EviSAM3: Evidence-Driven SAM3 for Referring Remote Sensing Image Segmentation

Yuhang Duan[1,2,*] **Xiaoshuai Wu**[1,*] **Chengqi Zhang**[3]
Jincheng Lu[4] Mingwei Zhang[5] Yu Liu[1,†]

[1]School of Software Technology, Dalian University of Technology, Dalian, China
[2]Shien-Ming Wu School of Intelligent Engineering,
South China University of Technology, Guangzhou, China
[3]Department of Energy and Power Engineering, Tsinghua University, Beijing, China
[4]Department of Engineering Physics, Tsinghua University, Beijing, China
[5]School of Computer Science and Technology,
Dalian University of Technology, Dalian, China
[*]Equal contribution. [†]Corresponding author.
wiyhduan@mail.scut.edu.cn wxs@mail.dlut.edu.cn
zhangcq23@mails.tsinghua.edu.cn lujc22@mails.tsinghua.edu.cn
yuliu@dlut.edu.cn

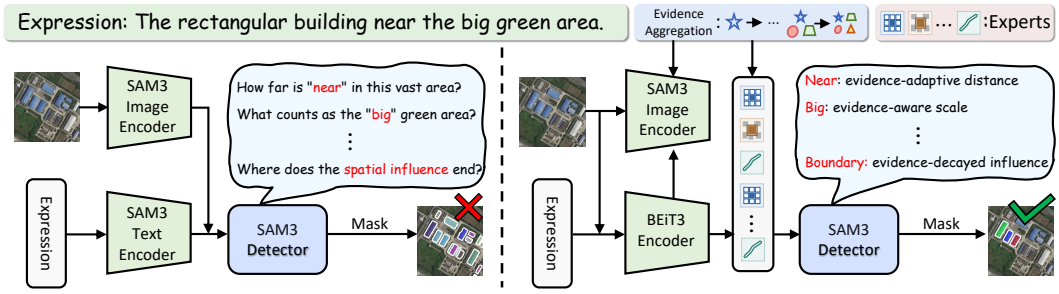
Abstract

Referring Remote Sensing Image Segmentation (RRSIS) aims to segment target objects from remote sensing images based on natural language descriptions, requiring precise vision–language alignment under significant ambiguity. Despite the strong generalization ability of Segment Anything Model 3 (SAM3), the large spatial extent and varying scales of targets in remote sensing images intensify the ambiguity of referring expressions. This poses a major challenge for fine-grained segmentation. To address the challenge, we propose **EviSAM3**, an evidence-driven parameter-efficient adaptation framework inspired by progressive evidence integration across backbone layers. Specifically, EviSAM3 progressively integrates cross-modal evidence through an evidence memory bank, where each incoming piece is aggregated via momentum-based updates. As evidence accumulates, the model performs iterative evidence rectification to resolve ambiguity under incomplete observations. Meanwhile, the contributions of different evidence are adaptively modulated based on their diagnostic relevance, allowing more informative evidence to dominate the reasoning process and guide the final prediction. Extensive experiments demonstrate that EviSAM3 consistently outperforms state-of-the-art methods, particularly in challenging scenarios with high ambiguity and significant scale variation. The code is available at `https://github.com/YaofireJinglian/EviSAM3`.

1 Introduction

Referring Remote Sensing Image Segmentation (RRSIS) is the task of segmenting target regions in remote sensing images based on natural language descriptions [1–3]. It can be viewed as a specialized subtask of Referring Image Segmentation (RIS) , focusing on the unique properties of remote sensing imagery [4–7]. This task holds significant practical potential and can be applied to various real-world scenarios, including disaster assessment [8], urban analysis [9], and environmental monitoring [10].

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



(a) Ambiguity in Vanilla SAM3 (b) Evidence-driven Adaptation (Ours)

Figure 1: Comparison between (a) the vanilla SAM3-based baseline and (b) our proposed EviSAM3.

Recently, Segment Anything Model 3 (SAM3) [11–13] has demonstrated strong generalization ability in generic image segmentation, benefiting from large-scale pre-training and its promptable design. Subsequent studies have further explored its adaptability across diverse domains, including medical image segmentation [14], camouflage object segmentation [15], and text-guided annotation systems, highlighting its strong transferability and extensibility [16]. Despite these advances, directly applying SAM3 to RRSIS remains challenging. Remote sensing imagery involves large spatial extents and extreme scale variations, which amplify the ambiguity of referring expressions and hinder precise vision–language alignment.

Despite its strong generalization ability, SAM3 fundamentally relies on concept-level prompts, which are typically represented as coarse noun phrases [13, 17]. This design prevents it from modeling structured referring expressions involving spatial relations, relative attributes, and contextual reasoning. As illustrated in Fig. 1a, spatial descriptions such as "near" and "big" lack explicit and consistent definitions in large-scale remote sensing imagery. As illustrated in Fig. 1(b), spatial terms such as "near" and "big" are inherently ambiguous, lacking consistent definitions across scenes and scales. Such ambiguity leads to inconsistent and unreliable prompt interpretation, causing SAM3 to produce inaccurate segmentation results. More critically, SAM3 performs category-level matching that identifies all instances of a category, rather than grounding a specific instance described by relational or contextual cues. This paradigm mismatch makes it fundamentally unsuitable for RRSIS, where the target is often defined by ambiguous and context-dependent descriptions rather than explicit semantic categories.

Inspired by progressive evidence integration in cognitive accounts of decision-making [18–20], we integrate complementary visual and linguistic cues across backbone depth. Coarse semantic cues narrow the candidate regions, while spatial relations, relative attributes, and scene context refine the target representation. Our computational mechanism maintains a persistent latent evidence state and uses it to rectify subsequent features and guide prompt generation. This is an architectural inspiration rather than a full model of human decision-making: it does not implement Bayesian uncertainty updates or a decision stopping rule.

Based on the above motivations, we propose Evidence-Driven SAM3 (**EviSAM3**), a framework that enables SAM3 to interpret ambiguous referring expressions in remote sensing imagery through progressive evidence reasoning. As illustrated in Fig. 1b, instead of relying on single-step prompt interpretation, EviSAM3 models grounding as the gradual integration of multiple cues, progressively refining target localization. We design an Evidence Integration Module (**EIM**) to model the integration of visual–linguistic evidence. EIM maintains a memory bank of intermediate evidence and iteratively infers new evidence through cross-modal interactions. The inferred evidence is then combined with historical evidence via a progressive update mechanism, enabling more reliable interpretation of ambiguous spatial descriptions and contextual cues. Inspired by Mixture-of-Experts architectures in large language models, where multiple experts specialize in different knowledge patterns [21–23], we further design an Evidence-Driven Mixture-of-Experts (**ED-MoE**) Prompting strategy to convert accumulated evidence into segmentation guidance. ED-MoE dynamically selects specialized experts for different visual patterns, such as structural details, cluttered scenes, and spatial configurations, and produces refined prompts to guide the SAM3 mask decoder for accurate segmentation. The contributions are summarized as follows:

- We analyze the limitations of the vanilla SAM3 when applied to RRSIS and reveal that its concept-level prompting paradigm oversimplifies referring segmentation as a category matching problem, which fails to handle ambiguous expressions involving spatial relations, relative attributes, and contextual cues.
- We propose **EviSAM3**, an evidence-driven framework that reformulates referring segmentation as a progressive evidence reasoning process, enabling SAM3 to gradually refine target localization under ambiguous descriptions.
- We design an Evidence Integration Module (**EIM**) to progressively integrate visual–linguistic evidence through a memory-based update mechanism, and design an Evidence-Driven Mixture-of-Experts (**ED-MoE**) Prompting strategy that dynamically routes evidence-aware features to specialized experts for generating refined segmentation prompts.
- Extensive experiments on RRSIS benchmarks demonstrate that EviSAM3 effectively resolves referential ambiguity and significantly outperforms state-of-the-art methods.

2 Related Work

Referring Remote Sensing Image Segmentation. RRSIS is an extension of Referring Image Segmentation (RIS) [4, 24–28]. Early work [3, 29, 30] established the RefSegRS dataset and introduced the language-guided cross-scale enhancement (LGCE) module to improve the detection of small and scattered objects. To address challenges such as large scale variations and complex orientations in aerial imagery, subsequent methods have explored more effective multi-scale and cross-modal interactions. For instance, RMSIN [31] incorporates adaptive rotated convolutions for scale-aware feature interaction, while SBANet [32] models scale-wise bidirectional alignment to better capture global and local information. Later approaches, including CroBIM [33] and DANet [34], further enhance cross-modal alignment through iterative interaction and explicit discrepancy reduction. More recently, the emergence of foundation models has driven new frameworks such as RSRefSeg [35], which leverages CLIP [36] and SAM [11] for improved generalization, and RS2-SAM2 [1], which adapts SAM2 [12] to remote sensing via hierarchical fusion strategies. While existing methods achieve promising results, they largely rely on one-step vision–language alignment, limiting their ability to handle ambiguity and complex spatial reasoning in remote sensing scenarios.

Segment Anything Model. The Segment Anything Model (SAM) [11] introduced a promptable segmentation paradigm with strong zero-shot transfer on the billion-mask SA-1B dataset. SAM 2 [12, 37] extended it to video via a streaming memory, while SAM 3 [13] further replaces geometric prompts with concept prompts, segmenting all instances matching a noun phrase or exemplar. Since SAM variants are pretrained on natural images, several works adapt them to remote sensing through task-specific prompting [38], domain adapters [39, 40], parameter-efficient tuning [41–43], or SAM 3-based open-vocabulary aerial segmentation [44]. HQ-SAM improves fine-grained mask quality, while PerSAM enables one-shot personalization for target-specific segmentation [45, 46]. However, these methods still treat textual prompts as deterministic signals, leaving the cross-modal ambiguity of aerial referring expressions unmodeled. This motivates a closer look at how uncertainty can be incorporated into SAM-based referring segmentation.

3 Method

3.1 Overview

Given a remote sensing image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a referring expression $\mathbf{E} = \{e_l\}_{l=1}^{L}$, the task is to predict a binary mask $\mathbf{M} \in \{0,1\}^{H\times W}$ for the referred target. As shown in Fig. 2, EviSAM3 follows four connected steps: (1) BEiT-3 prior extraction and SAM3 image encoding; (2) EIM cross-modal fusion, evidence-memory update, and feature rectification across backbone layers; (3) ED-MoE routing and dense/sparse prompt generation; and (4) SAM3 mask decoding [13].

We first extract compact multimodal priors from a frozen BEiT-3 vision–language encoder [47, 1], using a low-resolution image (224×224) and tokenized text. This encoder outputs a multimodal class token $\mathbf{v}_{\text{cls}}$, visual embeddings $\mathbf{V}$, and text embeddings $\mathbf{T}$. The class token is fused with visual tokens by cross-attention followed by element-wise interaction: $\mathbf{v}'_{\text{cls}} = \mathbf{v}_{\text{cls}} \odot \text{MHCA}(\mathbf{v}_{\text{cls}}, \mathbf{V}, \mathbf{V})$,

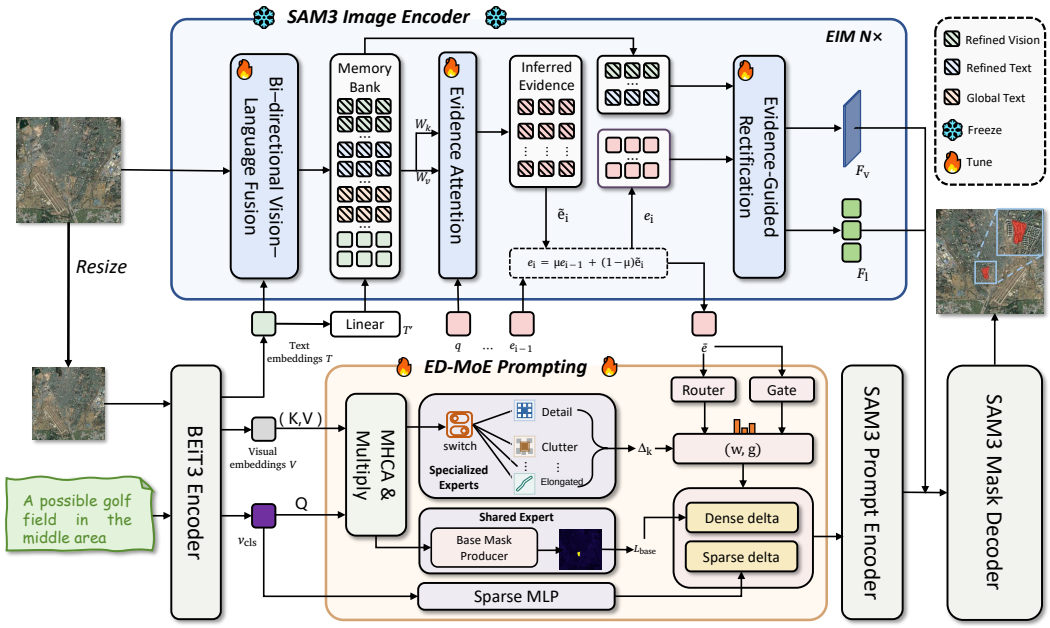


Figure 2: EviSAM3 pipeline. BEiT-3 supplies visual tokens $\mathbf{V}$, text tokens $\mathbf{T}$, and the class token $\mathbf{v}_{\text{cls}}$; the projected text $\mathbf{T}'$ enters EIM. Learned queries $\mathbf{q}$ extract new evidence $\tilde{\mathbf{e}}_i$, which updates $\mathbf{e}_{i-1}$ across depth and rectifies features. $\mathbf{F}_v$ and $\mathbf{F}_l$ denote the final visual and text features. ED-MoE uses the final evidence $\bar{\mathbf{e}}$ to predict weights $\mathbf{w}$ and gate g for expert corrections $\Delta_k = (\Delta\mathbf{L}_k, \Delta\mathbf{s}_k)$. These refine base logits $\mathbf{L}_{\text{base}}$ and sparse tokens into dense and sparse prompts for SAM3 decoding.

which provides the base signal for downstream prompt construction. Meanwhile, $\mathbf{T}$ is projected to $\mathbf{T}' \in \mathbb{R}^{N_t \times d}$ (d=256) and fed into adapter blocks for progressive cross-modal evidence reasoning. The final dense and sparse prompts are then decoded by the SAM3 mask decoder [13].

3.2 Evidence Integration Module

While the multimodal prior extractor provides a joint vision–language representation, the SAM3 backbone operates on high-resolution images (1008×1008) without direct access to textual inputs. To enable progressive cross-modal reasoning within the mostly frozen backbone, we inject an *Evidence Rectified Adapter* (ERA) into each of the $N_b = 32$ ViT blocks.

Bidirectional Vision–Language Fusion. At backbone block i, let $\mathbf{F}_i \in \mathbb{R}^{H_i W_i \times C}$ and $\mathbf{T}_i \in \mathbb{R}^{N_t \times d}$ denote visual and text tokens, respectively; H_i, W_i are the feature-map dimensions and C is the visual channel width. The first text input is $\mathbf{T}_1 = \mathbf{T}'$. Learned projections map both branches to width $D_h = C/4$, yielding $\mathbf{F}_{h,i}$ and $\mathbf{T}_{h,i}$. Cross-attention first updates vision from text and then text from the updated vision:

$$\hat{\mathbf{F}}_i = \mathbf{F}_{h,i} + \text{MHCA}(\mathbf{F}_{h,i}, \mathbf{T}_{h,i}, \mathbf{T}_{h,i}), \quad \hat{\mathbf{T}}_{h,i} = \mathbf{T}_{h,i} + \text{MHCA}(\mathbf{T}_{h,i}, \hat{\mathbf{F}}_i, \hat{\mathbf{F}}_i). \tag{1}$$

Here MHCA takes query, key, and value tokens, in that order.

Progressive Evidence Accumulation. The fused features supply new cross-modal evidence to a persistent token memory. Learnable queries $\mathbf{q} \in \mathbb{R}^{N_e \times d}$ attend to the input text, projected fused text, and pooled visual tokens:

$$\mathbf{m}_i = [\mathbf{T}_i;\ \phi_t(\hat{\mathbf{T}}_{h,i});\ \phi_v(\text{Pool}(\hat{\mathbf{F}}_i))], \quad \tilde{\mathbf{e}}_i = \text{MHCA}(\mathbf{q}, \mathbf{m}_i, \mathbf{m}_i). \tag{2}$$

The projections ϕ_t, ϕ_v map tokens to width $d = 256$; concatenation is along the token dimension. With $N_e = 8$ evidence tokens, the state is updated across backbone depth as

$$\mathbf{e}_1 = \tilde{\mathbf{e}}_1, \qquad \mathbf{e}_i = \mu\mathbf{e}_{i-1} + (1-\mu)\tilde{\mathbf{e}}_i \quad (i > 1), \qquad \mu = 0.6. \tag{3}$$

This update integrates partial cues across layers within one forward pass; it is not a Bayesian posterior update.

Evidence-Guided Residual Rectification. Let $\mathbf{E}_i = \psi(\mathbf{e}_i)$ denote the accumulated evidence projected to the hidden width. We define the text-branch evidence-guided update operator explicitly as

$$U(\mathbf{X}, \mathbf{E}_i) = \mathrm{LN}\big(\mathrm{MHCA}(\mathbf{X}, \mathbf{E}_i, \mathbf{E}_i)\big). \tag{4}$$

Thus $U(\mathbf{T}_i, \mathbf{E}_i)$ denotes a text-shaped correction after projection to the shared hidden width, with text queries attending to evidence keys and values; in the implementation it is evaluated on the fused hidden text $\hat{\mathbf{T}}_{h,i}$. The visual branch uses an analogous operator U_v with separate parameters. Both branches are rectified before passing to the next block:

$$\mathbf{F}_i^{\mathrm{rect}} = \hat{\mathbf{F}}_i + \sigma(g_{v,i})\big(\boldsymbol{\lambda}_v \odot U_v(\hat{\mathbf{F}}_i, \mathbf{E}_i)\big), \quad \mathbf{T}_{h,i}^{\mathrm{rect}} = \hat{\mathbf{T}}_{h,i} + \sigma(g_{t,i})\big(\boldsymbol{\lambda}_t \odot U(\hat{\mathbf{T}}_{h,i}, \mathbf{E}_i)\big). \tag{5}$$

Here LN denotes layer normalization, σ is sigmoid, and $g_{v,i}, g_{t,i}$ are learned gates conditioned on pooled evidence. Gate biases start at -6 and channel-wise LayerScale vectors $\boldsymbol{\lambda}_v, \boldsymbol{\lambda}_t$ at 10^{-3}. Output projections restore the original feature widths, with a visual residual connection and a residual blend of the input and updated text. The resulting features and evidence state feed the next block; the final evidence guides ED-MoE below.

3.3 Evidence-Driven Mixture-of-Experts Prompting and Decoding

After all N_b blocks, the accumulated evidence $\bar{\mathbf{e}}$ is aggregated from the available backbone output layers (in our default setting, this is primarily the last output layer). The base mask logits are computed by element-wise multiplying the fused class token $\mathbf{v}'_{\text{cls}}$ with the reshaped visual embeddings $\mathbf{V}_{2D}$, followed by linear projection: $\mathbf{L}_{\text{base}} = \mathrm{Linear}(\mathbf{V}_{2D} \odot \mathbf{v}'_{\text{cls}})$.

Expert Routing and Specialization. ED-MoE routes the evidence to three specialized experts: (1) a **Detail Expert** using 3×3 depthwise convolution for fine-grained boundaries; (2) an **Elongated Expert** using Adaptive Rotated Convolution (ARC) [31] for oriented structures (*e.g.*, roads, rivers); and (3) a **Clutter Expert** using GroupNorm + 1×1 convolution for cluttered backgrounds. The routing weights are:

$$\mathbf{r} = [\bar{\mathbf{e}}_{\text{pool}};\ \beta \cdot \mathbf{t}_{\text{cls}}], \quad \mathbf{w} = \mathrm{softmax}\big(\mathrm{MLP}(\mathbf{r})/\tau\big), \tag{6}$$

where $\bar{\mathbf{e}}_{\text{pool}}$ is the token-averaged evidence, $\mathbf{t}_{\text{cls}}$ is the projected text class embedding, β is a learnable coefficient (initialized to 0), and τ=1.5 is the temperature. Each expert produces dense logit deltas $\Delta\mathbf{L}_k$ and token deltas $\Delta\mathbf{s}_k$, which are weighted-summed and gated to form the final outputs:

$$\mathbf{L} = \mathbf{L}_{\text{base}} + g \cdot \gamma \textstyle\sum_k w_k \Delta\mathbf{L}_k, \quad \mathbf{S} = [\mathbf{s}_{\text{cls}};\ g \cdot \gamma' \textstyle\sum_k w_k \Delta\mathbf{s}_k], \tag{7}$$

where $g = \sigma(\mathrm{Linear}(\mathbf{r}))$ is a sigmoid gate (initialized near zero), and γ, γ' are LayerScale parameters (10^{-3}). A balance loss $\mathcal{L}_{\text{bal}} = K \sum_k \bar{w}_k^2$ [21] encourages balanced expert utilization, where $K = 3$ and $\bar{w}_k$ is the batch-mean routing weight for expert k.

The final logits $\mathbf{L}$ are converted into dense mask prompts via sigmoid and a lightweight projection head, then combined with sparse prompt tokens $\mathbf{S}$ and decoded by the SAM3 mask decoder [13]. The decoder performs bidirectional interaction between prompts and multi-scale features (three resolutions when high-resolution features are enabled). Training uses

$$\mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{bal}}\mathcal{L}_{\text{bal}}, \tag{8}$$

with λ_{ce}=0.9, λ_{dice}=0.1, and λ_{bal}=0.01 [48, 21]. The BEiT-3 prior extractor and most SAM3 backbone parameters are frozen; adapters and prompt-generation modules, including ED-MoE, are optimized for the downstream task.

4 Experiment

We conduct a comprehensive evaluation of EviSAM3 to address the following research questions: **RQ1)** how does EviSAM3 perform compared to state-of-the-art methods? **RQ2)** what is the contribution of each component? **RQ3)** how effective is the evidence accumulation mechanism under ambiguous expressions? **RQ4)** what are its limitations in challenging scenarios?

Table 1: Comparison with state-of-the-art methods on the RRSIS-D dataset. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Pr@0.5		Pr@0.6		Pr@0.7		Pr@0.8		Pr@0.9		oIoU		mIoU	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
CMPC [50]	57.93	55.83	48.85	47.40	38.50	36.94	25.28	25.45	9.31	9.19	70.15	69.22	50.41	49.24
BRINet [51]	58.79	56.90	49.54	48.77	39.65	39.12	28.21	27.03	9.19	8.73	70.73	69.88	51.14	49.65
CMPC+ [52]	59.19	57.65	49.36	47.51	38.67	36.97	25.91	24.33	8.16	7.78	70.14	68.64	51.41	50.24
LGCE [3]	68.10	67.65	60.52	61.53	52.24	51.45	42.24	39.62	23.85	23.33	76.68	76.34	60.16	59.37
RIS-DMMI [6]	70.40	68.74	63.05	60.96	54.14	50.33	41.95	38.38	23.85	21.63	77.01	76.20	60.72	60.12
LAVT [53]	69.54	69.52	63.51	63.63	53.16	53.29	43.97	41.60	24.25	24.94	77.59	77.19	61.46	61.04
EVF-SAM [47]	73.51	72.16	67.87	66.50	58.33	56.59	46.15	43.92	25.92	25.48	76.32	76.77	64.03	62.75
FIANet [54]	-	74.46	-	66.96	-	56.31	-	42.83	-	24.13	-	76.91	-	64.01
RMSIN [31]	74.66	74.26	68.22	67.25	57.41	55.93	45.29	42.55	24.43	24.53	78.27	77.79	65.10	64.20
RS2-SAM2 [1]	79.25	77.56	77.56	72.34	63.85	63.85	50.57	47.92	30.40	29.73	80.16	78.99	68.81	66.72
SAM3-Adapter [15]	72.31	73.65	65.19	66.69	55.50	56.08	43.06	40.80	23.28	22.59	74.47	74.44	63.71	63.39
SAM3-LoRA [55]	67.26	67.14	60.95	60.26	51.09	50.37	40.54	38.99	22.13	22.16	72.63	73.03	59.99	59.81
EviSAM3 (Ours)	**83.45**	**82.81**	**77.82**	**77.65**	**66.95**	**66.37**	**53.33**	**50.56**	**32.36**	**30.82**	**81.38**	**80.27**	**71.25**	**70.65**

Table 2: Comparison with state-of-the-art methods on the RefSegRS dataset. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Pr@0.5		Pr@0.6		Pr@0.7		Pr@0.8		Pr@0.9		oIoU		mIoU	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
BRINet [51]	36.86	20.72	35.53	14.26	19.93	9.87	10.66	2.98	2.84	1.14	61.59	58.22	38.73	31.51
EVF-SAM [47]	57.77	35.17	37.59	22.34	16.24	9.36	4.87	2.86	1.86	0.39	59.61	55.51	46.98	36.64
CMPC [50]	46.09	32.36	26.45	14.14	12.76	6.55	7.42	1.76	1.39	0.22	63.55	55.39	42.08	40.63
CMPC+ [52]	56.84	49.19	37.59	28.31	20.42	15.31	10.67	8.12	2.78	2.55	70.62	66.53	47.13	43.65
CARIS [56]	68.45	45.40	47.10	27.19	25.52	15.08	14.62	8.87	3.71	1.98	75.79	69.74	54.30	42.66
LAVT [53]	80.97	51.84	58.70	30.27	31.09	17.34	15.55	9.52	4.64	2.09	78.50	71.86	61.53	47.40
RIS-DMMI [6]	86.17	63.89	74.71	44.30	38.05	19.81	18.10	6.49	3.25	1.00	74.02	68.58	65.72	52.15
LGCE [3]	90.72	73.75	86.31	61.14	71.93	39.46	32.95	16.02	10.21	5.45	83.56	76.81	72.51	59.96
RMSIN [31]	93.97	79.20	89.33	65.99	74.25	42.98	29.70	16.51	7.89	3.25	82.41	75.72	73.84	62.58
FIANet [54]	-	84.09	-	77.05	-	61.86	-	33.41	-	7.10	-	78.32	-	68.67
RS2-SAM2 [1]	95.36	84.31	94.90	79.42	92.58	70.89	83.76	55.70	36.66	21.19	88.03	80.87	80.87	73.90
SAM3-Adapter [15]	81.21	66.70	74.48	56.74	63.57	43.20	42.23	20.25	9.05	4.40	72.52	66.62	69.09	58.54
SAM3-LoRA [55]	86.77	67.58	78.89	56.52	63.11	41.50	37.82	16.46	8.12	4.07	79.28	73.89	70.35	58.11
EviSAM3 (Ours)	**98.14**	**89.93**	**96.29**	**85.25**	**94.90**	**79.58**	**93.27**	**68.63**	**83.53**	**45.95**	**92.44**	**86.51**	**91.32**	**80.56**

4.1 Experimental Settings

Datasets. We evaluate our method on two representative RRSIS benchmarks, RefSegRS [3] and RRSIS-D [31], capturing both semantic and spatial challenges in remote sensing imagery, as well as the inherent ambiguity in referring expressions. RefSegRS consists of 512Œ512 samples with diverse scenes and compositional expressions, emphasizing fine-grained alignment under appearance similarity and scale variation. RRSIS-D, with higher-resolution images (800Œ800), introduces dense instances, arbitrary orientations, and elongated structures, along with complex background clutter, enabling a comprehensive evaluation of model robustness.

Implementation Details. All experiments are conducted on 4 NVIDIA H200 GPUs with a batch size of 4 per GPU, 512GB system memory, and the Ubuntu operating system. We train for 60 epochs on both RefSegRS and RRSIS-D. The AdamW optimizer [49] is employed with a total batch size of 16. The base learning rates are set to 5×10^{-5} for RefSegRS and 1×10^{-5} for RRSIS-D. In the final 10 epochs, all learning rates are decayed to 0.1 of their initial values.

Metrics. Following the evaluation protocol of RS2-SAM2 [1], we measure performance using multiple criteria, including precision under different IoU thresholds (Pr@0.5–Pr@0.9), mean Intersection-over-Union (mIoU), and overall Intersection-over-Union (oIoU). All metrics are reported on the validation and test splits of both RefSegRS and RRSIS-D.

4.2 Comparison with State-of-the-Art Methods (RQ1)

RRSIS-D dataset. As shown in Table 1, EviSAM3 consistently outperforms all prior methods. On the test set, it surpasses the strongest baseline RS2-SAM2 [1] by +5.25 at Pr@0.5 and maintains gains at stricter thresholds (e.g., +1.09 at Pr@0.9), indicating more precise localization under high-overlap criteria. In terms of region-level metrics, EviSAM3 achieves 80.27 oIoU and 70.65

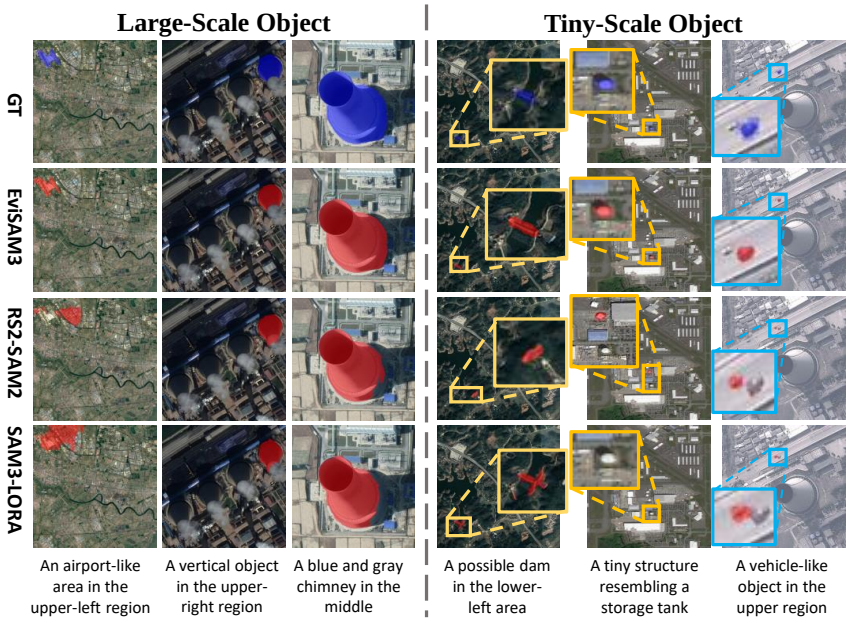


Figure 3: Qualitative comparison results on multi-scale targets under ambiguous referring expressions.

mIoU, improving over RS2-SAM2 [1] by +1.28 and +3.93, respectively, with the larger gain in mIoU reflecting better overall mask quality. Compared with earlier RIS methods such as LAVT [53] and RMSIN [31], the margin is substantial, highlighting the difficulty of handling complex spatial layouts. Furthermore, EviSAM3 also consistently outperforms SAM-based adaptations, including SAM3-Adapter [15] and SAM3-LoRA [55], suggesting that simple parameter-efficient tuning is insufficient for resolving ambiguity in RRSIS.

RefSegRS dataset. As shown in Table 2, EviSAM3 significantly outperforms all prior methods. Compared with RS2-SAM2 [1], it improves Pr@0.5 by +5.62 and achieves larger gains at higher thresholds, reaching +12.93 at Pr@0.8 and +24.76 at Pr@0.9, indicating markedly better high-precision segmentation. In terms of region-level metrics, EviSAM3 surpasses RS2-SAM2 [1] by +5.64 oIoU and +6.66 mIoU, reflecting more consistent mask quality. It also clearly outperforms earlier methods such as LAVT [53], LGCE [3], and RMSIN [31], as well as SAM-based adaptations [15, 55], demonstrating its effectiveness in handling ambiguous and complex referring scenarios.

As illustrated in Fig. 3, EviSAM3 produces more accurate and complete segmentation masks across targets of varying scales under ambiguous referring expressions. In contrast, competing methods often exhibit imprecise boundaries or miss small regions, highlighting the advantage of our progressive evidence integration.

4.3 Ablation Study (RQ2)

Component-wise Ablation Analysis. Table 3 shows that vanilla SAM3 (without EIM and ED-MoE) performs poorly on both datasets, highlighting its limitation in handling ambiguous referring expressions. Adding EIM brings substantial gains across all metrics, demonstrating the effectiveness of progressive evidence integration for improving cross-modal alignment. Further introducing ED-MoE yields consistent improvements, especially at higher IoU thresholds, indicating enhanced refinement of fine-grained details under ambiguity. The larger gains on RefSegRS further suggest its advantage in handling semantically ambiguous expressions in the RRSIS task. Notably, the improvements become more pronounced at stricter IoU thresholds, reflecting more accurate boundary delineation enabled by accumulated evidence.

Contribution of the BEiT-3 Prior. Table 4 compares variants with and without the frozen BEiT-3 prior on the official RefSegRS test split. Both use seed 0 and identical 60-epoch training and evaluation settings. Without BEiT-3, EviSAM3 achieves 79.42 mIoU and 85.63 oIoU; adding the

Table 3: Component-wise ablation analysis on RRSIS-D and RefSegRS datasets.

		EIM	ED-MoE	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	oIoU	mIoU
RRSIS-D	val	-	-	11.95	10.69	8.91	6.72	3.97	15.56	12.82
		✓	-	79.01	72.99	61.70	48.51	27.64	77.63	67.70
		✓	✓	**83.45**	**77.82**	**66.95**	**53.33**	**32.36**	**81.38**	**71.25**
	test	-	-	11.17	9.82	8.24	6.43	3.88	14.66	11.64
		✓	-	79.04	73.60	62.04	47.22	27.44	77.59	67.57
		✓	✓	**82.81**	**77.65**	**66.37**	**50.56**	**30.82**	**80.27**	**70.65**
RefSegRS	val	-	-	30.16	20.19	9.28	2.09	1.16	26.02	24.71
		✓	-	91.18	89.10	84.69	76.33	24.83	79.51	79.47
		✓	✓	**98.14**	**96.29**	**94.90**	**93.27**	**83.53**	**92.44**	**91.32**
	test	-	-	26.53	18.00	7.59	2.42	0.77	28.03	25.69
		✓	-	80.02	73.97	64.45	46.23	10.68	75.46	68.65
		✓	✓	**89.93**	**85.25**	**79.58**	**68.63**	**45.95**	**86.51**	**80.56**

prior improves these by 1.14 and 0.88 percentage points, respectively. Removing BEiT-3 reduces total parameters by 318.98M without changing the 119.64M trainable parameters. This comparison indicates a complementary contribution from BEiT-3, while the evidence-driven adaptation remains effective without it.

Table 4: BEiT-3 ablation on RefSegRS test under matched 60-epoch settings (seed 0). IoU metrics are percentages; parameter counts are in millions.

Variant	mIoU ↑	oIoU ↑	Trainable	Total
With frozen BEiT-3	80.56	86.51	119.64	1008.97
Without BEiT-3	79.42	85.63	119.64	689.99

Transfer to the SAM2 Backbone. To examine whether the gains depend solely on SAM3, we also implement the evidence-driven modules on SAM2, denoted EviD-SAM2. Table 5 reports test-set results on both datasets. EviD-SAM2 improves mIoU over RS2-SAM2 by 2.23 points on RRSIS-D and 5.00 points on RefSegRS. These results support the applicability of evidence-driven adaptation beyond SAM3. EviD-SAM2 is evaluated in our experiments, whereas RS2-SAM2 results are taken from its publication [1]; this is a comparison under the same backbone family, not a fully controlled retraining comparison.

Variation Across Random Seeds. We repeat the full EviSAM3 experiment on RefSegRS with seeds 0, 1, and 2 using the paper's evaluation protocol. Table 6 reports each run and the mean $\pm$ sample standard deviation, computed with denominator $n-1$ for $n=3$. Test mIoU is 80.56 ± 0.18 and oIoU is 86.51 ± 0.15, indicating low variation across these runs. These statistics quantify random-seed variability on RefSegRS; they do not establish significance against baselines or characterize variability on RRSIS-D.

Table 5: SAM2-backbone test results (%). RS2-SAM2 is paper-reported; EviD-SAM2 is evaluated by us.

Dataset	Method	oIoU	mIoU
RRSIS-D	RS2-SAM2	78.99	66.72
	EviD-SAM2	79.68	68.95
RefSegRS	RS2-SAM2	80.87	73.90
	EviD-SAM2	85.30	78.90

Table 6: RefSegRS test results (%) over three random seeds.

Seed	mIoU	oIoU
0	80.56	86.51
1	80.38	86.36
2	80.74	86.66
Mean $\pm$ std	80.56 ± 0.18	86.51 ± 0.15

Design-level Ablation Analysis. We further analyze key design choices in EviSAM3. As shown in Table 7, momentum-based updates are essential: μ=0.0 performs poorly, while a moderate value (μ=0.6) achieves the best overall performance by balancing stability and adaptability in evidence accumulation. A larger momentum (μ=0.9) improves certain metrics but degrades others, indicating over-smoothing that weakens responsiveness to new evidence. Table 8 examines the capacity of the evidence representation. Performance improves up to E_t=8 tokens and E_d=256, but declines with larger settings, suggesting that excessive evidence introduces redundancy and noise, which hinders

Table 7: EMA momentum coefficient analysis on the RefSegRS dataset.

μ	Pr@0.5	Pr@0.7	Pr@0.9	oIoU	mIoU
μ=0.0	86.50	83.23	71.83	80.51	79.53
μ=0.3	97.22	94.05	82.53	91.23	90.34
μ=0.6 **(Ours)**	**98.14**	94.90	**83.53**	**92.44**	**91.32**
μ=0.9	97.81	**96.79**	81.96	90.84	89.83

Table 8: Ablation study of evidence length and dimension on the RefSegRS dataset.

	E_t	E_d	Pr@0.5	Pr@0.7	Pr@0.9	oIoU	mIoU
Tokens	4	256	96.06	90.95	70.30	89.61	88.66
	8	256	**98.14**	**94.90**	**83.53**	**92.44**	**91.32**
	16	256	96.54	92.81	74.64	90.07	89.70
	32	256	95.79	91.88	71.23	89.89	88.95
Dims	8	128	95.82	91.42	72.62	90.16	88.47
	8	256	**98.14**	**94.90**	**83.53**	**92.44**	**91.32**
	8	512	97.91	**94.90**	80.28	91.84	90.43
	8	1024	96.06	93.04	75.02	90.89	89.87

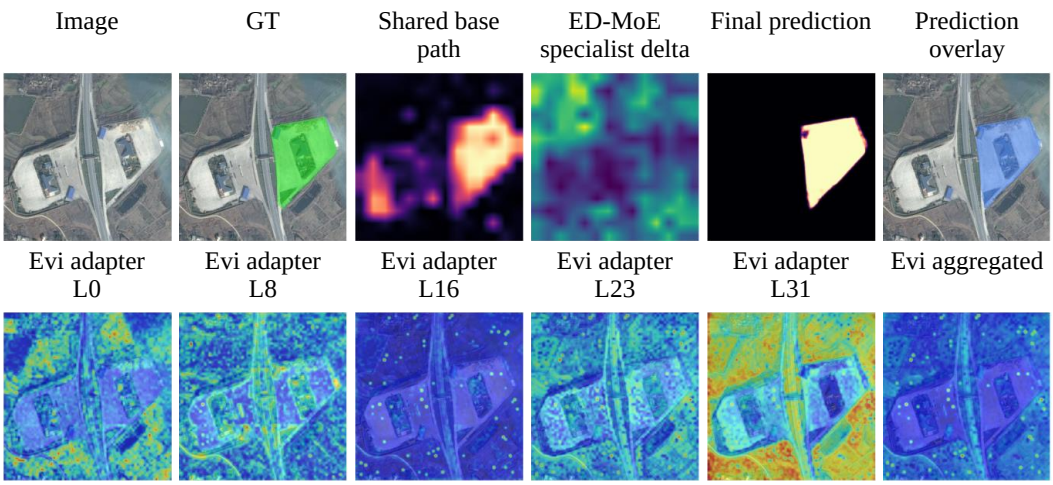


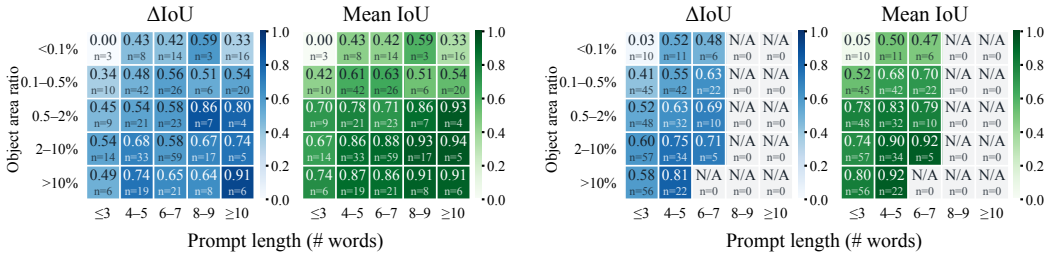
Figure 4: Visualization of evidence accumulation in EviSAM3 for the query "a service-area-like region toward the right", showing progressively refined and more concentrated responses.

effective aggregation. These results indicate that a moderate evidence capacity is crucial for reliable reasoning.

4.4 Evidence Integration Analysis (RQ3)

Figure 4 illustrates the progressive evidence accumulation process in EviSAM3. While the base path initially produces coarse and spatially dispersed responses, the ED-MoE branch progressively refines predictions by integrating complementary multi-granularity evidence from both visual and linguistic cues. As the inference depth increases, the evidence responses become increasingly concentrated and better aligned with the semantic intent of the referring expression. This progressive refinement effectively suppresses irrelevant background regions while enhancing target-consistent activations. Ultimately, the aggregated evidence map exhibits a stable and compact focus on the object region, clearly demonstrating the effectiveness of the proposed evidence-driven reasoning mechanism for fine-grained referring remote sensing segmentation.

As shown in Fig. 5a, the heatmaps summarize 400 uniformly sampled RRSIS-D validation cases over a 5×5 grid of object area ratio and prompt length, where the left panel reports mean ΔIoU and the right shows mean IoU. The **consistently positive** ΔIoU **across all bins** demonstrates that the proposed **Evidence Integration** mechanism reliably improves the baseline, with gains most pronounced for small-to-medium objects ($a \in [0.1\%, 2\%]$) and longer prompts ($\ell \geq 6$), where progressive integration of multiple evidences is most beneficial. In contrast, large objects ($a > 10\%$) with short prompts ($\ell \leq 3$) show smaller Δ but already achieve high IoU ($\gtrsim 0.7$), indicating a saturation effect, and **no negative** Δ **is observed**, suggesting consistent improvement without degrading the baseline. Similar trends are observed on RefSegRS in Fig. 5b.



(a) ΔIoU (left) and Mean IoU (right) on RRSIS-D

(b) ΔIoU (left) and Mean IoU (right) on RefSegRS

Figure 5: Conditional performance analysis across prompt length and object area ratio. Each cell reports IoU (or ΔIoU), with n indicating the number of samples in the bin.

4.5 Failure Case Analysis (RQ4)

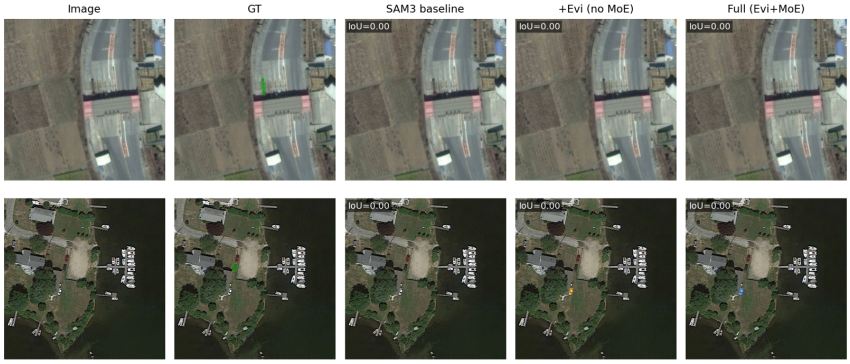


Figure 6: Failure cases of EviSAM3 on ambiguous queries. Two representative low-IoU examples illustrate backbone-limited failure and MoE mis-routing, corresponding to the queries "a gray vehicle" (top) and "the vehicle slightly larger than the one in the upper-left" (bottom).

As shown in Fig. 6, we present representative failure cases of EviSAM3 selected from a 200-sample validation pool with $\mathrm{IoU}_{\mathrm{Full}} \leq 0.30$. In the backbone-limited case, all variants fail to localize the target due to weak visual responses for small or occluded objects, resulting in the absence of a reliable base map. In contrast, the MoE mis-routing case shows that although +Evi partially captures the target, the Full model is distracted by a visually similar region, suggesting misaligned routing weights $\mathbf{w}$. These failure modes are complementary, indicating limitations in both upstream visual representation and downstream routing, and point to potential directions for further improvement.

5 Conclusion

We present **EviSAM3**, an evidence-driven framework for referring remote sensing image segmentation that addresses the mismatch between concept-level prompting in SAM3 and the need for fine-grained referring grounding. By modeling grounding as a progressive evidence accumulation process, EviSAM3 incrementally integrates visual–linguistic cues through an evidence memory mechanism and refines predictions via evidence-aware expert prompting. This design enables the model to progressively suppress irrelevant regions while strengthening semantically consistent evidence across iterative reasoning stages.

Limitation and Future Work. EviSAM3 may fail when the backbone cannot provide reliable initial evidence, especially for small or occluded targets. In addition, the expert routing in ED-MoE can be suboptimal in highly confusing scenes. Future work will explore stronger visual encoders, more robust evidence routing strategies, and tighter coupling between evidence modeling and foundation models. Extending the framework to temporal settings, such as video-based referring segmentation with dynamic evidence accumulation, is another promising direction.

Acknowledgments and Disclosures

This work was supported by the Fundamental Research Funds for the Central Universities under Grant DUTZD25239. The authors declare no competing interests.

References

[1] F. Rong, M. Lan, Q. Zhang, and L. Zhang, "Rs2-sam2: Customized sam2 for referring remote sensing image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 11, 2026, pp. 8751–8759.

[2] J. Zhang, L. Li, L. Jiao, X. Liu, F. Liu, W. Ma, and S. Yang, "A multiscale vision–text collaborative dual encoder for referring rs image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 64, pp. 1–15, 2026.

[3] Z. Yuan, L. Mou, Y. Hua, and X. X. Zhu, "Rrsis: Referring remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[4] Z. Mao, Y. Yang, C. Ma, D. Jiang, J. Yao, Y. Zhang, and Y. Wang, "Safire: Saccade-fixation reiteration with mamba for referring image segmentation," *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[5] Y. Liu, Z. Ma, J. Pu, Z. Qi, Y. Wu, S. Ying, and C. W. Chen, "Unipixel: Unified object referring and segmentation for pixel-level visual reasoning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[6] Y. Hu, Q. Wang, W. Shao, E. Xie, Z. Li, J. Han, and P. Luo, "Beyond one-to-one: Rethinking the referring image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4067–4077.

[7] T. Liu and S. 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*, 2025, pp. 29 634–29 643.

[8] W. Mi, J. Wang, F. Zhuang, and N. Zheng, "Grrsis: Generalized referring remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–17, 2025.

[9] T. Wellmann, A. Lausch, E. Andersson, S. Knapp, C. Cortinovis, J. Jache, S. Scheuer, P. Kremer, A. Mascarenhas, R. Kraemer *et al.*, "Remote sensing in urban planning: Contributions towards ecologically sound policies?" *Landscape and urban planning*, vol. 204, p. 103921, 2020.

[10] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multiscale location attention network for building and water segmentation of remote sensing image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–19, 2023.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[12] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," in *The Thirteenth International Conference on Learning Representations*, 2025.

[13] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. S. Coll-Vinent, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. HAZRA, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollar, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer, "SAM 3: Segment anything with concepts," in *The Fourteenth International Conference on Learning Representations*, 2026.

[14] A. Liu, R. Xue, X. R. Cao, Y. Shen, Y. Lu, X. Li, Q. Chen, and J. Chen, "Medsam3: Delving into segment anything with medical concepts," *arXiv preprint arXiv:2511.19046*, 2025.

[15] T. Chen, R. Cao, X. Yu, L. Zhu, C. Ding, D. Ji, C. Chen, Q. Zhu, C. Xu, P. Mao *et al.*, "Sam3-adapter: Efficient adaptation of segment anything 3 for camouflage object segmentation, shadow detection, and medical image segmentation," *arXiv preprint arXiv:2511.19425*, 2025.

[16] S. Seok, H.-S. Park, S. Kim, Y. Lee, and T. Kim, "Development of a text-guided annotation system for object segmentation training data in crop images based on the segmentation anything model 3," *Journal of Bio-Environment Control*, vol. 35, no. 1, pp. 26–39, 2026.

[17] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "The sam2-to-sam3 gap in the segment anything model family: Why prompt-based expertise fails in concept-driven image segmentation," *arXiv preprint arXiv:2512.06032*, 2025.

[18] M. D. Lee and T. D. Cummins, "Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models," *Psychonomic bulletin & review*, vol. 11, no. 2, pp. 343–352, 2004.

[19] J. D. Davidson and A. El Hady, "Foraging as an evidence accumulation process," *PLoS computational biology*, vol. 15, no. 7, p. e1007060, 2019.

[20] N. J. Evans, E.-J. Wagenmakers *et al.*, "Evidence accumulation models: Current limitations and future directions," *The Quantitative Methods for Psychology*, vol. 16, no. 2, pp. 73–90, 2020.

[21] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu *et al.*, "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1280–1297.

[22] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts in large language models," *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[23] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li, "Towards understanding the mixture-of-experts layer in deep learning," *Advances in neural information processing systems*, vol. 35, pp. 23 049–23 062, 2022.

[24] Z. Xia, D. Han, Y. Han, X. Pan, S. Song, and G. Huang, "Gsva: Generalized segmentation via multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3858–3869.

[25] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vlt: Vision-language transformer and query generation for referring segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7900–7916, 2022.

[26] J. Wu, X. Li, X. Li, H. Ding, Y. Tong, and D. Tao, "Toward robust referring image segmentation," *IEEE Transactions on Image Processing*, vol. 33, pp. 1782–1794, 2024.

[27] Y. X. Chng, H. Zheng, Y. Han, X. Qiu, and G. Huang, "Mask grounding for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 573–26 583.

[28] N. Liu, X. Xu, Y. Su, H. Zhang, and H.-C. Li, "Pointsam: Pointly-supervised segment anything model for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.

[29] D. Peng, X. Liu, Y. Zhang, H. Guan, Y. Li, and L. Bruzzone, "Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges," *International Journal of Applied Earth Observation and Geoinformation*, vol. 136, p. 104282, 2025.

[30] X. Li, F. Xu, A. Yu, X. Lyu, H. Gao, and J. Zhou, "A frequency decoupling network for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–21, 2025.

[31] S. Liu, Y. Ma, X. Zhang, H. Wang, J. Ji, X. Sun, and R. Ji, "Rotated multi-scale interaction network for referring remote sensing image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 658–26 668.

[32] K. Li, G. Vosselman, and M. Y. Yang, "Scale-wise bidirectional alignment network for referring remote sensing image segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 226, pp. 350–363, 2025.

[33] Z. Dong, Y. Sun, T. Liu, W. Zuo, and Y. Gu, "Cross-modal bidirectional interaction model for referring remote sensing image segmentation," *arXiv preprint arXiv:2410.08613*, 2024.

[34] Y. Pan, R. Sun, Y. Wang, T. Zhang, and Y. Zhang, "Rethinking the implicit optimization paradigm with dual alignments for referring remote sensing image segmentation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2031–2040.

[35] K. Chen, J. Zhang, C. Liu, Z. Zou, and Z. Shi, "Rsrefseg: Referring remote sensing image segmentation with foundation models," in *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2025, pp. 1070–1074.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[37] H. Wang, L. Qiao, Z. Jie, Z. Huang, C. Feng, Q. Zheng, L. Ma, X. Lan, and X. Liang, "X-sam: From segment anything to any segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 31, 2026, pp. 26 187–26 196.

[38] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[39] J. Zhang, X. Yang, R. Jiang, W. Shao, and L. Zhang, "RSAM-Seg: A SAM-based model with prior knowledge integration for remote sensing image semantic segmentation," *Remote Sensing*, vol. 17, no. 4, p. 590, 2025.

[40] H. Li, R. Zhang, H. Yao, X. Zhang, Y. Hao, X. Song, X. Li, Y. Zhao, L. Li, and Y. Chen, "Da-ada: Learning domain-aware adapter for domain adaptive object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 103 574–103 598, 2024.

[41] Z. Zhong, Z. Tang, T. He, H. Fang, and C. Yuan, "Convolution meets LoRA: Parameter efficient finetuning for segment anything model," in *International Conference on Learning Representations (ICLR)*, 2024.

[42] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[43] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-efficient fine-tuning for large models: A comprehensive survey," *arXiv preprint arXiv:2403.14608*, 2024.

[44] K. Li, S. Zhang, Y. Deng, Z. Wang, D. Meng, and X. Cao, "Segearth-ov3: Exploring sam 3 for open-vocabulary semantic segmentation in remote sensing images," *arXiv preprint arXiv:2512.08730*, 2025.

[45] L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, and F. Yu, "Segment anything in high quality," in *Advances in Neural Information Processing Systems*, 2023.

[46] R. Zhang, Z. Jiang, Z. Guo, C. Yan, Z. Dong, Y. Qiao, and P. Gao, "Personalize segment anything model with one shot," *arXiv preprint arXiv:2305.03048*, 2023.

[47] Y. Zhang, T. Cheng, L. Zhu, R. Hu, L. Liu, H. Liu, L. Ran, X. Chen, W. Liu, and X. Wang, "Evf-sam: Early vision-language fusion for text-prompted segment anything model," *arXiv preprint arXiv:2406.20076*, 2024.

[48] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[49] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[50] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li, "Referring image segmentation via cross-modal progressive comprehension," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 488–10 497.

[51] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4424–4433.

[52] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, "Cross-modal progressive comprehension for referring segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4761–4775, 2021.

[53] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 155–18 165.

[54] S. Lei, X. Xiao, T. Zhang, H.-C. Li, Z. Shi, and Q. Zhu, "Exploring fine-grained image-text alignment for referring remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–11, 2025.

[55] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *Iclr*, vol. 1, no. 2, p. 3, 2022.

[56] S.-A. Liu, Y. Zhang, Z. Qiu, H. Xie, Y. Zhang, and T. Yao, "Caris: Context-aware referring image segmentation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 779–788.

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 claims in the abstract and introduction are consistent with the proposed methods and are fully supported by the experimental results and ablation studies.

   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: The limitations of the proposed method are discussed in the failure case analysis (Section 4) and further summarized in the conclusion section (Section 5).

   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: The paper presents a method and empirical evaluations, but does not state 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: Sections 3 and 4 describe the model architecture, training settings, datasets, and evaluation protocols; implementation and reproduction notes are available in the public repository at `https://github.com/YaofireJinglian/EviSAM3`.

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: The code is publicly available at `https://github.com/YaofireJinglian/EviSAM3`. The repository provides environment setup, dataset preparation notes, and training and evaluation commands; its reproduction guide documents the current limitations of the released package.

   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 paper provides comprehensive experimental details, including data splits, training configurations, optimization settings, and hyperparameter choices, which are sufficient to understand and reproduce the reported results.

   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: [Yes]

   Justification: Table 6 reports RefSegRS test results for seeds 0, 1, and 2, together with the mean and sample standard deviation (denominator $n - 1$). These statistics capture variation across random seeds under the same evaluation protocol; repeated-run statistics are not reported for RRSIS-D or for all baselines.

   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: The paper provides detailed descriptions of the computational resources used for all experiments, including hardware configurations and training settings, which are sufficient for reproduction of the reported results.

   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: The work complies with the NeurIPS Code of Ethics, and no ethical concerns are involved in data usage, experiments, or model design.

   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: [No]

Justification: The introduction describes potential benefits for disaster assessment, urban analysis, and environmental monitoring, but this version does not include a discussion of potential negative societal impacts.

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: [N/A]

Justification: The proposed model does not involve high-risk applications such as generative content or sensitive data release, and therefore no specific safeguards are required.

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: All datasets, models, and code used in this work are properly cited, and their licenses and terms of use are respected in accordance with their original sources.

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: [Yes]

    Justification: We release our source code at `https://github.com/YaofireJinglian/EviSAM3`, with a README and reproduction guide covering environment setup, dependencies, dataset preparation, training and evaluation commands, and known limitations. License information is included in the repository.

    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 experiments or human subject studies.

    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 human subjects research and therefore does not require IRB approval.

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 LLM was only used for writing assistance such as language polishing and does not affect the methodology, experiments, or results.

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.