DYNAMIC REASONING ROUTING FOR EFFICIENT ONLINE VIDEO OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online reasoning video object segmentation aims to segment targets in an incoming video stream based on complex or implicit language queries. Maintaining accurate target grounding requires timely adaptation to evolving visual evidence. However, existing invocation schedules incur substantial reasoning costs without consistent accuracy gains. To address this, we propose *Dynamic Reasoning Routing (DRR)*, a training-free framework that adaptively schedules semantic reassessment using motion-compensated residuals. By monitoring changes in the target and its context at low cost, DRR reduces redundant reasoning while remaining responsive to evolving visual evidence. To limit error accumulation during propagation, DRR further introduces confidence-weighted mask encoding, which softens masks from low-confidence reassessments. Extensive experiments on five reasoning and referring VOS benchmarks show an improved accuracy–efficiency trade-off. On ReasonVOS, DRR improves accuracy at 6.2–$7.0\times$ the throughput of matched periodic baselines, and surpasses the prior art by 13.0 $\mathcal{J}\&\mathcal{F}$ with a $1.4\times$ speedup. A paired-replay analysis further demonstrates that DRR provides greater average benefit for subsequent segmentation than periodic scheduling. Code will be publicly available.

1 INTRODUCTION

Online reasoning video object segmentation aims to segment targets specified by complex or implicit language queries in each incoming frame of a video stream (Liu et al., 2026). This task is essential for streaming applications such as robotics and embodied assistance (Jiang et al., 2024; Ahn et al., 2022; Khani et al., 2020), where systems must interpret user instructions and act on relevant objects with low latency. Resolving such instructions requires reasoning about spatial relations, temporal events, and world knowledge. In Figure 1 (a), the query "who is the parent in the scene" is resolved only when an adult enters and steadies the child, revealing the relevant interaction.

Most prior reasoning VOS methods have focused on the offline setting, using an MLLM to interpret the query on selected frames and a propagation module to track the resulting masks (Yan et al., 2024; Gong et al., 2025; Wang et al., 2025; Zheng et al., 2025). These methods, however, rely on access to the full video for keyframe selection and global reasoning. Such access is unavailable in online processing, where frames are observed sequentially and new evidence may emerge, disappear, or contradict earlier grounding. Mask propagation offers efficient tracking of a target yet remains agnostic to semantic changes, whereas MLLM-based reassessment can update the target identity at the expense of substantial latency. A key challenge is therefore the scheduling of reassessment to maintain accurate grounding with low latency.

To investigate this trade-off, we examine how existing online methods schedule reassessment and identify a scheduling inefficiency. Specifically, frequent MLLM invocation can incur substantial latency without consistent gains in segmentation accuracy. As shown in Figure 1 (b), StreamingR-VOS (Zhang et al., 2026) triggers semantic updates based on mask quality and temporal consistency, invoking the MLLM on 61.1–66.6% of ReasonVOS frames and 69.7–75.8% in the longer, more motion-intensive egocentric settings (Liang et al., 2025). This computation can concentrate on persistent failures, with the 1B model invoking reasoning on more than 95% of frames for 52% of expressions yet achieving only 20.1 $\mathcal{J}\&\mathcal{F}$ on those expressions, as illustrated in Figure 1 (c). The training-free CoT-RVS pipeline (Kao et al., 2026) instead schedules semantic reassessment at fixed

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.

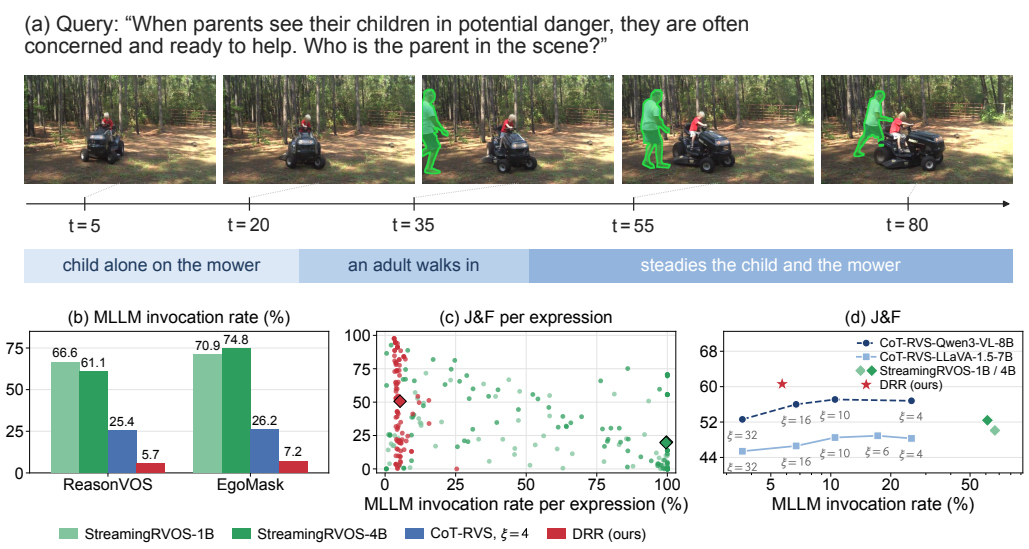

Figure 1: **Online reasoning segmentation and semantic invocation.** **(a)** Temporal interactions reveal the queried target. **(b)** MLLM invocation rates on ReasonVOS and EgoMask. **(c)** Per-expression $\mathcal{J}\&\mathcal{F}$ versus invocation rate on the 100 longest ReasonVOS expressions; diamonds mark means for DRR and the $> 95\%$-invocation subset of StreamingRVOS-1B. **(d)** ReasonVOS results for CoT-RVS at fixed strides ξ, StreamingRVOS, and DRR.

intervals. However, increasing reassessment frequency raises processing latency without monotonically improving segmentation accuracy. In Figure 1 (d), reducing the interval from six to four frames lowers LLaVA-based $\mathcal{J}\&\mathcal{F}$ from 48.9 to 48.3. Increasing the interval reduces invocation frequency but may delay semantic updates when new observations invalidate earlier grounding.

To address these issues, we propose *Dynamic Reasoning Routing (DRR)*, a training-free framework for efficient and responsive semantic reassessment. Inspired by predictive coding in video codecs (Wiegand et al., 2003; Sullivan et al., 2012), DRR uses motion-compensated residuals relative to a grounded reference frame for low-cost, frame-wise monitoring of visual changes. Specifically, foreground residuals track target-region changes, and local target-context contrast captures relative contextual changes. These low-cost cues proxy changes in grounding evidence, so conditioning invocation on them reduces redundant reasoning during stable periods and enables timely reassessment when new evidence may invalidate earlier grounding. Additionally, to mitigate error accumulation from noisy intermediate frames, DRR further retains earlier grounding references and introduces confidence-weighted mask encoding, which softens masks from low-confidence reassessments to preserve temporal consistency during memory updates.

Experiments on five reasoning and referring VOS benchmarks demonstrate improved accuracy–efficiency trade-offs under online inference. On ReasonVOS, DRR improves segmentation accuracy at 6.2–7.0$\times$ the throughput of matched periodic baselines. DRR-Ovis2.5-9B achieves 65.4 $\mathcal{J}\&\mathcal{F}$ at 6.1 FPS, exceeding StreamingRVOS-4B by 13.0 in $\mathcal{J}\&\mathcal{F}$ with 1.4$\times$ the throughput, while invoking reassessment on only 5.5% of frames. DRR-Seg-Zero-7B achieves the highest throughput among the compared methods on ReasonVOS at 6.8 FPS, with 60.6 $\mathcal{J}\&\mathcal{F}$ and a 5.7% invocation rate. Paired-replay analysis further shows that DRR schedules reassessment with greater average benefit to subsequent segmentation than periodic scheduling, while confidence-weighted mask encoding reduces the downstream degradation caused by unreliable reassessed masks. Our contributions are summarized as follows:

- We identify scheduling inefficiencies in existing online reasoning VOS methods and introduce *Dynamic Reasoning Routing (DRR)*, a training-free framework for efficient and responsive semantic reassessment in online reasoning VOS.
- Our residual-based routing mechanism combines foreground residuals within the tracked target with local target-context contrast to guide semantic reassessment. We further introduce confidence-weighted mask encoding to limit error accumulation during propagation.

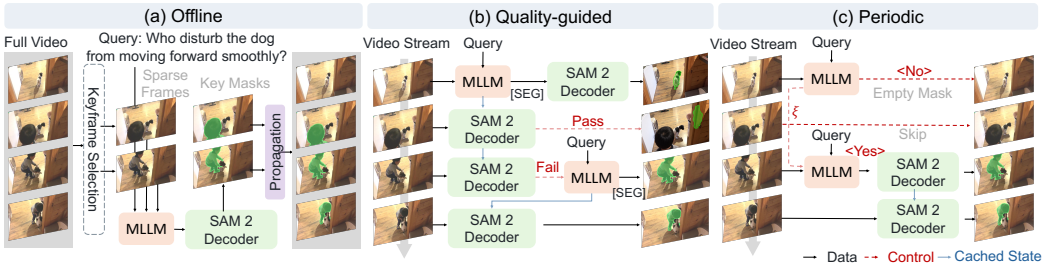

Figure 2: **Semantic invocation strategies.** **(a)** Offline methods select grounding frames with full-video access. **(b)** Quality-guided methods trigger reassessment based on mask quality and temporal consistency. **(c)** Periodic methods evaluate the selector every ξ frames, with affirmative verdicts triggering grounding. Online methods propagate masks between reassessments.

- *DRR* improves accuracy–efficiency trade-offs across five benchmarks. On ReasonVOS, it improves accuracy at 6.2–7.0$\times$ the throughput of matched periodic baselines. Paired replays show greater average segmentation benefits from DRR-selected reassessments than from periodic reassessments.

2 Related Work

Reasoning video object segmentation. Given a video and a language query, referring video object segmentation (RVOS) predicts masks for an explicitly described object (Seo et al., 2020; Botach et al., 2022; Wu et al., 2022), whereas reasoning VOS infers the target from implicit instructions using world knowledge and video context (Lai et al., 2024; Bai et al., 2024; Yan et al., 2024; Kao et al., 2026). Building on image reasoning segmentation (Lai et al., 2024), a common architectural paradigm connects an MLLM with a mask decoder, using the hidden representation of a special segmentation token to condition pixel-level mask prediction on visual and language inputs (Yan et al., 2024; Yuan et al., 2025). Temporal modeling is further improved through dynamic token aggregation and keyframe selection in VRS-HQ (Gong et al., 2025), adaptive context selection and spatiotemporal token interaction in ViLLa (Zheng et al., 2025), and segmentation-aware feature fusion with dynamic anchor updates in VIRST (Hong & Do, 2026).

Online reasoning VOS requires a target mask for each incoming frame using only current and past observations (Liu et al., 2026). The ORVOS baseline (Liu et al., 2026) combines continually updated segmentation prompts with a temporal token reservoir, invoking the MLLM at each time step to incorporate newly observed evidence. StreamingRVOS (Zhang et al., 2026) reuses semantic embeddings across frames and adaptively invokes the MLLM based on mask quality and temporal consistency. The online variant of the training-free CoT-RVS pipeline (Kao et al., 2026) instead evaluates a keyframe selector at fixed intervals; accepted frames are passed to a reasoning image segmenter to obtain new grounding masks. However, these policies rely on mask-level feedback or fixed intervals rather than directly monitoring changes in visual evidence. We address this inefficiency through DRR, a complementary scheduling method that adaptively schedules semantic reasoning based on motion-compensated residuals between visual inputs.

Video codecs. Video codecs such as H.264/Advanced Video Coding (AVC) (Wiegand et al., 2003) and High Efficiency Video Coding (HEVC) (Sullivan et al., 2012) reduce redundancy through spatial prediction, motion-compensated temporal prediction, and residual coding. Motion vectors and prediction residuals also provide useful representations for video analysis. CoViAR (Wu et al., 2018) combines reference-frame appearance, motion vectors, and residuals for action recognition, while CoVOS (Xu & Yao, 2022) uses motion vectors for mask propagation and residuals to identify regions requiring correction. Recent multimodal models use codec-derived cues to allocate visual computation. OneVision-Encoder (Tang et al., 2026) introduces codec-aligned patch sparsity, and LLaVA-OneVision-2 (An et al., 2026) combines adaptive temporal grouping with motion–residual-guided spatial selection. DRR applies the principle of motion-compensated prediction to target-specific semantic invocation in an online stream.

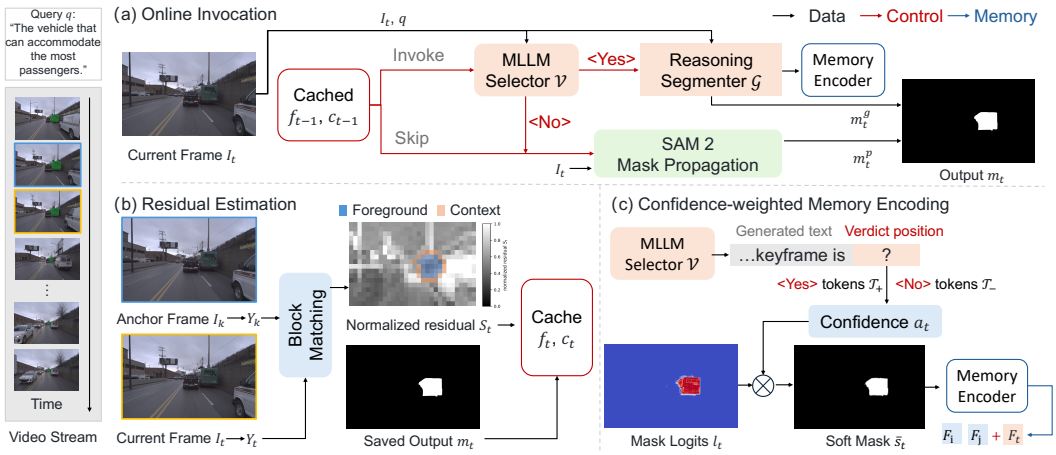

Figure 3: Overview of Dynamic Reasoning Routing (DRR). **(a)** Cached residuals guide semantic invocation. **(b)** Motion-compensated residuals quantify foreground changes and local target-context contrast. **(c)** Verdict-token confidence controls mask-logit scaling for memory encoding.

3 METHOD

3.1 PRELIMINARIES

Task and pipeline. Online reasoning VOS requires identifying targets from complex or implicit language queries and segmenting them in each incoming frame using only current and past observations. Given a query q and a video stream $\{I_t\}_{t=1}^{T}$, the system produces

$$\mathbf{m}_t = \mathcal{F}_\theta(q, I_{1:t}) \in \{0,1\}^{H\times W}, \qquad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{m}_t$ is the final output mask, $I_{1:t}$ denotes the observed video prefix, and H and W are the frame height and width.

We instantiate $\mathcal{F}_\theta$ with a training-free modular pipeline (Kao et al., 2026; Jin et al., 2026), comprising an MLLM selector $\mathcal{V}$, a reasoning image segmenter $\mathcal{G}$ (Liu et al., 2025), and a video propagation module Ψ (Ravi et al., 2024). The invocation policy determines when $\mathcal{V}$ evaluates (I_t, q) for suitable grounding evidence. A `Yes` verdict invokes $\mathcal{G}$ to produce a grounding mask $\mathbf{m}_t^{\mathrm{g}}$, which prompts Ψ. If $\mathcal{V}$ is skipped or returns `No`, Ψ propagates from available memory to produce $\mathbf{m}_t^{\mathrm{p}}$; before the first accepted grounding, the output is empty. We refer to this scheduled selector evaluation and conditional grounding as *semantic reassessment*. The online CoT-RVS baseline (Kao et al., 2026) invokes $\mathcal{V}$ at a fixed interval $\xi \in \mathbb{N}_+$:

$$b_t^{\mathrm{fix}} = \mathbf{1}[(t-1) \bmod \xi = 0], \tag{2}$$

where $\mathbf{1}[\cdot]$ is the indicator function. This yields $\lceil T/\xi \rceil$ selector evaluations independently of visual changes, while calls to $\mathcal{G}$ depend on the verdicts. Prompt formats are provided in Appendix A.

Memory-Based Mask Propagation. Memory-based mask propagation conditions current-frame segmentation on target representations encoded from previously segmented frames (Cheng et al., 2024; Ravi et al., 2024). These representations are retained in a memory bank, which in SAM 2 (Ravi et al., 2024) comprises conditioning entries $\mathcal{C}_{t-1}$ associated with externally prompted frames and non-conditioning entries $\mathcal{N}_{t-1}$ generated during propagation. Each entry stores spatial memory features $\mathbf{F}_i$ and an object pointer $\mathbf{p}_i$. Given image features $\mathbf{X}_t$ and a preprocessed mask $\mathbf{a}_t$, the spatial memory encoder computes

$$\mathbf{F}_t = \mathrm{Proj}_{\mathrm{out}}\big(\mathrm{Fuse}\big(\mathrm{Proj}_{\mathrm{img}}(\mathbf{X}_t) + \mathrm{Down}(\mathbf{a}_t)\big)\big), \tag{3}$$

where Down, Fuse, and Proj denote mask downsampling, feature fusion, and projection, respectively. This encoding combines visual features with the target's spatial support to guide subsequent mask propagation. Memory selection and retrieval are detailed in Appendix C.

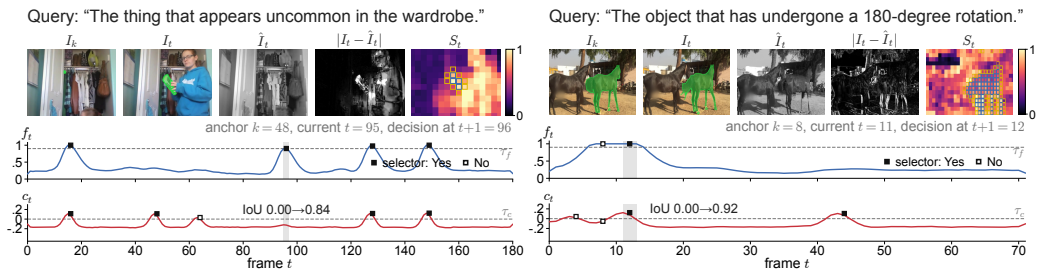

Figure 4: **Residual-guided invocation.** Foreground residuals and local target–context contrast guide semantic reassessment in response to visual changes. The shaded intervals highlight the corresponding invocations and subsequent mask IoU gains.

3.2 Residual-Guided Invocation

Semantic reassessment can incur redundant computation during visually stable periods, while fixed intervals may delay grounding updates when new visual evidence emerges. Inspired by predictive video coding (Wiegand et al., 2003; Sullivan et al., 2012), DRR computes motion-compensated residuals relative to a grounded anchor for low-cost, frame-wise monitoring of visual changes. Foreground residuals quantify changes within the tracked target, while local target-context contrast captures surrounding changes that foreground monitoring alone may miss. Together, these cues guide semantic reassessment without requiring an additional learned model.

Residual Estimation. Let I_k denote the most recent frame with an accepted grounding from $\mathcal{G}$, serving as the anchor available at frame t. After producing $\mathbf{m}_t$, we downsample I_t and I_k and convert them to luma using BT.601 channel weights (ITU-R, 2011), yielding Y_t and Y_k. For each square patch P, block matching searches displacements $\delta \in \mathcal{D}_{\text{search}}$ in the anchor image and retains the minimum mean absolute difference:

$$E_t(P) = \min_{\delta \in \mathcal{D}_{\text{search}}} \frac{1}{|P|} \sum_{x \in P} \left| Y_t(x) - Y_k(x+\delta) \right|, \tag{4}$$

where $|P|$ is the number of pixels in P. Matching compensates for local translation within the search range, while the remaining error measures appearance mismatch after alignment. Such mismatch may reflect deformation, occlusion, illumination changes, or motion beyond the search range. We compress and normalize the patch residuals to obtain bounded routing scores:

$$S_t(P) = \text{clip}\left(\frac{\log(1 + E_t(P))}{\max(\nu_t, \epsilon)}, 0, 1 \right), \tag{5}$$

where ν_t is an exponential moving average of the spatial 99th percentile of $\log(1 + E_t(P))$, and $\epsilon > 0$ prevents division by zero. Logarithmic compression moderates large residuals, while the running percentile adapts the normalization scale and smooths its variation across frames.

Regional Contrast Aggregation. DRR characterizes anchor-relative visual changes through foreground residual aggregation and local target-context contrast. The former quantifies changes within the predicted target, while the latter measures residual changes in its surroundings relative to the foreground. Let $\mathcal{P}_t$ denote the foreground patches identified from the resized output mask $\mathbf{m}_t$. Its context ring $\mathcal{R}_t$ consists of valid four-neighbor patches outside $\mathcal{P}_t$. For any nonempty patch set $\mathcal{U}$, define the mean residual as

$$\langle S_t \rangle_{\mathcal{U}} = \frac{1}{|\mathcal{U}|} \sum_{P \in \mathcal{U}} S_t(P). \tag{6}$$

The foreground residual and local target-context contrast are computed as

$$f_t = \langle S_t \rangle_{\mathcal{P}_t}, \qquad c_t = \langle S_t \rangle_{\mathcal{R}_t} - f_t. \tag{7}$$

Here, $f_t \in [0, 1]$ measures change within the predicted target, while $c_t \in [-1, 1]$ measures the excess residual change in its surroundings. A positive c_t can therefore capture surrounding changes even when the target remains comparatively stable. These residual measurements provide visual proxies for changes in target appearance or local interactions that may alter the evidence supporting query-dependent grounding, motivating semantic reassessment as new observations become available.

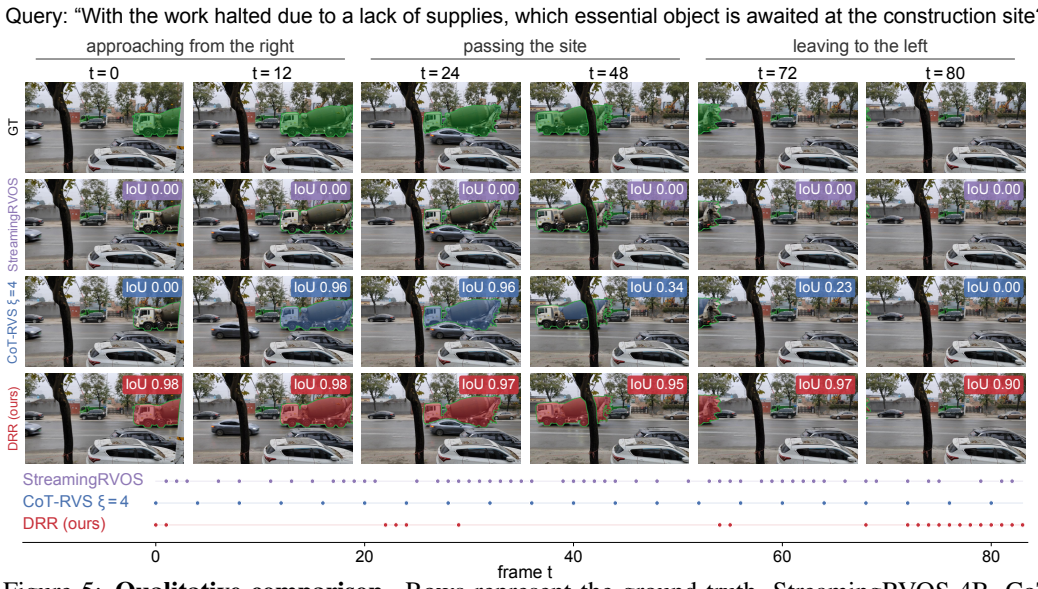

Figure 5: **Qualitative comparison.** Rows represent the ground truth, StreamingRVOS-4B, CoT-RVS (ξ=4), and DRR, with per-frame IoU. Lanes below mark reasoning reassessments. StreamingRVOS never finds the mixer and CoT-RVS loses it while it passes the site; DRR keeps it to the end, with 21 calls over 84 frames.

Causal invocation policy. DRR computes the cues after processing frame t and caches them for the invocation decision at frame $t+1$. For valid foreground and context supports, the residual trigger is

$$r_{t+1}^{\text{res}} = \mathbf{1}[(f_t \geq \tau_f) \vee (c_t \geq \tau_c)], \tag{8}$$

where τ_f and τ_c are the foreground-residual and local-contrast thresholds. A positive trigger requests evaluation of $\mathcal{V}$ on the next frame; a new grounding mask is produced only if the selector returns an affirmative verdict. This policy allows visual changes to request reassessment without waiting for a mask-quality failure. Block matching uses a reduced luma grid and a bounded displacement set, while regional aggregation requires only sums and averages. For N grid pixels covered by non-overlapping patches and D displacement candidates per patch, direct matching requires $O(ND)$ pixel comparisons. Thus, the routing cues can be evaluated frame by frame without an additional semantic-model invocation. Figure 4 qualitatively illustrates how changes in the target and its local context provide residual evidence for revisiting semantic grounding.

3.3 Confidence-Weighted Memory Encoding

Reassessed masks supply new target evidence, but an unreliable grounding can affect later predictions after it is encoded into propagation memory. Previous method (Kao et al., 2026) resets propagation state before inserting a new accepted grounding. DRR instead retains earlier references and regulates the mask supplied to the memory encoder, allowing historical evidence to remain available alongside new grounding evidence. Let $\mathcal{T}_+$ and $\mathcal{T}_-$ be disjoint token sets representing single-token positive and negative verdicts. At the verdict position, we compute

$$p_t = \frac{\sum_{j \in \mathcal{T}_+} \exp(z_{t,j})}{\sum_{j \in \mathcal{T}_+ \cup \mathcal{T}_-} \exp(z_{t,j})}, \tag{9}$$

where $z_{t,j}$ is the next-token logit conditioned on the current frame, query, and preceding generated text. The normalized positive-verdict probability p_t provides a frame-level confidence proxy for encoding an accepted grounding. For an accepted grounding, we map p_t to an encoding weight

$$\alpha_t = \text{clip}\left(\frac{p_t - 0.5}{\tau_{\text{m}} - 0.5}, 0, 1\right), \tag{10}$$

where $\tau_{\text{m}} \in (0.5, 1)$ is the confidence threshold above which the weight equals one. Given grounding-mask logits $\boldsymbol{\ell}_t^{\text{g}}$, we obtain soft mask probabilities as

$$\bar{\mathbf{s}}_t = \sigma(\alpha_t \boldsymbol{\ell}_t^{\text{g}}), \tag{11}$$

Table 1: Main results on ReasonVOS and ReVOS. Bold and underlined values indicate the best and second-best results, respectively.

Method	ReasonVOS			ReVOS-Overall		
	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑
SFT-based						
ORVOS (Liu et al., 2026)	–	–	–	54.3	–	–
StreamingRVOS-1B (Zhang et al., 2026)	50.1	66.6	5.0	55.9	17.9	5.6
StreamingRVOS-4B (Zhang et al., 2026)	52.4	61.1	4.4	**59.5**	15.3	5.4
Training-free						
CoT-RVS-LLaVA-1.5-7B (Kao et al., 2026)	52.0	25.4	1.1	46.0	26.5	1.0
DRR-LLaVA-1.5-7B	57.1	**5.2**	6.7	52.7	**6.8**	6.0
DRR-Seg-Zero-7B	60.6	5.7	**6.8**	52.5	9.0	**6.3**
DRR-Ovis2.5-9B	**65.4**	5.5	6.1	57.3	8.1	5.4

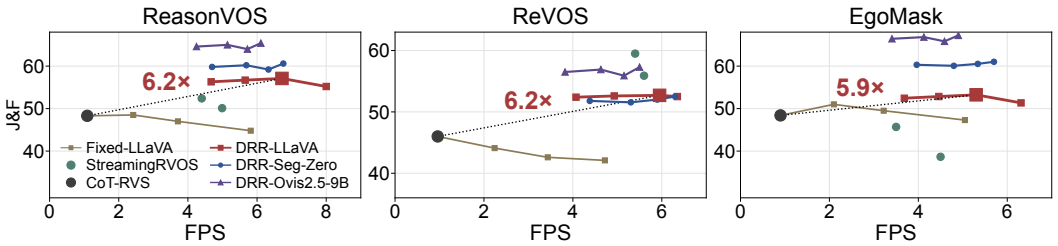

Figure 6: **Accuracy–efficiency trade-off**. Fixed-$*$ uses $\xi \in \{4, 10, 16, 32\}$ throughout, whereas DRR uses the corresponding ξ only for bootstrap. DRR consistently achieves a more favorable accuracy–efficiency trade-off than the fixed schedules across all three datasets. The dotted connector reads the speed-up between the two configurations of the same selector.

where σ denotes the sigmoid function. SAM 2's native memory preprocessing maps the soft mask probabilities to the encoder input:

$$\mathbf{a}_t = \gamma_{\text{mem}} \bar{\mathbf{s}}_t + \beta_{\text{mem}}, \tag{12}$$

where γ_{mem} and β_{mem} denote the mask scaling and bias parameters, respectively. The preprocessed mask $\mathbf{a}_t$ and image features $\mathbf{X}_t$ are jointly encoded according to Equation (3). The resulting spatial features form the new conditioning memory entry.

4 Experiments

4.1 Benchmarks and Implementation Details

Datasets. We evaluate DRR on five benchmarks covering online reasoning and referring VOS. For reasoning VOS, ReasonVOS (Bai et al., 2024) and ReVOS (Yan et al., 2024) include queries requiring world knowledge and contextual reasoning. For referring VOS, the evaluation comprises MeViS (Ding et al., 2023) for motion-dependent expressions, Ref-DAVIS17 (Khoreva et al., 2019) for challenging object motion and occlusion, and EgoMask (Liang et al., 2025) for fine-grained spatiotemporal grounding in egocentric videos, including periods of target absence.

Metrics. We evaluate segmentation accuracy using region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$. For EgoMask (Liang et al., 2025), we additionally report temporal recall T_{recall}, the fraction of target-present frames with non-empty predictions. The metrics IoU_{all}, IoU_{gold}, and $\text{IoU}_{\text{gold_pred}}$ average IoU over all frames, target-present frames, and frames with a non-empty ground-truth or predicted mask, respectively. We assess efficiency using throughput in frames per second (FPS) and MLLM invocation rate, defined as the percentage of processed frames on which semantic reasoning is invoked. For DRR and CoT-RVS, invocation rate counts evaluations of the selector $\mathcal{V}$, including negative verdicts.

Implementation details. We evaluate four selectors $\mathcal{V}$: LLaVA-1.5-7B (Liu et al., 2024), Seg-Zero-7B (Liu et al., 2025), Qwen3-VL-8B (Bai et al., 2025), and Ovis2.5-9B (Lu et al., 2025), denoting each variant by DRR-X according to its selector. The reasoning segmenter $\mathcal{G}$ is instantiated

Table 2: Results on MeViS, Ref-DAVIS17, and EgoMask. EgoMask reports means across duration subsets. Bold and underlined values indicate the best and second-best results, respectively.

Method	MeViS			Ref-DAVIS17			EgoMask (Avg.)		
	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑	$\text{IoU}_{gp}\uparrow$	Inv. (%)↓	FPS↑
SFT-based									
OnlineRefer	32.3	–	–	64.8	–	–	–	–	–
StreamingRVOS-1B	49.7	–	–	73.0	8.7	7.2	20.4	72.0	4.5
StreamingRVOS-4B	<u>50.9</u>	–	–	71.3	6.9	7.4	25.3	73.2	3.5
Training-free									
CoT-RVS-LLaVA-1.5-7B	45.9	25.5	0.9	<u>73.9</u>	25.5	0.6	26.3	26.2	0.9
DRR-LLaVA-1.5-7B	46.4	**5.4**	**6.8**	71.6	**5.1**	7.3	27.9	**7.1**	<u>5.3</u>
DRR-Seg-Zero-7B	48.0	6.8	<u>6.4</u>	72.4	<u>5.3</u>	**8.1**	<u>34.2</u>	<u>7.2</u>	**5.8**
DRR-Ovis2.5-9B	**51.2**	<u>5.7</u>	6.0	**75.5**	<u>5.3</u>	<u>7.5</u>	**39.5**	7.5	4.9

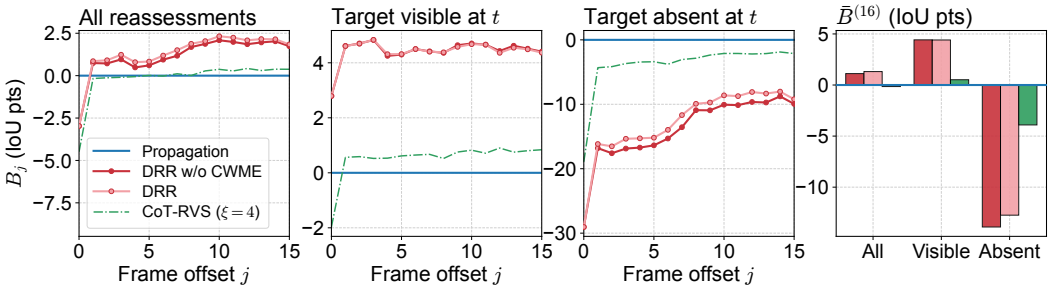

Figure 7: **Paired replay of semantic reassessment on ReasonVOS.** Each invocation is replayed against continued propagation for up to 16 frames. Curves show the per-frame IoU benefit B_j and bars show the window mean $\bar{B}^{(16)}$, grouped by target presence at invocation. Residual-guided reassessment yields a higher return per invocation than the fixed schedule comparison, and confidence-weighted encoding reduces the loss incurred on absent targets.

with Ovis2.5-9B for DRR-Ovis2.5-9B and Seg-Zero-7B for the remaining variants. All variants use SAM2.1-Hiera-L (Ravi et al., 2024) for propagation. In DRR, we set the foreground-residual and local-contrast thresholds to $\tau_f = 0.90$ and $\tau_c = 0$, respectively, and use a confidence threshold of $\tau_{\text{m}} = 0.75$ for mask encoding. All experiments run on a single NVIDIA H200 GPU.

4.2 Main Results

Reasoning VOS. Table 1 demonstrates that DRR advances the accuracy-efficiency Pareto frontier on ReasonVOS and ReVOS. With LLaVA-1.5-7B, DRR improves $\mathcal{J}\&\mathcal{F}$ by 5.1 and 6.7 over fixed-period CoT-RVS, respectively. DRR-Ovis2.5-9B attains 65.4 $\mathcal{J}\&\mathcal{F}$ on ReasonVOS, the best among the compared methods and 13.0 above StreamingRVOS-4B. On ReVOS, it achieves 57.3 overall $\mathcal{J}\&\mathcal{F}$, the best among the evaluated training-free methods, without dataset-specific instruction tuning. These gains are achieved with substantially fewer reasoning calls and increased throughput. The LLaVA-1.5-7B variant attains approximately 6.2× the throughput of CoT-RVS, while invocation decreases from 25.4% to 5.2% on ReasonVOS and from 26.5% to 6.8% on ReVOS. On ReVOS, DRR-Ovis2.5-9B maintains throughput comparable to that of StreamingRVOS-4B at a lower invocation rate of 8.1% versus 15.3%. DRR-Seg-Zero-7B achieves the highest throughput on both benchmarks at 6.8 and 6.3 FPS, with 60.6 and 52.5 $\mathcal{J}\&\mathcal{F}$, respectively. Figure 5 further illustrates that DRR identifies the mixer while it passes the site and maintains its segmentation.

Referring VOS. The efficiency gains extend to referring VOS on MeViS and Ref-DAVIS17, as shown in Table 2. On MeViS, DRR-LLaVA-1.5-7B improves $\mathcal{J}\&\mathcal{F}$ from 45.9 to 46.4 over CoT-RVS, while increasing throughput from 0.9 to 6.8 FPS and reducing invocation from 25.5% to 5.4%. On Ref-DAVIS17, the same configuration increases throughput from 0.6 to 7.3 FPS, with $\mathcal{J}\&\mathcal{F}$ of 71.6 compared with 73.9 for CoT-RVS. DRR also improves localization accuracy and efficiency on EgoMask, whose egocentric videos feature rapid camera motion and intermittent target visibility. DRR-Ovis2.5-9B achieves the highest mean IoU_{gp} of 39.5, compared with 26.3 for CoT-RVS and

Table 3: Component ablation with LLaVA-1.5-7B selector. CWME denotes confidence-weighted mask encoding.

Policy	f_t	c_t	CWME	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑
Fixed-period	–	–	–	52.0	25.4	1.1
Random	–	–	–	51.4	5.2	7.0
Adaptive	✓	–	–	53.0	6.4	6.0
Adaptive	–	✓	–	55.2	5.2	6.8
Adaptive	✓	✓	–	56.7	5.2	6.7
Full DRR	✓	✓	✓	**57.1**	5.2	6.7

Table 4: Parameter sensitivity analysis with the LLaVA-1.5-7B selector.

Param.	Value	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑
τ_c	0.00	**57.1**	5.2	6.7
	0.03	56.0	5.0	7.1
	0.10	55.2	4.7	8.0
τ_f	0.75	56.7	6.1	6.0
	0.90	**57.1**	5.2	6.7
	0.95	56.5	4.9	7.0
τ_{m}	0.60	56.9	5.2	6.7
	0.75	**57.1**	5.2	6.7
	0.90	56.8	5.2	6.7
α_t	Fixed 0.5	56.4	5.2	6.7
	Adaptive	**57.1**	5.2	6.7
	Fixed 1.0	56.7	5.2	6.7

Table 5: Per-frame latency (ms) with the Seg-Zero-7B selector. ξ denotes the fixed CoT-RVS interval.

Schedule	$\xi = 4$	$\xi = 10$	$\xi = 16$	$\xi = 32$	DRR
Latency	1037	480	314	200	**148**

25.3 for StreamingRVOS-4B. DRR-Seg-Zero-7B achieves the highest mean throughput of 5.8 FPS with 34.2 IoU_{gp}. Full results by duration are provided in Supplementary Section F.3.

4.3 Paired Replay Analysis

We replay each invocation after initialization against continued propagation from the same state for up to 16 frames without further reassessment; Appendix D details the protocol. Figure 7 shows a mean window-level IoU benefit of $+1.1$ for DRR and -0.2 for CoT-RVS. Despite invoking reassessment $7.4\times$ more often, CoT-RVS yields beneficial updates in only 8.5% of invocations, compared with 15.6% for DRR. These averages reflect each policy's invocation frames and propagation states. Confidence-weighted encoding preserves the target-present benefit of $+4.4$ while reducing target-absent losses from -13.9 to -12.7, limiting the downstream effects of unreliable grounding.

4.4 Ablation Analysis

Component analysis. Table 3 demonstrates complementary gains from residual-based routing and confidence-weighted encoding. Full DRR increases $\mathcal{J}\&\mathcal{F}$ from 52.0 to 57.1 over periodic invocation, while reducing invocation from 25.4% to 5.2% and increasing throughput from 1.1 to 6.7 FPS. At the same reported invocation rate of 5.2%, local contrast achieves 55.2 versus 51.4 for random routing, supporting content-dependent allocation of reasoning. Foreground residuals further increase $\mathcal{J}\&\mathcal{F}$ to 56.7, and confidence-weighted encoding yields 57.1. For the Seg-Zero-7B configuration, Table 5 reports a latency decrease from 1037 to 200 ms as the periodic interval increases from $\xi = 4$ to $\xi = 32$. DRR reduces latency to 148 ms, achieving a 26% reduction relative to $\xi = 32$ despite the additional residual computation.

Routing and encoding sensitivity. Table 4 characterizes the trade-off controlled by routing thresholds. Increasing τ_c from 0 to 0.10 reduces invocation from 5.2% to 4.7% and raises throughput from 6.7 to 8.0 FPS, while $\mathcal{J}\&\mathcal{F}$ decreases from 57.1 to 55.2. Lowering τ_f to 0.75 increases invocation to 6.1% but reduces $\mathcal{J}\&\mathcal{F}$ to 56.7. Encoding is less sensitive to the tested τ_{m} values, with $\mathcal{J}\&\mathcal{F}$ ranging from 56.8 to 57.1. Confidence-dependent weighting achieves 57.1, exceeding unscaled logits ($\alpha_t \equiv 1$, 56.7) and fixed softening ($\alpha_t \equiv 0.5$, 56.4).

5 Conclusion

We introduced *Dynamic Reasoning Routing (DRR)*, a training-free framework for efficient and responsive semantic reassessment in online reasoning VOS. Residual-based routing adapts reasoning invocation to changes in the target and its context, while confidence-weighted mask encoding moderates uncertain grounding updates to propagation memory. Experiments on five reasoning and referring VOS benchmarks demonstrate improved accuracy–efficiency trade-offs. Paired replays further show greater average segmentation benefits from DRR-selected reassessments than periodic. Future work will explore reasoning schedules under explicit latency budgets and confidence calibration for propagation-memory updates.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026. URL https://arxiv.org/abs/2605.25979.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

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. *arXiv preprint arXiv:2409.19603*, 2024. URL https://arxiv.org/abs/2409.19603.

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.

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.

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.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29183–29192, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gong_The_Devil_is_in_Temporal_Token_High_Quality_Video_Reasoning_CVPR_2025_paper.html.

Jihwan Hong and Jaeyoung Do. VIRST: Video-instructed reasoning assistant for SpatioTemporal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

ITU-R. Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. Technical Report Recommendation BT.601-7, International Telecommunication Union, 2011. URL https://www.itu.int/rec/R-REC-BT.601-7-201103-I/en.

Chen Jiang, Yuchen Yang, and Martin Jagersand. CLIPUNetr: Assisting human-robot interface for uncalibrated visual servoing control with CLIP-driven referring expression segmentation. In *IEEE International Conference on Robotics and Automation*, pp. 6620–6626, 2024.

Woojeong Jin, Jaeho Lee, Heeseong Shin, Seungho Jang, Junhwan Heo, and Seungryong Kim. AgentRVOS: Reasoning over Object Tracks for Zero-Shot Referring Video Object Segmentation. *arXiv preprint arXiv:2603.23489*, 2026. URL https://arxiv.org/abs/2603.23489.

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

Mehrdad Khani, Pouya Hamadanian, Arash Nasr-Esfahany, and Mohammad Alizadeh. Real-time video inference on edge devices via adaptive model streaming. *arXiv preprint arXiv:2006.06628*, 2020.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision – ACCV 2018*, pp. 123–141. Springer, 2019.

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.

Shuo Liang, Yiwu Zhong, Zi-Yuan Hu, Yeyao Tao, and Liwei Wang. Fine-grained spatiotemporal grounding on egocentric videos. *arXiv preprint arXiv:2508.00518*, 2025. URL https://arxiv.org/abs/2508.00518.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2310.03744.

Jinyuan Liu, Yang Wang, Zeyu Zhao, Weixin Li, Song Wang, and Ruize Han. Online reasoning video object segmentation. *arXiv preprint arXiv:2604.11411*, 2026. URL https://arxiv.org/abs/2604.11411.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-Zero: Reasoning-Chain Guided Segmentation via Cognitive Reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. URL https://arxiv.org/abs/2503.06520.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

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 (ECCV)*, 2020.

Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012. doi: 10.1109/TCSVT.2012.2221191.

Feilong Tang, Xiang An, Yunyao Yan, Yin Xie, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Chunyuan Li, Shikun Feng, Changrui Chen, Huajie Tan, Ming Hu, Manyuan Zhang, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. OneVision-Encoder: Codec-aligned sparsity as a foundational principle for multimodal intelligence. *arXiv preprint arXiv:2602.08683*, 2026. URL https://arxiv.org/abs/2602.08683.

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.

Thomas Wiegand, Gary J. Sullivan, Gisle Bjøntegaard, and Ajay Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003. doi: 10.1109/TCSVT.2003.815165.

Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Compressed video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6026–6035, 2018.

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.

Kai Xu and Angela Yao. Accelerating video object segmentation with compressed video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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 Model. In *European Conference on Computer Vision (ECCV)*, 2024.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos. *arXiv preprint arXiv:2501.04001*, 2025. URL https://arxiv.org/abs/2501.04001v1.

Wenkang Zhang, Kaicheng Yang, Xiang An, Qiang Li, Ziyong Feng, Wankou Yang, and Jiankang Deng. Towards streaming referring video segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24598–24607, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_Towards_Streaming_Referring_Video_Segmentation_via_Large_Language_Model_CVPR_2026_paper.html.

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.

Jiawen Zhu, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Huchuan Lu, Yifeng Geng, and Xuansong Xie. Tracking with human-intent reasoning. *arXiv preprint arXiv:2312.17448*, 2023. URL https://arxiv.org/abs/2312.17448.

A Pipeline and Prompts

Following training-free modular frameworks (Kao et al., 2026; Jin et al., 2026), our pipeline comprises an MLLM selector $\mathcal{V}$, a reasoning image segmenter $\mathcal{G}$ (Liu et al., 2025), and a video propagation module Ψ (Ravi et al., 2024), as defined in Section 3.1. This section specifies the selector interface and the Seg-Zero-based grounding configuration, including prompt templates, image preprocessing, and truncated decoding.

Selector prompt. We adopt the keyframe-assessment prompt from CoT-RVS (Kao et al., 2026). The original frame is processed by the selector's backbone-specific image processor, and `{query}` is replaced with the unmodified language query.

> Consider the query "{query}" for an object tracking task. First analyze what can be seen in the input image and then output a simple answer with Yes or No to justify whether the input image is suitable as a keyframe. A good keyframe is an image where the target object can be seen. Please follow the output format "The image contains .... Therefore, the justification of using this image as keyframe is <Yes./No.>".

Segmenter prompt. The reasoning MLLM receives the current frame resized to 840×840 pixels, while SAM 2 video propagation uses a 1024×1024 input. The query is converted to lowercase and its trailing period is removed. We adapt the Seg-Zero (Liu et al., 2025) template to request a single bounding box, omitting the two inscribed-circle point prompts from both the instruction and the output example.

> Please find '{query}' with bbox.Compare the difference between objects and find the most closely matched one.Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.Keep the thinking process to one short sentence.Output the one bbox of the interested object in JSON format.i.e., <think> thinking process here </think> <answer>{'bbox': [10,100,200,210]}</answer>

Multi-referent prompt. ReVOS (Yan et al., 2024) includes queries referring to multiple instances, which constitute 15.7% of its validation queries. To accommodate spatially separated referents, the multi-referent configuration requests one bounding box per matching instance. The output is a list, including a single-element list when only one instance matches. Image preprocessing, query normalization, and the reasoning-token budget remain unchanged.

> Please find '{query}' with bbox.The expression may refer to more than one object; find every object it refers to.Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.Keep the thinking process to one short sentence.Output a JSON list with one bbox per matching object, in that list even if there is only one.i.e., <think> thinking process here </think> <answer>[{'bbox': [10,100,200,210]}, {'bbox': [230,90,320,240]}]</answer>

B Invocation Schedule

Bootstrap. Before the first accepted grounding, propagation memory is uninitialized and no propagated mask is available to define the regions in Equation (7). The selector is therefore invoked at a fixed bootstrap interval:

$$r_{t+1}^{\text{boot}} = \mathbf{1}[(t+1) \bmod \xi_0 = 0], \qquad \xi_0 = 4. \tag{13}$$

Bootstrap continues until an affirmative verdict produces a valid grounding mask. This first grounding initializes the propagation memory and the anchor used for residual estimation. After initialization, fixed schedules invoke the selector every ξ frames. DRR instead uses the residual trigger in Equation (8), subject to a minimum invocation gap of γ frames. A maximum gap of 32 frames forces selector invocation when no earlier request has occurred.

Empty-mask fallback. If the propagated mask $\mathbf{m}_t$ is empty, $\mathcal{P}_t$ and the corresponding residual cues are undefined. DRR then requests selector evaluation at frame $t+1$, subject to the same minimum gap γ. This fallback is independent of the residual trigger and does not update the cached residuals.

Evaluated schedules. All configurations use the default bootstrap interval $\xi_0 = 4$ before the first accepted grounding. The four DRR configurations in Figure 6 use $\tau_c = 0$ with $\gamma \in \{4, 8, 16\}$, or $\gamma = 16$ with $\tau_c = 0.10$. The fixed schedules use $\xi \in \{4, 10, 16, 32\}$. Configurations are ordered by invocation rate.

C Memory Encoding and Retrieval

SAM 2 (Ravi et al., 2024) conditions mask prediction on stored spatial features and object pointers. DRR modifies mask preprocessing and reference retention within this pipeline.

Memory encoding. Let $\mathbf{s}_t$ denote the mask probabilities supplied to the spatial memory encoder. SAM 2 applies its native scale and bias:

$$\mathbf{a}_t = \gamma_{\text{mem}}\mathbf{s}_t + \beta_{\text{mem}}. \tag{14}$$

For propagated predictions, $\mathbf{s}_t = \sigma(\boldsymbol{\ell}_t^{\text{p}})$. For accepted groundings, DRR uses $\mathbf{s}_t = \sigma(\alpha_t \boldsymbol{\ell}_t^{\text{g}})$, where $\boldsymbol{\ell}_t^{\text{g}}$ denotes the grounding-mask logits supplied for memory encoding and α_t is defined in Section 3.3. The probabilities remain continuous, without binarization or an additional sigmoid before mask downsampling.

The preprocessed mask $\mathbf{a}_t$ is integrated with image features $\mathbf{X}_t$ to construct the spatial memory:

$$\mathbf{F}_t = \text{Proj}_{\text{out}}\big(\text{Fuse}\big(\text{Proj}_{\text{img}}(\mathbf{X}_t) + \text{Down}(\mathbf{a}_t)\big)\big), \tag{15}$$

where Down, Fuse, and Proj denote mask downsampling, feature fusion, and projection, respectively. Together with an object pointer $\mathbf{p}_t$, $\mathbf{F}_t$ forms a memory entry. DRR's confidence weighting modifies the mask input, without directly scaling the encoded feature or object pointer.

Reference retention and selection. Conditioning entries $\mathcal{C}_{t-1}$ originate from externally prompted frames, whereas non-conditioning entries $\mathcal{N}_{t-1}$ originate from propagation. The online CoT-RVS implementation (Kao et al., 2026) resets propagation state upon accepting a new grounding; DRR retains earlier conditioning references. For prediction at frame t, SAM 2 selects spatial memories and eligible object pointers from previously available entries. Flattened spatial features and object-pointer tokens form the memory sequence $\mathbf{Z}_{t-1}$. Retention therefore determines reference availability, while selection determines participation in each prediction.

Memory attention. Within each attention layer, self-attention with a residual connection and subsequent normalization produces the cross-attention input $\mathbf{X}_t'$. For one attention head, omitting positional terms and layer indices,

$$\mathbf{Q}_t = \mathbf{X}_t' W_Q, \quad \mathbf{K}_t = \mathbf{Z}_{t-1} W_K, \quad \mathbf{V}_t = \mathbf{Z}_{t-1} W_V, \tag{16}$$

where W_Q, W_K, and W_V are learned projections. Memory retrieval is computed as

$$\text{Attn}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) = \text{softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_t^{\mathsf{T}}}{\sqrt{d}}\right)\mathbf{V}_t, \tag{17}$$

where d is the key dimension and softmax operates over memory tokens. Residual connections and feed-forward transformations integrate the retrieved information into memory-conditioned features $\widetilde{\mathbf{X}}_t$.

Mask decoding. On propagation frames, the decoder $\mathcal{D}$ combines $\widetilde{\mathbf{X}}_t$ with high-resolution image features $\mathbf{H}_t$:

$$\begin{aligned}(\boldsymbol{\ell}_t^{\text{p}}, \mathbf{u}_t) &= \mathcal{D}(\widetilde{\mathbf{X}}_t, \mathbf{H}_t),\\ \mathbf{p}_t &= \text{Proj}_{\text{ptr}}(\mathbf{u}_t).\end{aligned} \tag{18}$$

Here $\boldsymbol{\ell}_t^{\text{p}}$ contains foreground logits and $\mathbf{u}_t$ is the decoder token used to construct the object pointer. The predicted mask and pointer supply the next memory entry, completing the encoding–retrieval cycle. These schematic expressions omit positional encodings, prompt embeddings, and native object-presence handling.

D Paired Replay Protocol

For each analyzed invocation at frame t, we initialize two replay branches from the same propagation state. The reassessment branch evaluates $\mathcal{V}$ and conditionally invokes $\mathcal{G}$ according to its verdict. The propagation-only branch skips reassessment. Both branches subsequently propagate without further semantic reassessment.

For a video of T frames, the evaluation horizon is $h_t = \min(16, T - t + 1)$. We define the mean reassessment benefit as

$$\bar{B}_t^{(16)} = \frac{1}{h_t} \sum_{j=0}^{h_t-1} \Big[\mathrm{IoU}(\mathbf{m}_{t+j}^{\mathrm{check}}, \mathbf{m}_{t+j}^{\mathrm{gt}}) - \mathrm{IoU}(\mathbf{m}_{t+j}^{\mathrm{skip}}, \mathbf{m}_{t+j}^{\mathrm{gt}}) \Big], \tag{19}$$

where $\mathbf{m}^{\mathrm{check}}$, $\mathbf{m}^{\mathrm{skip}}$, and $\mathbf{m}^{\mathrm{gt}}$ denote the reassessment-branch prediction, propagation-only prediction, and ground-truth mask, respectively. Including $j = 0$ captures both the immediate correction and its effect on subsequent propagation. We restrict the analysis to initialized tracks and stratify invocations by whether the ground-truth mask is nonempty at frame t. The curves in Figure 7 average the IoU differences at each offset over replays with an available observation. The bars average $\bar{B}_t^{(16)}$ over invocations within each group, assigning equal weight to each invocation.

E Additional Qualitative Examples

Figures 8–10 compare StreamingRVOS-4B, CoT-RVS with ξ=4, and DRR on ReasonVOS and EgoMask clips whose referent is fixed by an action, an interaction, or a change of viewpoint rather than by a single frame. Each figure shows the ground truth and the three predictions over six frames with the per-frame IoU, and the lanes below mark the frames on which each method invokes its reasoning model. The last example is a failure case, in which the referent enters late and DRR binds to the wrong object before its first reassessment.

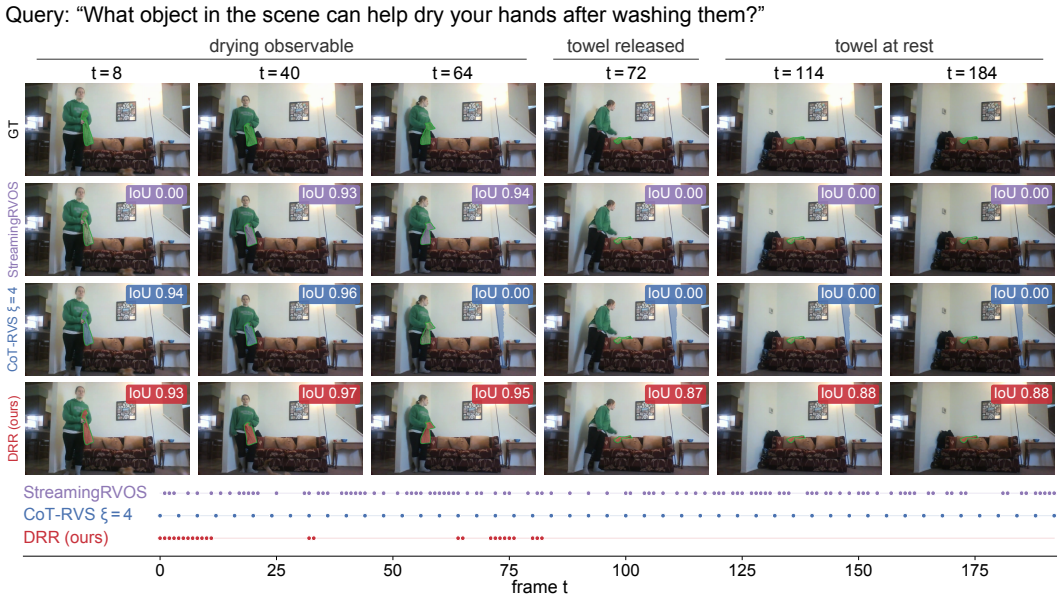

Figure 8: Both baselines lose the towel after it is released; DRR keeps it with 19 calls over 190 frames.

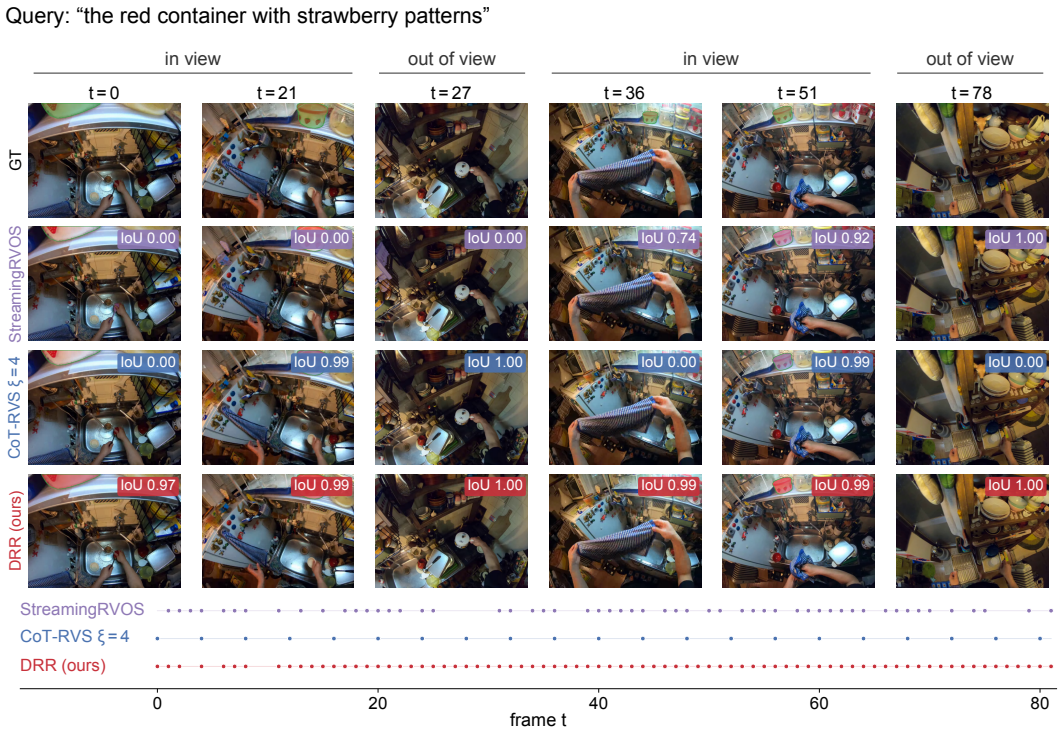

Figure 9: Both baselines miss the container at first and CoT-RVS still draws a mask after it has left the view; DRR tracks it in view and stays empty out of view, with 78 calls over 82 frames.

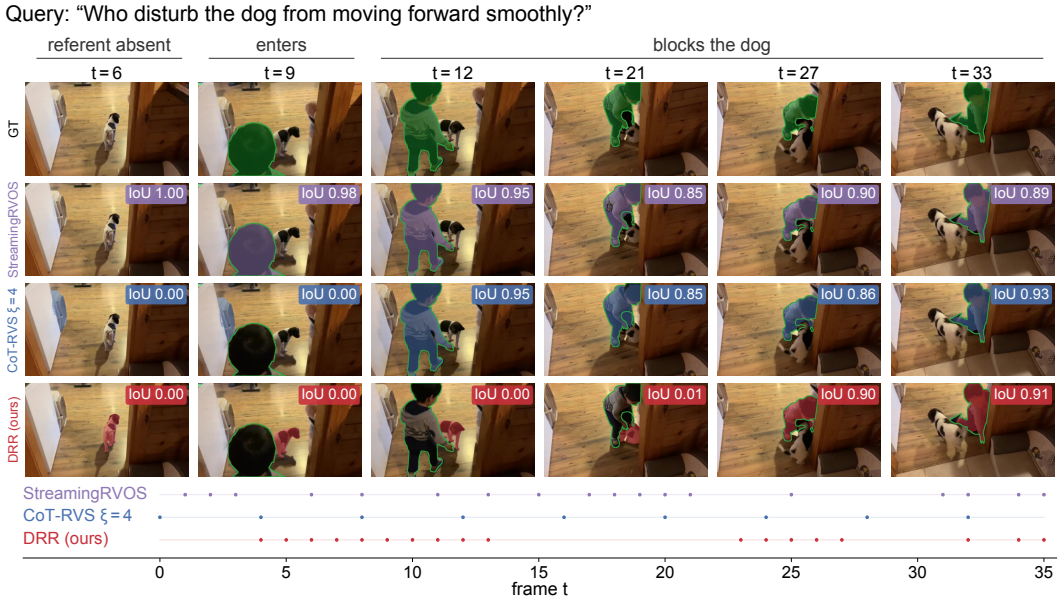

Figure 10: **A failure case:** the child is absent until t=9; DRR binds the dog first and re-binds to the child only at t=27, while both baselines recover by t=12; 18 calls over 36 frames.

F Additional Results

We extend the main evaluation with comparisons across model configurations, query types, and video durations.

F.1 Comparison across Model Configurations

Table 6 compares four DRR configurations on ReasonVOS, ReVOS, MeViS, and Ref-DAVIS17. All configurations use SAM2.1-Hiera-L for propagation. DRR-Ovis2.5-9B uses Ovis2.5-9B for both selection and grounding; the remaining variants use Seg-Zero-7B for grounding. These comparisons assess complete configurations rather than the isolated effect of the selector. The Ovis configuration achieves the highest accuracy on all four benchmarks. DRR-Seg-Zero-7B achieves the highest inference speed on ReasonVOS, ReVOS, and Ref-DAVIS17, while DRR-LLaVA-1.5-7B is fastest on MeViS. Across all configurations and benchmarks in this table, selector invocation rates do not exceed 9.0%.

Table 6: Comparison of DRR configurations across four benchmarks. Inv. denotes selector invocation rate, and FPS measures inference speed. Bold and underlined values denote the best and second-best results in each column.

DRR variant	ReasonVOS			ReVOS-Overall		
	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑
DRR-LLaVA-1.5-7B	57.1	**5.2**	<u>6.72</u>	52.7	**6.8**	<u>5.95</u>
DRR-Seg-Zero-7B	<u>60.6</u>	5.7	**6.76**	52.5	9.0	**6.30**
DRR-Qwen3-VL-8B	60.4	**5.2**	6.20	<u>53.4</u>	<u>8.1</u>	5.79
DRR-Ovis2.5-9B	**65.4**	<u>5.5</u>	6.11	**57.3**	<u>8.1</u>	5.41
DRR variant	**MeViS**			**Ref-DAVIS17**		
	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑	$\mathcal{J}\&\mathcal{F}\uparrow$	Inv. (%)↓	FPS↑
DRR-LLaVA-1.5-7B	46.4	**5.4**	**6.8**	71.6	**5.1**	7.3
DRR-Seg-Zero-7B	<u>48.0</u>	6.8	<u>6.4</u>	72.4	5.3	**8.1**
DRR-Qwen3-VL-8B	47.9	<u>5.7</u>	6.1	<u>72.9</u>	<u>5.2</u>	7.1
DRR-Ovis2.5-9B	**51.2**	<u>5.7</u>	6.0	**75.5**	5.3	<u>7.5</u>

F.2 ReVOS Results

Table 7 reports accuracy separately for referring and reasoning queries. DRR-LLaVA-1.5-7B improves over CoT-RVS-LLaVA-1.5-7B by 3.4 and 10.1 $\mathcal{J}\&\mathcal{F}$ points on referring and reasoning queries, respectively. StreamingRVOS-4B achieves the highest accuracy in both categories, while the Ovis configuration performs best among the compared training-free methods.

Table 7: Segmentation accuracy ($\mathcal{J}\&\mathcal{F}$) on ReVOS by query type. Bold and underlined values denote the best and second-best results in each column.

Method	Referring	Reasoning
SFT-based		
LISA-13B (Lai et al., 2024)	46.6	36.7
TrackGPT-7B (Zhu et al., 2023)	48.2	39.0
TrackGPT-13B (Zhu et al., 2023)	49.5	40.5
StreamingRVOS-1B (Zhang et al., 2026)	59.1	52.7
StreamingRVOS-4B (Zhang et al., 2026)	**62.1**	**56.9**
Training-free		
CoT-RVS-LLaVA-1.5-7B (Kao et al., 2026)	54.6	37.3
DRR-LLaVA-1.5-7B	58.0	47.4
DRR-Seg-Zero-7B	59.3	45.7
DRR-Qwen3-VL-8B	58.7	48.0
DRR-Ovis2.5-9B	<u>61.7</u>	<u>52.9</u>

F.3 EgoMask Results

We evaluate accuracy and efficiency on the short-, medium-, and long-duration subsets of EgoMask.

Table 8 reports the available accuracy metrics, with $\mathrm{IoU}_{\mathrm{gold_pred}}$, $\mathrm{IoU}_{\mathrm{gold}}$, and T_{recall} abbreviated as IoU_{gp}, IoU_{g}, and T_{rec}, respectively. The Ovis configuration achieves the highest IoU_{gp} and IoU_{g} in every subset, whereas StreamingRVOS achieves the highest temporal recall.

Table 8: Segmentation accuracy on the EgoMask subsets. Bold and underlined values denote the best and second-best results for each metric within each subset.

Method	IoU_{gp}	IoU_{g}	T_{rec}	$\mathrm{IoU}_{\mathrm{all}}$	$\mathcal{J}\&\mathcal{F}$
Short					
StreamingRVOS-1B	28.7	33.2	<u>92.9</u>	32.6	34.7
StreamingRVOS-4B	31.6	36.5	**95.8**	35.9	38.1
CoT-RVS-LLaVA-1.5-7B	40.7	43.9	77.9	48.2	50.1
DRR-LLaVA-1.5-7B	42.4	43.6	77.5	50.7	52.7
DRR-Seg-Zero-7B	46.5	47.6	80.8	55.2	57.4
DRR-Qwen3-VL-8B	<u>46.8</u>	<u>48.0</u>	81.3	<u>55.9</u>	<u>58.3</u>
DRR-Ovis2.5-9B	**51.0**	**52.0**	81.8	**60.1**	**62.8**
Medium					
StreamingRVOS-1B	18.6	35.3	**86.7**	44.0	44.3
StreamingRVOS-4B	25.1	36.6	<u>84.2</u>	55.9	56.3
CoT-RVS-LLaVA-1.5-7B	22.5	36.7	76.6	46.5	47.0
DRR-LLaVA-1.5-7B	24.4	32.4	69.8	54.2	54.9
DRR-Seg-Zero-7B	29.4	33.6	66.3	64.8	65.6
DRR-Qwen3-VL-8B	<u>34.5</u>	<u>37.5</u>	68.7	<u>72.2</u>	<u>72.9</u>
DRR-Ovis2.5-9B	**36.4**	**41.2**	68.4	**72.2**	**73.0**
Long					
StreamingRVOS-1B	13.9	32.8	**85.4**	43.0	43.2
StreamingRVOS-4B	19.2	34.2	<u>84.0</u>	54.5	54.8
CoT-RVS-LLaVA-1.5-7B	15.7	32.9	75.3	44.0	44.4
DRR-LLaVA-1.5-7B	16.8	26.3	64.4	51.6	52.1
DRR-Seg-Zero-7B	26.8	34.4	68.3	65.9	66.3
DRR-Qwen3-VL-8B	<u>28.4</u>	<u>34.8</u>	67.5	<u>71.1</u>	<u>71.6</u>
DRR-Ovis2.5-9B	**31.1**	**40.0**	66.7	**72.9**	**73.4**