STEERSEG: ATTENTION STEERING FOR REASONING VIDEO SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video reasoning segmentation requires localizing objects across video frames from natural language expressions, often involving spatial reasoning and implicit references. Recent approaches leverage frozen large vision-language models (LVLMs) by extracting attention maps and using them as spatial priors for segmentation, enabling training-free grounding. However, these attention maps are optimized for text generation rather than spatial localization, often resulting in diffuse and ambiguous grounding signals. In this work, we introduce **SteerSeg**, a lightweight framework that identifies attention misalignment as the key bottleneck in attention-based grounding and proposes to *steer attention at its source* through input-level conditioning. SteerSeg combines learnable soft prompts with reasoning-guided Chain-of-Thought (CoT) prompting. The soft prompts reshape the attention distribution to produce more spatially concentrated maps, while CoT-derived attributes resolve ambiguity among similar objects by guiding attention toward the correct instance. The resulting attention maps are converted into point prompts across keyframes to guide a segmentation model, while candidate tracklets are ranked and selected using correlation-based scoring. Our approach freezes the LVLM and segmentation model parameters and learns only a small set of soft prompts, preserving the model's pretrained reasoning capabilities while significantly improving grounding. Despite being trained only on Ref-YouTube-VOS, SteerSeg generalizes well across diverse benchmarks, significantly improving the spatial grounding capability of LVLMs.

1 INTRODUCTION

Video reasoning segmentation aims to localize and segment objects in video frames given natural language queries requiring reasoning over spatial, temporal, and semantic relationships. While some queries can be resolved from direct visual cues, more complex expressions, such as *"the object the child is reaching for"*, require understanding interactions and intent over time. These scenarios demand models that can both interpret language and reason about visual content consistently across frames, making video reasoning segmentation a challenging testbed for evaluating visual reasoning capabilities. Addressing this problem naturally calls for large vision-language models (LVLMs), which capture rich semantic and contextual relationships across modalities Khoreva et al. (2018); Seo et al. (2020); Ding et al. (2023); Yan et al. (2024); Bai et al. (2024).

A common approach couples LVLMs with segmentation foundation models such as SAM Kirillov et al. (2023) in a cascaded framework, where the LVLM provides semantic guidance and the segmentation model predicts pixel-level masks. Methods such as LISA Lai et al. (2024), VISA Yan et al. (2024), and VideoLISA Bai et al. (2024) train these cascaded pipelines end-to-end using large-scale annotated datasets. Although effective, such approaches require expensive supervised optimization, substantial labeled data, and re-training for each LVLM, even with parameter-efficient tuning methods such as LoRA Hu et al. (2022).

Recent advances in LVLMs Liu et al. (2023); Bai et al. (2025); Chen et al. (2024) have shown that their reasoning capabilities implicitly encode localization cues within attention maps. Emerging training-free approaches Han et al. (2026); Kang et al. (2025a) leverage these attention maps as grounding signals to generate spatial prompts for segmentation models such as SAM2 Ravi et al. (2025). However, LVLM attention maps are not optimized for precise localization and are often

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

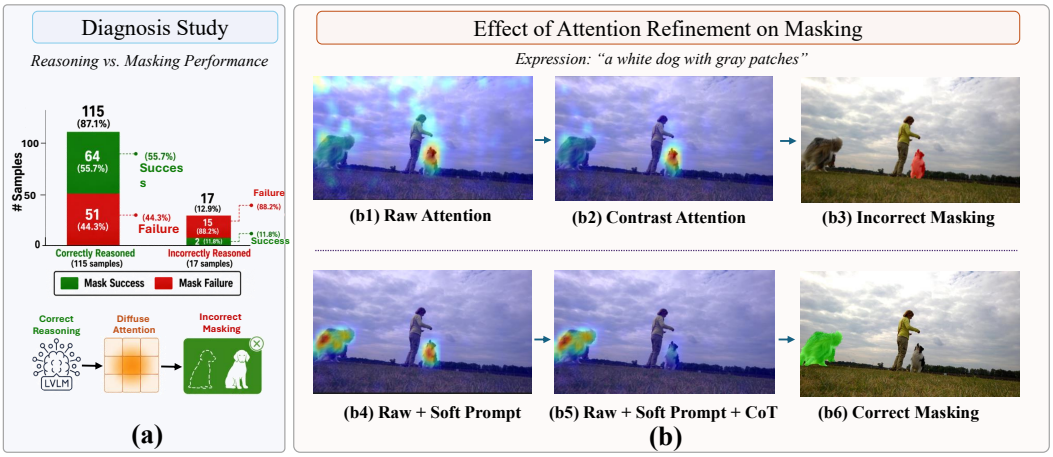


Figure 1: **(a)** Diagnostic study on reasoning and grounding. While the LVLM often identifies the correct target object, the corresponding attention remains poorly localized, leading to inaccurate masks. **(b)** Effect of attention refinement on segmentation. Raw and contrast-based attention produce ambiguous localization, while soft prompts and CoT reasoning progressively improve attention quality and segmentation accuracy.

diffuse, noisy, and affected by attention sinks Kang et al. (2025b), leading to ambiguous grounding (Fig. 1(b1)). Although prior methods attempt to mitigate this issue through contrastive prompting Han et al. (2026), the resulting attention often remains imprecise and produces inaccurate masks (Fig. 1(b2)).

To better understand this gap, we conduct a diagnostic study on 132 samples randomly selected from the ReasonVOS Bai et al. (2024) dataset. For each query, we prompt Qwen2.5-VL Bai et al. (2025) to generate distinguishing attributes describing the referred object. We then evaluate the responses through human annotation using a custom-designed GUI interface (see Appendix A). The evaluation shows that the model correctly identifies the target object in 115 cases (87.1%), indicating strong semantic reasoning capability (Fig. 1 (a)). However, when applying the state-of-the-art attention-based segmentation method DecAF Han et al. (2026), only 55.7% of these correctly understood cases produce accurate masks, while only 2 out of 17 incorrectly reasoned cases are successfully segmented. These results reveal a clear discrepancy between reasoning and grounding: correct reasoning is necessary but not sufficient for accurate localization. These results suggest that the main limitation lies in attention localization, motivating the need for explicit attention steering.

To this end, we adopt an input-level attention steering strategy that operates directly at the source of attention formation. Specifically, we prepend a small set of learnable soft prompts to a frozen LVLM, allowing them to participate in self-attention at every layer. By influencing attention throughout the network, these prompts reshape how the generated response token attends to visual tokens, yielding more concentrated and spatially aligned attention maps (Fig. 1(b4)). To further resolve ambiguity among visually similar objects, we incorporate a lightweight single-step Chain-of-Thought (CoT) reasoning module. Given the input query, the model first generates discriminative attributes (e.g., color, position, motion), which are appended to the original prompt. This attribute-augmented prompt guides attention toward the correct instance through explicit semantic cues, producing more discriminative attention maps (Fig. 1(b5)). Finally, from the refined attention maps, we generate point prompts by selecting high-confidence locations across sampled keyframes. These points are fed to a segmentation model to produce candidate masks, which are propagated into tracklets. We then introduce a correlation-based scoring mechanism that measures consistency between each candidate and the attention signals, enabling robust tracklet selection under challenging conditions.

Importantly, our approach requires no fine-tuning of the LVLM and preserves its pretrained reasoning capabilities, while effectively adapting its attention for accurate and reliable grounding. Building upon these insights, we summarize the primary contributions of our work as follows. **First**, we identify attention misalignment as the primary bottleneck in attention-based grounding and show that attention rollout Abnar & Zuidema (2020), traditionally used as a post-hoc visualization tool, can be effectively controlled through input-level conditioning. **Second**, we propose a principled

attention steering framework that combines learnable soft prompts for spatial attention concentration with Chain-of-Thought (CoT) reasoning for ambiguity resolution, enabling accurate grounding without modifying LVLM parameters. **Third**, we show that explicit attention steering enables transferable object-centric grounding in LVLMs, yielding consistent cross-dataset generalization without task-specific fine-tuning.

2 Related work

Large Vision-Language Models. LLMs have demonstrated strong reasoning capabilities Brown et al. (2020); Dubey et al. (2024); Yang et al. (2024), motivating their extension to vision through large vision-language models (LVLMs). Early models such as Flamingo Alayrac et al. (2022), BLIP-2 Li et al. (2023), and LLaVA Liu et al. (2023) established the paradigm of connecting a frozen vision encoder Radford et al. (2021) to an LLM through a lightweight projector, later refined by LLaVA-1.5 Liu et al. (2024), InternVL Chen et al. (2024), Cambrian-1 Tong et al. (2024), Molmo Deitke et al. (2025), and InternVL3 Zhu et al. (2025). Built on Transformers Vaswani et al. (2017), these models face the quadratic cost of self-attention for video inputs. Many video LVLMs therefore compress visual tokens into fixed-length representations Maaz et al. (2024); Jin et al. (2024); Li et al. (2024b); Song et al. (2024); Shen et al. (2025); Zhang et al. (2025b), often sacrificing fine-grained spatial information. In contrast, LLaVA-style models Liu et al. (2023); Zhang et al. (2025c) retain dense patch features through simple projection, while Qwen2-VL Wang et al. (2024), Qwen2.5-VL Bai et al. (2025), and Oryx Liu et al. (2025) further improve high-resolution processing while preserving spatial detail. We build on these models to explore the localisation capability of LVLMs.

Attention-Guided Grounding in LVLMs. Recent work shows that LVLM self-attention encodes strong visual-textual correspondences for grounding without retraining. Building on attention rollout Abnar & Zuidema (2020), VL-SAM Lin et al. (2024) aggregates attention across heads and layers to prompt SAM, while Loc-Head Kang et al. (2025a) shows that a few frozen attention heads suffice for competitive grounding. F-LMM Wu et al. (2025) converts word–pixel attention into segmentation masks, and ControlMLLM Wu et al. (2024) steers attention via test-time optimization for training-free prompting. "LVLMs Know Where to Look" Zhang et al. (2025a) further shows that internal attention can localize objects even when textual predictions are incorrect.

Video Reasoning Segmentation. Video reasoning segmentation localizes and tracks objects from complex, world-knowledge-dependent queries Yan et al. (2024); Bai et al. (2024), extending traditional referring video object segmentation Khoreva et al. (2018); Seo et al. (2020); Ding et al. (2023). Existing methods follow two main paradigms. *Training-based methods* fine-tune LVLMs with mask decoders: LISA Lai et al. (2024) links an LVLM to SAM Kirillov et al. (2023) via a <SEG> token; GLaMM Rasheed et al. (2024) and PixelLM Ren et al. (2024) extend this to grounded conversation and multi-target segmentation; and VISA Yan et al. (2024), VideoLISA Bai et al. (2024), VRS-HQ Gong et al. (2025b), and GLUS Lin et al. (2025) adapt it to videos. *Training-free methods* exploit localization signals in frozen LVLMs: VL-SAM Lin et al. (2024), Loc-Head Kang et al. (2025a), and F-LMM Wu et al. (2025) derive image prompts or masks from attention, while CoT-RVS Kao et al. (2026), SDAM Zhu et al. (2026), PARSE-VOS Anonymous (2026), and DecAF Han et al. (2026) extend this to video reasoning. Our method lies between these regimes, using *lightweight training on limited data* to steer LVLM attention while preserving pretrained reasoning and achieving accuracy competitive with fully trained systems.

3 Method

Problem Formulation. We study *video reasoning segmentation*, where the goal is to localize and segment a target object across video frames given a natural language query that may require reasoning. Formally, we are given a video $\mathcal{V} = \{\mathbf{I}_t\}_{t=1}^{T}$, where each frame $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$, and a natural language expression $\mathcal{E}$ (e.g., "the person who served the ball"). The objective is to predict a sequence of segmentation masks:

$$\mathcal{M} = \{\mathbf{M}_t\}_{t=1}^{T}, \quad \mathbf{M}_t \in [0,1]^{H\times W}, \tag{1}$$

where $\mathbf{M}_t$ denotes the spatial extent of the target object at frame t. Unlike standard referring segmentation, the expression $\mathcal{E}$ may require multi-step reasoning, temporal understanding, or relational

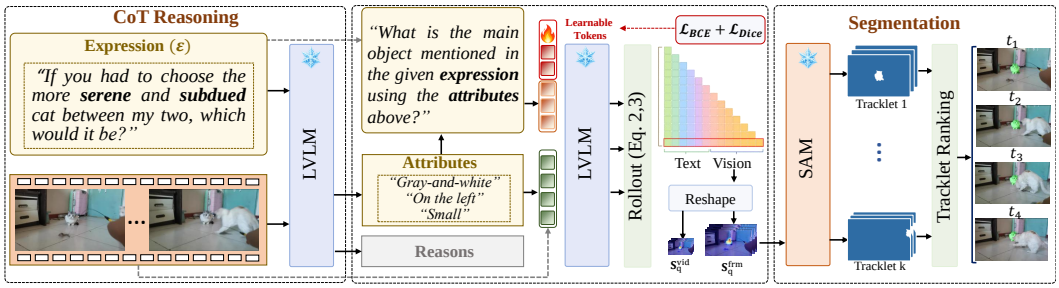


Figure 2: Given a expression and video frames, the frozen LVLM first performs CoT reasoning to extract discriminative attributes of the target object. These attributes are incorporated into a prompt for response-token generation, together with learnable soft prompts. Attention rollout from the response token produces spatial attention maps, which are reshaped into frame-level and video-level localization cues. The resulting attention maps are used to guide SAM for mask propagation and candidate tracklet generation, followed by tracklet ranking to obtain the final segmentation mask.

inference. We assume access to a pretrained LVLM and a segmentation foundation model (e.g., SAM2 Ravi et al. (2025)), both kept frozen. Our goal is to exploit the intrinsic cross-modal attention of the LVLM to derive spatio-temporal grounding signals for accurate segmentation.

3.1 Response-Token Attention Extraction

Given a video-expression pair $(\mathcal{V}, \mathcal{E})$, we prompt the frozen LVLM with a concise instruction to identify the main object referred to in $\mathcal{E}$ and respond with a single-word descriptor, which we denote by q. We treat this token as the query and analyze its attention over the token sequence to understand how the model grounds the target concept in the visual input.

At transformer layer ℓ, let $\mathbf{A}^{(\ell)} \in \mathbb{R}^{H\times N\times N}$ denote the attention tensor, where H is the number of heads and $N = N_v + N_t$ is the total number of tokens, consisting of N_v visual tokens and N_t non-visual tokens. Each slice $\mathbf{A}_h^{(\ell)} \in \mathbb{R}^{N\times N}$ represents the pairwise attention weights between tokens for head h. Following Abnar & Zuidema (2020), we first average across heads and incorporate residual connections and compute the attention rollout from layer L_0 to L as:

$$\mathbf{R} = \prod_{\ell=L_0}^{L} \tilde{\mathbf{A}}^{(\ell)}, \qquad \tilde{\mathbf{A}}^{(\ell)} = \tfrac{1}{2}\left(\tfrac{1}{H}\sum_{h=1}^{H} \mathbf{A}_h^{(\ell)} + \mathbf{I}\right). \tag{2}$$

where $\mathbf{R} \in \mathbb{R}^{N\times N}$ captures the aggregated token-to-token influence across layers. Given the response token q, let i_q denote its index in the token sequence. The attention corresponding to q is obtained from the rollout matrix $\mathbf{R}$ as:

$$\mathbf{r}_q = \mathbf{R}[i_q, :] \in \mathbb{R}^{1\times N}, \quad \mathbf{r}_q^{\text{vis}} = \mathbf{r}_q[:, \mathcal{I}_{\text{vis}}] \in \mathbb{R}^{1\times N_v}, \tag{3}$$

where $\mathcal{I}_{\text{vis}}$ denotes the index set of visual tokens. This provides a generic query-conditioned visual attention vector, which can be reshaped according to the visual-token layout to form a spatial grounding map.

Dual-Granularity Rollout Attention.

The rollout formulation above is agnostic to the visual-token layout. We therefore instantiate it at two complementary granularities to balance spatial precision and temporal context. For the *frame modality*, we uniformly sample $T_f = 16$ key frames and compute rollout attention independently for each frame, yielding attention maps $\{\mathbf{S}_q^{(t)}\}_{t=1}^{T_f}$, where $\mathbf{S}_q^{(t)} \in \mathbb{R}^{H_v\times W_v}$. We denote their stack by $\mathbf{S}_q^{\text{frm}} \in \mathbb{R}^{T_f\times H_v\times W_v}$. This modality preserves high spatial resolution and provides precise localization cues. For the *video modality*, we process $T_v = 8$ key frames jointly to capture temporal consistency across the sequence. To reduce the computational cost of 3D attention, we apply token merging and spatial downsampling by a factor of 2, producing a video-level rollout tensor $\mathbf{S}_q^{\text{vid}} \in \mathbb{R}^{T_v\times \tilde{H}_v\times \tilde{W}_v}$, where $\tilde{H}_v = H_v/2$ and $\tilde{W}_v = W_v/2$.

3.2 ATTENTION STEERING

The rollout map provides grounding cues but is often diffuse, noisy, and prone to attention sinks Kang et al. (2025b), reflecting coarse semantic relevance rather than precise localization (Fig. 1 (b1)). To address this, we keep the model fixed and steer attention through input-level conditioning, aiming to *align attention directly to the target object* and obtain more focused and discriminative attention (Fig. 1(b4–b5)). Specifically, we introduce two complementary mechanisms: (1) prompt-based steering to concentrate attention, and (2) reasoning-based steering to resolve ambiguity.

Prompt-Based Steering. We introduce a set of learnable prompt tokens to guide the attention formation process while keeping the LVLM frozen. Let $\mathbf{P} \in \mathbb{R}^{N_p \times d}$ denote the N_p learnable prompt embeddings, where d is the embedding dimension. These prompt tokens are appended to the input sequence used for generating the response token. Consequently, the sequence length increases from N to $N + N_p$. This also changes the attention tensors at layer ℓ to $\mathbf{A}^{(\ell)} \in \mathbb{R}^{H \times (N+N_p) \times (N+N_p)}$. The resulting attention rollout matrix is therefore given by $\mathbf{R} \in \mathbb{R}^{(N+N_p) \times (N+N_p)}$.

Since the grounding map $\mathbf{S}_q$ is obtained from the rollout matrix $\mathbf{R}$, optimizing the soft prompts $\mathbf{P}$ directly influences $\mathbf{S}_q$. We therefore train $\mathbf{P}$ to make $\mathbf{S}_q$ align with the target object. To this end, we supervise the rollout-based attention ($\mathbf{S}_q$) using ground-truth segmentation masks. We train the soft prompts $\mathbf{P}$ on the training set of Ref-YouTube-VOS dataset Seo et al. (2020), which provides pixel-level annotations for objects in videos. Let $\hat{\mathbf{G}} \in [0,1]^{H_v \times W_v}$ denote the resized ground-truth mask, and let $\mathbf{S}_q \in [0,1]^{H_v \times W_v}$ denote the normalized rollout attention map for the query token q. We optimize the soft prompts using a combination of binary cross-entropy and Dice loss:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(\mathbf{S}_q, \hat{\mathbf{G}}) + \mathcal{L}_{\text{Dice}}(\mathbf{S}_q, \hat{\mathbf{G}}). \tag{4}$$

The binary cross-entropy term enforces pixel-wise alignment, while the Dice loss encourages overlap between the predicted attention map and the ground-truth mask, improving robustness to class imbalance. Together, these losses guide the soft prompts to produce attention maps that are both localized and structurally consistent with the target object.

Reasoning-Guided Steering. Despite the attention concentration achieved by soft prompts, ambiguity remains a challenge for expressions requiring deeper reasoning. As shown in Fig. 1(b4), when multiple objects share similar appearance, the attention map may localize an incorrect instance. This typically occurs when the model fails to reason correctly about the expression, indicating that accurate reasoning is necessary for reliable grounding.

To address this, we incorporate a lightweight, single-round CoT reasoning step to guide attention. Given the expression $\mathcal{E}$, we prompt the LVLM (without soft prompts) to briefly reason about the target object and output a small set of distinguishing attributes $a = \{a_1, \ldots, a_K\}$ (e.g., color, position, motion) that help disambiguate it from similar instances; the full prompt template is provided in Appendix B. We append these attributes to the original prompt used for response token generation, forming an attribute-augmented prompt (the full prompt template is provided in Appendix C). The input sequence is further prepended with the soft prompts $\mathbf{P}$, resulting in a unified prompt that combines $\mathcal{E}$, a, and $\mathbf{P}$ (see Fig. 2).

Intuitively, this design improves grounding in two complementary ways. First, the model performs an explicit reasoning step that extracts attributes capturing distinguishing properties of the target object. Second, these attribute tokens are jointly processed with visual tokens and soft prompts, enabling richer interactions across tokens. This enriched context reduces ambiguity by directing the query token's attention to the correct instance, improving grounding accuracy and reliability.

3.3 MASK TRACKLETS GENERATION AND SELECTION

We use the frame-level attention maps to generate candidate masks, as they provide higher spatial resolution and more precise localization cues compared to the video-level maps. For each key frame t, we select the location with the maximum attention value,

$$(x_t, y_t) = \arg\max_{(x,y)} \mathbf{S}_q^{(t)}(x, y), \tag{5}$$

as a point prompt and feed it to the segmentation model (SAM2 Ravi et al. (2025)), producing one mask per frame. This results in T_f candidate masks. To remove redundant or overlapping

predictions, we apply non-maximum suppression (NMS) with an IoU threshold of 0.7. Finally, we construct tracklets by propagating the selected masks forward and backward across frames using the segmentation model, producing temporally consistent object tracks.

Best Tracklet Selection. Given a set of N_c candidate tracklets, we select the most reliable one by measuring their consistency with the attention signals. Intuitively, a correct tracklet should align well with the rollout attention maps, both spatially (frame-level) and temporally (video-level). For each candidate tracklet, we first sample its predicted masks at the selected key frames and stack them into a tensor $\mathbf{M} \in \mathbb{R}^{T_f \times H_v \times W_v}$. We transform these masks into the logit space to better capture correlation in a continuous domain. We then compute the Pearson correlation between $\mathbf{M}$ and the rollout attention tensors. For the frame modality, we directly compare $\mathbf{M}$ with the frame-level rollout tensor $\mathbf{S}_q^{\mathrm{frm}} \in \mathbb{R}^{T_f \times H_v \times W_v}$. For the video modality, let $\mathbf{S}_q^{\mathrm{vid}} \in \mathbb{R}^{T_v \times \tilde{H}_v \times \tilde{W}_v}$ denote the video-level rollout tensor. We spatially and temporally align the the candidate mask volume $\mathbf{M}$ to the resolution of the video-level attention map using an alignment operator $\mathcal{A}_{\mathrm{vid}}(\cdot)$, yielding $\mathbf{M}^{\mathrm{vid}} = \mathcal{A}_{\mathrm{vid}}(\mathbf{M}) \in \mathbb{R}^{T_v \times \tilde{H}_v \times \tilde{W}_v}$. The two agreement scores are computed as

$$s_{\mathrm{frm}} = \mathrm{corr}(\mathbf{M}, \mathbf{S}_q^{\mathrm{frm}}), \qquad s_{\mathrm{vid}} = \mathrm{corr}(\mathbf{M}^{\mathrm{vid}}, \mathbf{S}_q^{\mathrm{vid}}), \tag{6}$$

where $\mathrm{corr}(\cdot, \cdot)$ denotes the Pearson correlation computed over all elements. We combine the two scores as $s = \alpha s_{\mathrm{frm}} + (1 - \alpha) s_{\mathrm{vid}}$ and select the tracklet with the highest score.

The selected tracklet is then propagated over all frames and resized to the original image resolution to obtain the predicted mask sequence $\mathcal{M}$. This design leverages the complementary strengths of the two modalities: frame-level attention provides fine-grained spatial alignment, while video-level attention captures temporal consistency.

4 Experiments

Datasets and Evaluation Metrics. We evaluate on five referring video object segmentation benchmarks, including three conventional referring VOS datasets: Ref-DAVIS Khoreva et al. (2018), Ref-YouTube-VOS Seo et al. (2020), and MeViS Ding et al. (2023), as well as two reasoning-intensive datasets: ReasonVOS Bai et al. (2024) and ReVOS Yan et al. (2024). ReasonVOS mainly focuses on complex reasoning and compositional understanding, while ReVOS additionally contains challenging reasoning-based referring expressions. Following prior works, we report the standard video object segmentation metrics: region similarity ($\mathcal{J}$), contour accuracy ($\mathcal{F}$), and their average ($\mathcal{J}\&\mathcal{F}$).

Implementation Details. We use publicly available LLaVA-OV-7B Li et al. (2024a), InternVL3-8B Zhu et al. (2025), Qwen2VL-7B Wang et al. (2024), and Qwen2.5VL-7B Bai et al. (2025), with SAM2-hiera-large Ravi et al. (2025) for segmentation. Attention rollout starts from the middle transformer layers (L_0 for Qwen2.5VL-7B; see Appendix D). We use 64 learnable soft prompt tokens, trained only on the Ref-YouTube-VOS training split Seo et al. (2020) with BCE and Dice losses, while all LVLM parameters remain frozen during training and inference. Training uses a learning rate of 5×10^{-4} for 6500 steps, an effective batch size of 4, and two NVIDIA A6000 GPUs. Soft prompts for the frame and video branches are initialized from fixed natural-language seed prompts using repeat-and-truncate rather than random initialization. Unless stated otherwise, $\alpha = 0.3$ combines frame- and video-level Pearson correlation scores for tracklet ranking, selected through hyperparameter analysis in the ablation study.

4.1 Results

Comparison with Existing Methods. Tab. 1 compares SteerSeg with both fully trained and frozen-LVLM text-conditioned VOS approaches across multiple LVLM backbones and benchmarks. The upper gray rows (LISA Lai et al. (2024), VISA Yan et al. (2024), VideoLISA Bai et al. (2024), GLUS Lin et al. (2025), VRS-HQ Gong et al. (2025b), Veason-R1 Gong et al. (2025a)) correspond to methods that jointly optimize the LVLM and segmentation model in an end-to-end cascade training pipeline, typically requiring extensive training on large-scale datasets. In contrast, the lower rows use frozen LVLMs, where methods marked with * are fully training-free. Our method occupies a lightweight middle ground: both the LVLM and SAM2 remain completely frozen, and only a small set of soft prompt tokens is learned.

Table 1: Comparison of LVLM-based text-conditioned VOS methods on Ref-DAVIS Khoreva et al. (2018), ReasonVOS Bai et al. (2024), and ReVOS Yan et al. (2024) benchmarks. The upper gray rows correspond to methods trained end-to-end jointly with the LVLM and segmentation model, while the lower colored rows use frozen LVLMs. Methods marked with ⋆ are fully training-free.

Method	LVLM	Ref-DAVIS			ReasonVOS			ReVOS$_{(Overall)}$			ReVOS$_{(Referring)}$			ReVOS$_{(Reasoning)}$		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Fully Trained Methods																
LISA$_{[CVPR'24]}$	LLaVA-7B	64.8	62.2	67.3	31.1	29.1	33.1	40.9	39.1	42.7	45.7	44.3	47.1	36.1	33.8	38.4
VISA$_{[ECCV'24]}$	ChatUniVi-7B	69.4	66.3	72.5	–	–	–	46.9	44.9	49.0	50.9	49.2	52.6	43.0	40.6	45.4
VideoLISA$_{[NeurIPS'24]}$	LLaVA-Phi-3-V	68.8	64.9	72.7	47.5	45.1	49.9	–	–	–	–	–	–	–	–	–
GLUS$_{[CVPR'25]}$	LLaVA-7B	–	–	–	49.9	47.5	52.4	54.9	52.4	57.3	58.3	56.0	60.7	51.4	48.8	53.9
VRS-HQ$_{[CVPR'25]}$	ChatUniVi-7B	76.0	72.6	79.4	–	–	–	59.1	56.6	61.6	62.1	59.8	64.5	56.1	53.5	58.7
Veason-R1$_{[arxiv'25.08]}$	Qwen2.5VL-7B	–	–	–	59.9	56.0	63.8	61.3	58.2	64.4	63.6	60.7	66.5	59.0	55.8	62.2
Frozen-LVLM Methods																
Loc-Head⋆$_{[CVPR'25]}$	LLaVA-7B	56.3	52.1	60.5	33.6	29.3	38.0	32.5	28.2	36.9	36.9	32.5	41.3	28.1	23.8	32.5
DecAF⋆$_{[ICLR'26]}$	LLaVA-OV-7B	59.4	54.8	64.0	52.8	49.3	56.3	40.0	35.8	44.1	43.4	39.1	47.6	36.6	32.6	40.7
SteerSeg$_{[Ours]}$	LLaVA-OV-7B	**70.0**	**65.7**	**74.3**	**58.6**	**55.7**	**61.5**	**49.2**	**45.6**	**52.8**	**51.9**	**48.4**	**55.5**	**47.0**	**43.4**	**50.6**
Loc-Head⋆$_{[CVPR'25]}$	InternVL3-8B	66.3	62.4	70.2	44.3	41.0	47.5	43.7	39.9	47.5	46.7	42.9	50.6	43.2	39.5	46.8
DecAF⋆$_{[ICLR'26]}$	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
SteerSeg$_{[Ours]}$	InternVL3-8B	**66.1**	**62.0**	**70.1**	**63.3**	**60.6**	**66.1**	**52.5**	**48.8**	**56.2**	**55.6**	**51.9**	**59.3**	**49.5**	**45.8**	**53.1**
Loc-Head⋆$_{[CVPR'25]}$	Qwen2VL-7B	61.9	58.0	65.8	34.0	31.8	36.2	44.0	40.8	47.2	52.7	49.1	56.2	35.4	32.6	38.2
DecAF⋆$_{[ICLR'26]}$	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
SteerSeg$_{[Ours]}$	Qwen2VL-7B	**77.8**	**74.2**	**81.4**	**63.6**	**60.8**	**66.4**	**53.8**	**50.4**	**57.1**	**59.2**	**56.2**	**62.3**	**48.3**	**44.6**	**52.0**
Loc-Head⋆$_{[CVPR'25]}$	Qwen2.5VL-7B	64.6	60.2	68.9	41.1	37.9	44.3	47.0	43.3	50.7	53.1	49.3	56.9	40.8	37.2	44.4
DecAF⋆$_{[ICLR'26]}$	Qwen2.5VL-7B	75.2	70.9	79.5	63.9	60.5	67.2	54.2	50.1	58.2	58.7	54.8	62.6	49.7	45.4	53.9
SteerSeg$_{[Ours]}$	Qwen2.5VL-7B	**81.4**	**78.0**	**84.8**	**65.9**	**63.1**	**68.7**	**56.6**	**53.5**	**59.8**	**61.1**	**58.2**	**63.9**	**52.4**	**48.9**	**55.9**

Compared with frozen-LVLM methods such as Loc-Head⋆ Kang et al. (2025a) and DecAF⋆ Han et al. (2026), SteerSeg consistently achieves strong improvements across datasets and backbones. The gains are especially notable on reasoning-intensive benchmarks such as ReasonVOS and ReVOS (Reasoning), demonstrating the effectiveness of attention steering and CoT-guided attribute conditioning for resolving ambiguity among visually similar objects. For instance, with Qwen2.5VL-7B, SteerSeg improves over DecAF⋆ by 6.2 points on Ref-DAVIS, 2.0 points on ReasonVOS, and 2.7 points on ReVOS (Reasoning) in terms of $\mathcal{J}\&\mathcal{F}$.

Despite freezing both the LVLM and segmentation model, SteerSeg also achieves performance competitive with or superior to fully trained approaches. On Ref-DAVIS, SteerSeg with Qwen2.5VL-7B achieves 81.4 $\mathcal{J}\&\mathcal{F}$, surpassing VRS-HQ by a large margin. Similar trends are observed on ReasonVOS and ReVOS, suggesting that improving attention alignment can be more effective than expensive end-to-end cascade training.

Another key observation is the strong consistency across different LVLM backbones. SteerSeg consistently improves localization quality for LLaVA-OV, InternVL3, Qwen2VL, and Qwen2.5VL. Moreover, although the soft prompts are trained only on Ref-YouTube-VOS, the method generalizes well to unseen datasets and reasoning-intensive benchmarks, indicating that the proposed attention steering mechanism learns transferable alignment behavior rather than dataset-specific heuristics.

Generalization to Additional Datasets. Tab. 2 further evaluates SteerSeg on MeViS Ding et al. (2023) and Ref-YouTube-VOS Seo et al. (2020) to study its generalization across datasets and LVLM backbones. Similar to previous experiments, the upper rows correspond to fully trained approaches, while the lower rows correspond to frozen-LVLM methods.

Compared with existing frozen-LVLM methods such as Loc-Head⋆ and DecAF⋆, SteerSeg consistently achieves substantial improvements on both datasets. The best SteerSeg variant achieves 53.1 $\mathcal{J}\&\mathcal{F}$ on MeViS and 67.9 on Ref-YouTube-VOS, exceeding DecAF⋆ by 5.0 and 8.0 points, respectively. These results further validate the effectiveness of the proposed attention steering mechanism for improving localization quality.

Despite freezing both the LVLM and SAM2, SteerSeg also achieves performance competitive with or superior to several fully trained approaches. On MeViS, SteerSeg with Qwen2VL-7B achieves the best overall performance, while on Ref-YouTube-VOS it remains competitive with recent end-to-end trained methods without requiring expensive cascade optimization.

Table 2: Comparison on additional datasets.

Method	LVLM	MeViS			Ref-YouTube-VOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
VISA$_{[ECCV'24]}$	Chat-UniVi-7B	44.5	41.8	47.1	61.5	**59.8**	63.2
VideoLISA$_{[NeurIPS'24]}$	LLaVA-Phi-3-V	44.4	41.3	47.6	63.7	61.7	65.7
GLUS$_{[CVPR'25]}$	LLaVA-7B	51.3	**48.5**	54.2	67.3	65.5	69.0
VRS-HQ$_{[CVPR'25]}$	Chat-UniVi-7B	50.6	47.6	53.7	**70.4**	**68.3**	**72.5**
Veason-R1$_{[arxiv'25.08]}$	Qwen2.5VL-7B	**52.2**	48.4	**56.0**	–	–	–
Loc-Head*$_{[CVPR'25]}$	Qwen2.5VL-7B	39.4	35.2	43.6	51.0	46.8	55.2
DecAF*$_{[ICLR'26]}$	Qwen2.5VL-7B	48.1	44.0	52.1	59.9	56.2	63.5
SteerSeg$_{[Ours]}$	LLaVA-OV-7B	47.8	43.8	51.9	58.0	55.4	60.6
SteerSeg$_{[Ours]}$	InternVL3-8B	48.2	43.8	52.5	56.4	53.4	59.4
SteerSeg$_{[Ours]}$	Qwen2VL-7B	**53.1**	**49.4**	**56.8**	**67.9**	**65.9**	**70.0**
SteerSeg$_{[Ours]}$	Qwen2.5VL-7B	51.5	47.8	55.2	67.2	65.3	69.1

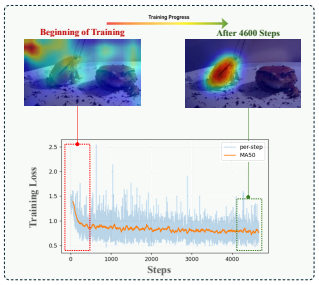


Figure 3: Attention alignment and training loss evolution.

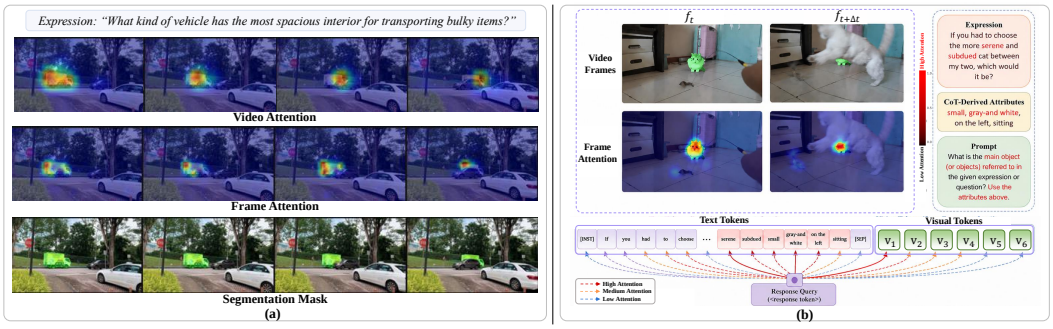


Figure 4: Qualitative visualization of our framework. (a) Video- and frame-level attention with resulting masks across sampled frames. (b) Response-token attention between visual tokens and CoT-derived attributes, with red indicating higher attention.

4.2 Qualitative Analysis

Fig. 3 qualitatively illustrates the training dynamics of the proposed attention steering mechanism through both the training loss evolution and the resulting attention maps. As training progresses, the loss steadily decreases while the attention becomes significantly more focused and aligned with the target object. Notably, this example is drawn from a cross-dataset setting, demonstrating that the learned soft prompts generalize beyond the training distribution and still produce coherent object-centric grounding. Fig. 4(a) presents qualitative examples of the proposed framework. The video-level branch provides temporally consistent localization, while the frame-level branch preserves finer spatial details. Their combination produces accurate and stable segmentation masks under challenging appearance and motion variations. Furthermore, Fig. 4(b) visualizes the response-token attention mechanism, where the response query attends jointly to visual tokens and CoT-derived attributes. Higher attention weights are highlighted in red, showing that the response token increasingly focuses on semantically relevant attributes and object-centric visual cues.

5 Ablation Study

Effect of Individual Components. Table 3 analyzes the contribution of CoT-derived attributes and learnable soft prompts. Using CoT alone provides only a modest improvement over the raw attention baseline, suggesting that reasoning alone is not the primary bottleneck in grounding quality. This observation is consistent with our earlier diagnostic analysis, which indicated that the main limitation arises from weak attention alignment rather than missing semantic reasoning. In contrast, introducing learnable soft prompts leads to a significant improvement in both segmentation performance and attention alignment. Specifically, we report the average Pearson correlation between the rollout attention tensors and the resized ground-truth masks across key frames for both the frame-level and video-level modalities over the entire dataset. Combining both components yields the best performance across all datasets and metrics, with the correlation trends remaining consistent with the segmentation results. Overall, the results suggest that CoT guidance is beneficial but not sufficient by itself, while attention steering through soft prompts plays the dominant role.

Components		J&F		Correlation	
Soft	CoT	DAVIS	ReasonVOS	DAVIS	ReasonVOS
✗	✗	65.7	60.6	0.19	0.25
✗	✓	66.6	63.6	0.21	0.30
✓	✗	80.9	64.1	0.41	0.44
✓	✓	**81.4**	**65.9**	**0.47**	**0.45**

Table 3: Effect of different components on segmentation performance and attention alignment.

N_p	DAVIS	ReasonVOS
4	78.3	63.8
8	80.4	65.1
16	80.6	65.4
32	80.8	65.5
64	**81.4**	**65.9**
128	81.3	65.9

Table 4: Effect of the number of soft prompt tokens N_p.

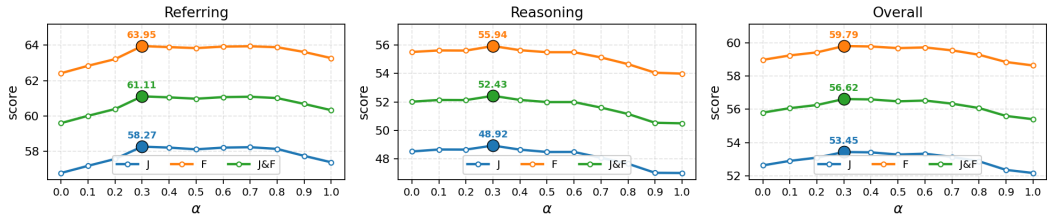


Figure 5: Effect of fusion weight α on ReVOS. Frame-only attention ($\alpha = 1$) is better for referring queries, while video-only ($\alpha = 0$) benefits reasoning tasks requiring temporal context. Performance peaks at $\alpha \approx 0.3$, showing the advantage of combining both.

Effect of the Number of Soft Prompt Tokens. We analyze the effect of the number of learnable soft prompt tokens N_p on segmentation performance. As shown in Table 4, increasing the number of soft tokens leads to consistent but relatively modest improvements on both DAVIS and ReasonVOS. Performance saturates quickly after small prompt sizes, indicating that the proposed attention steering mechanism is not highly sensitive to the exact number of soft tokens. The best performance is achieved at $N_p = 64$, which is used in all experiments.

Effect of fusion weight α on different query types. We further analyze the effect of the fusion weight α on different query types in ReVOS. Comparing the two extremes, frame-only attention ($\alpha = 1$) performs better on referring queries, highlighting the importance of spatial precision, whereas video-only attention ($\alpha = 0$) performs better on reasoning queries, emphasizing the role of temporal context. Nevertheless, the best performance is consistently achieved at an intermediate value ($\alpha \approx 0.3$), showing that neither modality alone is sufficient. Combining frame-level precision with video-level temporal consistency yields the most effective grounding across all query types.

6 Limitations

Our method may miss briefly appearing objects not captured in sampled key frames, which adaptive temporal sampling could mitigate. Additionally, the empirically chosen fusion weight may vary across datasets and LVLM backbones.

7 Conclusion

We presented SteerSeg, a lightweight attention steering framework for text-conditioned video object segmentation using frozen LVLMs. We showed that the main limitation of attention-based grounding is not the lack of semantic understanding, but the misalignment and diffusion of attention maps. To address this, SteerSeg combines learnable soft prompts for attention concentration with a lightweight CoT-guided attribute extraction mechanism for ambiguity resolution. The resulting attention maps provide more accurate and discriminative localization cues, which are further integrated with SAM2 for robust tracklet generation and segmentation. Extensive experiments across multiple benchmarks and LVLM backbones demonstrate that SteerSeg consistently improves grounding quality and achieves performance competitive with or superior to fully trained approaches, despite freezing both the LVLM and segmentation model. Moreover, the strong cross-dataset generalization suggests that the proposed attention steering mechanism learns transferable attention alignment behavior rather than dataset-specific heuristics.

AI USE STATEMENT

In this work, we used generative AI tools solely to edit and polish the manuscript for clarity, grammar, and readability. Generative AI tools were not used to develop the research methodology, formulate hypotheses, design experiments, analyze or interpret results, or generate scientific claims. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the manuscript.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure the reproducibility of SteerSeg. The full pipeline is described in Sec. 3: response-token attention extraction via attention rollout, dual-granularity (frame- and video-level) rollout, soft-prompt and reasoning-guided attention steering, the training objective, and mask tracklet generation and correlation-based selection. All LVLM backbones (LLaVA-OV-7B, InternVL3-8B, Qwen2VL-7B, Qwen2.5VL-7B) and the segmentation model (SAM2-hiera-large) are publicly available checkpoints and are kept frozen throughout. The soft prompts are trained only on the public Ref-YouTube-VOS training split, and all evaluation benchmarks (Ref-DAVIS, Ref-YouTube-VOS, MeViS, ReasonVOS, and ReVOS) are publicly available and evaluated with the standard $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$ metrics (Sec. 4). Training and inference settings are given in Sec. 4, and the complete hyperparameter configuration appears in Appendix D (Table 5). This includes the number of soft-prompt tokens, optimizer, learning-rate schedule, training steps, batch size, rollout layers, frame sampling strategy, point-selection strategy, and fusion weight. The exact prompt templates for CoT attribute extraction and query-token generation are provided verbatim in Appendices B and C. The protocol of the diagnostic study, including sample selection, the annotation interface, and the statistical tests, is described in Appendices A and E. The experimental settings for the LoRA comparison, supervision-regime analysis, and training-efficiency comparison are detailed in Appendices H, I, and J. To facilitate reproducibility, we will publicly release the source code for training and inference, the learned soft-prompt weights, and the evaluation scripts upon acceptance.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Anonymous. Unleashing hierarchical reasoning: An LLM-driven framework for training-free referring video object segmentation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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)*, 2024.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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. In *International Conference on Learning Representations (ICLR)*, 2026.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=7uDI7w5RQA.

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.

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*, 2024a.

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.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision (ECCV)*, 2024b.

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.

Zhiwei Lin, Yongtao Wang, and Zhi Tang. Training-free open-ended object detection and segmentation via attention as prompts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel grounding large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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 Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

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

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.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. ControlMLLM: Training-free visual prompt learning for multimodal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-LMM: Grounding frozen large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *International Conference on Learning Representations (ICLR)*, 2025a.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. LLaVA-Mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations (ICLR)*, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. In *International Conference on Learning Representations (ICLR)*, 2025c.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zhengtong Zhu, Jiaqing Fan, Zhixuan Liu, and Fanzhang Li. Training-free spatio-temporal decoupled reasoning video segmentation with adaptive object memory. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Appendix

A Diagnostic Study and Annotation Interface

To better understand the relationship between semantic reasoning and spatial grounding in LVLMs, we conducted a diagnostic study on the ReasonVOS Bai et al. (2024) dataset. The goal of this study was to evaluate whether the LVLM can correctly identify the referred object at a semantic level, independent of the quality of the resulting attention maps or segmentation masks.

For each sampled video-expression pair, we prompted Qwen2.5VL-7B Bai et al. (2025) to generate a set of distinguishing attributes describing the referred object, including appearance, position, and motion-related cues. We then manually evaluated whether the generated attributes correctly corresponded to the target object referred to in the expression.

To facilitate efficient and consistent annotation, we developed a lightweight graphical user interface, shown in Fig. 6. The interface displays the sampled video clip as an animated GIF together with the referring expression and the attributes generated by the LVLM. Annotators were asked to mark each sample as `True` if the generated attributes correctly identified the referred object, or `False` otherwise. A `Skip` option was also provided for ambiguous cases.

The GUI additionally tracks annotation progress and allows rapid navigation between samples. Using this interface, we annotated 132 randomly sampled examples from ReasonVOS, which formed the basis of the diagnostic analysis presented in the main paper.

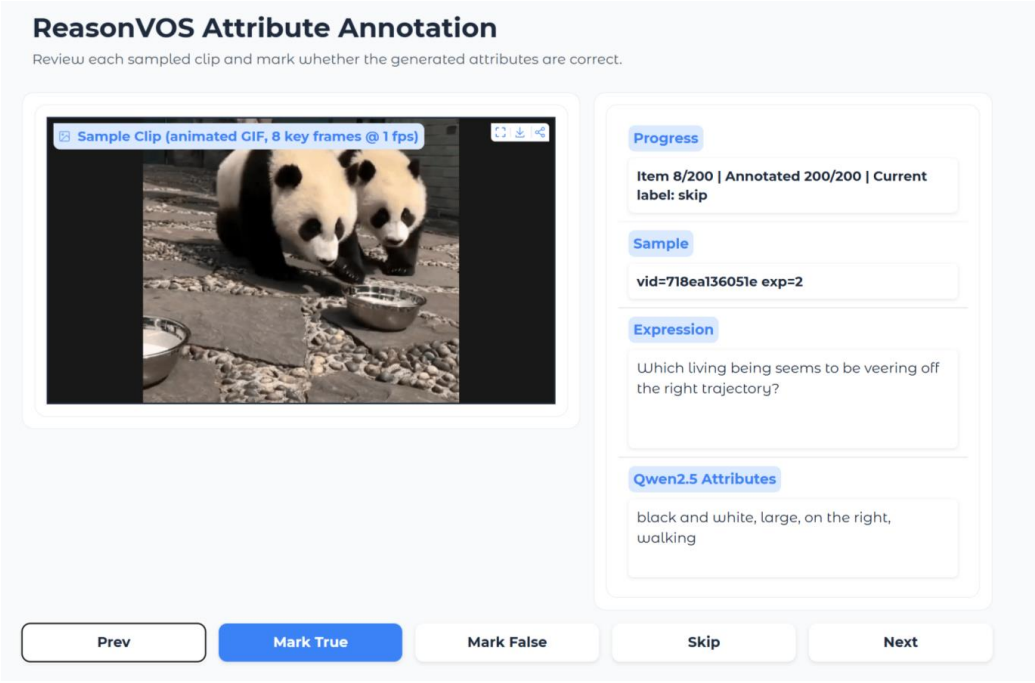


Figure 6: GUI interface used for human evaluation of the generated reasoning attributes on ReasonVOS samples.

B CoT Prompt for Attribute Extraction

CoT Prompt for Attribute Extraction

Expression: {sent}

First, briefly reason about which object in the scene best satisfies this expression — consider the role, function, or intent the expression implies, not just visual similarity to the words. Then list distinguishing attributes (color, size, position, shape, motion) of THAT object only.

Respond in EXACTLY this format, nothing else:
Reasoning: *¡one or two short sentences¿*
Attributes: *¡comma-separated list, max ∼10 words, e.g., 'large, white, on the right, parked'¿*

During inference, we employ a lightweight Chain-of-Thought (CoT) prompting step to extract discriminative attributes that guide attention. Given an input expression $\mathcal{E}$, the LVLM is prompted to first reason about the target object and then produce a concise set of attributes describing it.

The prompt is designed to encourage the model to move beyond superficial keyword matching and instead identify the target object based on its semantic role and context within the scene. This is particularly important for reasoning-based queries, where the correct object may not be directly described by appearance alone.

The extracted attributes serve as a compact and structured representation of the target, which is appended to the original expression and used together with learnable soft prompts to steer attention. Enforcing a strict output format improves the stability of the extracted attributes and reduces noisy or irrelevant responses, leading to more reliable attention guidance.

C Query Prompt for Attention Extraction

Query Token Extraction Prompt

Expression: {sent}

Distinguishing attributes of the target: {attrs}.

What is the main object (or objects) referred to in the given expression or question?
Use the attributes above to disambiguate from other similar objects. Respond with a single word (e.g., 'cat', 'person', 'dog') that best describes the target object(s).

After obtaining discriminative attributes via the CoT stage (Sec. B), we prompt the LVLM to generate a single query token that represents the target object. This token is subsequently used to extract attention maps via attention rollout.

This prompt is designed to produce a concise semantic representation of the target object in the form of a single word. By incorporating the CoT-derived attributes, the model is encouraged to resolve ambiguities and select the correct object category even in the presence of visually similar instances.

The generated token serves as the query in the cross-modal attention mechanism. Specifically, we extract the attention between this token and all visual tokens, and propagate it through attention rollout to obtain a spatio-temporal attention map. This design ensures that the attention map is explicitly grounded in the model's semantic understanding of the target.

Constraining the output to a single word improves stability and consistency across samples, and prevents the model from generating verbose or ambiguous descriptions that would degrade the quality of the extracted attention.

D Implementation Details

Table 5 summarizes the main hyperparameters used for soft-prompt training, attention-rollout extraction, and SAM2-based mask propagation. Unless otherwise stated, all settings are shared across different LVLM backbones.

Hyperparameter	Value
Soft prompt tokens N_p	64 (canonical)
Soft prompt learning rate	5×10^{-4}
Training steps	6500
Effective batch size	4
Optimizer	AdamW
Learning-rate schedule	Cosine + warmup
Training loss	BCE + Dice
Input frames	16
Rollout layers	14–27
Video rollout frames T_v	8
Frame selection strategy	Uniform sampling
Point selection strategy	Argmax
SAM2 backbone	`sam2_hiera_large`
Fusion metric	Pearson correlation
Best fusion weight α	0.3 (default)

Table 5: Hyperparameters used for soft-prompt training, rollout extraction, and SAM2 propagation.

E Extended Diagnostic Analysis

The diagnostic experiment in Sec. 1 examines whether failures in attention-based video segmentation primarily originate from incorrect semantic reasoning or inaccurate spatial grounding. We provide here a statistical analysis of the original 132-sample study.

Among the 132 randomly sampled ReasonVOS examples, Qwen2.5VL-7B correctly identifies the referred object in 115 cases, corresponding to a reasoning accuracy of 87.1% with a 95% Wilson confidence interval of $[80.3, 91.8]$. Among these 115 correctly reasoned cases, DecAF produces a successful grounding result for only 64 examples, corresponding to a grounding success rate of 55.7% with a 95% Wilson confidence interval of $[46.5, 64.4]$.

To quantify the relationship between reasoning correctness and grounding success, we additionally perform Fisher's exact test on the corresponding 2×2 contingency table. Among correctly reasoned samples, 64 are successfully grounded and 51 fail, whereas among incorrectly reasoned samples only 2 are successfully grounded and 15 fail. Fisher's exact test yields an odds ratio of 9.4 with $p = 0.0012$, demonstrating a strong association between correct semantic reasoning and successful grounding.

Importantly, however, more than 44% of correctly reasoned examples still fail at grounding. Therefore, although correct reasoning is strongly associated with successful localization, semantic understanding alone is not sufficient for accurate spatial grounding. This observation supports the central motivation of SteerSeg: a substantial remaining bottleneck lies in how the LVLM's internal semantic understanding is expressed through spatial attention.

F Attention Concentration Analysis

To more directly characterize the quality of LVLM attention, we evaluate three complementary measurements: attention entropy, the fraction of total attention mass contained inside the ground-truth target mask, and Pearson correlation between the attention map and the ground-truth mask. Lower entropy indicates a more concentrated attention distribution, while higher in-mask attention mass and attention–mask correlation indicate stronger spatial alignment with the referred object.

As shown in Table 6, DecAF improves all three measurements relative to raw LVLM attention, indicating that its attention-processing strategy suppresses part of the irrelevant response. Nevertheless, a substantial fraction of its attention remains distributed over contextual regions and competing objects.

Method	Entropy ↓	Mass in GT ↑	Attn.–mask corr. ↑
Raw attention	6.82	0.31	0.25
DecAF Han et al. (2026)	6.21	0.42	0.34
SteerSeg (ours)	**5.47**	**0.56**	**0.45**

Table 6: Attention-quality analysis on the diagnostic ReasonVOS samples using Qwen2.5VL-7B. We report mean attention entropy, the fraction of attention mass inside the ground-truth mask, and Pearson correlation between attention and the ground-truth mask.

In contrast, SteerSeg produces lower attention entropy (5.47), places substantially more attention mass inside the target region (0.56), and achieves a higher attention–mask correlation (0.45). These results quantitatively support the qualitative behavior shown in Fig. 1: SteerSeg does not merely amplify the original attention response, but instead reshapes its spatial distribution toward the referred object.

Why does diffuse attention persist? We attribute diffuse LVLM attention primarily to a mismatch between the model's training objective and the localization behavior required for spatial grounding. Standard autoregressive LVLM training directly supervises generated tokens rather than the spatial regions attended to during generation. Consequently, the model only needs to aggregate sufficient visual information to produce the correct response, with no explicit objective requiring its attention distribution to coincide tightly with the referred object's spatial support.

Attention sinks Kang et al. (2025b) can further attract attention toward semantically uninformative tokens, while head averaging and layer composition during rollout may additionally spread the final signal.

Importantly, our experiments suggest that diffuse attention should not be interpreted as an absence of localization information in the pretrained LVLM. With all model parameters frozen, optimizing only the input soft prompts substantially improves attention–mask alignment. For example, on Ref-DAVIS the attention–mask correlation increases from 0.19 with raw attention to 0.47 with the complete method, as reported in Table 3. This indicates that useful localization information is already present within the pretrained representation, but is not optimally expressed through the default attention distribution.

G Role and Robustness of Reasoning-Guided Attributes

The component analysis in Table 3 reveals complementary roles for soft-prompt steering and CoT-derived attributes. Starting from raw attention, CoT guidance improves $\mathcal{J}\&\mathcal{F}$ from 65.7 to 66.6 on Ref-DAVIS and from 60.6 to 63.6 on ReasonVOS. In comparison, soft prompts alone increase performance to 80.9 and 64.1, respectively, while combining soft prompts and CoT gives the strongest results of 81.4 and 65.9.

This pattern suggests that attention concentration is the dominant factor in improving spatial grounding, while CoT-derived attributes primarily act as an *instance-disambiguation* mechanism. A highly concentrated attention map can still produce an incorrect mask if it concentrates on the wrong object instance. The CoT stage mitigates this failure mode by explicitly extracting discriminative properties of the referred object, such as appearance, location, and motion.

Consistent with this interpretation, CoT provides a substantially larger gain on the reasoning-intensive ReasonVOS benchmark ($+3.0$ $\mathcal{J}\&\mathcal{F}$ over raw attention) than on Ref-DAVIS ($+0.9$). This indicates that explicit reasoning guidance becomes particularly useful when several visually similar candidates must be distinguished.

Sensitivity to incorrect attributes. We additionally analyze whether errors from the attribute-generation stage propagate to the final segmentation output. Under the constrained generation format described in Appendix B, the measured attribute error rate is 1.7%. On this incorrect-attribute subset, the ReasonVOS $\mathcal{J}\&\mathcal{F}$ decreases from 65.9 to 64.2.

Method	Trainable params.	Train time	DAVIS	ReasonVOS
LoRA ($r = 8$)	4M	113 h	77.3	61.3
Soft prompts	**0.46M**	**10 h**	**81.4**	**65.9**

Table 7: Comparison between soft-prompt attention steering and rank-8 LoRA on Qwen2.5VL-7B. Both approaches use the same attention-supervision objective, training data, and optimization-step budget. Performance is reported as $\mathcal{J}\&\mathcal{F}$.

Method	Supervision tier	Mask decoder	Trainable params.	Training data
Loc-Head Kang et al. (2025a)	Training-free	SAM2	0	None
DecAF Han et al. (2026)	Training-free	SAM2	0	None
VRS-HQ Gong et al. (2025b)	End-to-end trained	SAM2	$\approx$ 290M	15 datasets, $\approx$ 0.96M samples
SteerSeg (ours)	Frozen + prompt-tuned	SAM2	$\approx$ 0.46**M**	Ref-YT-VOS train, $\approx$ 12K samples

Table 8: Supervision tier, segmentation backbone, trainable parameters, and training data of representative approaches.

The relatively small aggregate degradation suggests that occasional attribute-generation errors do not dominate overall performance. Nevertheless, the result confirms the complementary roles of semantic reasoning and spatial steering: incorrect semantic guidance can direct otherwise concentrated attention toward an incorrect instance.

H Comparison with LoRA Adaptation

Our approach keeps all LVLM parameters frozen and optimizes only a small set of input soft prompts. To compare this design against a conventional parameter-efficient fine-tuning strategy, we additionally evaluate LoRA Hu et al. (2022) under the same attention-supervision setting.

Specifically, we train a rank-8 LoRA variant on the query and value projections of all transformer layers of Qwen2.5VL-7B. Both LoRA and soft-prompt tuning use the same Ref-YouTube-VOS training data, the same attention-supervision objective, and the same optimization-step budget.

As shown in Table 7, soft-prompt steering uses approximately $8.7\times$ fewer trainable parameters than LoRA while requiring substantially less training time. It also improves performance by 4.1 $\mathcal{J}\&\mathcal{F}$ points on Ref-DAVIS and 4.6 points on ReasonVOS.

The two adaptation mechanisms also differ conceptually. LoRA changes the effective projection weights used by the LVLM during inference. In contrast, SteerSeg leaves the pretrained model parameters unchanged and performs adaptation entirely through additional input embeddings. Consequently, the steering module can be inserted or removed without modifying the underlying LVLM checkpoint, allowing a single frozen model to support both its original inference behavior and spatially steered grounding.

I Supervision Regime

SteerSeg should be distinguished from fully training-free attention-grounding methods. We categorize the compared approaches into three supervision regimes: *training-free*, *frozen-LVLM prompt-tuned*, and *end-to-end trained*. SteerSeg belongs to the second category: both the LVLM and SAM2 remain frozen, while a small set of soft-prompt embeddings is optimized using segmentation supervision from the Ref-YouTube-VOS training split.

This distinction is important when interpreting the results. Loc-Head and DecAF require no task-specific optimization, whereas SteerSeg deliberately introduces lightweight mask-supervised prompt tuning. Indeed, removing the learned soft prompts reduces the performance of our pipeline to 65.7 $\mathcal{J}\&\mathcal{F}$ on Ref-DAVIS and 60.6 on ReasonVOS.

Therefore, we do not characterize SteerSeg as training-free. Instead, the method occupies a lightweight middle ground between training-free attention extraction and end-to-end task-specific

Method	Trainable params.	Training compute	Training data	Ref-DAVIS $\mathcal{J}\&\mathcal{F}$
VRS-HQ Gong et al. (2025b)	$\approx$ 290M	1100 GPU-h (4×A800)	15 datasets, $\approx$ 0.96M samples	76.0
SteerSeg (ours)	$\approx$ 0.46**M**	**10 GPU-h (1×A100)**	**1 dataset, $\approx$ 12K samples**	**81.4**

Table 9: Training-efficiency comparison between SteerSeg and a representative end-to-end trained method. SteerSeg adapts only a small set of soft prompts while keeping both the LVLM and SAM2 frozen.

fine-tuning: the LVLM and SAM2 parameters remain completely frozen, while only approximately 0.46M input prompt parameters are optimized on a single training dataset.

This categorization also clarifies the interpretation of comparisons with end-to-end trained systems, whose segmentation decoders and LVLM components may be jointly optimized using substantially more supervision.

J Training Efficiency

We further compare the training cost of SteerSeg with VRS-HQ, a representative end-to-end trained video reasoning segmentation method. The goal of this comparison is not to claim identical training protocols, but to illustrate the difference in adaptation scale between directly steering a frozen LVLM's attention and jointly training a large LVLM–segmentation architecture.

SteerSeg optimizes approximately $630\times$ fewer parameters than VRS-HQ and uses substantially less training data and compute. VRS-HQ uses approximately 290M trainable parameters and around 0.96M samples from 15 datasets, whereas SteerSeg optimizes only approximately 0.46M soft-prompt parameters using around 12K samples from a single dataset.

Despite this substantially lighter adaptation regime, SteerSeg with Qwen2.5VL-7B achieves 81.4 $\mathcal{J}\&\mathcal{F}$ on Ref-DAVIS, compared with 76.0 for VRS-HQ.

These results highlight the practical advantage of adapting the internal attention representation of a frozen LVLM rather than jointly optimizing a substantially larger LVLM–segmentation pipeline. At the same time, the comparison should be interpreted in the context of the different model architectures, optimization procedures, and supervision regimes of the two methods.

K Discussion of Methodological Positioning

The primary novelty of SteerSeg does not lie in introducing a new segmentation loss. Instead, our contribution concerns identifying and directly addressing a bottleneck in attention-based grounding.

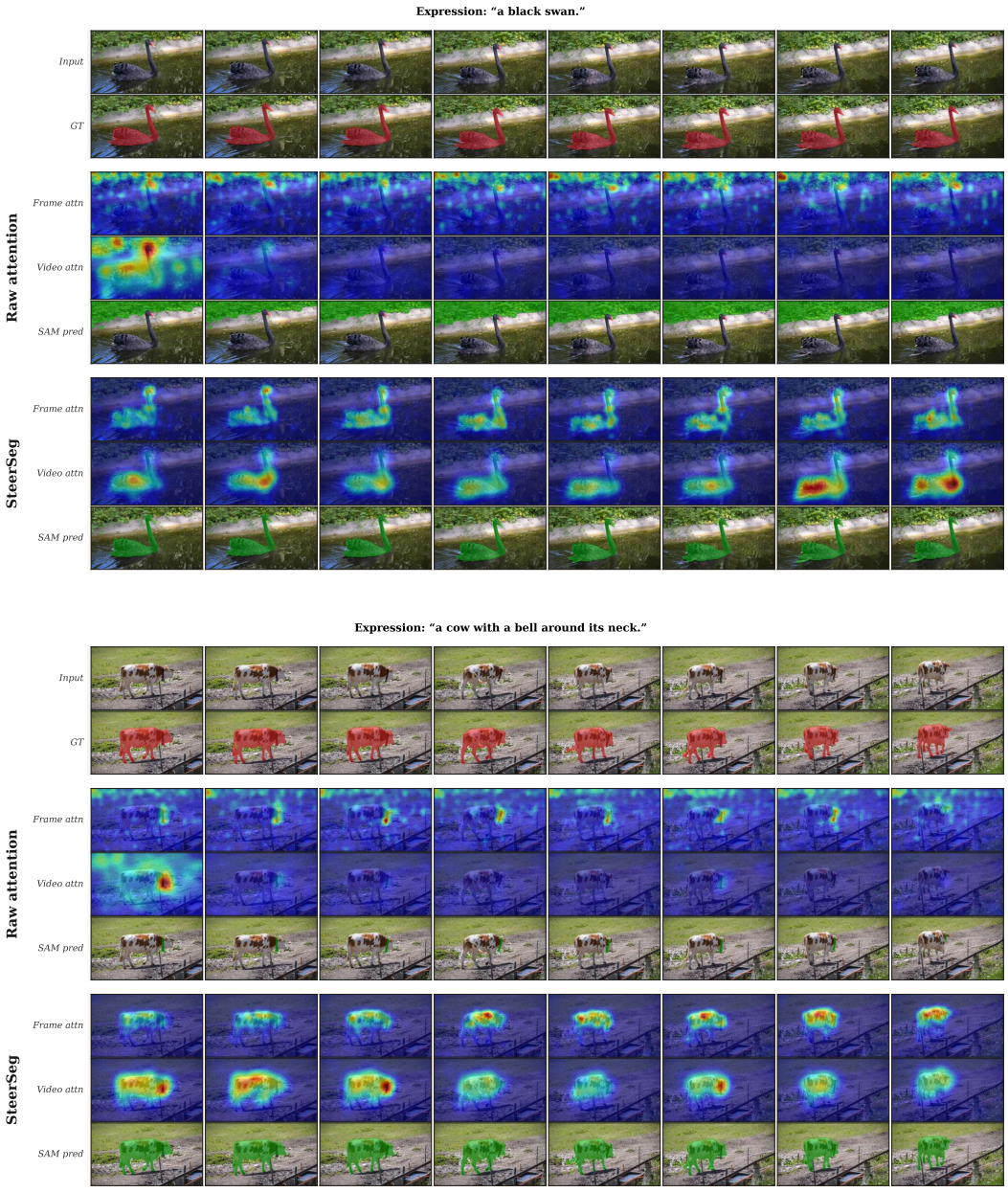
First, the diagnostic study shows that semantic reasoning and spatial grounding are distinct failure modes. Although the LVLM correctly identifies the referred object in 87.1% of the diagnostic examples, successful grounding is obtained for only 55.7% of these correctly reasoned cases. The statistical analysis in Sec. E further confirms a strong association between reasoning and grounding, while also demonstrating that correct reasoning alone remains insufficient.

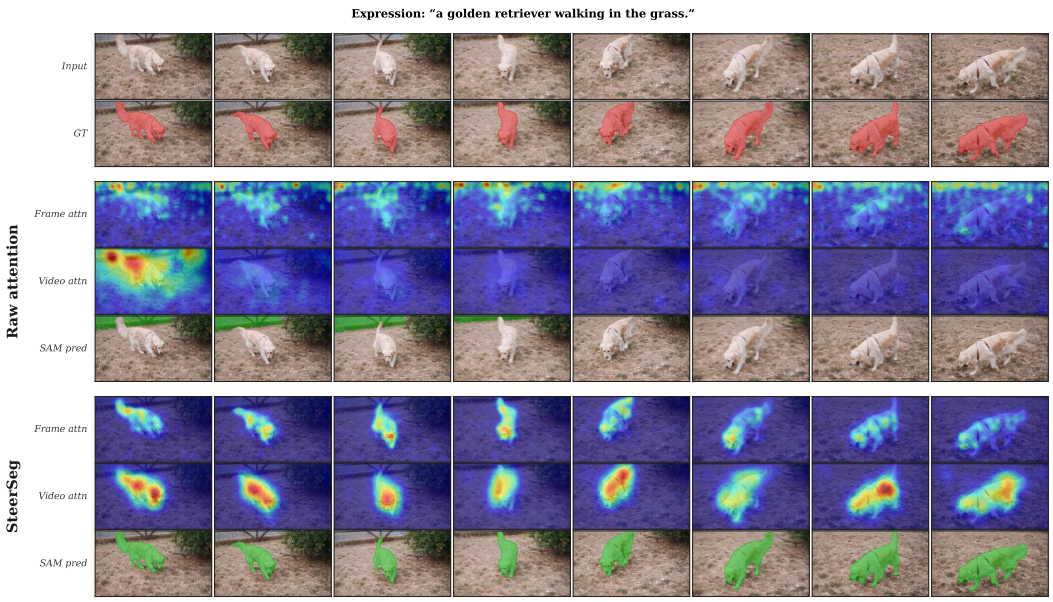
Second, we show that attention rollout, traditionally used as a post-hoc interpretability mechanism, can itself become a controllable representation. Rather than modifying the pretrained model weights, SteerSeg uses input-level soft prompts and directly supervises the resulting internal attention map against segmentation masks. The training objective therefore acts on an internal grounding representation rather than directly optimizing the LVLM's language output.

Third, the attention-quality results in Table 6 show that this steering mechanism changes the spatial structure of attention: entropy decreases, the amount of attention contained inside the target mask increases, and attention–mask correlation improves.

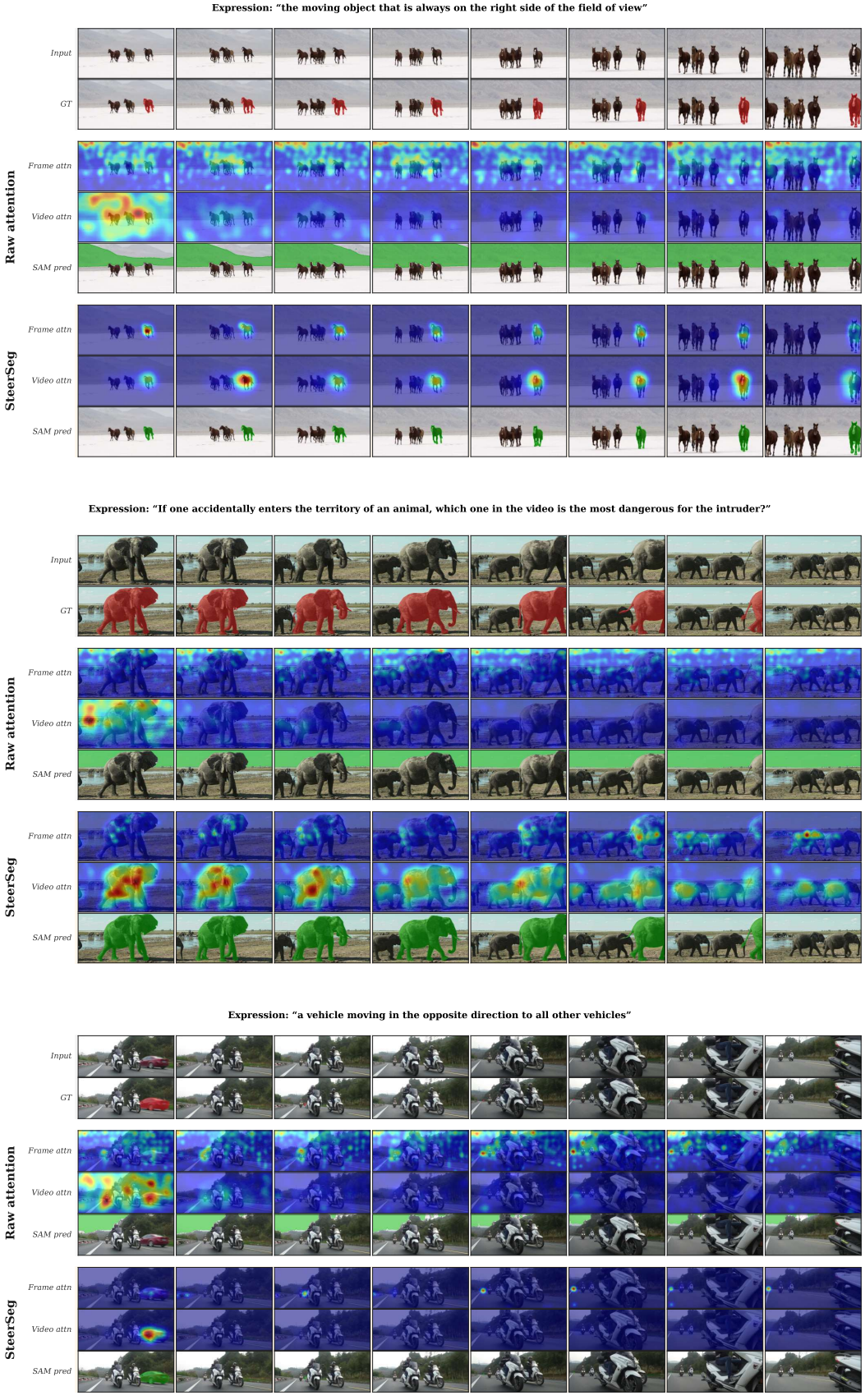
Finally, the comparisons in Tables 7, 8, and 9 characterize the practical trade-off of the proposed formulation. SteerSeg is not training-free, but it requires only lightweight prompt optimization while keeping both foundation models frozen. This provides a middle ground between fully training-free attention extraction and large-scale end-to-end adaptation.

L QUALITATIVE RESULTS




Expression: “a golden retriever walking in the grass.”
Input
GT
Raw attention
Frame attn
Video attn
SAM pred
SteerSeg
Frame attn
Video attn
SAM pred


Expression: “a brown camel in the front.”
Input
GT
Raw attention
Frame attn
Video attn
SAM pred
SteerSeg
Frame attn
Video attn
SAM pred


Expression: "the moving object that is always on the right side of the field of view"
Input
GT
Raw attention
Frame attn
Video attn
SAM pred
SteerSeg
Frame attn
Video attn
SAM pred
Expression: "If one accidentally enters the territory of an animal, which one in the video is the most dangerous for the intruder?"
Input
GT
Raw attention
Frame attn
Video attn
SAM pred
SteerSeg
Frame attn
Video attn
SAM pred
Expression: "a vehicle moving in the opposite direction to all other vehicles"
Input
GT
Raw attention
Frame attn
Video attn
SAM pred
SteerSeg
Frame attn
Video attn
SAM pred