eSAM: Editing SAM3 Attention for Training-Free Referring Segmentation

Yaoting Wang[1], **Yun Zhou**[1], **Hengrui Hu**[1], **Chang Liu**[2], **Henghui Ding**[1*]
[1]Institute of Big Data, College of Computer Science and Artificial Intelligence,
Fudan University, China
[2]Shanghai University of Finance and Economics
https://github.com/FudanCVL/eSAM

Abstract

Referring expression segmentation (RES) aims to segment objects described by natural language in images (RIS) or videos (RVOS). Existing zero-shot methods reduce reliance on mask-language annotations but still require auxiliary training or multi-model pipelines. SAM 3 enables native text-prompted segmentation within a single foundation model, yet performs poorly when directly applied to RES. We trace this gap to an attention sink in its multimodal fusion encoder, where the start-of-text token absorbs a substantial portion of each image patch's cross-attention mass, limiting the influence of content tokens. Suppressing the sink alone redistributes attention indiscriminately, making spatial guidance necessary. We therefore introduce eSAM, a fully training-free framework that addresses this issue by editing SAM 3's cross-attention. Attention Mask Editing (AME) suppresses the sink and uses a CLIPSeg-derived, patch-dependent spatial prior to redirect attention toward relevant content tokens. Value Tensor Editing (VTE) rescales value tensors to amplify content-token contributions to the fused visual features. On RIS (RefCOCO/+/g) and RVOS (Ref-DAVIS17, Ref-YouTube-VOS, MeViS) benchmarks, eSAM achieves state-of-the-art performance among training-free methods without auxiliary LLMs.

1 Introduction

Referring Expression Segmentation (RES) [4, 8, 9, 15, 19, 20, 31] aims to segment a target object specified by free-form natural language, in either images (RIS) [15, 20, 24] or videos (RVOS) [4, 6, 16, 31]. RES requires joint reasoning over visual content and unrestricted language. This fine-grained multimodal grounding makes RES important for interactive perception, embodied agents, and human-AI collaboration. However, progress in RES has been constrained by the cost of supervision: each training sample typically requires both a pixel-level mask and a corresponding referring expression, making large-scale annotation expensive and limiting open-world scalability.

To reduce this dependence on mask-language supervision, recent zero-shot and training-free RES methods have repurposed pre-trained foundation models [17, 25, 35, 36, 44, 45]. A common strategy is to combine a vision-language model such as CLIP or a generative VLM with an external segmentation model such as SAM. Although effective, these pipelines often rely on multiple independently trained components. As a result, the language-grounding signal and the segmentation mechanism are only loosely coupled, and errors in localization, proposal generation, or prompt conversion can propagate to the final mask. More recently, SAM 3 introduced native support for text-prompted segmentation within a single segmentation foundation model, making it a promising starting point for

*Corresponding author (henghui.ding@gmail.com).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

RES. However, our experiments show that directly applying SAM 3 to standard RES benchmarks leaves substantial room for improvement.

We investigate this gap by probing the cross-attention layers in SAM 3's multimodal fusion encoder, where text embeddings are injected into the visual stream. We observe a severe attention-sink phenomenon: the <start_of_text> (SOT) token absorbs a large fraction of the cross-attention mass across image patches, often dominating both foreground and background regions. Since the SOT token primarily represents global sequence context rather than the content words that describe the referent, this imbalance weakens the influence of informative tokens on the fused visual features. Because the fused features are subsequently passed to SAM 3's mask decoder, this attention imbalance can degrade localization quality.

This observation motivates us to intervene in the cross-attention computation at inference time. We view cross-attention as two coupled stages: (1) *routing*, which determines how each image patch distributes its attention over text tokens; and (2) *readout*, which determines how strongly the attended token values contribute to the fused feature. Based on this decomposition, we propose eSAM, a training-free framework that edits SAM 3's cross-attention without updating model parameters.

During the routing stage, we first suppress the SOT logit, which releases the attention mass for the content tokens to absorb. Our experiments show that this operation alone could significantly improve the RES performance of the SAM 3 foundation model. In addition, to specify where the released attention budget should go, we inject a spatial prior into the cross-attention. This prior guides where attention shifts from the SOT token toward content tokens, supplying the spatial guidance that pure sink suppression lacks. Combining the two gives the most effective routing-side intervention: sink suppression frees the attention budget, and the spatial prior directs it to the right content tokens. We refer to this combined intervention as **Attention Mask Editing (AME)**. Furthermore, for the readout stage, we rescale the value tensors so that content tokens contribute more strongly to the fused output. We term this readout-side intervention **Value Tensor Editing (VTE)**. Together, AME and VTE form eSAM, a training-free framework that edits SAM 3's attention without any parameter updates.

Our contributions are summarized as follows:

- We diagnose a pronounced attention sink in SAM 3's multimodal fusion encoder and show that closing the resulting performance gap requires not only suppressing the sink but also providing directional spatial guidance for the released attention.
- We propose **eSAM**, a training-free framework that edits cross-attention at both the routing stage (Attention Mask Editing, AME) and the readout stage (Value Tensor Editing, VTE) of SAM 3's frozen multimodal fusion encoder, requiring no gradient updates.
- We evaluate eSAM on RIS (RefCOCO/+/g, gRefCOCO) and RVOS (Ref-DAVIS17, Ref-YouTube-VOS, MeViS), where it achieves the best performance among the compared training-free methods that do not rely on auxiliary MLLM controllers.

2 Related Work

Referring Expression Segmentation. Referring Expression Segmentation (RES) aims to predict a pixel-level mask for the object described by a free-form natural language expression, with image and video variants known as RIS [3, 8, 15, 20, 24] and RVOS [4, 6, 16, 31, 37]. Fully-supervised RIS methods mostly adopt Transformer-based vision-language fusion, in which language tokens query image features through cross-attention. Representative works include LAVT [42], which performs early multimodal fusion via language-aware visual transformers, and PolyFormer [21], which casts mask prediction as sequential polygon generation. RVOS extends RIS to the temporal axis by enforcing target-consistent segmentation across challenging video frames [5, 7]. ReferFormer [39] treats expressions as object queries; OnlineRefer [38] introduces an efficient online formulation; HTML [11] models hybrid temporal-scale interactions; and MUTR [41] unifies multiple referring modalities within a temporal Transformer. MeViS [4, 6] further pushes RVOS toward motion-centric expressions. While effective, these methods rely on task-specific architectures and large-scale mask supervision, limiting their scalability to open-vocabulary scenarios.

Training-Free RES via Foundation Models. Training-free RES methods leverage pre-trained foundation models without task-specific mask-language supervision. Global-Local [44] and HybridGL [22] combine SAM mask proposals with global-local CLIP features, while AL-Ref-

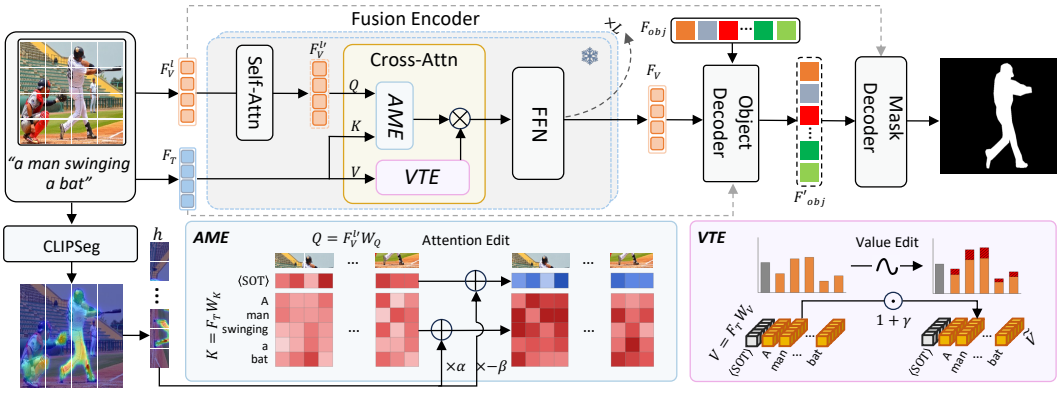

Figure 1: Overview of eSAM. We keep SAM 3 frozen and edit its multimodal fusion encoder's cross-attention through two inference-time interventions guided by a CLIPSeg-derived spatial prior h. **Attention Mask Editing (AME)** edits the attention map by suppressing the SOT logit and injecting h as an additive bias on content tokens. **Value Tensor Editing (VTE)** rescales the value tensor, amplifying content tokens' readout magnitude. The fused features pass through SAM 3's native decoders without modification.

SAM 2 [14] integrates GPT-4 reasoning with GroundedSAM [28] and SAM 2 [27] to generate segmentation prompts [26]. Most closely related to our work, RefAM [17] identifies attention sinks in a frozen diffusion transformer, uses appended stop-words to absorb excess attention, and feeds the resulting heatmaps to SAM 2. Inference-time attention editing has also been explored in language models [10, 18, 40, 46], vision-language models [32], and diffusion models [2, 13]. Recent MLLM-assisted methods further incorporate semantic reasoning. RESAnything [34] uses attribute descriptions for mask selection, EVOL-SAM3 [43] performs evolutionary prompt search, and Tarot-SAM3 [47] combines MLLM-guided prompting with DINOv3-based mask refinement. Unlike these methods, eSAM directly edits SAM 3's internal attention routing and value tensors, using a CLIPSeg-derived spatial prior to guide attention toward relevant content tokens.

3 Method: eSAM

This section details **eSAM**, a training-free framework for adapting SAM 3 [1] to referring expression segmentation. We first introduce notation and characterize the attention-sink phenomenon that motivates our design (Section 3.1). We then present the open-vocabulary spatial prior used to guide our intervention (Section 3.2), followed by the two complementary edits at the core of our method: Attention Mask Editing (AME) at the routing stage of cross-attention (Section 3.3) and Value Tensor Editing (VTE) at the readout stage (Section 3.4).

3.1 Preliminaries

3.1.1 Cross-Attention as Routing and Readout

Given an input image, SAM 3 partitions it into a flattened grid of L patches with queries $\mathbf{Q} \in \mathbb{R}^{L\times d}$. The referring expression is tokenised into a sequence of N tokens, including an SOT token at the start of the sequence and content tokens corresponding to the words of the expression, which produce keys $\mathbf{K} \in \mathbb{R}^{N\times d}$ and values $\mathbf{V} \in \mathbb{R}^{N\times d}$. We denote the index sets of the SOT and content tokens by $\mathcal{I}_S$ and $\mathcal{I}_C$ respectively, with $\mathcal{I} = \mathcal{I}_S \cup \mathcal{I}_C$. The fused output $\mathbf{O} \in \mathbb{R}^{L\times d}$ of the cross-attention is computed as

$$\mathbf{O} = \underbrace{\mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)}_{\mathbf{A}\in\mathbb{R}^{L\times N}\text{ (routing)}} \underbrace{\mathbf{V}}_{\text{readout}}. \tag{1}$$

The computation can be decomposed into two functionally distinct stages. The *routing* stage, governed by the attention matrix $\mathbf{A}$, decides how each image patch distributes its attention over the N text tokens, with $\sum_j \mathbf{A}_{i,j} = 1$. The *readout* stage, governed by $\mathbf{V}$, supplies the value vector that each routed token contributes to the fused output.

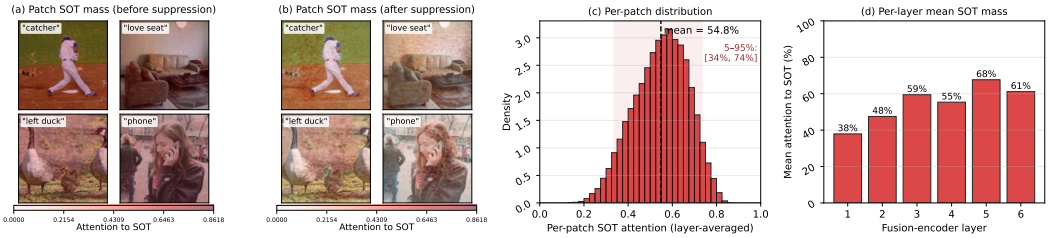

Figure 2: **Attention sink in SAM 3 fusion encoder. (a)** Per-patch SOT mass on four examples from the RefCOCO val split; high mass is observed at both referent and background patches. **(b)** The same patches after sink suppression. **(c)** Per-patch distribution of SOT mass over 500 samples from the RefCOCO val split. **(d)** Per-layer mean SOT mass shows that the sink happens at each layer.

3.1.2 Attention Sink in SAM 3

Probing a pre-trained SAM 3 fusion encoder reveals that the routing distribution $\mathbf{A}$ collapses onto the SOT token. As shown in Figure 2(a), each image patch routes most of its cross-attention mass to the SOT, irrespective of whether the patch lies on the referent. The per-patch distribution in Figure 2(c) quantifies this collapse, with the SOT absorbing 54.8% of the mass on average. Figure 2(d) further shows that the collapse persists across all six fusion-encoder layers, with per-layer means ranging from 38% to 68%. Since the SOT token aggregates global sequence context rather than the linguistic content describing the referent, content tokens are left with negligible influence on the fused features.

3.2 Open-Vocabulary Spatial Prior

To guide the routing intervention without relying on additional training signals, we use an external language-aligned localization cue from the lightweight open-vocabulary segmenter CLIPSeg [23]. Given an input image and a referring expression, CLIPSeg produces a dense logit map. We resize this map to match the SAM 3 patch grid using bilinear interpolation and then flatten it into $\boldsymbol{\ell} \in \mathbb{R}^L$. The spatial prior $\mathbf{h} \in [0,1]^L$ is computed as

$$\mathbf{h} = \mathrm{MinMax}(\sigma(\boldsymbol{\ell})), \tag{2}$$

where $\sigma(\cdot)$ is the element-wise sigmoid function, and $\mathrm{MinMax}(\cdot)$ rescales the responses of each image to the range $[0,1]$. The sigmoid maps the unbounded logits to a probability scale, but the resulting values tend to lie within a narrow interval. The subsequent min–max normalization expands this range on a per-image basis, preventing $\mathbf{h}$ from becoming nearly uniform. The final $\mathbf{h}$ serves as a coarse localization signal aligned with the input text, indicating which patches are more relevant to the content tokens.

3.3 Routing Intervention: Attention Mask Editing

The first component of eSAM, Attention Mask Editing (AME), adjusts the routing stage by adding a bias matrix $\mathbf{B} \in \mathbb{R}^{L \times N}$ to the pre-softmax logits. For each patch index $i \in \{0, \ldots, L-1\}$ and token index $j \in \mathcal{I}$,

$$\mathbf{B}_{i,j} = \begin{cases} -\beta\, \mathbf{h}_i, & j \in \mathcal{I}_S, \\ \alpha\, \mathbf{h}_i, & j \in \mathcal{I}_C, \end{cases} \tag{3}$$

where $\beta > 0$ controls the suppression of the SOT logit and $\alpha > 0$ controls the amplification of content tokens. The resulting routing weights are

$$\mathbf{A}_{\text{edit}} = \mathrm{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B}\Big). \tag{4}$$

Applying the modification at the logit level preserves the normalization $\sum_j \mathbf{A}_{i,j} = 1$ without requiring an additional renormalization step. A key property of Equation 3 is that the spatial prior $\mathbf{h}_i$ jointly scales both the suppression of the SOT token and the promotion of content tokens using the same patch-specific weight. In contrast, a uniform variant that replaces $-\beta\mathbf{h}_i$ with $-\beta$ in the SOT column would suppress the sink at every patch indiscriminately, leaving the released mass to be redistributed without external guidance. The prior-modulated form $-\beta\mathbf{h}_i$ instead aligns

the redistribution of attention with the same signal $\mathbf{h}$ that determines where suppression occurs. When $\mathbf{h}_i$ is large, the SOT is strongly suppressed and content tokens receive a corresponding boost, so the released mass is directed toward semantically relevant patches. When $\mathbf{h}_i$ is small, both effects diminish and the original SAM 3 routing remains largely unchanged. As a result, AME acts selectively, modifying routing only where the prior provides useful information and avoiding unguided redistribution.

3.4 Readout Intervention: Value Tensor Editing

AME shifts routing weights toward content tokens, but the fused output remains a weighted sum of value vectors, $\mathbf{O}_i = \sum_j \mathbf{A}_{i,j}\mathbf{V}_j$, where the contribution of token j at patch i depends on both the routing weight $\mathbf{A}_{i,j}$ and the magnitude $\|\mathbf{V}_j\|$. Modifying routing weights alone leaves the readout magnitudes unchanged. To address this limitation, we introduce Value Tensor Editing (VTE), a multiplicative intervention on the value tensor. Given $\mathbf{V} \in \mathbb{R}^{N\times d}$, VTE constructs a modified tensor $\widetilde{\mathbf{V}} \in \mathbb{R}^{N\times d}$ as

$$\widetilde{\mathbf{V}}_j = \begin{cases} \mathbf{V}_j, & j \in \mathcal{I}_S, \\ (1+\gamma)\,\mathbf{V}_j, & j \in \mathcal{I}_C, \end{cases} \tag{5}$$

where $\gamma > 0$ increases the magnitude of content tokens. This scaling changes the contribution of content tokens to the fused output while keeping the direction of $\mathbf{V}_j$ unchanged.

The combined operation $\mathbf{A}_{\text{edit}}\widetilde{\mathbf{V}}$ biases the fused output toward content tokens through two separate mechanisms: AME modifies the routing distribution and VTE rescales the value magnitudes, addressing the routing and readout stages of cross-attention introduced in Section 3.1.1. The edited cross-attention output at any chosen fusion-encoder layer is

$$\mathbf{O}_{\text{edit}} = \mathbf{A}_{\text{edit}}\,\widetilde{\mathbf{V}}, \tag{6}$$

while all other layers and components (backbone, decoders, parameters) remain unchanged.

4 Experiments

4.1 Datasets and Metrics

For RIS, we evaluate on four standard benchmarks: RefCOCO [15], RefCOCO+ [15], RefCOCOg [24], and gRefCOCO [20]. We report overall IoU (oIoU) as well as mean IoU (mIoU). For RVOS, we evaluate on three benchmarks: Ref-DAVIS17 [16], Ref-YouTube-VOS [31], and MeViS [4]. We report region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their combined $\mathcal{J}\&\mathcal{F}$ score.

4.2 Results on RIS Benchmarks

Table 1 presents comparisons with both training-free methods and zero-shot methods that rely on additional training, evaluated on RefCOCO, RefCOCO+, RefCOCOg, and gRefCOCO. eSAM achieves the best performance on all 11 splits under oIoU and on 10 out of 11 splits under mIoU. When averaging across all splits, eSAM improves over the SAM 3 zero-shot baseline by $+8.51$ in oIoU and outperforms RefAM, the strongest prior training-free method, by $+12.20$. CLIPSeg, which serves as the source of spatial priors, consistently performs worse than eSAM across all settings, suggesting that the improvements are not driven by CLIPSeg alone but by the proposed cross-attention-based refinement. eSAM also compares favorably with zero-shot methods that use additional supervision. On the RefCOCO validation set, it surpasses VLM-VG by $+8.22$ in oIoU while remaining fully training-free.

4.3 Results on RVOS Benchmarks

Table 2 presents comparisons on three RVOS benchmarks, namely Ref-DAVIS17, Ref-YouTube-VOS, and MeViS. Among training-free methods, eSAM achieves the best performance across all three benchmarks and all evaluation metrics. Averaged over the three benchmarks, eSAM improves $\mathcal{J}\&\mathcal{F}$ over the SAM 3 zero-shot baseline by $+13.64$, and over RefAM by $+8.09$.

Table 1: Main results on RIS benchmarks (RefCOCO/+/g and gRefCOCO). Bold and underlined numbers indicate the best and second-best results across the listed methods.

Metric	Method	RefCOCO			RefCOCO+			RefCOCOg		gRefCOCO		
		val	testA	testB	val	testA	testB	val	test	val	testA	testB
oIoU	*zero-shot w/ additional training*											
	Pseudo-RIS [45]	37.33	43.43	31.90	40.19	46.43	33.63	41.63	43.52	-	-	-
	VLM-VG [35]	45.40	48.00	41.40	37.00	40.70	30.50	42.80	44.10	-	-	-
	zero-shot training-free											
	Global-Local [44]	21.71	24.48	20.51	23.70	28.12	21.86	26.57	28.21	-	-	-
	Ref-Diff [25]	35.16	37.44	34.50	35.56	38.66	31.40	38.62	37.50	-	-	-
	HybridGL [22]	41.81	44.52	38.50	35.74	41.43	30.90	42.47	42.97	-	-	-
	CLIPSeg [23]	30.19	31.99	28.86	31.84	35.04	30.40	35.93	35.86	29.35	27.46	33.96
	SAM 3 [1]	47.23	48.05	48.75	37.43	42.17	33.10	44.43	43.88	26.45	28.95	36.43
	RefAM [17]	42.67	46.21	40.89	34.43	38.70	30.53	39.58	41.13	21.67	32.20	28.18
	eSAM (Ours)	**53.62**	**57.94**	**51.33**	**43.15**	**50.19**	**36.06**	**54.44**	**54.47**	**39.73**	**47.21**	**42.31**
mIoU	*zero-shot w/ additional training*											
	Pseudo-RIS [45]	41.05	48.19	33.48	44.33	51.42	35.08	45.99	46.67	-	-	-
	VLM-VG [35]	49.90	53.10	46.70	42.70	47.30	36.20	48.00	48.50	-	-	-
	zero-shot training-free											
	Global-Local [44]	33.12	36.52	29.58	35.29	39.58	31.89	40.08	40.74	-	-	-
	CaR [33]	33.57	35.36	30.51	34.22	36.03	31.02	36.67	36.57	-	-	-
	Ref-Diff [25]	37.21	38.40	37.19	37.29	40.51	33.01	44.02	44.51	-	-	-
	HybridGL [22]	49.48	53.37	45.19	43.40	49.13	37.17	51.25	51.59	-	-	-
	CLIPSeg [23]	26.91	28.18	26.00	28.65	31.06	27.84	33.95	33.95	57.60	36.07	40.02
	SAM 3 [1]	46.72	46.42	49.43	37.57	41.00	34.78	46.27	45.25	**65.62**	44.46	49.40
	RefAM [17]	52.17	55.60	49.64	43.06	48.58	37.55	48.11	48.57	20.52	34.85	30.03
	eSAM (Ours)	**58.56**	**61.48**	**56.97**	**48.19**	**54.10**	**41.83**	**61.39**	**61.27**	65.22	**56.35**	**52.99**

Table 2: Results on RVOS benchmarks (Ref-DAVIS17, Ref-YouTube-VOS, and MeViS). Bold and underlined numbers indicate the best and second-best results. Comparison results follow RefAM [17] and prior literature.

Method	Ref-DAVIS17			Ref-YouTube-VOS			MeViS		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
fully-supervised methods									
ReferFormer [39]	58.10	64.10	61.10	62.80	67.00	64.90	29.80	32.20	31.00
OnlineRefer [38]	61.60	67.70	64.65	61.60	65.50	63.55	–	–	–
HTML [11]	59.20	65.10	62.15	61.50	65.20	63.35	–	–	–
MUTR [41]	62.80	70.00	66.40	65.40	69.60	67.50	–	–	–
Dshmp [12]	61.70	68.10	64.90	65.00	69.10	67.05	43.00	49.80	46.40
ReferDINO [19]	65.10	72.90	68.90	67.00	71.50	69.25	44.70	53.90	49.30
MPG-SAM2 [29]	68.80	76.00	72.40	71.70	76.10	73.90	50.70	56.70	53.70
training-free methods									
GL CLIP + SAM2 [44]	37.60	43.60	40.60	24.30	29.70	27.00	20.40	30.00	25.20
GL SAM + SAM2 [44]	44.00	49.70	46.85	29.90	37.30	33.60	22.70	30.50	26.60
RefAM [17]	54.50	60.60	57.55	37.60	47.80	42.70	24.70	36.60	30.65
SAM 3 [1]	41.02	44.18	42.60	41.40	43.46	42.43	26.89	31.56	29.23
eSAM (Ours)	**55.89**	**60.65**	**58.27**	**55.59**	**59.41**	**57.50**	**36.07**	**42.73**	**39.40**

4.4 Effect of eSAM on Cross-Attention

Following the routing-and-readout decomposition in Section 3.1.1, we compare SAM 3 and eSAM at the edited fusion layer on 500 RefCOCO validation samples (Figure 3). For routing (a, b), eSAM reduces SOT attention from 47.5% to 27.1% while increasing content-token attention from 27.8% to 42.8%. For readout (c, d), the SOT contribution decreases from 36.8% to 16.8%, while the content-token contribution increases from 58.4% to 71.4%. These results demonstrate that eSAM shifts both attention allocation and output contributions toward content tokens.

4.5 Ablation Study

4.5.1 Prior Source and Resolution

Table 3 shows that CLIPSeg, MaskCLIP [48], and CLIP Grad-CAM [30] consistently outperform SAM 3 across three RIS datasets, with CLIPSeg performing best overall. On RefCOCO val, reducing the CLIPSeg prior resolution to 44×44 and 22×22 yields 53.59 and 53.57 oIoU, compared with 53.62 at the original resolution, indicating low sensitivity to prior resolution.

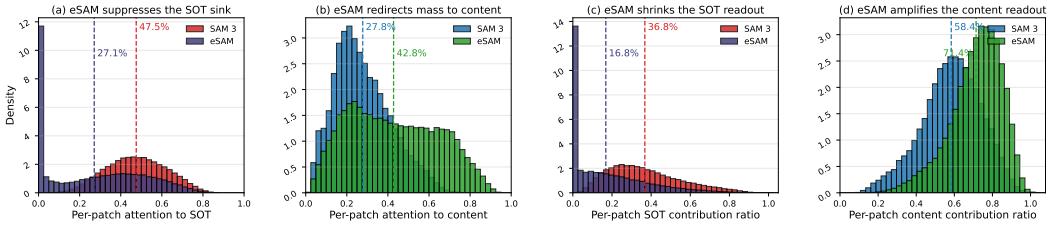

Figure 3: Per-patch effect of eSAM on cross-attention at the edited fusion-encoder layer.

Table 3: **oIoU on alternative spatial-prior sources**.

Method	RefCOCO	RefCOCOg	gRefCOCO
eSAM (CLIPSeg)	**53.62**	**54.44**	**39.73**
eSAM (MaskCLIP)	52.91	54.06	39.69
eSAM (Grad-CAM)	52.65	53.98	39.41

Table 4: **SOT suppression strategies** on RefCOCO val.

SOT bias	oIoU	mIoU
None (boost only)	53.55	58.53
$-\beta$ (unconditional)	53.38	58.33
$-\beta\mathbf{h}_i$ (prior-aware)	**53.62**	**58.56**

4.5.2 Prior-Aware Sink Suppression

Table 4 compares three SOT suppression strategies while keeping the content-token boost $\alpha\mathbf{h}_i$ fixed. Unconditional suppression ($-\beta$) slightly degrades performance, whereas prior-aware suppression ($-\beta\mathbf{h}_i$) achieves the best results, outperforming unconditional suppression by 0.24 oIoU and 0.23 mIoU points.

4.5.3 Cascaded Prompting

We examine whether the CLIPSeg prior can achieve comparable improvements through standard SAM 3 prompting without internal attention editing. Specifically, we convert CLIPSeg predictions into mask, box, and point prompts for SAM 3. As shown in Table 5, all three cascades underperform vanilla SAM 3 on RefCOCO val, whereas eSAM improves oIoU by 6.39 points. These results indicate that conventional prompting with the same spatial prior does not reproduce the benefits of internal editing.

4.5.4 Prior Modulation

We examine different modulation functions for prior-aware SOT suppression, defined as $\mathbf{B}_{i,\mathcal{I}_S} = -\beta g(\mathbf{h}_i)$. Table 6 shows that all prior-aware variants outperform unconditional suppression by 0.21–0.34 oIoU on RefCOCO val. Similar results hold on RefCOCOg val, where linear, square, and hard-threshold modulation achieve 54.44, 54.49, and 54.48 oIoU, respectively. We adopt linear modulation for its simplicity and absence of additional hyperparameters.

Table 5: **Cascaded prompting** on RefCOCO val.

Method	oIoU	Δ vs. SAM 3
CLIPSeg	30.19	-17.04
CLIPSeg $\rightarrow$ SAM 3 (mask)	30.91	-16.32
CLIPSeg $\rightarrow$ SAM 3 (box)	28.14	-19.09
CLIPSeg $\rightarrow$ SAM 3 (point)	29.50	-17.73
SAM 3	47.23	—
eSAM	**53.62**	$+$**6.39**

Table 6: **Prior-modulation** on RefCOCO val.

Modulation	$g(\mathbf{h}_i)$	oIoU	Δ
Unconditional	1	53.38	—
Linear (used)	$\mathbf{h}_i$	53.62	$+0.24$
Norm-peak	$\mathbf{h}_i / \max(\mathbf{h})$	53.60	$+0.22$
Square	$\mathbf{h}_i^2$	53.61	$+0.23$
Cube	$\mathbf{h}_i^3$	53.59	$+0.21$
Hard τ=0.3	$\mathbf{1}[\mathbf{h}_i > 0.3]$	53.72	$+0.34$
Hard τ=0.5	$\mathbf{1}[\mathbf{h}_i > 0.5]$	53.70	$+0.32$

4.5.5 Layer Sensitivity Analysis

We select the intervention layer on RefCOCO val and keep it fixed across all RIS and RVOS benchmarks. Table 7 shows that L1 consistently achieves the highest oIoU across three RIS datasets, while joint L1+L5 editing provides no further gains. This may be because L1 operates after initial image-text interactions while leaving sufficient fusion layers to propagate the edited features. We also

Table 7: **eSAM layer sweep** across three RIS datasets (oIoU). L1+L5 edits both layers.

Dataset	SAM 3	L0	L1	L2	L3	L4	L5	L1+L5
RefCOCO	47.23	34.28	**53.62**	42.93	43.15	46.59	48.63	52.54
RefCOCOg	44.43	42.23	**54.44**	46.76	48.66	48.90	51.23	54.02
gRefCOCO	26.45	31.42	**39.73**	34.26	35.08	35.31	36.70	39.38

Table 8: **Performance and efficiency comparison.** We report single-image inference cost on an NVIDIA A6000 GPU at 1008×1008 resolution.

Method	RefCOCO val oIoU↑	MeViS $\mathcal{J}\&\mathcal{F}$ ↑	Latency (s↓)	FLOPs (G↓)	Params (B↓)
RefAM	42.67	<u>30.65</u>	3.97	150,500	17.5
SAM 3	<u>47.23</u>	29.23	**0.146**	**2,551**	**0.84**
eSAM (Ours)	**53.62**	**39.40**	<u>0.170</u>	<u>2,590</u>	<u>0.99</u>

measure the L2 distance between edited and unedited patch features on RefCOCOg val. The target-to-background change ratio increases from 1.641 at L0 to 1.976 at L1 (20.4%), with L1 achieving a higher ratio on 82.4% of samples. These results indicate more target-focused feature changes at L1, consistent with its superior segmentation performance.

4.6 Efficiency and MLLM Compatibility

Inference Efficiency. Table 8 compares the performance and computational cost of eSAM with RefAM and SAM 3. Compared with SAM 3, eSAM improves RefCOCO val oIoU from 47.23 to 53.62 and MeViS $\mathcal{J}\&\mathcal{F}$ from 29.23 to 39.40, while introducing only 1.5% additional FLOPs (2,551G to 2,590G). The AME and VTE operations require no additional parameters, negligible FLOPs, and only 1.59 ms of editing latency. The CLIPSeg spatial prior accounts for the additional 0.15B parameters and 39G FLOPs. Overall, eSAM increases per-image inference latency from 0.146 s to 0.170 s on an NVIDIA A6000.

Compatibility with MLLM-assisted Methods. Recent MLLM-assisted methods incorporate additional reasoning and prompt refinement into the segmentation pipeline. RESAnything [34] employs MLLM-assisted attribute prompting and selection, reporting 12.1 s per image. EVOL-SAM3 [43] uses MLLM-based prompt evolution (23.2 s), while Tarot-SAM3 [47] combines MLLM-assisted prompting with mask selection (15.2 s). eSAM instead edits SAM 3's internal attention without requiring an external MLLM. It can also complement MLLM-based prompting: integrating eSAM into EVOL-SAM3 improves RefCOCO val oIoU from 68.7 to 69.4 ($+0.7$), with negligible additional inference latency relative to the full pipeline.

4.7 Hyperparameters and Sensitivity Analysis

AME uses $\alpha = 5$ for content-token boosting and $\beta = 10$ for prior-aware SOT suppression, while VTE uses $\gamma = 0.5$ for content-value amplification. We select this configuration on RefCOCO val and keep it fixed across all reported RIS and RVOS benchmarks. We also evaluate the sensitivity on RefCOCOg val by varying one coefficient at a time, with the other two fixed at their default values. Figure 4 shows the sweeps for $\alpha \in \{1, \ldots, 10\}$, $\beta \in \{1, \ldots, 10\}$, and $\gamma \in \{0.1, 0.2, \ldots, 1.0\}$. The oIoU ranges are 53.83–54.44 for α, 54.27–54.51 for β, and 53.65–54.45 for γ, corresponding to spreads of 0.61, 0.24, and 0.80 points, respectively. Performance varies by less than one oIoU point within each tested range, indicating that the method does not depend on narrowly tuned coefficients.

4.8 Qualitative Analysis

Figure 5 compares the segmentation results of RefAM, SAM 3, and eSAM on RefCOCO. Despite its strong segmentation capabilities, SAM 3 can struggle with complex referring expressions, consistent with the attention-sink issue analyzed above. For instance, in the third sample, SAM 3 incorrectly segments a bottle with a yellow cap, failing to capture the spatial relation and negated attribute in

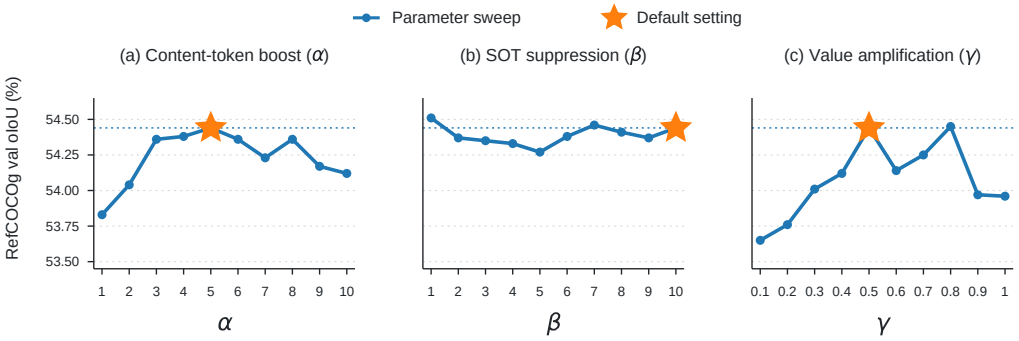

Figure 4: **Hyperparameter sensitivity of full prior-aware eSAM on RefCOCOg val.** We vary (a) content-token boosting α, (b) prior-aware SOT suppression β, and (c) value amplification γ, changing one parameter at a time while keeping the others fixed. Stars indicate the default configuration $(\alpha, \beta, \gamma) = (5, 10, 0.5)$, which achieves 54.44 oIoU. The dotted horizontal line marks this default performance; all three panels use the same vertical scale.

"clear bottle between candles without yellow cap". In contrast, eSAM improves visual grounding by correcting attention routing through AME and enhancing content-token contributions through VTE. In the second and last samples, eSAM redirects attention from irrelevant regions toward the target objects, yielding more accurate segmentation. In the fifth sample, eSAM captures the key semantic cue "blue umbrella" to identify the intended person. These results demonstrate the benefits of attention editing for accurate referring segmentation.

4.9 Generalization Analysis

Table 9 shows that eSAM consistently outperforms SAM 3 on natural paraphrases, lexical negation, and multi-target expressions, with oIoU gains of 6.72, 6.22, and 18.88 points, respectively. For 721 paraphrases of 250 fixed RefCOCO targets, prediction consistency increases from 0.50 to 0.60, while the mean IoU standard deviation decreases from 0.2254 to 0.1789.

Table 9: **oIoU on referring expressions**.

Subset	N	SAM 3	eSAM	Δ
Natural paraphrases	721	49.34	56.06	+6.72
Lexical negation	145	16.92	23.14	+6.22
Multi-target	250	24.12	43.00	+18.88

5 Conclusion

In this work, we identify the attention sink in the multimodal fusion encoder of SAM 3, where the SOT token receives disproportionately high cross-attention on the complete RefCOCOg validation set, potentially limiting the influence of referential information on visual features. To address this issue, we introduce eSAM, a training-free framework that edits cross-attention at two complementary stages: Attention Mask Editing (AME) suppresses SOT dominance and redirects attention toward content tokens, while Value Tensor Editing (VTE) enhances their contribution to the attention output. Our experiments demonstrate that internal attention editing improves referring segmentation even without an external spatial prior, with CLIPSeg providing further refinement. Without updating model parameters, eSAM achieves the strongest performance among the compared training-free methods without auxiliary MLLM controllers across the evaluated RIS and RVOS benchmarks.

6 Future Work and Limitations

Our findings suggest several directions for future work. First, an important next step is to examine why L1 is particularly effective and whether this preference extends to other vision foundation models. Second, routing-stage editing relies on spatial priors to guide attention redistribution. While alternative prior sources yield consistent gains, stronger or task-adaptive priors, including learned priors and more capable off-the-shelf localization models, may further improve the guidance signal. Finally, eSAM mitigates the attention-sink phenomenon through inference-time editing of a frozen

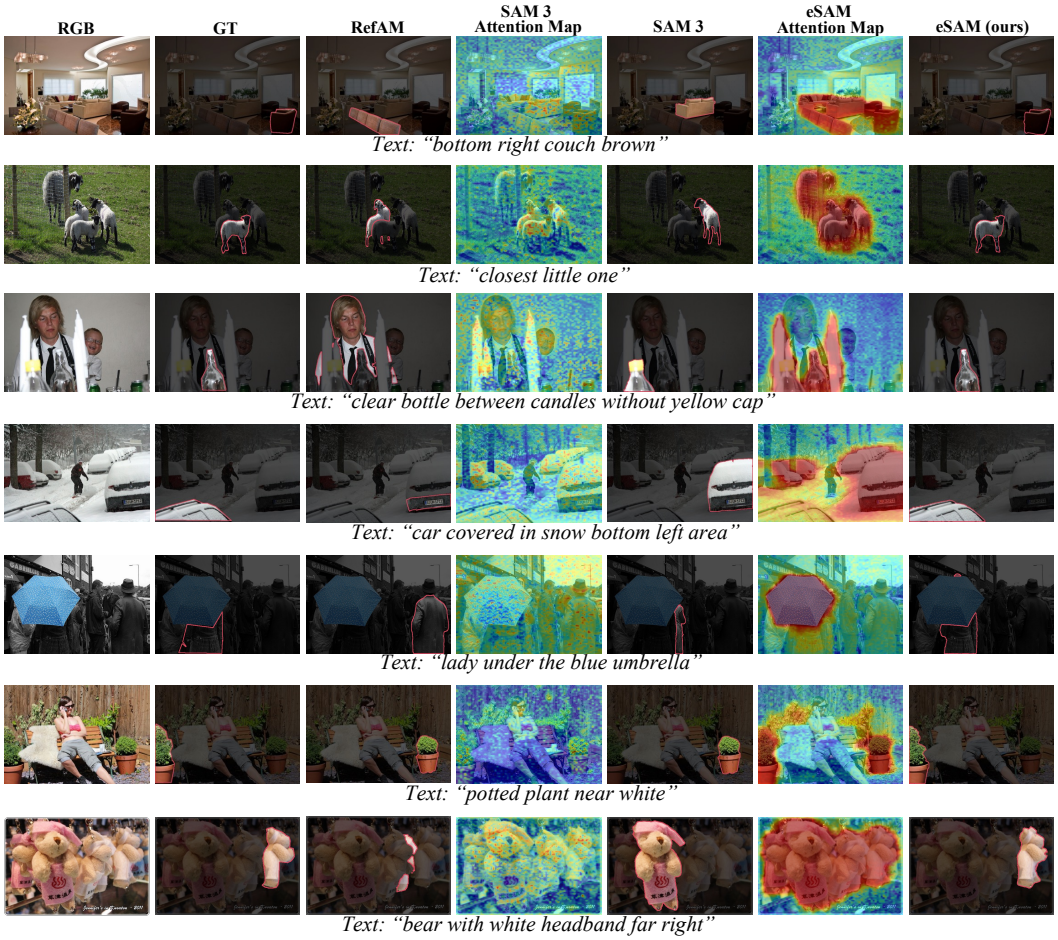

Figure 5: **Qualitative comparison on RefCOCO.** For each expression, we show the RGB image, ground-truth mask, RefAM prediction, SAM 3 attention map and prediction, and eSAM attention map and prediction.

model. Monitoring routing and readout signals during training may enable auxiliary objectives that discourage excessive attention concentration during optimization.

7 Broader Impacts

This work advances training-free referring segmentation, potentially benefiting applications such as assistive visual interfaces, human-robot interaction, and language-guided video understanding. By improving text-based object localization without task-specific training, eSAM may facilitate the efficient adaptation of foundation models to downstream applications while reducing annotation and computational costs. However, more accurate referring segmentation could also enable privacy-invasive object tracking or surveillance, particularly in large-scale video analysis. In addition, segmentation errors may affect the reliability of downstream systems, especially in safety-critical applications. Responsible deployment should therefore consider privacy protection, human oversight, and evaluation under real-world conditions.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62472104 and the Science and Technology Commission of Shanghai Municipality under Grant No. 25511103600, and the Shanghai Pujiang Program 2025PJA201. Henghui Ding was supported by the Xiaomi Young Scholars Program.

References

[1] 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, pages 138846–138923, 2026.

[2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[3] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16321–16330, 2021.

[4] 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*, pages 2694–2703, 2023.

[5] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20224–20234, 2023.

[6] Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. MeViS: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[7] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025.

[8] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Yu-Gang Jiang. GREx: Generalized referring expression segmentation, comprehension, and generation. *International Journal of Computer Vision*, 134(2):79, 2026.

[9] Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. Multimodal referring segmentation: A survey. *International Journal of Computer Vision*, 134(9):405, 2026.

[10] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024.

[11] Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, and Yu Qiao. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13414–13423, 2023.

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

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[14] Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3715–3723, 2025.

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

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

[17] Anna Kukleva, Enis Simsar, Alessio Tonioni, Muhammad Ferjad Naeem, Federico Tombari, Jan Eric Lenssen, and Bernt Schiele. Refam: Attention magnets for zero-shot referral segmentation. *arXiv preprint arXiv:2509.22650*, 2025.

[18] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

[19] Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. Referdino: Referring video object segmentation with visual grounding foundations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20009–20019, 2025.

[20] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023.

[21] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18653–18663, 2023.

[22] Ting Liu and Siyuan Li. Hybrid global-local representation with augmented spatial guidance for zero-shot referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29634–29643, 2025.

[23] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022.

[24] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[25] Minheng Ni, Yabo Zhang, Kailai Feng, Xiaoming Li, Yiwen Guo, and Wangmeng Zuo. Ref-diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*, 2023.

[26] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[27] 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.

[28] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[29] Fu Rong, Meng Lan, Qian Zhang, and Lefei Zhang. Mpg-sam 2: Adapting sam 2 with mask priors and global context for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23979–23989, 2025.

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[31] 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*, pages 208–223. Springer, 2020.

[32] Anushka Sivakumar, Andrew Zhang, Zaber Hakim, and Chris Thomas. Steervlm: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23640–23665, 2025.

[33] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13171–13182, 2024.

[34] Ruiqi Wang and Hao Zhang. Resanything: Attribute prompting for arbitrary referring segmentation. *Advances in Neural Information Processing Systems*, 38:66893–66944, 2026.

[35] Shijie Wang, Dahun Kim, Ali Taalimi, Chen Sun, and Weicheng Kuo. Learning visual grounding from generative vision and language model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8057–8067. IEEE, 2025.

[36] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5669–5677, 2024.

[37] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. In *European conference on computer vision*, pages 196–213. Springer, 2024.

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

[39] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022.

[40] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pages 21875–21895, 2024.

[41] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6449–6457, 2024.

[42] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18155–18165, 2022.

[43] Kai Ye, Xiaotong You, Jianghang Lin, Jiayi Ji, Pingyang Dai, and Liujuan Cao. Evolving, not training: Zero-shot reasoning segmentation via evolutionary prompting. *arXiv preprint arXiv:2512.24702*, 2025.

[44] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *CVPR*, 2023.

[45] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Pseudo-ris: Distinctive pseudo-supervision generation for referring image segmentation. In *European Conference on Computer Vision*, pages 18–36. Springer, 2024.

[46] Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for llms. *arXiv preprint arXiv:2311.02262*, 2023.

[47] Weiming Zhang, Dingwen Xiao, Songyue Guo, Guangyu Xiang, Shiqi Wen, Minwei Zhao, Lei Chen, and Lin Wang. Tarot-sam3: Training-free sam3 for any referring expression segmentation, 2026. URL `https://arxiv.org/abs/2604.07916`. Withdrawn preprint.

[48] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pages 696–712. Springer, 2022.

A Extended Attention Sink Statistics

Figure 2 visualizes the SOT attention sink on 500 RefCOCO val samples. We extend the statistics to the complete RefCOCOg validation set (4,896 expressions). The SOT token receives 44.81% of cross-attention mass on average, with a 5th–95th percentile range of 24.40%–66.08%. The per-layer mean SOT masses at L0–L5 are 28.06%, 35.59%, 46.82%, 44.57%, 59.89%, and 53.93%, respectively. We further group RefCOCOg expressions by content-token length (Table 10). Let A_S denote the attention percentage assigned to the SOT token and A_C the average attention percentage assigned to an *individual* content token. As expressions become longer, the absolute SOT attention share decreases, but the SOT token still receives 8.30–9.50 times the mean attention of a single content token. eSAM improves oIoU in all four groups, with larger gains on longer expressions in this analysis. The per-token attention imbalance therefore persists across expression lengths.

Table 10: **Expression-length analysis** on the complete RefCOCOg validation set. A_S and A_C are attention percentages; Δ oIoU denotes the improvement of eSAM over SAM 3.

Content tokens	Samples	A_S	A_C	A_S/A_C	SAM 3 oIoU	Δ oIoU
≤ 6	1,642	55.57	5.85	9.50	54.50	$+6.98$
7–8	988	45.62	5.13	8.89	42.69	$+8.65$
9–11	1,243	40.25	4.63	8.69	40.91	$+9.37$
≥ 12	1,023	32.30	3.89	8.30	34.63	$+17.62$

B Mechanistic Analysis of AME and VTE

AME: direct intervention on attention routing. For image patch i, let $a_{S,i}$ be the pre-edit mass assigned to the SOT token and $a_{C,i} = 1 - a_{S,i}$ the aggregate mass on content tokens, using the partition in Section 3.1.1. To isolate the mechanism, consider AME without a spatial prior: it adds $-\beta$ to the SOT logit and $+\alpha$ to each content-token logit. Softmax normalization gives

$$a'_{S,i} = \frac{a_{S,i}e^{-\beta}}{a_{S,i}e^{-\beta} + a_{C,i}e^{\alpha}} = \frac{a_{S,i}}{a_{S,i} + a_{C,i}e^{\alpha+\beta}}, \tag{7}$$

$$\frac{a'_{C,i}}{a'_{S,i}} = e^{\alpha+\beta}\frac{a_{C,i}}{a_{S,i}}. \tag{8}$$

For $\alpha, \beta > 0$ and $a_{C,i} > 0$, $a'_{S,i} < a_{S,i}$. AME therefore reduces the SOT token's *relative routing mass* by construction. In eSAM, the spatial prior replaces $\alpha + \beta$ in the ratio with $(\alpha + \beta)h_i$, so the intervention is stronger at patches with larger h_i and vanishes where $h_i = 0$.

VTE: direct intervention on value readout. For fixed attention weights $\mathbf{A}$, VTE leaves the SOT value unchanged and scales each content-token value by $1 + \gamma$ (Equation 5). The resulting output is

$$\widetilde{\mathbf{O}}_i = \mathbf{A}_{i,S}\mathbf{V}_S + (1+\gamma)\sum_{j\in\mathcal{I}_C}\mathbf{A}_{i,j}\mathbf{V}_j. \tag{9}$$

Thus, a VTE intervention amplifies the aggregate content-token readout without changing routing weights. AME and VTE address routing and readout, respectively. Figure 3 shows the corresponding shifts under the combined method.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims and contributions are summarized in the Abstract and Introduction and supported by the experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It can be found in Section Future Work and Limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper presents an empirical method and does not introduce formal theoretical results requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The Datasets and Metrics section specifies the evaluation datasets and metrics, while the Hyperparameters and Sensitivity Analysis appendix details the method's hyperparameters and their selection.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release all code related to the data and models. All datasets used in this paper are publicly available and subject to their respective licenses.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The experimental settings are described in the Datasets and Metrics section and the Hyperparameters and Sensitivity Analysis appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report extensive ablation studies and hyperparameter sensitivity analyses, but do not provide error bars or formal statistical significance tests.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Table 6 shows performance and efficiency comparison.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide impacts of the work in the paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers of all datasets and methods used in our work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce or release new datasets, model weights, or other standalone research assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.