mask IoU. The labels summarize manual inspection of the predicted masks.

B.2 Support-Aware Adaptation

The 20,000 adaptation examples comprise 8,930 from MeViS and 11,070 from Long-RVOS. The 2:1:1 mixture contains 10,000 target-supported cases, 5,000 observation-missing cases, and 5,000 target-absent cases. In observation-missing cases, the target is present in the full video but absent from the sampled observations. Each example uses up to 32 frames. Short videos may supply fewer distinct frames.

We apply rank-32 LoRA (Hu et al., 2022) with alpha 64, dropout 0.05, one epoch, learning rate 7×10^{-6}, and effective batch size 64. The vision tower, SAMTok, and SAM3 remain frozen. Supervision concerns proposing or abstaining and does not directly optimize mask overlap or final video $\mathcal{J}\&\mathcal{F}$. No validation or test example or mask is used for optimization.

Table 8: Controlled target-proposal policies. Candidate limits specify the maximum number of supported outputs.

Policy	Required decision
Direct	Return up to three easy-to-segment candidate frames and local descriptions. Return none if no plausible target is visible.
Presence-checked	Decide whether observations visibly support the referred target before returning up to three candidates. Abstain when category resemblance is the only support.
Evidence-first diagnostic	Resolve target scope and distinguishing temporal relations, summarize supporting observations, and cover distinct visibility periods before returning candidates.

Table 9: The decrease at the reference checkpoint remains largest across matched diagnostic thresholds. Final counts condition on passing the local-IoU threshold in that row.

Local IoU	**Qualified** N	**Ref. loss**	**Final** ≥ 0.3	**Final** ≥ 0.4	**Final** ≥ 0.5	**Final** ≥ 0.6	**Final** ≥ 0.7
0.3	536	227	529	520	502	450	419
0.4	525	238	520	511	498	450	419
0.5	501	262	497	489	481	444	417
0.6	472	291	471	463	457	428	406
0.7	424	339	423	420	415	386	368

B.3 Frozen Proposal Policies

Table 8 summarizes the three policies. Appendix E reproduces their complete text. Each receives the expression, candidate limit, and a mapping from sampled observations to original frame indices. Runtime values fill these fields without changing the template.

Direct produces compact proposals. Presence-checked implements support-aware proposal control. Evidence-first produces the verbose diagnostic output. The compact-versus-verbose comparison in Table 4 fixes components, observations, and candidate budget. Its policies differ in both output structure and instructions, including evidence-sufficiency reasoning and episode-aware allocation. The comparison therefore evaluates the complete proposal policies.

C Supplementary Diagnostic Results

C.1 Threshold Sensitivity

Table 9 supports the checkpoint analysis in Fig. 2(a). It uses the same compact route and selected reference. Across matched local-IoU and final-$\mathcal{J}\&\mathcal{F}$ thresholds from 0.3 to 0.7, more queries fail to form an accurate reference than fail during propagation after forming one. Final counts in each row condition on passing that row's local threshold.

C.2 Support-Aware State Migration

Table 10 supplements the support-aware control comparison in Section 5.3. Qualified means that any decoded candidate reaches local IoU 0.5. Failed means that candidates are produced but none qualifies. Abstained means that no candidate is produced. Over all 793 expressions, adaptation recovers 49 base-model failures while moving 43 qualified cases to failure and 27 to abstention. The higher mean quality on proposed queries thus reflects both corrected predictions and changes in which queries receive proposals.

Table 10: Base-to-support-aware state migration over all 793 expressions. Rows are Base states and columns are adapted states.

	Qualified	**Failed**	**Abstained**
Qualified	401	43	27
Failed	49	195	30
Abstained	0	5	43

Table 11: Higher reference-mask IoU is associated with higher final video quality in the 763-expression compact route, although some propagation failures remain.

Selected local IoU	N	**Mean final $\mathcal{J}\&\mathcal{F}$**	**Failure rate**
$[0, 0.3)$	227	0.0865	100.00%
$[0.3, 0.5)$	35	0.5188	40.00%
$[0.5, 0.7)$	77	0.6792	14.29%
$[0.7, 0.9)$	236	0.8386	1.69%
$[0.9, 1]$	188	0.9115	2.66%

C.3 Reference Quality and Paired Redundancy Results

Table 11 gives the selected-reference quality bins behind Fig. 2(b). The failure rate is the fraction with final video $\mathcal{J}\&\mathcal{F}$ below 0.5 within each bin. Higher reference IoU is associated with higher final video $\mathcal{J}\&\mathcal{F}$, although some accurate references still lead to failed video predictions.

The redundancy comparison uses the same 566 quality-first outputs. Increasing propagated references from one to three changes mean final $\mathcal{J}\&\mathcal{F}$ from 0.6667 to 0.7009, with 286 wins, 31 ties, and 249 losses. On this paired covered set, native target-absent recall changes from 0.5221 to 0.4743. The 0.6946 value in Table 2 includes queries for which the controller abstains in the all-query evaluation and has a different denominator. The paired comparison measures the effect of adding references on the same queries. The all-query score also includes empty predictions caused by abstention.

D Supplementary Resource Results

D.1 Proposal-Model Profiles and Paired Changes

Table 12 expands Table 7(b) with outcome and runtime profiles. The 30B-A3B variant uses 32 uniform observations and the base model without support-aware LoRA, while the 4B/8B variants use 32 relevance-guided plus 32 uniform observations. Table 13 gives the paired 4B-to-8B analysis with the same 32 relevance-guided plus 32 uniform observations, top-3 candidate budget, frozen SAMTok and SAM3, up to three propagated references, and no additional proposal pass. The comparison changes both model size and its trained adapter. Small changes in the overall mean combine improvements on some queries with regressions on others.

D.2 Computational Cost Accounting

Table 14 sums worker runtimes for all 793 MeViS `valid_u` expressions on NVIDIA RTX PRO 6000D GPUs under Ubuntu Linux, excluding CPU evaluation and profiling. GPU-hours measure aggregate device time. DsHmp uses author-reported timing, while other components use measured logs.

The 6.01 GPU-hour 4B control in Table 12(b) uses one pass with 32 relevance-guided plus 32 uniform observations. Full 4B Ours starts with 32 uniform observations: its estimated cost combines the measured 10.67 GPU-hour quality-first pass with recovery time scaled to 227 of 793 queries. The recovery increment is approximately 1.72 GPU-hours (16%), yielding the 12.38 total in Table 1. The complete two-pass configuration was not timed directly. Stage shares use the quality-first total. Refer-Agent uses 39.61 proposal and 1.06 SAM2 GPU-hours. Totals use unrounded measurements, so rounded entries may not sum exactly.

Table 12: Outcome and runtime profiles for the single-pass proposal-model variants in Table 7(b). The 30B-A3B variant uses 32 uniform observations and no support-aware LoRA, while 4B/8B use 32 relevance-guided plus 32 uniform observations.

(a) Reliability profile

Model	Interfaces		High-$\mathcal{J}$			Low-$\mathcal{J}$		$\mathcal{J}\&\mathcal{F}$		Absent
	Output cov. ↑	Acc. ref. ↑	@70 ↑	@85 ↑	@90 ↑	@30 ↓	@50 ↓	Median ↑	Mean ↑	Recall ↑
4B	89.41%	65.70%	53.59%	38.71%	31.65%	29.63%	36.95%	0.7772	0.6100	**0.5887**
8B	94.07%	71.63%	50.82%	37.70%	28.88%	30.26%	36.95%	0.7659	0.6114	0.4799
30B-A3B	**98.11%**	**78.69%**	**54.85%**	**42.24%**	**31.90%**	**23.71%**	**31.53%**	**0.7991**	**0.6669**	0.4581

(b) Single-pass runtime breakdown

Model	Four-GPU wall time (h)	Proposal GPU-h	SAMTok GPU-h	SAM3 GPU-h	Total GPU-h
4B	**1.54**	**1.94**	**1.23**	2.83	**6.01**
8B	1.94	2.82	2.16	**2.54**	7.52
30B-A3B	3.41	8.09	2.28	2.99	13.37

The 4B row uses one pass. Full 4B Ours adds conditional recovery (Table 1). See Table 14 for its cost breakdown.

Table 13: Changing from 4B to 8B improves queries with weak references and degrades queries with strong references, leaving the overall mean nearly unchanged. SAMTok and SAM3 remain frozen. Intervals are paired bootstrap 95% confidence intervals.

4B source condition	N	Δ final $\mathcal{J}\&\mathcal{F}$	95% CI
No candidate	84	$+0.1197$	$[+0.0252, +0.2117]$
Best local IoU < 0.5	188	$+0.1178$	$[+0.0814, +0.1547]$
Best local IoU in $[0.5, 0.7)$	54	-0.0358	$[-0.1025, +0.0239]$
Best local IoU ≥ 0.7	467	-0.0624	$[-0.0866, -0.0393]$
All expressions	793	$+0.0014$	–

Table 14: Aggregate GPU-hours on the complete MeViS `valid_u` split. System totals and quality-first stage costs use aggregate device time. Ours includes an estimated recovery increment. The stage shares apply to quality-first.

System or stage	GPU-hours	Share
Complete systems		
Sa2VA-1B	1.10	–
Sa2VA-4B	1.15	–
DsHmp	1.65	–
Veason-R1-3B	2.95	–
VideoLISA-3.8B	4.59	–
Ours quality-first	10.67	–
Ours	12.38	–
Refer-Agent-9B	40.66	–
Ours quality-first stage breakdown		
Target resolution	2.32	21.7%
Reference formation	1.52	14.2%
Video propagation	6.83	64.0%

E EXACT PROPOSAL PROMPTS

The following blocks reproduce the complete prompts for the policies in Appendix B.3. At inference, the expression, candidate limit, and sampled-frame index table replace {expression}, {top_k}, and {frame_index_table}, respectively. All other prompt text remains fixed. Field names, including the seed terminology in the evidence-first prompt, are retained exactly as used in the experiments.

E.1 DIRECT TARGET PROPOSAL

```
You are a candidate-frame proposer for referring video object segmentation.

Expression:
{expression}

Maximum candidate count:
{top_k}

Sample/frame index table:
{frame_index_table}

Use only the supplied frames. Find the object described by the expression and select the
↪  easiest frame or frames for single-image segmentation. Do not perform an explicit
↪  evidence-sufficiency analysis. Do not invent frame indices. Return no more than {top_k}
↪  candidates. If no supplied frame contains a plausible target, return an empty candidate
↪  list. Each candidate prompt must describe only what is visible in that frame and must
↪  distinguish nearby similar objects.

Return valid JSON only:
{
  "schema_version": "paper_direct_target",
  "target_visible_in_sampled_frames": true,
  "sampled_evidence_confidence": "high | medium | low",
  "target": {
    "head": "",
    "category": "",
    "cardinality": "single | multiple | group | uncertain",
    "scope": "whole_object | object_part | group | uncertain",
    "attributes": [],
    "expected_mask_target": "",
    "context_objects_to_avoid": [],
    "target_scope_note": ""
  },
  "target_identity_summary": "",
  "candidate_frames": [
    {
      "rank": 1,
      "sample_index": 0,
      "frame_index": 0,
      "single_frame_segmentation_prompt": "Segment the target. Do not segment nearby
      ↪  distractors."
    }
  ],
  "failure_reason": null
}

When candidate_frames is empty, set target_visible_in_sampled_frames=false and provide a
↪  non-empty failure_reason. When it is non-empty, set target_visible_in_sampled_frames=true.
↪  The maximum candidate count is a hard limit.
```

E.2 PRESENCE-CHECKED PROPOSAL

```
You are a target-presence verifier and candidate-frame proposer for referring video object segmentation.

Expression:
{expression}

Maximum candidate count:
{top_k}

Sample/frame index table:
{frame_index_table}

Use only the supplied frames. First decide whether the referred target is visibly supported by the observations. Category resemblance alone is not sufficient when the expression identifies one instance among similar objects. If the target is not supported, abstain instead of guessing. If it is supported, select up to {top_k} distinct, easy-to-segment candidate frames. Do not invent frame indices or copy the same frame under different ranks. Each candidate prompt must be image-local and must state what to segment and what important nearby objects to exclude.

Return valid JSON only:
{
  "schema_version": "paper_presence_checked",
  "target_visible_in_sampled_frames": true,
  "sampled_evidence_confidence": "high | medium | low",
  "target": {
    "head": "",
    "category": "",
    "cardinality": "single | multiple | group | uncertain",
    "scope": "whole_object | object_part | group | uncertain",
    "attributes": [],
    "expected_mask_target": "",
    "context_objects_to_avoid": [],
    "target_scope_note": ""
  },
  "target_identity_summary": "",
  "candidate_frames": [
    {
      "rank": 1,
      "sample_index": 0,
      "frame_index": 0,
      "single_frame_segmentation_prompt": "Segment the target. Do not segment nearby distractors."
    }
  ],
  "failure_reason": null
}

When evidence is insufficient, set target_visible_in_sampled_frames=false, return candidate_frames=[], and provide a non-empty failure_reason. The maximum candidate count is a hard limit, not a requirement to fabricate candidates.
```

E.3 EVIDENCE-FIRST PROPOSAL

```
You are an evidence-first target interpreter and temporal seed-candidate proposer for referring video object segmentation.

Expression:
{expression}

Maximum candidate count:
{top_k}

Sample/frame index table:
{frame_index_table}

Use only the supplied frames. Apply the following procedure before emitting candidates:
1. Resolve the grammatical target, cardinality, whole/part/group scope, and important distractors.
2. Identify the motion, interaction, relation, disappearance/reappearance, or other evidence that distinguishes the expression from a category-only match.
3. Decide whether the observations sufficiently support the referred target. If not, abstain instead of guessing.
4. Partition the observed target appearances into distinct visibility episodes. A disappearance followed by reappearance or a shot discontinuity starts a new episode.
5. Emit up to {top_k} distinct candidates. Cover different supported visibility episodes before adding another frame from an already covered episode. Never fabricate a candidate merely to reach the maximum.
6. Each candidate prompt must be image-local: describe only visible attributes and relations in that frame, and explicitly exclude important nearby distractors.

Return valid JSON only:
{
  "schema_version": "paper_evidence_first_episode",
  "target_visible_in_sampled_frames": true,
  "sampled_evidence_confidence": "high | medium | low",
  "target": {
    "head": "",
    "category": "",
    "cardinality": "single | multiple | group | uncertain",
    "scope": "whole_object | object_part | group | uncertain",
    "attributes": [],
    "expected_mask_target": "",
    "context_objects_to_avoid": [],
    "target_scope_note": ""
  },
  "target_identity_summary": "",
  "evidence_summary": "",
  "visibility_episodes": [
    {
      "episode_id": "episode_1",
      "sampled_start_index": 0,
      "sampled_end_index": 0,
      "identity_evidence": ""
    }
  ],
  "candidate_frames": [
    {
      "episode_id": "episode_1",
      "rank": 1,
      "sample_index": 0,
      "frame_index": 0,
      "single_frame_segmentation_prompt": "Segment the target. Do not segment nearby distractors."
    }
  ],
  "failure_reason": null
}

When evidence is insufficient, set target_visible_in_sampled_frames=false, return visibility_episodes=[] and candidate_frames=[], and provide a non-empty failure_reason. The maximum candidate count is a hard limit.
```