REVEAL: ELICITING ROLE-AWARE LATENT VISUAL EVIDENCE FROM MLLMS FOR REASONING VIDEO OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning video object segmentation (ReVOS) requires segmenting the target specified by an implicit query throughout a video. Existing MLLM-based methods typically require costly cross-model alignment, tying MLLM training to a downstream segmentation model. We present **ReVEAL**, a decoupled **thinking-with-video** framework for ReVOS that elicits role-aware latent visual evidence from MLLM reasoning to directly guide segmentation. ReVEAL introduces **Visual Evidence Tokens** (VETs) into video reasoning to associate necessary visual regions or entities with target, background, or distractor roles, while a Dynamic Attention Layer Router adaptively aggregates VET-to-video attention across layers into role-specific grounding evidence. We train the MLLM to perform VET-augmented reasoning on our constructed ReVEAL-47K dataset, with a Contrastive Attention Grounding Loss as auxiliary supervision to apply role-specific constraints to VET attention and optimize layer routing for reliable latent grounding. At inference, ReVEAL identifies the target through this thinking-with-video process and fuses the resulting role-specific evidence to guide a frozen SAM2 in predicting target masks. Experiments on five benchmarks show that ReVEAL outperforms alignment-free baselines and remains competitive with alignment-based methods.

1 INTRODUCTION

Reasoning Video Object Segmentation (ReVOS) aims to segment target objects throughout a video according to free-form and often implicit language expressions. Unlike conventional referring video object segmentation (RVOS), the target may not be identifiable from a single frame or an explicit category name. Instead, it relies on the model's ability to perceive and understand the video content in relation to the query, such as identifying *"the object used to keep the drink cold"*. This makes Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Bai et al., 2023) a natural backbone for ReVOS. However, the challenge is not only to understand the expression, but also to connect such understanding to the visual evidence that supports a consistent object mask across time.

Since MLLMs do not directly produce pixel masks, existing MLLM-based ReVOS methods typically rely on cross-modal alignment with an external segmenter, making MLLM training dependent on the downstream segmentation model. Existing approaches often introduce a dedicated segmentation token, such as `<SEG>`, and align its hidden representation with the prompt space of an external segmenter (Kirillov et al., 2023; Ravi et al., 2024). Although straightforward for mask generation, such design incurs additional training costs and ties the framework to a specific MLLM–segmenter pair, complicating adaptation to alternative backbones and downstream settings.

Recent reasoning MLLMs adopt a thinking-with-images paradigm that incorporates visual primitives into intermediate chain-of-thoughts (CoT) steps, using boxes and points (OpenAI, 2025; Liu et al., 2025b; Wang et al., 2025b) or special tokens (Qin et al., 2025; Su et al., 2025; Zhou et al., 2026). This paradigm extends to thinking-with-video, where spatio-temporal cues serve as intermediate references for reasoning across frames (Zhang et al., 2025; Meng et al., 2025). Interactive video segmentation models such as SAM2 (Ravi et al., 2024) likewise accept spatial primitives as prompts for object identification and mask propagation, providing a natural interface through which visual evidence produced during MLLM reasoning can directly guide segmentation (Kao et al., 2025).

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.

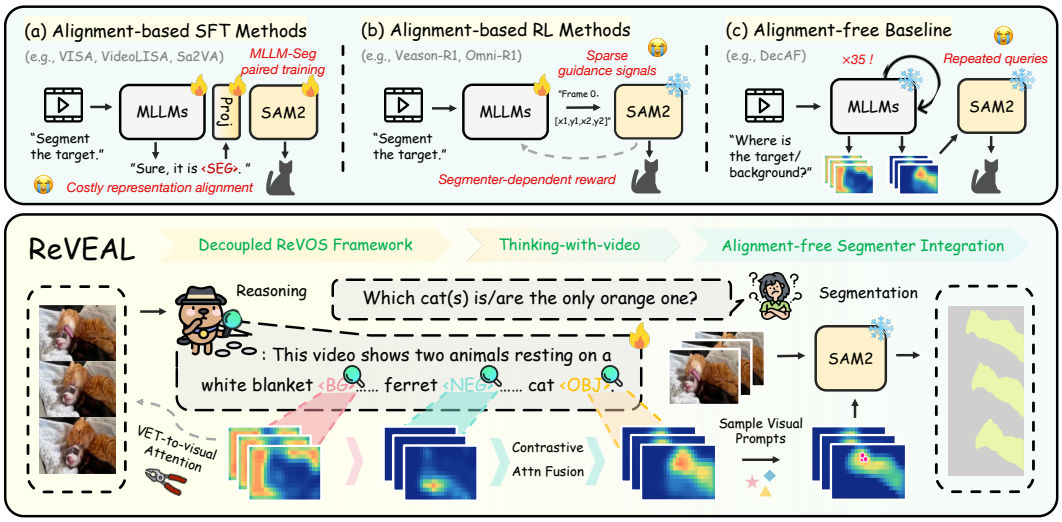

Figure 1: **Overview of ReVEAL.** Previous methods mainly involve the segmenter during training or rely on multiple MLLM queries at inference to obtain grounding signals. In contrast, ReVEAL adopts a decoupled ReVOS framework that elicits internal visual evidence from the MLLM's own video reasoning, enabling alignment-free integration with an external segmenter.

Yet these primitives are typically limited to selected keyframes (Gong et al., 2025a; Zhong et al., 2026), providing only sparse grounding evidence across the video. Causal self-attention, on the other hand, offers a denser alternative (Vaswani et al., 2017): text tokens attend to visual tokens across preceding frames, yielding video-wide attention maps with spatial grounding cues (Li et al., 2025), although these cues may not reliably localize the corresponding regions (Jang et al., 2026). DecAF (Han et al., 2025) relies on a tailored multi-rollout procedure to obtain object and background attention maps for segmentation, incurring substantial inference cost without leveraging video reasoning itself. This motivates us to explore whether an MLLM can produce the grounding evidence needed for segmentation as part of its own video reasoning within a single response.

To this end, we present **ReVEAL**, a thinking-with-video framework for ReVOS that elicits role-aware latent visual evidence from MLLM reasoning to guide segmentation. ReVEAL introduces Visual Evidence Tokens (VETs) to mark necessary visual regions or entities as target (<OBJ>), background (<BG>), or distractors (<NEG>) during reasoning. A Dynamic Attention Layer Router (DALR) learns to adaptively aggregate their attention over preceding visual tokens across Transformer layers into role-specific evidence maps. To enable the MLLM to elicit faithful role-aware latent visual evidence through video reasoning, we perform supervised fine-tuning (SFT) for VET-augmented reasoning on our constructed ReVEAL-47K dataset, using a Contrastive Attention Grounding Loss (CAGL) as auxiliary supervision to optimize layer routing and apply role-specific constraints to VET attention over video tokens for reliable latent grounding.

At inference, ReVEAL identifies the target through VET-augmented video reasoning. DALR then aggregates VET-to-video attention from the CoT across Transformer layers into role-specific evidence. These attention maps are then contrastively fused, and point prompts are sampled from high-confidence regions to guide a frozen external SAM2 in predicting the target masks, forming an alignment-free interface between MLLM reasoning and video object segmentation.

In a nutshell, our contributions are summarized as follows:

- We present **ReVEAL**, a decoupled thinking-with-video framework that elicits role-aware latent visual evidence during MLLM reasoning to guide segmentation without cross-model alignment.
- We equip the MLLM with this capability by combining reasoning supervision on **ReVEAL-47K** with **CAGL** for attention grounding over layers adaptively selected and weighted by **DALR**.
- We demonstrate that internal signals from MLLM reasoning can provide **effective grounding cues** for video object segmentation through experiments on five RVOS and ReVOS benchmarks.

2 Related Work

2.1 Multimodal Large Language Models

Building on the success of LLMs in natural language processing, image MLLMs extend language models to visual inputs by bridging visual features into the language model through cross-attention resamplers (Alayrac et al., 2022), querying transformers (Li et al., 2023; Dai et al., 2023), or projection-based visual instruction tuning (Liu et al., 2023; Zhu et al., 2024). Subsequent systems further scale visual backbones and training data to improve general vision–language capabilities (Chen et al., 2024b; Bai et al., 2023; Li et al., 2024; Wang et al., 2024). Video MLLMs extend this paradigm to video by incorporating sampled-frame representations and temporal modeling (Zhang et al., 2023; Maaz et al., 2024; Lin et al., 2024; Li et al., 2026). Beyond general video understanding, recent work increasingly adopts video MLLMs as general-purpose backbones for downstream tasks that require fine-grained perception and reasoning, including dense video captioning (Chen et al., 2024a; Xu et al., 2024), temporal grounding (Ren et al., 2024a; Huang et al., 2024a), subject-specific video question answering (Yuan et al., 2025c; Lian et al., 2025), and embodied spatial reasoning (Yang et al., 2025). We extend this line of research to ReVOS by leveraging MLLMs' spatio-temporal reasoning capabilities to identify target objects specified by complex queries.

2.2 Referring and Reasoning Video Object Segmentation

Referring Video Object Segmentation (RVOS) segments an object in a video based on a natural language expression. Early methods focus on explicit expressions through cross-modal fusion or query propagation (Seo et al., 2020; Botach et al., 2022; Wu et al., 2022). Reasoning Video Object Segmentation (ReVOS) extends this task to implicit queries that identify targets through their functions, behaviors, or relations to other objects, requiring more complex reasoning (Yan et al., 2024). Most MLLM-based methods interpret implicit queries and learn cross-modal alignment by mapping the hidden state of a special <SEG> token to a downstream segmenter's prompt space for mask prediction (Lai et al., 2024; Yan et al., 2024; Bai et al., 2024; Jin et al., 2026; Yuan et al., 2025a; Lin et al., 2025; Gong et al., 2025b). Recent work promotes this alignment indirectly through reinforcement learning, using SAM2-based segmentation rewards to optimize key-frame selection and MLLM-generated bounding boxes (Gong et al., 2025a; Zhong et al., 2026). Both strategies tie MLLM training to a downstream segmenter, leaving the potential of native MLLM reasoning as an independent source of visual grounding evidence underexplored (Kao et al., 2025). In contrast, we train the MLLM under a thinking-with-video paradigm to elicit latent grounding evidence from video reasoning and use this evidence to guide an external segmenter at inference.

2.3 Visual Evidence in MLLM Reasoning

OpenAI-o3 (OpenAI, 2025) introduces a thinking-with-images paradigm that incorporates references to visual regions into intermediate reasoning steps. Recent methods explicitly refer to these regions using boxes and points (Liu et al., 2025b; Wang et al., 2025b) or special tokens representing masks and image patches (Qin et al., 2025; Su et al., 2025; Yuan et al., 2025b). This idea also extends to thinking-with-video, where spatio-temporal cues serve as references for video reasoning (Cheng et al., 2026; Meng et al., 2025). However, generating such references throughout a video becomes cumbersome for dense video prediction. Instead, attention from output tokens to the preceding multimodal context provides a potential source of latent grounding evidence by highlighting visual regions associated with token semantics (Kang et al., 2025; Wu et al., 2025; Li et al., 2025; Hu et al., 2026; Vaswani et al., 2017). Despite this potential, raw attention is often noisy and susceptible to attention sinks (Huang et al., 2024b; Darcet et al., 2024), limiting its reliability as visual evidence. InterRVOS (Jin et al., 2026) constrains attention to relevant object regions to enrich semantic representations during cross-model alignment, whereas DecAF (Han et al., 2025) aggregates attention maps across multiple queries for visual grounding. In contrast, ReVEAL derives role-aware latent visual evidence directly from video reasoning and fuses those cues for more accurate target localization.

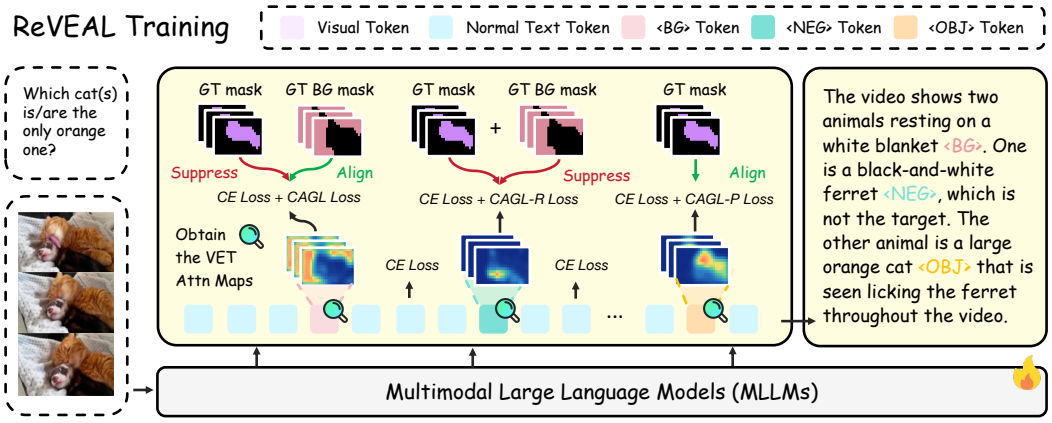

Figure 2: **Training of ReVEAL.** We train the MLLM to perform VET-augmented video reasoning using cross-entropy loss, with CAGL providing auxiliary role-specific constraints on VET-to-video attention for reliable latent grounding.

3 METHOD

Reasoning video object segmentation (ReVOS) aims to segment the target specified by an implicit query throughout a video. We propose ReVEAL, a decoupled thinking-with-video framework that uses role-aware latent evidence from MLLM reasoning to guide segmentation. During training, the MLLM learns to use VETs as role-specific semantic anchors for necessary visual regions or entities, with spatio-temporal grounding constraints applied directly to its own attention rather than through alignment with a downstream segmenter. At inference, ReVEAL derives role-specific latent visual evidence through VET-augmented video reasoning and fuses this evidence as grounding signals to guide a frozen SAM2 in predicting target masks.

Taxonomy of VETs. ReVEAL augments the MLLM vocabulary $\mathcal{V}_{\text{orig}}$ with three role-specific VETs: <OBJ> specifies the referred target, <NEG> identifies the foreground distractor to exclude, and <BG> indicates the background region to suppress:

$$\mathcal{V} = \mathcal{V}_{\text{orig}} \cup \mathcal{V}_{\text{VET}} \qquad \mathcal{V}_{\text{VET}} = \{\texttt{<OBJ>}, \texttt{<BG>}, \texttt{<NEG>}\} \tag{1}$$

Obtaining VET Attention Maps. We treat each VET's attention to the preceding visual tokens as a role-specific latent grounding cue. The corresponding attention map is obtained by aggregating these cues across Transformer layers, with later layers selected as candidates because they generally encode stronger semantic signals (Han et al., 2025). Let $\mathcal{L}_c$ denote the set of L_c candidate layers and $A \in [0,1]^{L_c \times N \times N}$ collect their post-softmax attention weights averaged across heads, where N is the whole sequence length. For the j-th VET occurrence with role r, we extract its attention weights over the preceding video tokens and reshape them into layer-wise attention maps $a^{r,j} \in [0,1]^{L_c \times T \times H_v \times W_v}$, where T is the number of frames and H_v and W_v denote the height and width of each frame's visual token grid.

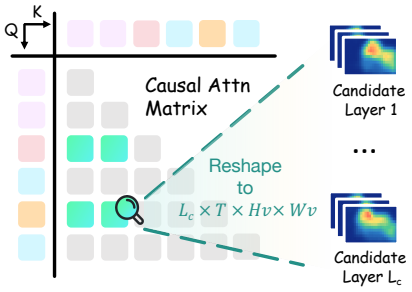

Figure 3: VET Attention Extraction.

However, the layers most relevant to target grounding may vary across different cases (Kang et al., 2025). Inspired by mixture-of-experts routing (Liu et al., 2024), we introduce a Dynamic Attention Layer Router (DALR) to dynamically select and weight candidate layers based on the <OBJ> attention maps. Given the <OBJ> attention maps for each video-query pair, DALR scores the candidate layers, selects the top-k layers as $\mathcal{S} \subset \mathcal{L}_c$, and normalizes their scores into routing weights w_l. These weights determine the contribution of each selected layer when aggregating VET attention during both training and inference. More details are provided in Appendix B.

3.1 LEARNING VIDEO REASONING WITH VETS

Supervising Thinking-with-video Reasoning. Through SFT, the MLLM learns to identify query-relevant visual entities or regions during video reasoning and place the corresponding role-specific VETs after their names or descriptions, conditioning each VET occurrence on an entity-specific context. Given a video V, a query q, and a VET-augmented response $\mathbf{u} = (u_1, \ldots, u_M)$, we optimize the model using the standard next token prediction objective:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{i=1}^{M} \log p(u_i \mid u_{<i}, V, q)\,. \tag{2}$$

Let $n_r(u)$ denote the number of occurrences of role r in response u. Each response contains $n_{\texttt{<OBJ>}}(u) = 1$ and $n_{\texttt{<BG>}}(u) \in \{0, 1\}$, while $n_{\texttt{<NEG>}}(u)$ varies with the number of distinct distractors described in the reasoning trace.

Aligning Attention Maps with Visual Regions. Although SFT teaches the model to perform VET-augmented video reasoning to identify relevant visual entities and their semantic roles, it does not ensure that VET attention provides reliable latent grounding evidence. Inspired by AML (Jin et al., 2026), we introduce the Contrastive Attention Grounding Loss (CAGL) to supervise the spatio-temporal distribution of VET attention using complementary positive and repulsive terms.

$$\mathcal{L}_{\mathrm{CAGL}} = \sum_{l \in \mathcal{S}} w_l \left(\mathcal{L}^l_{\mathrm{CAGL-P}} + \lambda_{\mathrm{rep}} \mathcal{L}^l_{\mathrm{CAGL-R}} \right). \tag{3}$$

Here, w_l is the routing weight assigned by DALR to layer l, and λ_{rep} balances the repulsive term.

For each occurrence j of a role $r \in \mathcal{P} = \{\texttt{<OBJ>}, \texttt{<BG>}\}$, the positive term applies BCE to align its attention distribution with the corresponding ground-truth (GT) region $m^r \in [0,1]^{T \times H_v \times W_v}$:

$$\mathcal{L}^l_{\mathrm{CAGL-P}} = \sum_{r \in \mathcal{P}} \sum_{j=1}^{n_r(u)} \mathrm{BCE}\Big(a_l^{r,j}, m^r\Big)\,, \tag{4}$$

During reasoning, the attention maps of the suppressive roles $r \in \mathcal{R} = \{\texttt{<BG>}, \texttt{<NEG>}\}$ provide suppressive evidence, but leakage into unintended regions may cause erroneous suppression. We therefore apply a repulsive term that penalizes the normalized attention mass within a role-specific forbidden region f^r. Specifically, we define $f^{\texttt{<BG>}} = m^{\texttt{<OBJ>}}$ and $f^{\texttt{<NEG>}} = m^{\texttt{<OBJ>}} \cup m^{\texttt{<BG>}}$:

$$\hat{a}_l^{r,j} = \frac{a_l^{r,j}}{\|a_l^{r,j}\|_1 + \epsilon}, \qquad \mathcal{L}^l_{\mathrm{CAGL-R}} = \sum_{r \in \mathcal{R}} \sum_{j=1}^{n_r(u)} -\log\left(1 - \left\langle \hat{a}_l^{r,j}, f^r \right\rangle + \epsilon\right). \tag{5}$$

Here, $\|a_l^{r,j}\|_1$ sums the attention over all spatio-temporal locations, and the inner product $\langle \hat{a}_l^{r,j}, f^r \rangle$ measures the normalized attention mass within the forbidden region.

Final Training Objective. The final objective combines SFT with CAGL:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SFT}} + \lambda_{\mathrm{CAGL}}\, \mathcal{L}_{\mathrm{CAGL}} \tag{6}$$

Here, λ_{CAGL} controls the strength of CAGL.

Training Data Construction. To provide supervision for ReVEAL, we construct **ReVEAL-47K**, a training dataset comprising VET-augmented CoT traces and region-level mask annotations, from multiple training sources (Lai et al., 2024; Khoreva et al., 2018; Seo et al., 2020; Ding et al., 2023; Deng et al., 2025; Yan et al., 2024). For each sample, we use GAR (Wang et al., 2025c) with the GT mask on a representative frame to obtain a target caption. Conditioned on this caption, the query, and the video frames, Gemini 2.5 Pro (Comanici et al., 2025) generates a CoT response and inserts VETs

at the corresponding positions. Qwen3VL (Bai et al., 2025) is used to verify and filter out unreliable samples. The mask annotations include the retained GT mask and a background mask, obtained by taking the complement of the union of the foreground regions predicted by BiRefNet (Zheng et al., 2024a) and the target region. More details are provided in Appendix A.

3.2 From VET Attention Maps to Segmentation Masks

At inference, ReVEAL identifies the target through VET-augmented video reasoning. DALR adaptively selects and weights Transformer layers to aggregate VET attention into role-specific latent evidence maps, which are then fused to guide mask prediction with a frozen, external SAM2.

Specifically, the <OBJ> attention is aggregated across the selected layers to form target evidence, whereas attention associated with all <BG> and <NEG> tokens is aggregated as suppressive evidence:

$$a^{\mathrm{obj}} = \sum_{l \in \mathcal{S}} w_l a_l^{\texttt{<OBJ>},1}, \; a^{\mathrm{sup}} \quad = \sum_{r \in \mathcal{R}} \alpha_r \sum_{j=1}^{n_r(u)} \sum_{l \in \mathcal{S}} w_l a_l^{r,j}, \tag{7}$$

where α_r controls the contribution of each suppressive role.

Contrastive Attention Fusion. We fuse the target and suppressive evidence frame by frame. For frame t, the corresponding slices $a_t^{\mathrm{obj}}, a_t^{\mathrm{sup}} \in \mathbb{R}^{H_v \times W_v}$ are combined by subtracting the suppressive evidence from the target evidence:

$$\tilde{P}_t = \max\big(a_t^{\mathrm{obj}} - a_t^{\mathrm{sup}}, 0\big) \qquad P_t = \frac{\tilde{P}_t - \min(\tilde{P}_t)}{\max(\tilde{P}_t) - \min(\tilde{P}_t) + \epsilon} \tag{8}$$

Here, the minimum and maximum are taken over spatial locations, and ϵ ensures numerical stability. We then sample points from the fused attention map P_t as visual prompts for the frozen SAM2.

SAM2 Prompting and Mask Generation. We apply a threshold τ_{pq} to the fused attention map and sample points from the retained regions according to their attention strengths as prompts for SAM2 to generate candidate mask tracklets. Following the SAM2 prompting pipeline of DecAF (Han et al., 2025), we remove redundant candidates using non-maximum suppression with an IoU threshold τ_{nms}, ranking them by point-attention and SAM2 confidence scores. We then filter the remaining tracklets by their temporal consistency with the retained attention regions using τ_{trk} and propagate them across all frames to produce the final segmentation masks.

4 Experiments

Datasets and metrics. We evaluate ReVEAL on five RVOS and ReVOS benchmarks: Ref-DAVIS (Khoreva et al., 2018), Ref-YouTube-VOS (Seo et al., 2020), MeViS (Ding et al., 2023), ReVOS (Yan et al., 2024), and ReasonVOS (Bai et al., 2024). Following standard practice, we report mask quality using region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$.

Implementation Details. ReVEAL uniformly samples six frames from each video as input. We finetune the MLLM using LLaMA-Factory (Zheng et al., 2024b), with the CAGL weight λ_{CAGL} set to 0.2 and the λ_{rep} in CAGL set to 0.1. The router top-k is set to 6. For inference and evaluation, we use SAM2-Hiera-Large (Ravi et al., 2024) as the frozen segmentation model. We set $\alpha_{<\mathrm{BG}>} = 0.5$ for Qwen2.5-VL and $\alpha_{<\mathrm{BG}>} = 0.1$ for Qwen3VL, while using $\alpha_{<\mathrm{NEG}>} = 1.0$ for both. All other settings follow the baseline (Han et al., 2025), using a threshold of $\tau_{\mathrm{pq}} = 0.8$, a NMS IoU threshold of $\tau_{\mathrm{nms}} = 0.7$, and a tracklet-consistency threshold of $\tau_{\mathrm{trk}} = 0.8$. Our experiments are conducted on NVIDIA H800 GPUs. More details are provided in Appendix B.

4.1 Main Results

Masks with Attn. Table 1 compares ReVEAL with attention-derived grounding methods under the same Qwen2.5-VL-7B backbone and six frame input setting. ReVEAL achieves the best performance across all benchmarks, outperforming DecAF by 5.7%, 3.0%, and 4.6% $\mathcal{J}\&\mathcal{F}$ on Ref-DAVIS,

Table 1: Comparison of MLLM-based ReVOS methods using attention-derived masks.

Method	MLLM	Ref-DAVIS			ReasonVOS			ReVOS (Overall)			ReVOS (Referring)			ReVOS (Reasoning)		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
TAM (Li et al., 2025)	Qwen2.5VL-7B	2.7	1.8	3.7	3.3	2.7	3.8	3.0	2.8	3.1	3.0	2.9	3.2	2.9	2.7	3.0
Loc-Head (Kang et al., 2025)	Qwen2.5VL-7B	13.0	16.8	9.2	12.2	17.1	7.3	13.3	18.9	7.7	14.1	20.3	7.9	12.5	17.4	7.5
DecAF (Han et al., 2025)	Qwen2.5VL-7B	23.7	30.4	17.1	19.1	24.1	14.1	19.4	25.2	13.5	21.1	27.9	14.3	17.7	22.6	12.7
ReVEAL (Ours)	Qwen2.5VL-7B	**29.4**	**39.5**	**19.2**	**22.1**	**29.5**	**14.8**	**24.0**	**33.6**	**14.4**	**25.4**	**35.6**	**15.3**	**22.5**	**31.6**	**13.5**

Table 2: Main comparison of MLLM-based ReVOS methods.

Method	MLLM	Ref-DAVIS			ReasonVOS			ReVOS (Overall)			MeViS			Ref-YT-VOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Cross-Model Alignment-Based Methods																
LISA (Lai et al., 2024)	LLaVA-7B	64.8	62.2	67.3	31.1	29.1	33.1	40.9	39.1	42.7	–	–	–	–	–	–
VISA (Yan et al., 2024)	ChatUniVi-7B	69.4	66.3	72.5	–	–	–	46.9	44.9	49.0	44.5	41.8	47.1	61.5	59.8	63.2
VideoLISA (Bai et al., 2024)	LLaVA-Phi-3-V	68.8	64.9	72.7	47.5	45.1	49.9	–	–	–	44.4	41.3	47.6	63.7	61.7	65.7
InstructSeg (Wei et al., 2025)	Mipha-3B	71.1	67.3	74.9	–	–	–	54.5	52.0	56.9	–	–	–	67.5	65.4	69.5
GLUS (Lin et al., 2025)	LLaVA-7B	–	–	–	49.9	47.5	52.4	54.9	52.4	57.3	51.3	48.5	54.2	67.3	65.5	69.0
VRS-HQ (Gong et al., 2025b)	ChatUniVi-7B	76.0	72.6	79.4	54.9	52.6	57.3	59.1	56.6	61.6	50.6	47.6	53.7	70.4	68.3	**72.5**
RGA3 (Wang et al., 2025a)	Qwen2.5VL-7B	72.8	68.3	77.3	53.6	51.3	56.0	58.0	55.9	60.0	50.1	47.4	52.8	68.5	66.8	70.1
UniPixel (Liu et al., 2025a)	Qwen2.5VL-7B	**76.4**	**72.7**	**80.1**	–	–	–	**63.7**	**61.7**	**65.7**	**55.8**	**53.2**	**58.3**	**71.0**	**69.5**	72.4
Veason-R1 (Gong et al., 2025a)	Qwen2.5VL-7B	–	–	–	**59.9**	**56.0**	**63.8**	61.3	58.2	64.4	52.2	48.4	56.0	–	–	–
Cross-Model Alignment-Free Methods																
Loc-Head (Kang et al., 2025)	Qwen2.5VL-7B	64.6	60.2	68.9	41.1	37.9	44.3	47.0	43.3	50.7	39.4	35.2	43.6	51.0	46.8	55.2
DecAF (Han et al., 2025)	Qwen2.5VL-7B	75.2	70.9	79.5	63.9	60.5	67.2	54.2	50.1	58.2	48.1	44.0	**52.1**	59.9	56.2	63.5
CoT-RVS (Kao et al., 2025)	Gemma3-12B	74.6	70.9	78.3	50.7	47.5	54.0	47.1	43.4	50.9	44.2	40.3	48.1	–	–	–
ReVEAL (Ours)	Qwen2.5VL-7B	77.7	74.1	81.3	65.1	61.9	68.3	59.0	56.3	61.6	47.7	44.7	50.6	69.2	67.2	71.1
ReVEAL (Ours)	Qwen3VL-8B	**80.7**	**77.3**	**84.0**	**66.1**	**62.9**	**69.2**	**61.6**	**58.9**	**64.3**	**49.1**	**46.2**	52.0	**71.3**	**69.3**	**73.3**

ReasonVOS, and ReVOS. These results indicate that ReVEAL produces attention maps with more complete spatio-temporal coverage of the target regions, providing stronger grounding evidence. Despite these gains, $\mathcal{F}$ remains relatively low because the coarse attention maps lack precise object boundaries and therefore still require a dedicated segmenter for accurate mask prediction.

Masks with SAM2. As shown in Table 2, among alignment-free methods, ReVEAL achieves the best performance across all benchmarks. With the same backbone, it outperforms DecAF by 4.8% and 9.3% $\mathcal{J}\&\mathcal{F}$ on ReVOS and Ref-YouTube-VOS, respectively, and consistently outperforms CoT-RVS despite using a smaller backbone. These results suggest that the high-attention regions identified by ReVEAL provide reliable visual prompts for mask generation. While alignment-based methods benefit from cross-model optimization between the MLLM and the segmentation module, ReVEAL approaches the performance of RGA-3 under the same Qwen2.5-VL-7B backbone without costly cross-model adapter training. With Qwen3-VL-8B, ReVEAL achieves the highest scores on Ref-DAVIS and ReasonVOS, surpassing the strongest alignment-based results by 4.3% and 6.2% $\mathcal{J}\&\mathcal{F}$, respectively, while substantially narrowing the gap on ReVOS, MeViS, and Ref-YouTube-VOS. Additional results are provided in the Appendix C.1

4.2 Ablation Study

In this section, we conduct comprehensive ablation studies. By default, results are reported with ReVEAL-Qwen3VL-8B. Additional results are provided in the Appendix C.2.

Thinking-with-video Paradigm for ReVOS. Table 3 evaluates the effect of the thinking-with-video paradigm. Without CoT, the model learns to emit VETs within a simple response (e.g., Sure, it is <BG>, <NEG> and <OBJ>.) With CoT, attention-derived mask $\mathcal{J}\&\mathcal{F}$ improves by 1.1% and 1.7% on Ref-DAVIS and ReasonVOS, respectively, while the corresponding SAM2 masks improve by 3.1% and 12.2%. These results show that VET-augmented video reasoning improves grounding beyond simple outputs, providing more effective guidance for downstream segmentation.

Training Components. Table 4 shows that SFT alone is insufficient for reliable VET grounding. CAGL-P improves target coverage in latent evidence, while CAGL-R makes attention scores in the fused maps more reliable for point selection. DALR further improves evidence aggregation over

Table 3: Effect of the paradigm.

Mask	Reason Type	$\mathcal{J}\&\mathcal{F}$	
		Ref-D	ReasV
Attn	w/o CoT	29.1	18.9
	w/ CoT	30.2	20.6
	Δ	**+1.1**	**+1.7**
SAM2	w/o CoT	77.6	53.9
	w/ CoT	80.7	66.1
	Δ	**+3.1**	**+12.2**

Table 4: Ablation of the training components. First three methods aggregate VET attention maps over six manually selected layers.

CAGL-P	CAGL-R	DALR	Attn Mask		SAM2 Mask	
			Ref-D	ReasV	Ref-D	ReasV
–	–	–	9.2	6.4	32.5	22.6
✓	–	–	29.5	**22.9**	76.4	61.5
✓	✓	–	29.9	21.0	79.3	64.9
✓	✓	✓	**30.2**	20.6	**80.7**	**66.1**

Table 5: Comparisons across more MLLM backbones.

Method	MLLM	Size	Ref-DAVIS			ReasonVOS			ReVOS (Overall)			ReVOS (Referring)			ReVOS (Reasoning)		
			$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
DecAF	InternVL3	8B	62.8	56.9	68.6	58.9	55.1	62.7	47.4	43.7	51.2	51.7	47.9	55.5	43.2	39.5	46.8
ReVEAL	InternVL3	8B	**77.8**	**74.2**	**81.4**	**63.6**	**60.9**	**66.3**	**57.4**	**54.7**	**60.1**	**62.1**	**59.6**	**64.6**	**52.7**	**49.8**	**55.6**
DecAF	Qwen2VL	7B	64.1	59.4	68.9	52.5	49.0	56.0	45.3	41.6	49.0	52.7	48.9	56.4	37.9	34.3	41.5
ReVEAL	Qwen2VL	7B	**74.9**	**71.3**	**78.5**	**61.2**	**58.4**	**64.1**	**55.9**	**53.2**	**58.6**	**60.7**	**58.0**	**63.3**	**51.1**	**48.4**	**53.8**
DecAF	Qwen2.5VL	3B	58.1	53.6	62.6	52.9	49.6	56.2	39.7	35.6	43.8	46.5	42.1	50.9	32.9	29.1	36.7
ReVEAL	Qwen2.5VL	3B	**74.5**	**70.2**	**78.8**	**54.0**	**51.3**	**56.8**	**51.6**	**48.6**	**54.7**	**56.9**	**54.1**	**59.6**	**46.4**	**43.0**	**49.7**
ReVEAL	Qwen3VL	4B	77.1	73.4	80.8	60.5	57.6	63.4	58.0	55.3	60.8	62.2	59.7	64.7	53.9	50.9	56.9
ReVEAL	Qwen3VL	8B	**80.7**	**77.3**	**84.0**	**66.1**	**62.9**	**69.2**	**61.6**	**58.9**	**64.3**	**65.0**	**62.5**	**67.5**	**58.3**	**55.4**	**61.1**

Table 6: Ablation across different segmentation backbones.

Model	Ref-D	ReasV	Ref-YT
SAM2-Hiera-T	79.1	64.0	67.5
SAM2-Hiera-S	78.7	64.3	69.0
SAM2-Hiera-B	79.9	67.0	70.3
SAM2-Hiera-L	**80.7**	66.1	71.3
SAM3	79.6	**67.6**	**71.4**

Table 7: Inference-time ablation of different VET attention map fusion strategies. Here, we report the $\mathcal{J}\&\mathcal{F}$ metric.

MLLM	<OBJ>	<BG>	<NEG>	Attn Mask		SAM2 Mask	
				Ref-D	ReasV	Ref-D	ReasV
Qwen3VL-8B	✓	–	–	27.2	18.8	78.4	63.2
	✓	–	✓	28.8	19.6	**80.7**	65.8
	✓	✓	–	**31.1**	**23.4**	78.5	64.4
	✓	✓	✓	30.2	20.6	**80.7**	**66.1**

fixed-layer selection schemes. Together, these components help ReVEAL elicit faithful role-aware visual evidence during video reasoning to guide segmentation.

VET Fusion at Inference. Table 7 compares different VET attention combinations from the same CoT at inference. Adding <BG> improves attention-mask scores by 3.9% and 4.6% on Ref-DAVIS and ReasonVOS, while adding <NEG> instead yields better SAM2 segmentation despite lower attention-mask scores. This suggests that background suppression refines the overall target coverage, whereas distractor suppression makes attention peaks more discriminative. Combining all three VET types yields the best overall performance, supporting the effectiveness of our role design.

MLLM Backbone Scalability. As shown in Table 5, ReVEAL consistently outperforms the baseline across diverse MLLM backbones, demonstrating its generalizability. We further observe that the performance generally improves with backbone size. For example, scaling Qwen3VL from 4B to 8B improves $\mathcal{J}\&\mathcal{F}$ by 3.6% on Ref-DAVIS and ReVOS and 5.6% on ReasonVOS, suggesting that larger MLLMs provide more reliable visual evidence for downstream segmentation.

Segmentation Backbone Scalability. ReVEAL uses latent visual evidence from MLLM reasoning to guide segmentation without requiring alignment to a specific segmenter. As shown in Table 6, the latent grounding evidence generalizes across smaller SAM2 variants and SAM3 (Carion et al., 2026) on all three benchmarks, with minor performance variations due to the segmenter's own capability.

Qualitative Analysis. We provide qualitative analysis of ReVEAL from two perspectives. First, Fig. 4 shows that ReVEAL reduces noise in the fused attention maps compared with DecAF, resulting in more precise focus on the target object and more accurate grounding for challenging queries

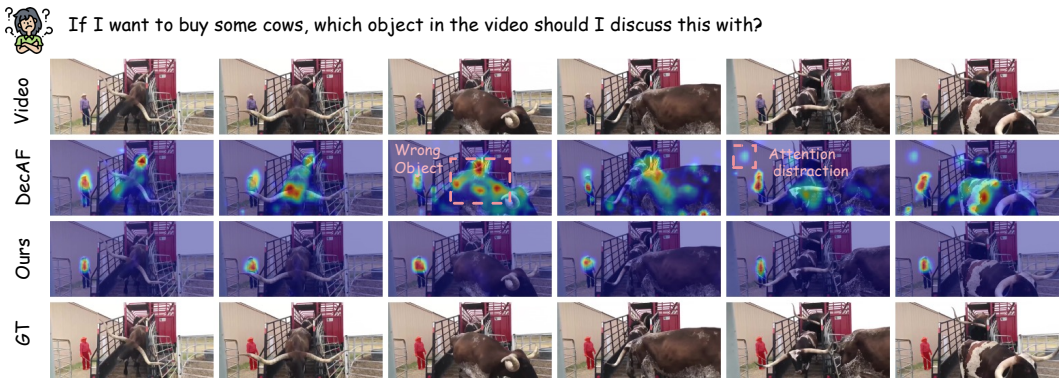

Figure 4: Qualitative comparison of ReVEAL and DecAF on a query requiring reasoning.

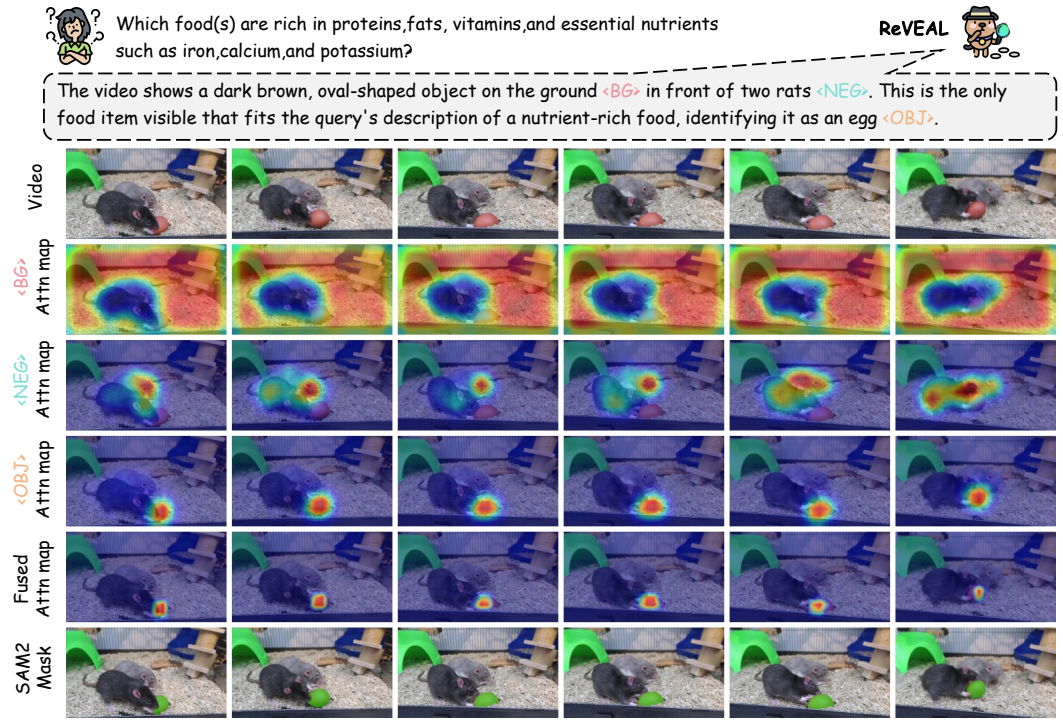

Figure 5: Visualization of ReVEAL's response and role-specific VET attention maps.

requiring implicit reasoning. Second, Fig. 5 shows that different VETs generally attend to regions consistent with their intended roles. Additional cases are provided in Appendix D.

5 Conclusion

We presented ReVEAL, a decoupled thinking-with-video framework for alignment-free ReVOS. ReVEAL enables the MLLM to identify targets through video reasoning and derive grounding cues from role-aware VET attention, without requiring cross-model alignment with the segmenter during training. At inference, ReVEAL derives role-specific evidence from VET-augmented video reasoning to guide a frozen SAM2 for target mask prediction. Experiments across five RVOS and ReVOS benchmarks demonstrate the effectiveness and generalizability of ReVEAL. Overall, our work highlights the potential of internal signals from MLLM reasoning to bridge high-level video understanding and dense visual prediction.

AI USE STATEMENT

We use generative AI tools in constructing the ReVEAL-47K training dataset, including video CoT annotation and formatting collected samples into VET-augmented CoT traces, as described in Section 3.1 and the Appendix. GPT is used to polish the manuscript, including language refinement and grammar checking of our original draft. We do not use AI tools for other research tasks requiring disclosure. We review all AI-assisted content and take full responsibility for the final paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

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

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

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9155–9164, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Andong Deng, Tongjia Chen, Shoubin Yu, Taojiannan Yang, Lincoln Spencer, Yapeng Tian, Ajmal Saeed Mian, Mohit Bansal, and Chen Chen. Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8625–8636, 2025.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Sitong Gong, Lu Zhang, Yunzhi Zhuge, Xu Jia, Pingping Zhang, and Huchuan Lu. Reinforcing video reasoning segmentation to think before it segments. *arXiv preprint arXiv:2508.11538*, 2025a.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Su Ho Han, Jeongseok Hyun, Pilhyeon Lee, Minho Shim, Dongyoon Wee, and Seon Joo Kim. Decomposed attention fusion in mllms for training-free video reasoning segmentation. *arXiv preprint arXiv:2510.19592*, 2025.

Ruina Hu, Chen Wang, Lai Wei, Jionghao Bai, Bin Yu, Weiran Huang, Kai Wang, and Yue Wang. Attend to evidence: Evidence-anchored spatial attention supervision for multimodal rlvr. *arXiv preprint arXiv:2605.30912*, 2026.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024b.

Jaehyun Jang, Eunseop Yoon, Hee Suk Yoon, SooHwan Eom, Mark A Hasegawa-Johnson, and Chang D Yoo. Token-level response-visual attention guidance for multimodal llms knowledge distillation. *arXiv preprint arXiv:2607.02593*, 2026.

Woojeong Jin, Seongchan Kim, Jaeho Lee, and Seungryong Kim. Interrvos: Interaction-aware referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10367–10376, 2026.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9339–9350, 2025.

Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Cot-rvs: Zero-shot chain-of-thought reasoning segmentation for videos. *arXiv preprint arXiv:2505.18561*, 2025.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian conference on computer vision*, pp. 123–141. Springer, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026.

Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 48–58, 2025.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21766–21777, 2025.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. *arXiv preprint arXiv:2509.18094*, 2025a.

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*, 2025b.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

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

Bo Miao, Mohammed Bennamoun, Yongsheng Gao, Mubarak Shah, and Ajmal Mian. Temporally consistent referring video object segmentation with hybrid memory. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11373–11385, 2024.

OpenAI. OpenAI-o3. `https://openai.com/index/introducing-o3-and-o4-mini/`, 2025.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26364–26373. IEEE, 2024b.

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

Yongyi Su, Haojie Zhang, Shijie Li, Nanqing Liu, Jingyi Liao, Junyi Pan, Yuan Liu, Xiaofen Xing, Chong Sun, Chen Li, Nancy F. Chen, Shuicheng Yan, Xulei Yang, and Xun Xu. Patch-as-decodable-token: Towards unified multi-modal vision tasks in mllms, 2025. URL `https://arxiv.org/abs/2510.01954`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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. *arXiv preprint arXiv:2507.19599*, 2025a.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025b.

Haochen Wang, Yuhao Wang, Tao Zhang, Yikang Zhou, Yanwei Li, Jiacong Wang, Jiani Zheng, Ye Tian, Jiahao Meng, Zilong Huang, et al. Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. *arXiv preprint arXiv:2510.18876*, 2025c.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24710–24721, 2025.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision (ECCV)*, 2024.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025a.

Haobo Yuan, Yueyi Sun, Yanwei Li, Tao Zhang, Xueqing Deng, Henghui Ding, Lu Qi, Anran Wang, Xiangtai Li, and Ming-Hsuan Yang. Visual reasoning tracer: Object-level grounded reasoning benchmark. *arXiv preprint arXiv:2512.05091*, 2025b.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18970–18980, 2025c.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 2023.

Jinglei Zhang, Yuanfan Guo, Rolandos Alexandros Potamias, Jiankang Deng, Hang Xu, and Chao Ma. Vtimecot: Thinking by drawing for video temporal grounding and reasoning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24203–24213. IEEE, 2025.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024a.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient finetuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024b.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *Advances in Neural Information Processing Systems*, 38:45266–45297, 2026.

Yikang Zhou, Tao Zhang, Dengxian Gong, Yuanzheng Wu, Ye Tian, Haochen Wang, Haobo Yuan, Jiacong Wang, Lu Qi, Hao Fei, et al. Samtok: Representing any mask with two words. *arXiv preprint arXiv:2601.16093*, 2026.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

APPENDIX OVERVIEW

This appendix provides supplementary details of ReVEAL.

A REVEAL-47K TRAINING DATASET 16
A.1 DATASET OVERVIEW 16
A.2 ANNOTATION PROMPTS 16

B MORE DETAILS OF REVEAL 18
B.1 DYNAMIC ATTENTION LAYER ROUTER 18
B.2 REVEAL SYSTEM PROMPT 19
B.3 TRAINING AND INFERENCE SETTINGS 19

C ADDITIONAL EXPERIMENTAL ANALYSIS 19
C.1 ADDITIONAL QUANTITATIVE RESULTS 19
C.2 ADDITIONAL ABLATION STUDIES 20

D MORE VISUALIZATIONS 22
D.1 MORE COMPARISONS WITH OTHER METHODS 22
D.2 MORE ROLE-TOKEN ATTENTION MAP VISUALIZATIONS 22
D.3 FAILURE CASE ANALYSIS AND LIMITATIONS 23

A REVEAL-47K TRAINING DATASET

A.1 DATASET OVERVIEW

To equip the model with dynamic role-aware VET reasoning and reliable latent visual grounding, we construct ReVEAL-47K training dataset, comprising 47K VET-augmented ReVOS CoT traces paired with target and background masks for attention grounding.

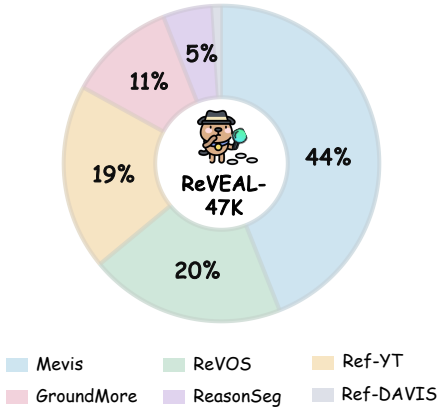

Figure 6: ReVEAL-47K Composition.

ReVEAL-47K is constructed from the training splits of five public RVOS and ReVOS datasets, including MeViS (Ding et al., 2023), ReVOS (Yan et al., 2024), GroundMoRe (Deng et al., 2025), Ref-DAVIS (Khoreva et al., 2018), and Ref-YouTube-VOS (Seo et al., 2020), together with the image reasoning segmentation dataset ReasonSeg (Lai et al., 2024), as shown in Fig. 6. We use six frames as the model input for all samples. For ReasonSeg, each sample is included twice, with its single image repeated across six frames and treated as a static video. The resulting dataset contains 47K VET-augmented CoT annotations. Each trace contains exactly one <OBJ> token for target evidence, even when the query refers to multiple target objects, as its attention map can cover multiple target regions, while suppressive VETs are introduced only when needed. Specifically, <BG> appears at most once to provide background evidence, whereas multiple <NEG> tokens can be used to distinguish different distractors. On average, each trace contains 1.00 <OBJ>, 0.42 <BG>, and 0.86 <NEG> tokens.

A.2 ANNOTATION PROMPTS

Stage 1: Mask-conditioned Target Description. We use GAR (Wang et al., 2025c) with the official inference setup to generate a mask-conditioned caption of the target object(s).

Stage 2: Original Video CoT Generation. This prompt is used for Gemini2.5 Pro (Comanici et al., 2025) to produce a naive CoT for the ReVOS query based on the video content:

Original Video CoT Generation Prompt

You are annotating video object question answering data.

Inputs

1. A set of video frames.
2. A referring query.
3. A detailed description of the target object(s).

Task

Write one concise reasoning paragraph explaining how the target described by the query can be identified and localized based on visual evidence in the video. Focus on observable cues such as appearance, position, movement, persistence across frames, and spatial relations to visible scene elements. If the target is trivially identifiable because it is unique, salient, and has no competing objects, one sentence naming the target is enough. Otherwise, include only the minimum cues needed to remove ambiguity, choosing from: background regions to ignore, concrete distractors that could be confused with the target, visible reference objects, temporal order, motion trajectory, persistence, or key-frame changes. Use only cues that are actually visible and helpful for this case. Do not pad, do not list absent elements, and do not invent distractors or references.

Constraints

- The target must follow the referring query exactly; do not re-interpret or substitute it.
- Keep the paragraph concise: 2–5 sentences, information-dense.
- Output plain text only. No JSON, no coordinates, no bounding boxes.

Stage 3: VET-Augmented Reasoning Trace Construction. Given the naive CoT traces, we prompt Gemini2.5 Pro (Comanici et al., 2025) to refine the reasoning by identifying key objects, background, and distractors, and inserting the corresponding role-specific VETs into the trace:

VET-Augmented Reasoning Trace Construction Prompt

You are rewriting reasoning traces for video object discovery annotations.

Inputs

1. A reasoning query.
2. An existing reasoning paragraph of the video.

Task

Rewrite the reasoning paragraph into one concise, natural visual reasoning trace that answers the reasoning query. Use the existing reasoning paragraph as evidence, but make the rewritten output focus on identifying and localizing the object(s) asked for by the query in the video. Do not merely polish the original text or insert tokens mechanically. The reasoning should proceed from visible video cues to the final answer: first explain which visual evidence supports the queried object(s), then conclude with the target object(s) marked by <OBJ>.

Visual Evidence Tokens

Insert Visual Evidence Tokens immediately after the phrase they label:

- <BG>: background scene regions to ignore, not foreground objects.
- <NEG>: concrete non-target foreground objects that could be confused with the target.
- <OBJ>: the final target object(s) to segment.

Reasoning Order

Prefer this order when applicable:

1. Interpret the query by focusing on the object(s) it asks to segment.
2. Use observable video cues to identify and localize that object(s), such as appearance, position, persistence across frames, motion, relation to stable scene anchors, or temporal changes.
3. Add <BG> only for background scene regions that help exclude irrelevant areas.
4. Add <NEG> only for visible non-target foreground objects that could plausibly be confused with the queried object.
5. End with the target that answers the query, marked once with <OBJ>.

The paragraph should provide a video-grounded response to the query, rather than a static image description or a list of labels.

Token Placement Rule

Each Visual Evidence Token must appear immediately after the object or region phrase it labels. Correct: the driveway <BG>, the soldier <NEG>, the man on the left <OBJ>.

Editing Rules

- The rewritten paragraph must answer the reasoning query, not just restate the existing paragraph.
- The object(s) marked with <OBJ> must be the final answer to the query.
- Any reasoning before <OBJ> should explain why that object(s), rather than another visible object(s) or region, satisfies the query.
- Do not mark an object(s) with <OBJ> only because it is visually salient; it must be the object(s) asked for by the query.
- Do not add new visual objects, distractors, references, background regions, actions, or attributes unless they are already stated or clearly implied by the reasoning paragraph.
- If the existing paragraph contains irrelevant details that do not help answer the query, omit them.
- Improve the reasoning so it sounds like it is derived from video evidence, not like a labeling checklist.
- Do not use meta-descriptions such as “initial context is established,” “the label matches,” or “this indicates the target.”
- <OBJ> is mandatory and must appear exactly once, attached to the final target.
- <BG> is optional and may appear at most once. <NEG> is optional and may appear multiple times, once for each distinct distractor. Use them only when useful and visually grounded.
- Use <BG> only for scene or background regions, never foreground objects.
- Use <NEG> only for concrete non-target foreground objects that could be mistaken for the target.
- Do not attach tokens to verbs, actions, relations, attributes, or abstract descriptions.

- Keep the output concise, fluent, and grounded in the video.
- Output only the rewritten reasoning paragraph. No JSON, no bullets, no headings, no explanation.

Stage 4: Quality control. Here is the prompt we use to filter invalid CoT annotations:

Quality-Control Prompt

You are validating video object reasoning annotations.

Inputs

1. Six uniformly sampled video frames.
2. A referring query.
3. A target description.
4. A candidate reasoning trace containing Visual Evidence Tokens.

Visual Evidence Tokens

- <OBJ>: the final target object.
- <NEG>: visible distractors or non-target objects.
- <BG>: safe background or regions to ignore.

Task

Determine whether the reasoning trace correctly identifies and localizes the target object(s). Use the referring query as the primary criterion and the target description only as supplementary guidance. Base the judgment solely on the visible frames.

Answer yes only if

- the target (s) is visible in the six frames;
- the trace follows the query and matches the target description;
- the trace marks the correct final target (s);
- any <BG> or <NEG> Visual Evidence Tokens are visually grounded and correctly used;
- the reasoning is clear, specific, and free of substantial irrelevant or redundant content.

Answer no if

- the target (s) is not visible in the six frames;
- the trace targets the wrong object or conflicts with the query;
- the trace conflicts with the target description;
- the trace contradicts the frames or invents key evidence;
- the trace is too ambiguous to localize the target;
- the trace contains substantial irrelevant, redundant, or generic content.

Constraints

Output exactly one lowercase word: yes or no.

B More Details of ReVEAL

B.1 Dynamic Attention Layer Router

Since target identification is central to ReVOS, DALR uses <OBJ> attention as the routing cue to select layers that provide informative target evidence. As these layers vary across inputs, DALR adaptively selects and weights candidate layers for each sample. The same layer selection and routing weights are applied to all VET roles, ensuring efficient aggregation of role-specific attention maps. As detailed in Algorithm 1, DALR employs noisy top-k routing during training to encourage exploration and deterministic routing at inference. We use $L_c = 14$ candidate layers for Qwen2.5VL-7B and $L_c = 18$ for Qwen3VL-8B, with a feature dimension of 256 and $k = 6$ by default.

Algorithm 1 Dynamic Attention Layer Router

Require: <OBJ> attention maps $a^{\texttt{<OBJ>},1}$ from candidate layers $\mathcal{L}_c$, feature dimension D, and selection size k

$\mathbf{X} \leftarrow \mathrm{ClampMin}\big(\mathrm{Flatten}(a^{\texttt{<OBJ>},1}), 0\big)$
$\mathbf{X} \leftarrow \mathrm{Normalize}(\mathbf{X})$
$\mathbf{F} \leftarrow \mathrm{AdaptiveAvgPool1D}(\mathbf{X}, D)$
$\mathbf{F} \leftarrow \mathrm{Normalize}(\mathbf{F})$
$\mathbf{F} \leftarrow \mathrm{StopGradient}(\mathbf{F})$
$\mathbf{s} \leftarrow g_{\mathrm{score}}(\mathbf{F})$
if training **then**
    $\boldsymbol{\sigma} \leftarrow \mathrm{Softplus}\big(g_{\mathrm{noise}}(\mathbf{F})\big)$
    $\tilde{\mathbf{s}} \leftarrow \mathbf{s} + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
else
    $\tilde{\mathbf{s}} \leftarrow \mathbf{s}$
end if
$\mathcal{S} \leftarrow \mathrm{TopK}_{l \in \mathcal{L}_c}(\tilde{s}_l, k)$
$w_l \leftarrow \dfrac{\exp(\tilde{s}_l)}{\sum_{m \in \mathcal{S}} \exp(\tilde{s}_m)}, \quad \forall l \in \mathcal{S}$
return selected layers $\mathcal{S}$ and routing weights $\{w_l\}_{l \in \mathcal{S}}$

B.2 ReVEAL System Prompt

We use the following system prompt during both training and inference:

ReVEAL System Prompt

You are a video object reasoner.

Task
Given a set of video frames and a referring query, your job is to answer the query by locating the target object(s) in the video. Output a reasoning paragraph that answers the query and contains the following three special tokens.

Visual Evidence Tokens

- <BG>: background regions to ignore, not foreground objects.
- <NEG>: distractor objects that look similar to the target.
- <OBJ>: the final target to segment.

Output
<BG> is optional and may appear at most once. <NEG> is optional and may appear multiple times, once for each distinct distractor; these tokens may appear in any order. End with a sentence that directly answers the query and names the target with <OBJ>. Output only the paragraph: no lists, coordinates, masks, or JSON.

B.3 Training and Inference Settings

Table 8 and 9 provide key hyperparameters in training and evaluating ReVEAL-Qwen3VL-8B.

C Additional Experimental Analysis

By default, results are reported with ReVEAL-Qwen3VL-8B and the $\mathcal{J}\&\mathcal{F}$ metric.

C.1 Additional Quantitative Results

Inference Efficiency. We compare the efficiency of MLLM-side attention production of ReVEAL with the baseline DecAF (Han et al., 2025) on Ref-DAVIS using the same Qwen2.5VL-7B. As shown in Table 10, ReVEAL requires only one MLLM response per query, compared with 35 for DecAF, and achieves over $4\times$ faster inference. Measurement is conducted on an NVIDIA H800 GPU.

Table 8: Training settings.

Hyperparameter	Value
Random Seed	42
Trainable Modules	MLLM, DALR
Optimizer	AdamW
Epoch	1
Learning Rate	2×10^{-5}
Weight Decay	0.05
LR Schedule	Cosine
Global Batch Size	16
Frame Max Pixels	$336 \times 32 \times 32$
Precision	bfloat16
DALR Top-k	6
CAGL Weight λ_{CAGL}	0.2
CAGL-R Weight λ_{rep}	0.1

Table 9: Inference settings.

Hyperparameter	Value
<BG> Sup Weight $\alpha_{\text{<BG>}}$	0.1
<NEG> Sup Weight $\alpha_{\text{<NEG>}}$	1.0
Max Point Prompts	3
Attention Threshold τ_{pq}	0.8
NMS IoU Threshold τ_{nms}	0.7
Tracklet-consistency Threshold τ_{trk}	0.8

Table 10: Per-query MLLM inference time on Ref-DAVIS using the same Qwen2.5-VL-7B.

Method	MLLM calls	Avg. time (s)
DecAF	35	23.5
ReVEAL (Ours)	**1**	**5.8** (4.1×)

Table 11: Comparison on GroundMoRe.

Method	Overall			Causal			Sequential			Counterfactual			Descriptive		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
None-MLLM RVOS Methods															
ReferFormer (Wu et al., 2022)	12.7	11.2	14.3	12.7	11.0	14.4	10.7	9.4	12.1	12.4	11.1	13.6	15.2	13.3	17.0
SgMg (Miao et al., 2023)	17.5	15.8	19.2	18.9	17.0	20.9	15.5	14.0	17.0	16.4	15.2	17.5	18.8	16.7	20.9
HTR (Miao et al., 2024)	14.5	12.7	16.2	15.7	13.8	17.5	12.5	10.8	14.1	13.5	12.2	14.9	15.9	13.8	18.1
LMPM (Ding et al., 2023)	13.3	12.7	14.0	13.7	13.1	14.3	11.3	10.6	12.0	13.2	12.6	13.7	14.3	13.6	15.1
Image Reasoning Segmentation Methods															
LISA (Lai et al., 2024)	6.1	4.9	7.3	5.5	4.4	6.7	5.2	4.1	6.3	5.9	4.8	7.0	7.6	6.1	9.0
PixelLM-7B (Ren et al., 2024b)	10.0	9.9	10.0	9.5	9.4	9.7	9.2	9.2	9.3	9.8	9.8	9.9	10.9	11.0	10.9
MLLM-Based ReVOS Methods															
VISA-7B (Yan et al., 2024)	5.3	4.7	5.9	6.1	5.5	6.6	4.5	3.8	5.2	5.1	4.3	5.9	5.4	5.0	5.8
MORA (Deng et al., 2025)	27.2	27.4	26.9	27.4	26.9	28.0	26.1	27.2	24.9	25.6	25.3	25.9	29.6	30.2	28.9
ReVEAL (Ours)	**37.1**	**35.8**	**38.4**	**37.7**	**36.4**	**39.1**	**34.0**	**32.7**	**35.3**	**31.5**	**30.3**	**32.7**	**44.8**	**43.3**	**46.2**

Motion-grounded Video Reasoning Performance. GroundMoRe (Deng et al., 2025) is a challenging benchmark that requires joint spatio-temporal reasoning to understand object motions and interactions from causal, sequential, counterfactual, and descriptive perspectives. The results in Table 11 show that ReVEAL achieves an overall $\mathcal{J}\&\mathcal{F}$ result of 37.1%, outperforming MORA (27.2%) by 9.9%. Notably, ReVEAL delivers substantial gains on both Causal and Descriptive questions, improving the scores from 27.4% to 37.7% and from 29.6% to 44.8%, respectively. These gains suggest that our thinking-with-video paradigm facilitates causal reasoning, while attention-derived latent evidence provides a video-wide perspective for tracking moving targets more accurately.

C.2 Additional Ablation Studies

Sensitivity to the CAGL Weights. Table 12 evaluates the sensitivity of CAGL to its loss weights. Overall, ReVEAL maintains relatively stable performance across different weight configurations, indicating robustness to these hyperparameters. We adopt $\lambda_{\text{CAGL}} = 0.2$ and $\lambda_{\text{rep}} = 0.1$, which achieves the best performance on both benchmarks.

VET Fusion Strategies on Qwen2.5VL-7B. Table 13 extends the VET fusion ablation to Qwen2.5VL-7B. Unlike Qwen3VL-8B, this backbone benefits most from background suppression, which improves attention-derived and SAM2 mask $\mathcal{J}\&\mathcal{F}$ by 4.6% and 1.9% on ReasonVOS, respectively. Fusing all three VET types achieves the best overall performance on both benchmarks, supporting the effectiveness of our role design and fusion strategy across MLLM backbones.

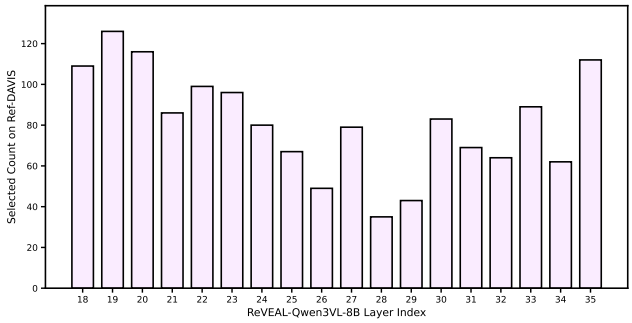

Figure 7: Layer selection distribution of DALR in ReVEAL-Qwen3VL-8B on Ref-DAVIS.

Table 12: λ_{CAGL} and λ_{rep} ablation.

λ_{CAGL}	λ_{rep}	SAM2 $\mathcal{J}\&\mathcal{F}$	
		Ref-D	ReasV
0.1	0.1	79.4	62.5
0.2	**0.1**	**80.7**	**66.1**
0.3	0.1	79.0	65.9
0.2	0.05	76.8	65.3
0.2	0.15	80.3	66.0

Table 13: Inference-time ablation of different VET combination strategies using ReVEAL-Qwen2.5VL-7B. ReVEAL uses all three types of VETs during evaluation.

MLLM	<OBJ>	<BG>	<NEG>	Attn Mask		SAM2 Mask	
				Ref-D	ReasV	Ref-D	ReasV
Qwen2.5-VL-7B	✓	–	–	24.8	17.2	77.3	62.2
	✓	–	✓	25.6	17.8	77.0	63.5
	✓	✓	–	28.6	21.8	77.4	64.1
	✓	✓	✓	**29.4**	**22.1**	**77.7**	**65.1**

DALR Top-k Ablation. We evaluate DALR with different top-k settings, each used during both training and inference. As shown in Table 14, selecting $k = 6$ layers achieves the best overall performance across the two benchmarks. Selecting more layers provides no consistent gains and may introduce additional noise into the grounding evidence.

DALR Selection Distribution. Figure 7 shows DALR's layer selection distribution on Ref-DAVIS with $k = 6$. The selected layer combinations vary across inputs, indicating input-dependent selection rather than reliance on a fixed subset.

Sensitivity to the Background Suppression Weight. Figure 8 analyzes the effect of the background suppression weight $\alpha_{\text{<BG>}}$ on attention-derived and SAM2 mask quality. Since <BG> attention often covers broad non- foreground regions, it helps suppress diffuse or spurious object attention. The required suppression strength varies across backbones due to differences in their object attention patterns. ReVEAL-Qwen3VL-8B and ReVEAL-Qwen2.5VL-7B achieve the best SAM2 performance at $\alpha_{\text{<BG>}} = 0.1$ and 0.5, respectively. While stronger suppression generally improves attention-derived masks, overly large weights can also remove useful target evidence and degrade final segmentation.

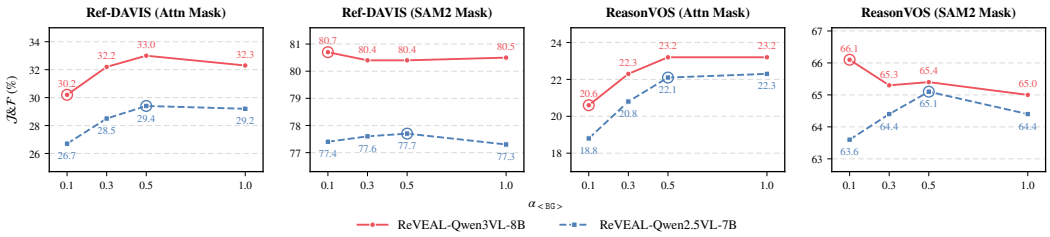

Figure 8: Effect of background suppression weight $\alpha_{\text{<BG>}}$ across two backbones.

Number of Sampled Point Prompts. Table 15 studies the maximum number of points sampled from the fused attention map as visual prompts for SAM2 at inference time. Across all three benchmarks, increasing the number of sampled points yields no consistent performance gains. The default setting provides the best overall trade-off between mask quality and prompt efficiency.

Table 14: DALR top-k ablation.

Router top-k	SAM2 $\mathcal{J}\&\mathcal{F}$	
	Ref-D	ReasV
$k = 4$	78.4	65.9
k = 6	**80.7**	66.1
$k = 8$	79.5	64.7
$k = 10$	80.1	64.2
$k = 12$	79.7	**66.6**

Table 15: Maximum points ablation.

Max Points	Ref-D	ReasV	Ref-YT
2	**81.0**	65.9	71.3
3	80.7	**66.1**	**71.3**
5	80.6	66.0	71.3
10	80.6	**66.1**	71.3

D More Visualizations

This section provides additional qualitative results of ReVEAL, including comparisons with other methods, more role-specific visualizations, and failure cases.

D.1 More Comparisons with Other Methods

Figure 9 compares a representative case from ReasonVOS between ReVEAL and the Without-CoT paradigm. In this example, ReVEAL improves attention-mask $\mathcal{J}\&\mathcal{F}$ by only 1.7% over Without-CoT, yet its VET attention consistently favors the true target despite some hallucinated responses, whereas Without-CoT favors a visually similar distractor. Attention-mask evaluation reflects the spatiotemporal coverage of regions retained after thresholding, whereas SAM2 uses point prompts sampled from high-confidence attention locations. A higher attention-mask score does not necessarily indicate more confident attention peaks on the target. Stronger attention peaks on the true target may therefore yield more reliable point prompts, helping explain the larger gain in SAM2 segmentation despite the modest improvement in attention-mask quality.

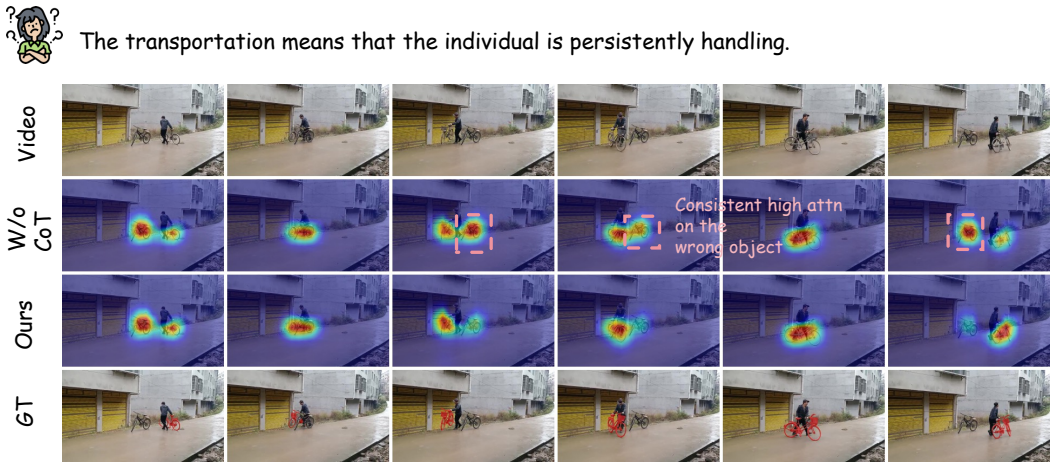

Figure 9: Qualitative comparison between ReVEAL and the Without-CoT paradigm on ReasonVOS.

Figure 10 presents additional qualitative comparisons of the fused attention maps produced by ReVEAL and DecAF (Han et al., 2025). In dynamic scenes requiring target tracking across frames, ReVEAL maintains consistent and clear attention on the target, whereas DecAF tends to produce spurious activations and lose focus on the target in some cases.

D.2 More Role-Token Attention Map Visualizations

Figure 11, Figure 12, and Figure 13 provide additional CoT responses and attention maps for the role-specific VETs, following Fig. 5. ReVEAL dynamically uses suppressive VETs (<BG> and <NEG>) according to the reasoning content, with their presence and number varying across examples. The CoT reasoning itself establishes semantic associations between visual entities and their corresponding VETs, enabling individual VETs to attend to relevant regions to some extent. This lightweight

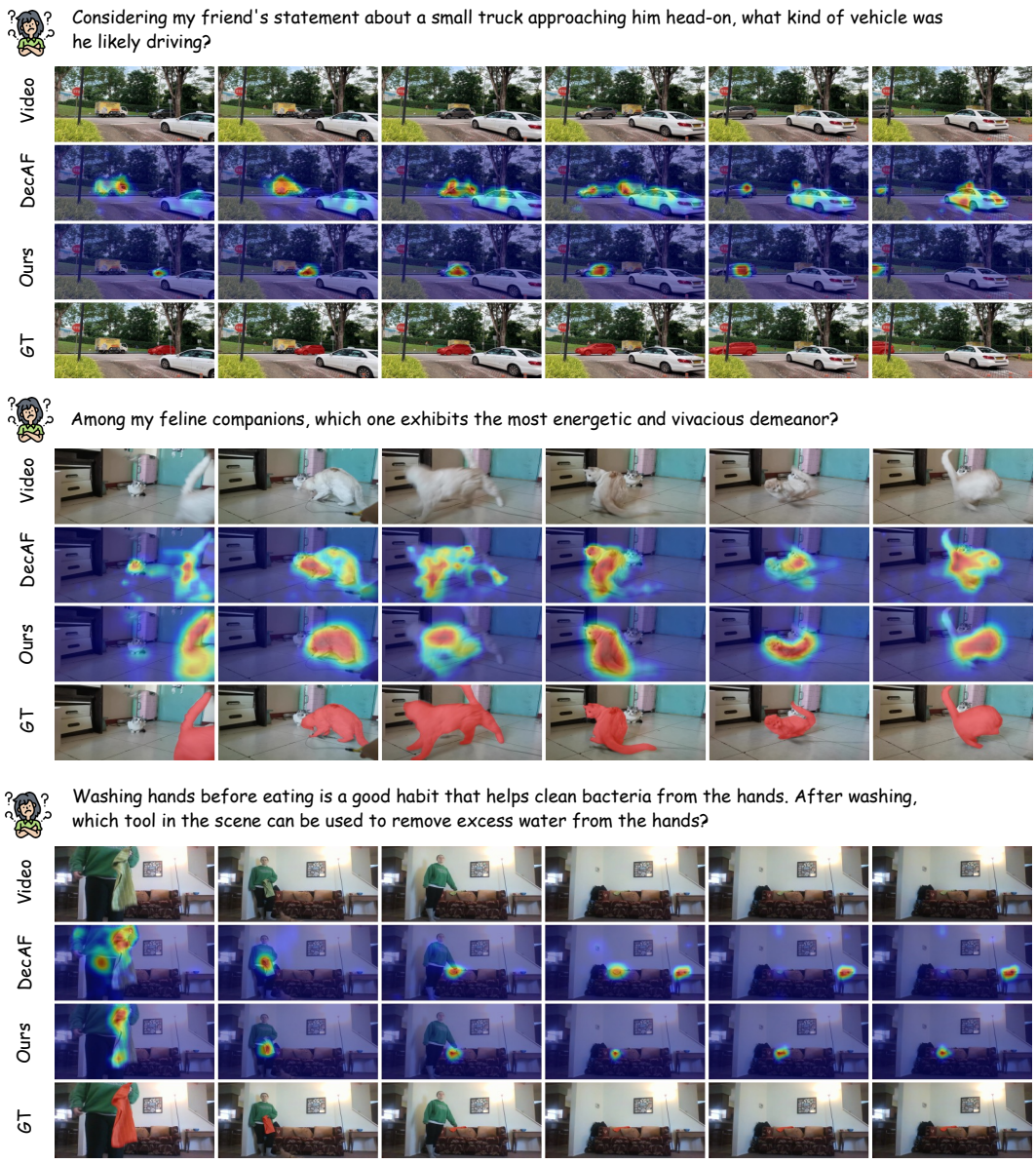

Figure 10: Additional qualitative comparisons between ReVEAL and DecAF.

association can reduce the reliance on precise region-level supervision for every <NEG> occurrence, which would otherwise incur additional annotation cost. As shown in Figure 12 and 13, different <NEG> occurrences tend to highlight distinct distractor regions.

D.3 Failure Case Analysis and Limitations

Failure cases. Figure 14, 15, and 16 show representative failures caused by limited target visibility and instance-level ambiguity. When the target is fully occluded, has not yet appeared, or is visible only briefly, limited visual and temporal context may lead to spurious attention activations in irrelevant regions. When multiple visually similar objects are present, the MLLM may struggle to identify the referred target during CoT reasoning, causing VET attention to focus on incorrect instances.

Limitations. Despite its effectiveness, ReVEAL has several limitations. First, the current <NEG> design may not adequately distinguish individual distractors when multiple visually similar objects

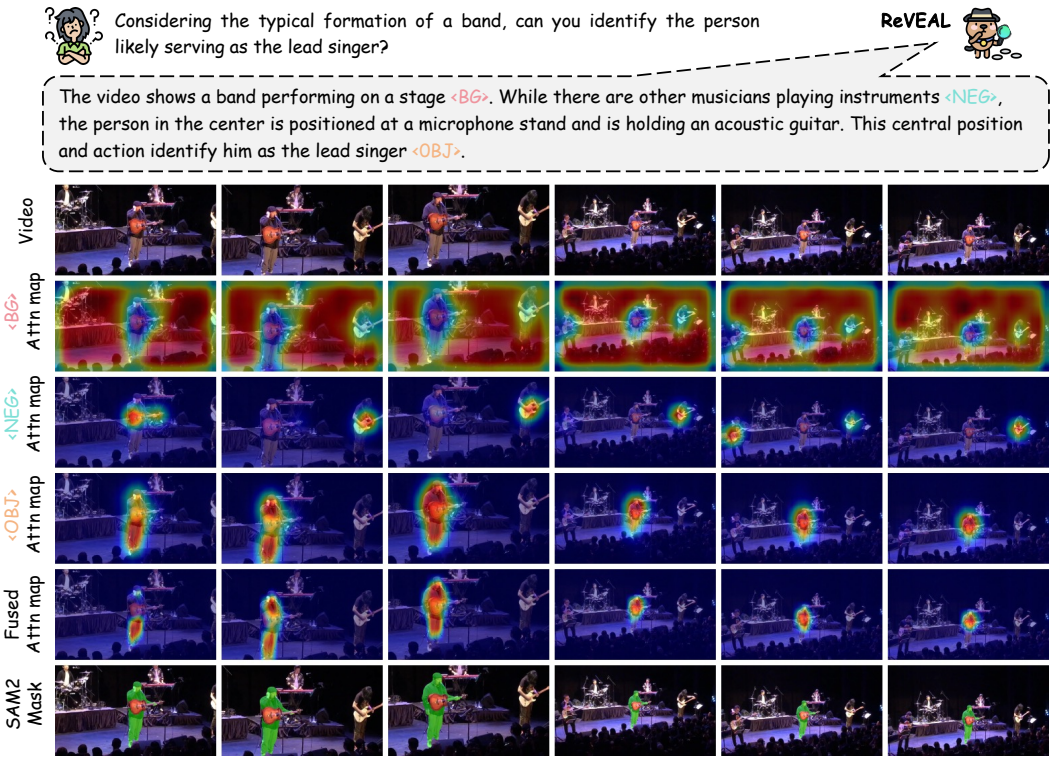

Figure 11: Additional qualitative result where all three VET roles are generated.

appear in more complex scenes. Second, VET grounding can become unreliable when MLLM attention is diffuse or focused on irrelevant regions, particularly under severe occlusion. Third, the current uniform sampling strategy may miss frames containing key visual entities, leaving them absent from the model input. The resulting lack of visual context may limit the model's ability to reason effectively and produce reliable latent visual evidence. Future work could explore lightweight optimizations centered on the MLLM's own video reasoning for ReVOS. For example, a frame selection module inspired by GLUS (Lin et al., 2025) and CoT-RVS (Kao et al., 2025) could help retain informative frames. More broadly, we will explore extending this efficient MLLM-centric design to omni scenarios.

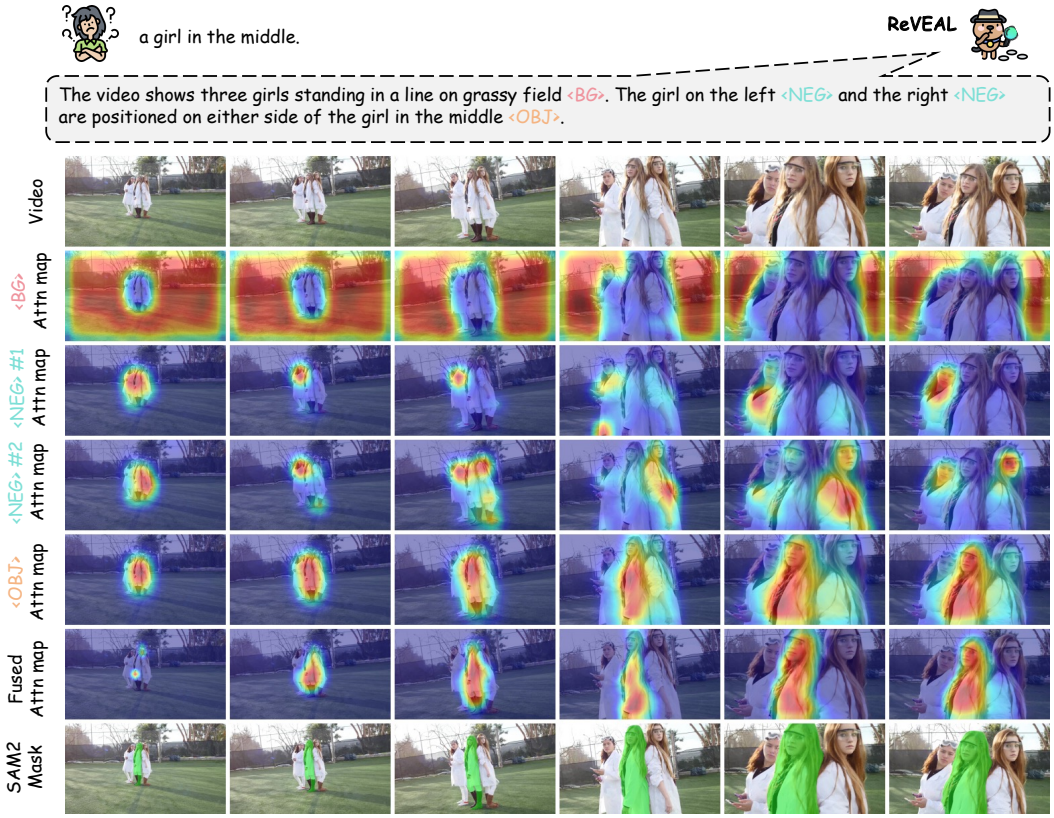

Figure 12: Qualitative result where multiple <NEG> tokens are generated to indicate different distactors.

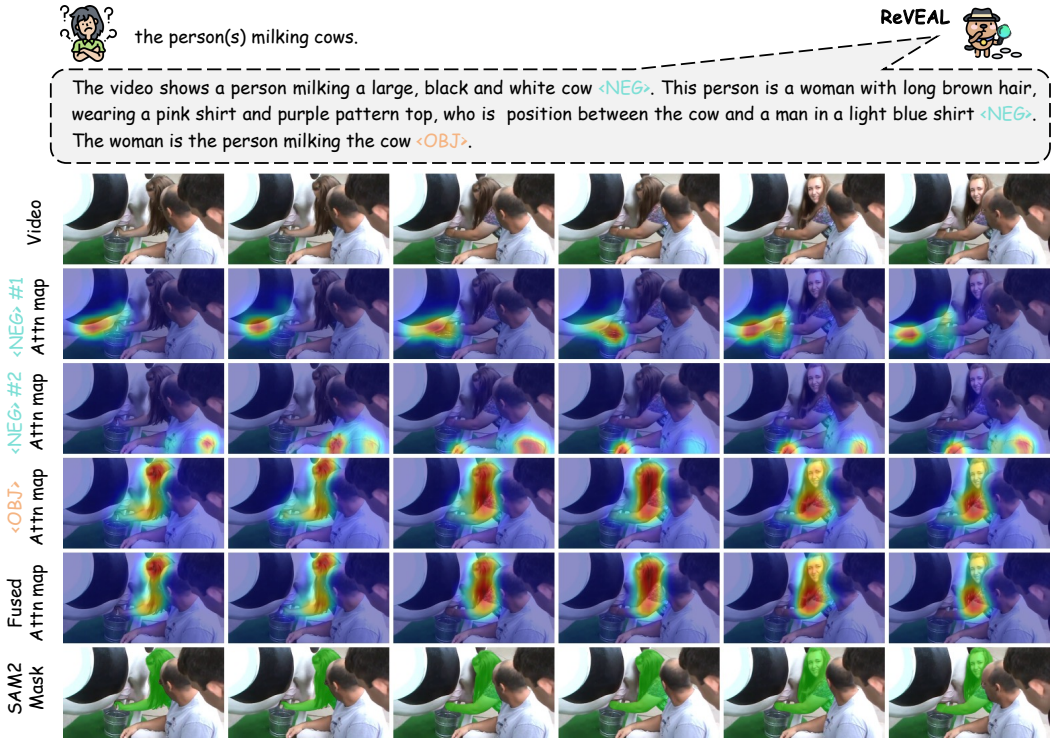

Figure 13: Qualitative result where only two type of VET roles are generated.

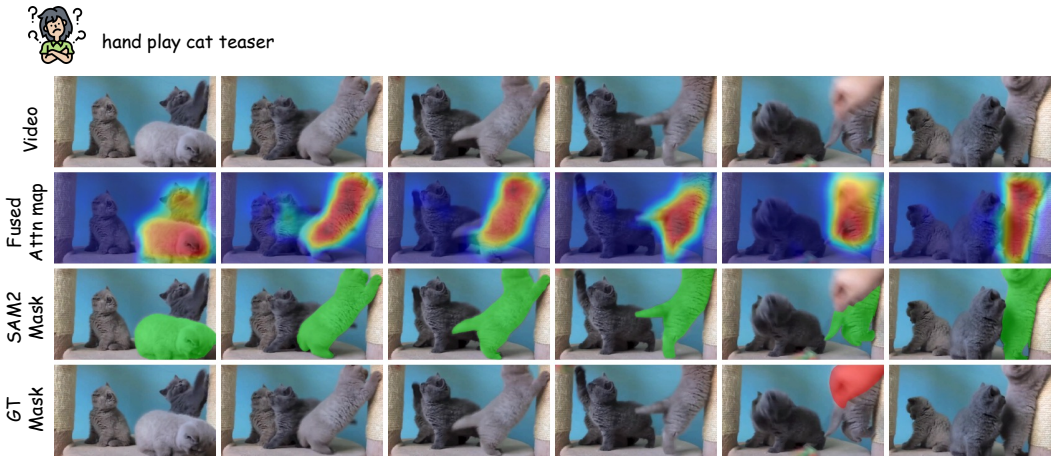

Figure 14: Representative failure case on MeViS_u, where the fast-moving hand appears in too few of the six uniformly sampled input frames to provide sufficient temporal and visual context.

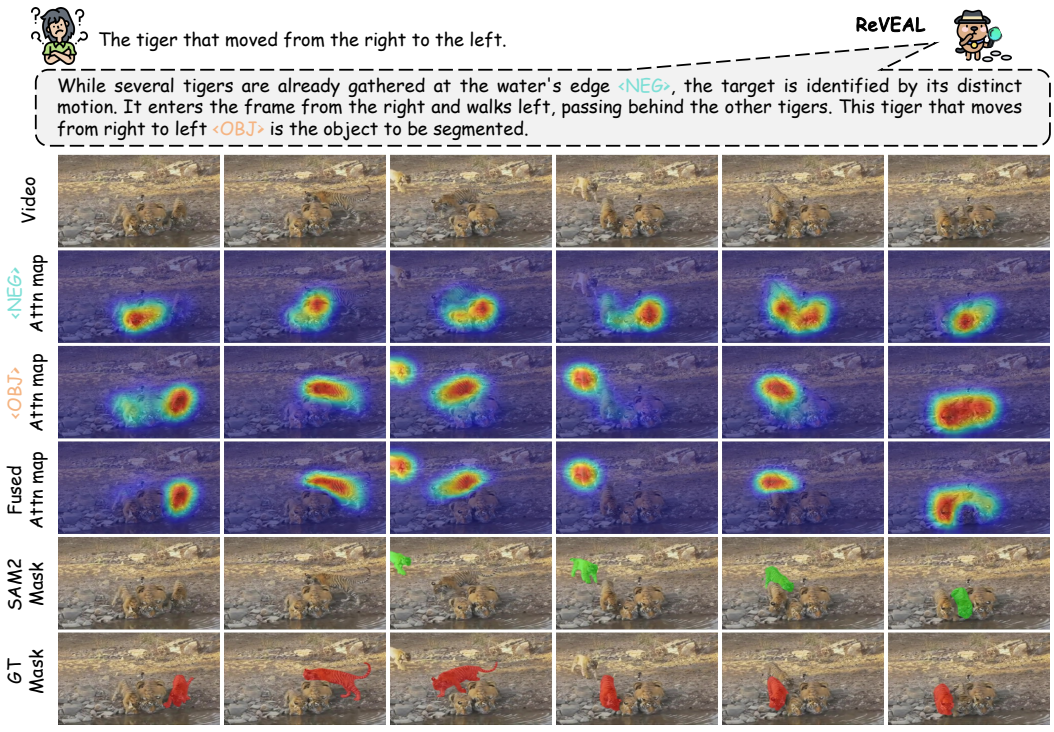

Figure 15: Representative failure case where the VET attention maps fail to distinguish the target among a group of similar tigers.



Figure 16: Representative failure case where ReVEAL fails to distinguish the target from textual reasoning when numerous visually similar distractors are present.