GeoReason: Bridging Logical Reasoning and Spatial Fidelity in Remote Sensing Segmentation

Cheng Fan[1] **Han Wu**[2] **Tingzhang Luo**[1]
Minjing Dong[1] **Yehui Tang**[3] **Jianyuan Guo**[1,*]

[1]City University of Hong Kong
[2]School of Computer Science, Peking University
[3]Samsung Research, Beijing

chengfan2-c@my.cityu.edu.hk
[*]Corresponding author: jianyguo@cityu.edu.hk

Abstract

Multimodal Large Language Models (MLLMs) have significantly advanced instruction-driven perception through the "Embedding-as-Mask" paradigm [17]. However, extending this capability to Remote Sensing (RS) reasoning segmentation remains exceptionally challenging under cluttered geographical contexts and extreme scale variations, often leading to *Spatial Fidelity Degradation* and *Target Misidentification*. We identify a critical yet underexplored bottleneck: the sequential alienation of the segmentation token `[SEG]` from its initial visual grounding within the autoregressive generation paradigm. Produced only at the end of long reasoning chains, `[SEG]` relies on visual evidence that has been progressively diluted by linguistic abstraction and autoregressive noise. This information decay is particularly catastrophic for RS targets with minuscule pixel footprints, where even marginal fidelity loss leads to localization failure. To address these bottlenecks, we propose **GeoReason**, a novel framework designed to fortify logical deduction and preserve spatial integrity. First, we introduce an *Anticipatory Prior* mechanism that injects the `[SEG]` token directly into the initial user query, shifting it from a terminal output to a primary condition. This paradigm shift transforms the token into a persistent spatial anchor, preventing information decay during long-form linguistic inference. Second, we enhance visual reasoning via a *Location-aware Chain-of-Thought (CoT)* that enforces coarse-to-fine localization using bounding boxes, coupled with a *Latent Patch-to-Token Alignment (PTA) Loss* that explicitly aligns the latent segmentation token with physical RS textures. Extensive experiments demonstrate that GeoReason consistently outperforms state-of-the-art methods on the Earthreason and RRSIS-D benchmarks. Code is available at `https://github.com/Fanc0126/GeoReason`.

1 Introduction

Multimodal Large Language Models (MLLMs) have shifted image segmentation from category level masks toward complex, instruction driven perception. A cornerstone of this evolution is the "Embedding as Mask" paradigm (*e.g.*, LISA [17]), which enables MLLMs to perform Reasoning Segmentation. In this framework, a dedicated token `[SEG]` is introduced into the language vocabulary and is generated autoregressively as part of the model's textual response. The hidden embedding corresponding to the generated `[SEG]` token is then extracted as a latent semantic query and projected into a mask decoder, such as the Segment Anything Model (SAM) [15], to obtain dense pixel-level outputs. Unlike Referring Expression Segmentation (RES) which relies on explicit object descriptions,

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



Figure 1: **Qualitative comparison of GeoReason against the baseline.** (Top) While the baseline identifies the viaduct, the resulting mask is fragmented and suffers from blurry boundaries. (Bottom) The baseline fails to interpret the functional implications of the query ("sports games or popular live music events"), leading to the segmentation of an irrelevant field.

reasoning segmentation necessitates interpreting implicit, complex instructions through sophisticated logical deduction and visual grounding.

Despite its success in natural scenes, this paradigm encounters unique hurdles in Remote Sensing (RS). Characterized by overhead perspectives, extreme scale variations, and cluttered geographical contexts [25], RS imagery demands not only robust logical inference to resolve functional attributes and spatial relationships, but also high-precision localization in dense environments. As illustrated in Fig. 1, current MLLM-based RS frameworks often fall into two traps:

- *Spatial Fidelity Degradation:* The predicted mask often exhibits fragmented or imprecise boundaries despite correctly identifying the target, indicating degraded spatial consistency.
- *Target Misidentification:* Under more severe information degradation, the segmentation may drift to an incorrect object, reflecting a failure to maintain the correct target during reasoning.

We attribute these failures to a critical yet underexplored bottleneck in current MLLM segmentation pipelines [17, 5, 21]: the sequential alienation of the segmentation token `[SEG]` from its initial visual grounding within the autoregressive generation paradigm. In existing frameworks, `[SEG]` is produced only at the end of a long reasoning chain. As the reasoning trajectory grows longer, the hidden states of later tokens become increasingly dominated by linguistic abstraction and autoregressive noise, progressively diluting the visual evidence required for accurate spatial grounding, as illustrated in Fig. 2. Although shortening or removing the reasoning trajectory can partially alleviate this degradation, it is not a principled remedy: difficult samples often rely on multi-step visual and semantic inference, for which CoT reasoning remains essential, as detailed in Sec 4.4. Consequently, the model may struggle to preserve precise spatial cues in the predicted mask and, in more severe cases, even lose track of the intended target during reasoning. This information decay is particularly detrimental for RS imagery, where accurate segmentation requires maintaining fine-grained structural details across cluttered geographical scenes and extreme scale variations. To address these challenges, we propose **GeoReason** to strengthen RS-oriented visual reasoning while preserving spatial fidelity in segmentation. Our approach is built upon two core pillars:

Anticipatory Prior via Prompt-side `[SEG]`. To mitigate the spatial fidelity degradation arising from the delayed generation of the segmentation token, we introduce an anticipatory prior mechanism

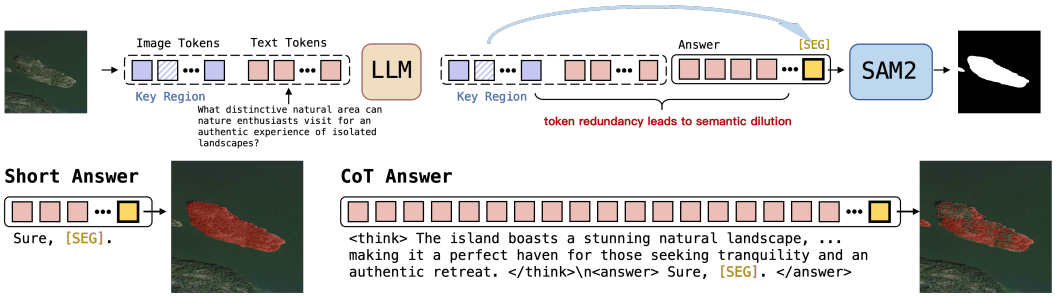


Figure 2: Spatial fidelity degradation of masks induced by long reasoning chains.

that injects one more specialized [SEG] token directly into the user query. Unlike tokens generated autoregressively in the answer, this prompt-side token acts as a deterministic input that remains unaffected by the cumulative linguistic abstraction and autoregressive noise. By functioning as a persistent spatial anchor throughout the entire forward pass, the token can pre-extract and maintain critical visual evidence from the outset. This design effectively mitigates the information decay inherent in long-form RS reasoning, ensuring the spatial fidelity of the resulting latent representations.

Enhanced Visual Reasoning for RS Domain. To further address target misidentification in complex remote sensing scenes, we enhance the model's visual reasoning through a dual strategy that combines explicit spatial grounding with implicit latent supervision. First, we introduce a *Location-aware Chain-of-Thought (CoT)* mechanism that decouples logical deduction from final mask extraction. By incorporating bounding box prediction as an intermediate reasoning step, this design enforces a coarse-to-fine localization process, allowing the model to verify candidate regions before final mask extraction and significantly reducing the risk of selecting incorrect targets. Complementing this explicit grounding, we propose a *Latent Patch-to-Token Alignment (PTA) Loss* to directly supervise the representation of the [SEG] token. Specifically, we align the hidden states of this token with the physical image patches corresponding to the ground-truth target. By grounding the latent reasoning token in RS-specific textures and multi-scale geometries, this alignment strengthens its discriminative capability and ensures reliable segmentation under cluttered geographical contexts.

Our main contributions are summarized as follows:

- We propose **GeoReason**, a novel framework for remote sensing reasoning segmentation that addresses two critical failure modes in MLLM-based segmentation: *Spatial Fidelity Degradation* and *Target Misidentification*.
- We introduce an *Anticipatory Prior* mechanism that injects the [SEG] token directly into the query, transforming it into a persistent spatial anchor and preventing visual evidence from being diluted during long-form reasoning.
- We develop a *Location-aware CoT* together with a *Latent PTA Loss*, which jointly improve logical grounding and strengthen the spatial discriminability of the [SEG] token in complex RS scenes.
- Extensive experiments demonstrate that GeoReason achieves state-of-the-art performance on RS reasoning segmentation benchmarks, significantly improving both reasoning accuracy and segmentation quality.

2 Related Work

Multimodal Large Language Models for Segmentation. MLLM-based segmentation has moved beyond VQA toward dense, language-conditioned prediction, and existing methods mainly follow two paradigms: token-based latent extraction (Embedding-as-Mask) and text-based spatial prompt generation [24, 29, 48, 61, 73, 9, 71]. LISA established the Embedding-as-Mask framework and the task of reasoning segmentation by appending a special [SEG] token whose hidden state prompts a mask decoder such as SAM [17, 15, 38, 3, 35]. This design has since become a standard recipe for integrating dense prediction into MLLMs [53, 37, 66, 50], with extensions to multi-target reasoning in PixelLM [39] and to video segmentation in SA2VA and related works [63, 56, 2, 7, 61, 8]. In contrast, text-only alternatives predict explicit spatial prompts such as boxes or points and then invoke an external segmentor [30, 62, 73]; representative examples include Seg-Zero [30], Seg-R1 [62], and the "text-as-mask" line explored by Text4Seg and Text4Seg++ [18, 19].

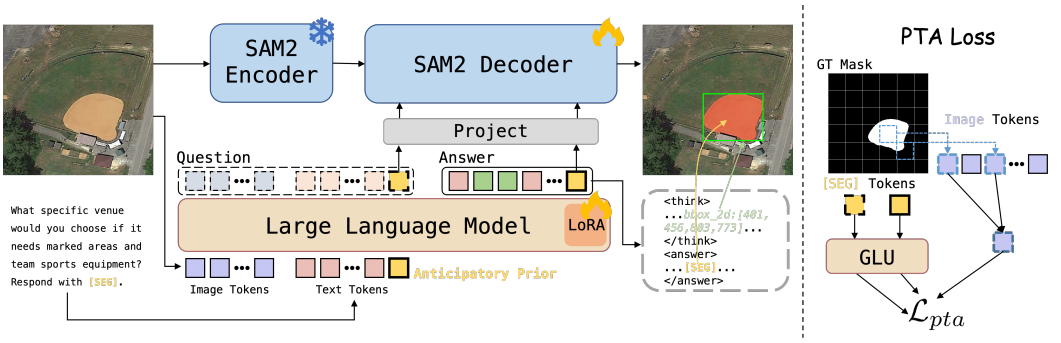


Figure 3: **Overview of the GeoReason framework.** (Left) Overall architecture of *Anticipatory Prior* mechanism, where a [SEG] token is injected into the user query to pre-anchor spatial evidence and mitigate information decay during long-chain reasoning. The MLLM is trained to output intermediate bounding boxes as spatial anchors (*Location-aware CoT*) before generating the final [SEG]. (Right) The *Latent Patch-to-Token Alignment (PTA)* loss aligns the hidden states of both prompt-side and answer-side [SEG] tokens with the visual patch embeddings corresponding to the GT target.

Language-Guided Remote Sensing Segmentation. In remote sensing, language-guided segmentation [26] has likewise evolved from referring expression segmentation [32] to more demanding reasoning segmentation settings [21, 54, 33, 60]. Early RS-RES methods, such as LGCE [64] and RMSIN [27], fuse lightweight language encoders with visual backbones, while SAM-based variants, including RS2-SAM2 [40], RSRefSeg [5], and RSRefSeg2 [4], improve localization by translating language into sparse segmentation prompts. These models are effective for referring-style grounding but remain limited on multi-step reasoning. MLLM-based RS segmentation addresses this gap [21, 54]: several methods adopt the LISA-style [SEG] token for mask generation in referring settings [34, 41, 28], while SegEarth-R1 and SegEarth-R2 extend the same paradigm to the Earthreason benchmark and show that logical deduction remains difficult in cluttered RS scenes [21, 54].

Visual Reasoning in MLLMs. CoT reasoning in MLLMs has progressed from purely textual deliberation [51, 55, 52, 45, 6] toward "thinking with images" [47, 67]. Recent methods strengthen reasoning with explicit visual interaction, including active perception through iterative cropping or zooming in DeepEyes, Thyme, and Pixel Reasoner [70, 68, 47], and spatial anchoring via intermediate box prediction in TreeVGR [46]. In parallel, latent reasoning methods argue that part of the reasoning process should remain in feature space rather than text [10]; for example, LVR aligns latent tokens with image representations to encode complex visual constraints [20]. These trends motivate our combination of explicit spatial grounding and token-level latent supervision for RS reasoning segmentation.

3 Our Method

GeoReason follows the embedding-as-mask paradigm and combines an MLLM with SAM2 [38] for reasoning segmentation in remote sensing. The method has two main goals: preserve spatial fidelity during long-form reasoning, and improve target identification in cluttered RS scenes. To this end, we introduce a prompt-side [SEG] anchor (*Anticipatory Prior*), a lightweight location-aware CoT, and a Latent Patch-to-Token Alignment (PTA) loss.

3.1 System Architecture

We use Qwen3VL [1] as the MLLM backbone and SAM2 [38] as the mask decoder. Given an RS image $\mathcal{I}$ and an instruction $\mathcal{T}$, the model processes a prompt that contains a prompt-side [SEG] token and generates an answer that may contain one or more answer-side [SEG] tokens. For any [SEG] token, its last-layer hidden state $\mathbf{h}_{seg} \in \mathbb{R}^d$ is treated as a latent mask query and projected by an MLP adapter f_ϕ into the SAM2 space:

$$\mathcal{M}_{pred} = \mathcal{D}(\mathcal{F}_{img}, f_\phi(\mathbf{h}_{seg})), \tag{1}$$

where $\mathcal{F}_{img}$ denotes the SAM2 image features. During training, segmentation losses are applied to both the prompt-side [SEG] and the answer-side [SEG] tokens. During inference, each of these

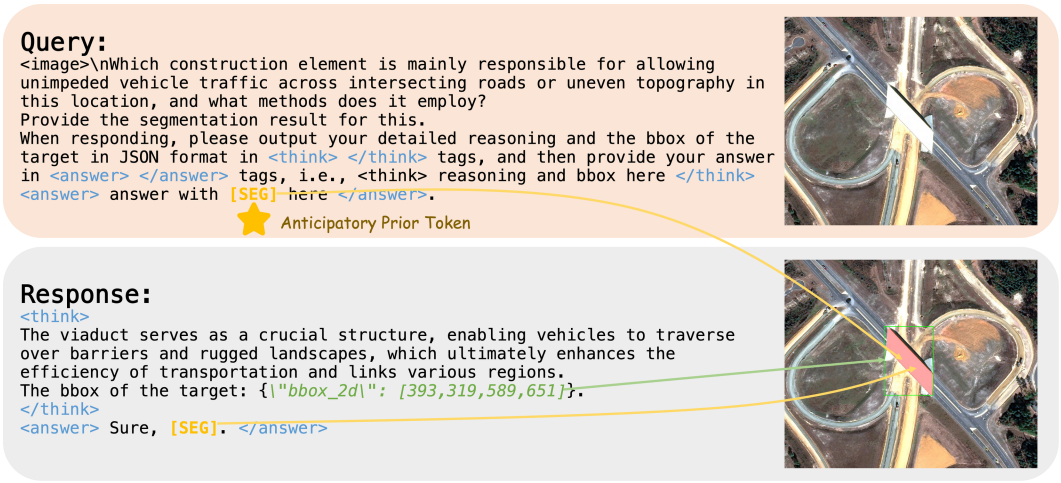


Figure 4: Illustration of the prompt template and model response format.

tokens predicts a candidate mask, and the final result is obtained by taking their union. Detailed prompt formats are provided in the appendix A.3.

3.2 Anticipatory Prior via Prompt-side [SEG] Token

In standard embedding-as-mask pipelines, the decisive [SEG] token is generated only after the reasoning chain, making its embedding vulnerable to attention dilution and semantic drift. This is especially problematic in RS, where fine boundaries and small structures must be preserved throughout long multimodal reasoning. We therefore place a dedicated [SEG] token on the prompt side, so that segmentation intent is established before autoregressive generation starts.

This prompt-side token serves as an *Anticipatory Prior*: it interacts with the image from the beginning of the forward pass and remains available to subsequent answer tokens as a stable spatial reference. As a result, the final answer-side [SEG] does not need to reconstruct segmentation cues from scratch after a long reasoning chain; instead, it can inherit a spatial prior that mitigates information decay.

3.3 Enhanced Visual Reasoning for Remote Sensing

We complement the prompt-side anchor with one explicit reasoning component and one latent alignment component.

Location-aware Chain-of-Thought. We ask the model to predict a bounding box $[x_{min}, y_{min}, x_{max}, y_{max}]$ in the <think> stage before producing the final [SEG]. This design is inspired by recent think-with-images reasoning strategies, but here it plays a lightweight role: providing an explicit coarse spatial hypothesis before dense mask decoding. It is simple but useful in RS scenes with severe clutter and scale variation. In our design, BBox reasoning provides coarse localization, while the [SEG] embedding remains the dense latent query used by SAM2 for final mask decoding.

Latent Vision Reasoning via PTA Loss. Mask supervision alone does not explicitly shape the [SEG] embedding to capture RS-specific textures and structures. We therefore introduce a Latent Patch-to-Token Alignment (PTA) loss that directly aligns each supervised [SEG] embedding with the target image patches. As shown in Fig. 3, PTA improves the latent quality of the segmentation token rather than only its decoded mask.

Given a set of visual patch tokens $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^{N}$ where $\mathbf{v}_i \in \mathbb{R}^D$, let $\mathcal{I}_{valid}$ denote the index set of patches overlapping with the ground-truth mask $\mathcal{M}_{gt}$. The target visual proxy $\mathbf{z}_{vis}$ is computed via average pooling:

$$\mathbf{z}_{vis} = \frac{1}{|\mathcal{I}_{valid}|} \sum_{i \in \mathcal{I}_{valid}} \mathbf{v}_i. \tag{2}$$

To achieve cross-modal alignment, we project the [SEG] token embedding $\mathbf{h}_{seg}$ into the visual representational space via a Gated Linear Unit (GLU) head $g_\theta(\cdot)$:

$$g_\theta(\mathbf{h}_{seg}) = \mathbf{W}_{down}\left(\sigma(\mathbf{W}_{gate}\mathbf{h}_{seg}) \odot (\mathbf{W}_{up}\mathbf{h}_{seg})\right), \tag{3}$$

where σ denotes the SiLU activation function. The PTA loss is then formulated as the cosine distance between the projected latent embedding and the visual proxy:

$$\mathcal{L}_{PTA} = 1 - \frac{g_\theta(\mathbf{h}_{seg}) \cdot \mathbf{z}_{vis}}{\|g_\theta(\mathbf{h}_{seg})\|\|\mathbf{z}_{vis}\|}. \tag{4}$$

By minimizing $\mathcal{L}_{PTA}$, the model is encouraged to encode foreground-specific visual evidence directly in the [SEG] embedding, which in turn improves both prompt-side anchoring and answer-side mask decoding. In this sense, the [SEG] token also acts as a latent reasoning carrier, it is to make the latent representation explicitly segmentation-oriented.

3.4 Optimization Objective

GeoReason is trained with language, mask, and latent-alignment supervision. Let $\hat{\mathcal{M}} \in \mathbb{R}^{H\times W}$ denote the predicted mask and $\mathcal{M}_{gt}$ the ground-truth mask. We use Binary Cross-Entropy and Dice losses for mask supervision:

$$\mathcal{L}_{bce} = -\frac{1}{HW}\sum_{i,j}\left[(\mathcal{M}_{gt})_{i,j}\log(\hat{\mathcal{M}}_{i,j}) + (1-(\mathcal{M}_{gt})_{i,j})\log\left(1-\hat{\mathcal{M}}_{i,j}\right)\right], \tag{5}$$

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_{i,j}\hat{\mathcal{M}}_{i,j}\cdot(\mathcal{M}_{gt})_{i,j} + \varepsilon}{\sum_{i,j}\hat{\mathcal{M}}^2_{i,j} + \sum_{i,j}(\mathcal{M}_{gt})^2_{i,j} + \varepsilon}. \tag{6}$$

We also apply $\mathcal{L}_{PTA}$ in Eq. (4). These segmentation-related losses are imposed on both the prompt-side and answer-side [SEG] tokens, encouraging consistent spatial semantics across the full reasoning trajectory. The overall objective is:

$$\mathcal{L} = \lambda_{text}\mathcal{L}_{text} + \lambda_{mask}(\mathcal{L}_{bce} + \mathcal{L}_{dice}) + \lambda_{pta}\mathcal{L}_{pta}, \tag{7}$$

where $\mathcal{L}_{text}$ is the standard next-token loss and $\lambda_{text}, \lambda_{mask}, \lambda_{pta}$ balance the three objectives.

4 Experiments

4.1 Experimental Setup

GeoReason uses Qwen3VL-4B [1] as the MLLM backbone and SAM2 [38] as the mask decoder, initialized from SA2VA [63]. To strengthen remote-sensing priors, we further fine-tune with filtered VQA and caption samples from GeoChatInstruct [16] and VRSBench [23], while removing any validation/test overlaps to avoid leakage. During training, the visual encoder is frozen, the LLM is optimized with LoRA [11], and the MLP adapter together with SAM2 are fully updated.

We evaluate on two representative RS segmentation benchmarks: EarthReason [21] for reasoning segmentation and RRSIS-D [27] for referring segmentation. Following prior works [5, 54, 17], we report gIoU and cIoU. Full training hyperparameters, data preprocessing, and prompt templates are provided in the appendix A.

4.2 Main Results

Tables 1 and 2 show that GeoReason performs strongly on both reasoning and referring segmentation, with the largest gains appearing in the reasoning setting that motivates our design.

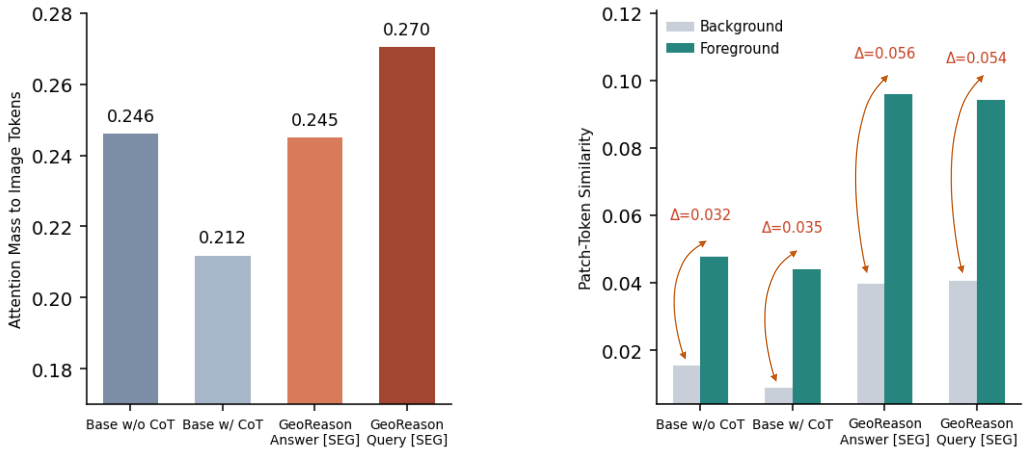
Reasoning Segmentation. GeoReason establishes a new SOTA on EarthReason across all metrics. The runner-up differs by metric: our backbone-matched baseline is second-best on gIoU, while UniGeoSeg is second-best on cIoU. Relative to these strongest alternatives, GeoReason improves val/test gIoU by $+2.57/+3.29$ and val/test cIoU by $+4.33/+5.99$. Notably, the baseline already benefits from a strong Qwen3VL+SAM2 backbone, yet still trails GeoReason by a large margin, especially on test cIoU (80.60% vs. 70.45%). This shows that the gains do not come from backbone

Table 1: Comparison on EarthReason. "Baseline" denotes our backbone-matched post-hoc [SEG] model without Anticipatory Prior or Enhanced Visual Reasoning. Best results are in **bold**; second-best are underlined.

Method	Val		Test	
	gIoU	cIoU	gIoU	cIoU
LISA [17]	61.04	57.39	60.88	59.10
PixelLM [39]	57.94	57.79	60.01	59.22
PSALM [69]	66.61	62.03	68.30	64.61
SegEarth-R1 [21]	68.60	64.13	70.75	68.25
RemoteReasoner [60]	69.00	67.80	71.00	69.10
Text4Seg++ [19]	71.90	69.80	73.00	65.60
SegEarth-R2 [54]	72.30	68.10	73.50	69.50
UniGeoSeg [33]	72.54	73.30	73.08	74.61
Baseline	75.04	70.33	75.32	70.45
GeoReason	**77.61**	**77.63**	**78.61**	**80.60**
Gain	*+2.57*	*+4.33*	*+3.29*	*+5.99*

Table 2: Comparison on RRSIS-D. "-" indicates that the metric was not reported by the original authors.

RRSIS-D	Val		Test	
	gIoU	cIoU	gIoU	cIoU
Traditional Methods				
LGCE [64]	60.16	76.68	59.37	76.34
RIS-DMMI [12]	60.72	77.01	60.12	76.20
LAVT [58]	61.46	77.59	61.04	77.19
RMSIN [27]	65.10	78.27	64.20	77.79
MLLM-based Methods				
LISA [17]	27.84	-	26.78	-
PixelLM [39]	33.89	-	31.65	-
NEXT-Chat [65]	26.98	-	24.98	-
GeoGround [72]	61.10	-	60.50	-
Text4Seg++ [19]	64.10	75.80	62.80	74.20
SegEarth-R1 [21]	67.56	78.92	66.40	78.01
GeoPixel [41]	67.99	**81.77**	67.30	**84.90**
GeoReason	**68.36**	79.87	**67.32**	79.39



(a) Direct attention from different [SEG] tokens to image tokens, averaged over the last four layers.

(b) Foreground/background patch alignment and token-only mask quality of different [SEG] tokens.

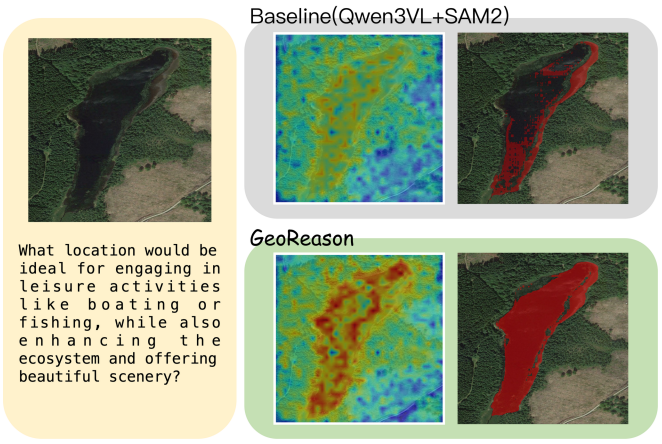
Figure 5: **Token-level mechanism analysis.** (a) Direct attention from different [SEG] token sources to image tokens. (b) Foreground/background alignment and token-only mask quality.

strength alone, but from stabilizing the segmentation token during long-form reasoning and improving its spatial grounding.

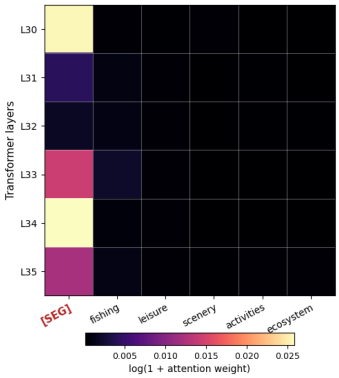
Referring Segmentation. On RRSIS-D, GeoReason achieves the best gIoU on both validation and test, and remains competitive in cIoU against specialized referring models. Since this benchmark involves little explicit reasoning, the gains mainly reflect better mask quality. This is consistent with the role of PTA: beyond helping reasoning segmentation, it also improves the discriminability of the [SEG] representation for standard referring scenarios.

4.3 Mechanistic Analysis of the Anticipatory Prior

To validate the role of the Anticipatory Prior, we compare four token sources on the EarthReason test set: the baseline answer-side [SEG] without CoT, the baseline answer-side [SEG] with CoT, the GeoReason answer-side [SEG], and the GeoReason prompt-side [SEG]. Fig. 5a shows that long autoregressive reasoning weakens direct visual access in the baseline: attention mass from the final [SEG] to image tokens drops from 0.246 without CoT to 0.212 with CoT. GeoReason largely removes this degradation. Its prompt-side [SEG] has the strongest image attention (0.270), and its answer-side [SEG] retains 0.245, nearly recovering the no-CoT level despite a longer reasoning



(a) Qualitative comparison on a representative case.



(b) Late-layer attention distribution of the **final answer-side** `[SEG]` token over the prompt-side `[SEG]` token and selected key words, with other tokens omitted for clarity.

Figure 6: **Case-level evidence.** (a) GeoReason preserves focus on the target region and produces a cleaner mask than the baseline. (b) In the same case, the final answer-side `[SEG]` focuses strongly on the prompt-side `[SEG]`.

Table 3: Ablation study of GeoReason components. AP: Anticipatory Prior, BBox: Location-aware CoT, PTA: Latent PTA Loss. g: gIoU, c: cIoU (%).

Components			Earthreason (Reasoning)				RRSIS-D (Referring)			
AP	BBox	PTA	Val g	Val c	Test g	Test c	Val g	Val c	Test g	Test c
✗	✗	✗	75.04	70.33	75.32	70.45	67.13	79.39	66.42	78.67
✗	✗	✓	76.02	72.26	76.32	73.01	67.98	79.71	66.99	79.32
✗	✓	✗	74.42	70.89	74.54	70.59	65.60	76.13	65.29	76.43
✓	✗	✗	76.24	77.32	78.02	78.87	67.54	79.45	66.13	78.32
✓	✗	✓	76.45	75.64	77.39	77.18	68.12	79.88	66.32	78.67
✓	✓	✗	76.66	76.24	77.25	76.62	67.58	**79.93**	66.27	78.28
✓	✓	✓	**77.61**	**77.63**	**78.61**	**80.60**	**68.36**	79.87	**67.32**	**79.39**

sequence. This is direct evidence that the prompt-side token acts as a persistent spatial anchor rather than an extra symbolic token.

We further ask whether the final answer-side `[SEG]` explicitly uses this anchor. After removing image and template tokens, we rank prompt tokens by their mean attention from the answer-side `[SEG]` over the last four layers. The prompt-side `[SEG]` has mean rank 2.01, appears in the top-3 attended prompt tokens for 99.69% of samples, and in the top-5 for 99.95%. The case heatmap in Fig. 6b shows the same behavior at the instance level: over the last six layers, the answer-side `[SEG]` attends much more strongly to the prompt-side `[SEG]` than to content words such as *fishing*, *leisure*, and *scenery*. Fig. 5b further shows that GeoReason also improves segmentation-oriented alignment: relative to the baseline CoT answer-side `[SEG]`, the GeoReason answer-side `[SEG]` raises foreground similarity from 0.044 to 0.096 and enlarges the foreground-background margin from 0.035 to 0.056. The qualitative case in Fig. 6a is consistent with these statistics, with GeoReason producing a more complete lake mask. Together, these results support our claim that AP preserves visual evidence while PTA makes the resulting `[SEG]` representation more discriminative for the target region.

4.4 Ablation Studies

Here we conduct ablation studies to disentangle the roles of the Anticipatory Prior (AP), the Location-aware CoT (BBox), and the Latent PTA loss (PTA). Table 3 reports the core component ablations on EarthReason and RRSIS-D, and Table 4 provides two targeted diagnostic variants.

Component contributions. AP is the dominant factor on reasoning segmentation: adding AP alone improves EarthReason test cIoU from 70.45% to 78.87%, confirming that early spatial anchoring is

Table 4: Diagnostic studies on EarthReason. **Shifted-SEG**: move the prompt-side [SEG] into the answer CoT. **Zero-Reasoning**: a short-answer baseline trained and evaluated without CoT. **Avg** is the mean of gIoU and cIoU across Val and Test.

Setting	Val		Test		Avg
	gIoU	cIoU	gIoU	cIoU	
Baseline	75.04	70.33	75.32	70.45	72.79
GeoReason	77.61	77.63	78.61	80.60	78.61
Shifted-SEG	75.65	73.60	76.22	71.07	74.14
Zero-Reasoning	75.64	72.59	72.86	77.43	74.63

Table 5: Oracle-referring diagnostics on EarthReason. **Oracle Ref.**: convert each reasoning query into a direct referring expression using Qwen3-30B. **Oracle Ref. + PTA**: additionally apply PTA in this simplified setting. **Avg** is the mean of gIoU and cIoU across Val and Test.

Setting	Val		Test		Avg
	gIoU	cIoU	gIoU	cIoU	
Baseline	75.04	70.33	75.32	70.45	72.79
GeoReason	77.61	77.63	78.61	80.60	78.61
Oracle Ref.	79.20	82.16	79.34	82.67	80.84
Oracle Ref. + PTA	79.94	82.70	80.33	84.14	81.78

critical under autoregressive reasoning. PTA alone also gives a clear gain ($70.45\% \rightarrow 73.01\%$) and is the main source of improvement on RRSIS-D, where mask quality matters more than long-form reasoning. By contrast, BBox alone is slightly below the baseline on EarthReason, indicating that longer reasoning does not help unless the segmentation token is first stabilized.

Complementarity. The full model performs best on both benchmarks. Partial variants such as AP+PTA or AP+BBox remain competitive but consistently trail the full design on EarthReason, showing that the three modules address different error sources: AP preserves spatial evidence, BBox improves coarse localization, and PTA sharpens the latent segmentation representation.

Diagnostic variants. *Shifted-SEG* keeps the additional [SEG] token but moves it from the prompt into the answer stream, reducing the average score to 74.14%. This confirms that the improvement does not come from simply adding another special token; what matters is prompt-side placement, which allows the token to condition the full forward pass. *Zero-Reasoning*, a short-answer variant without CoT, outperforms the full reasoning baseline (74.63% vs. 72.79% Avg), indicating that CoT can hurt segmentation if no spatial-preservation mechanism is provided. However, it still remains well below full GeoReason (78.61% Avg), showing that the goal is not to remove reasoning, but to make reasoning compatible with spatial fidelity. Fig. 7 gives a example: for the query *"In the context of this facility, what infrastructure do you consider vital for the safe storage of raw materials utilized in production?"*, GeoReason reasons that the protected silo is the correct target, whereas the Zero-Reasoning model drifts to the exposed open area where materials are merely placed.

Quantifying the reasoning burden in RS. To verify that reasoning itself is a major bottleneck, we construct an *Oracle Referring* setting by rewriting each EarthReason query into a direct referring expression with Qwen3-30B [57], thereby preserving the target semantics while removing the need for autonomous logical deduction. This converts the task from reasoning segmentation into a simplified referring-style segmentation problem and thus approximates the segmentation ceiling once reasoning difficulty is stripped away. As shown in Table 5, the baseline reasoning model (72.79% Avg) is 8.05 below Oracle Referring (80.84% Avg), quantifying a substantial *reasoning cost* in RS segmentation. Importantly, GeoReason reaches 78.61% Avg, recovering 5.82 of this gap and leaving only a 2.23 residual to the oracle ceiling. This supports our central claim: beyond preserving spatial fidelity, GeoReason also strengthens reasoning itself, bringing complex reasoning segmentation substantially closer to direct referring segmentation. We further apply PTA on top of Oracle Referring. Even after the reasoning burden is removed, PTA still raises performance to 81.78% Avg and 84.14% test cIoU. This mirrors the gains on RRSIS-D and indicates that PTA is not merely compensating for reasoning errors; it acts as a task-agnostic enhancer of the [SEG] token by improving its latent alignment with RS textures and structures.

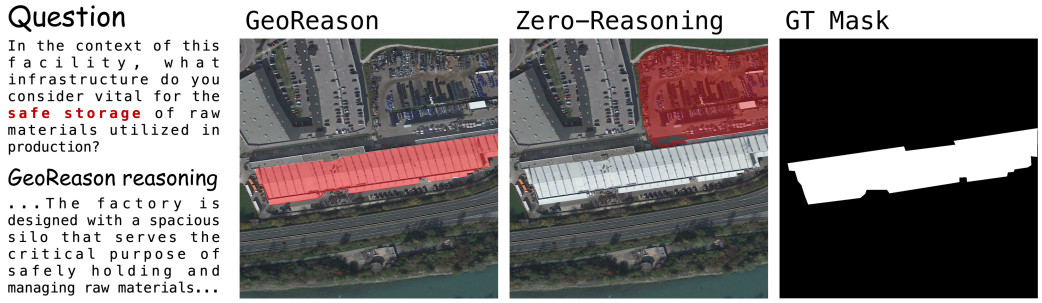


Figure 7: **Reasoning-sensitive case study on EarthReason.** GeoReason correctly selects the protected storage structure required for *safe* raw-material storage, whereas the Zero-Reasoning variant overfits to a visually plausible but functionally incorrect open yard.

5 Conclusion

We presented **GeoReason**, a framework for remote-sensing reasoning segmentation that addresses *Spatial Fidelity Degradation* and *Target Misidentification*. By combining a prompt-side `[SEG]` anchor, Location-aware CoT, and a PTA loss, GeoReason preserves visual evidence, improves grounding, and strengthens the segmentation-token representation. Experiments on EarthReason and RRSIS-D demonstrate state-of-the-art performance on both reasoning and referring segmentation.

Acknowledgments

This work was supported by the Hong Kong Research Grants Council (RGC) under Grant No. 21218026, and the City University of Hong Kong under Grants No. 9382010, 9048367 and 7020171.

References

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

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

[3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[4] Keyan Chen, Chenyang Liu, Bowen Chen, Jiafan Zhang, Zhengxia Zou, and Zhenwei Shi. Rsrefseg 2: decoupling referring remote sensing image segmentation with foundation models. *arXiv preprint arXiv:2507.06231*, 2025.

[5] Keyan Chen, Jiafan Zhang, Chenyang Liu, Zhengxia Zou, and Zhenwei Shi. Rsrefseg: Referring remote sensing image segmentation with foundation models. In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, pages 1070–1074. IEEE, 2025.

[6] Cheng Fan, Junyi Zhou, Tingzhang Luo, RongJian Xu, Qiyanhui Lu, Mingjian Zhu, Hanting Chen, and Jianyuan Guo. Reading is not reasoning: Bridging the agentic policy gap in vision-text compression, 2026. URL `https://arxiv.org/abs/2608.08960`.

[7] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. *Advances in neural information processing systems*, 37:57207–57239, 2024.

[8] Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29183–29192, 2025.

[9] Jianyuan Guo, Zhiwei Hao, Chengcheng Wang, Cheng Fan, Tingzhang Luo, Hongguang Li, Ying Gao, Hefei Mei, Jiankun Peng, Rongjian Xu, Minjing Dong, Han Wu, Mengyu Zheng, Kai Han, Shiqi Wang, Chang Xu, and Yunhe Wang. From question answering to task completion: A survey on agent system and harness design, 2026. URL `https://arxiv.org/abs/2606.20683`.

[10] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[12] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4067–4077, 2023.

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

[14] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[16] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27831–27840, 2024.

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024.

[18] Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*, 2024.

[19] Mengcheng Lan, Chaofeng Chen, Jiaxing Xu, Zongrui Li, Yiping Ke, Xudong Jiang, Yingchen Yu, Yunqing Zhao, and Song Bai. Text4seg++: Advancing image segmentation via generative language modeling. *arXiv preprint arXiv:2509.06321*, 2025.

[20] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025.

[21] Kaiyu Li, Zepeng Xin, Li Pang, Chao Pang, Yupeng Deng, Jing Yao, Guisong Xia, Deyu Meng, Zhi Wang, and Xiangyong Cao. Segearth-r1: Geospatial pixel reasoning via large language model. *arXiv preprint arXiv:2504.09644*, 2025.

[22] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

[23] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024.

[24] Jiazhen Liu, Mingkuan Feng, and Long Chen. Better, stronger, faster: Tackling the trilemma in mllm-based segmentation with simultaneous textual mask prediction. *arXiv preprint arXiv:2512.00395*, 2025.

[25] Ruizhong Liu, Tingzhang Luo, Shaoguang Huang, Yuwei Wu, Zhen Jiang, and Hongyan Zhang. Crossmatch: Cross-view matching for semi-supervised remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[26] Ruizhong Liu, Tingzhang Luo, Zaiyan Zhang, Jundong Chen, Hongruixuan Chen, Shaoguang Huang, and Hongyan Zhang. Geoseg-ov: Bridging geospatial gaps with structural guidance for open-vocabulary remote sensing segmentation, 2026. URL `https://arxiv.org/abs/2608.10426`.

[27] Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong 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*, pages 26658–26668, 2024.

[28] Xu Liu and Zhouhui Lian. Rsunivlm: A unified vision language model for remote sensing via granularity-oriented mixture of experts. *arXiv preprint arXiv:2412.05679*, 2024.

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

[30] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] Tingzhang Luo, Ruizhong Liu, Yichao Liu, Cheng Fan, Yu Liu, and Jianyuan Guo. Cross: Cascaded distillation and dual-constraint grounding for remote sensing referring segmentation. In *European Conference on Computer Vision*, pages 1–21. Springer, 2026.

[33] Shuo Ni, Di Wang, He Chen, Haonan Guo, Ning Zhang, and Jing Zhang. Unigeoseg: Towards unified open-world segmentation for geospatial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 34865–34876, 2026.

[34] Ruizhe Ou, Yuan Hu, Fan Zhang, Jiaxin Chen, and Yu Liu. Geopix: A multimodal large language model for pixel-level image understanding in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[35] Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how< seg> token works. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24722–24731, 2025.

[36] Jerome Quenum, Wen-Han Hsieh, Tsung-Han Patrick Wu, Ritwik Gupta, Trevor Darrell, and David Chan. Lisat: Language-instructed segmentation assistant for satellite imagery. *Advances in Neural Information Processing Systems*, 38, 2026.

[37] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[38] 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.

[39] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[40] Fu Rong, Meng Lan, Qian Zhang, and Lefei Zhang. Rs2-sam2: Customized sam2 for referring remote sensing image segmentation. *arXiv preprint arXiv:2503.07266*, 2025.

[41] Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad S Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. *arXiv preprint arXiv:2501.13925*, 2025.

[42] Yan Shu, Bin Ren, Zhitong Xiong, Danda Pani Paudel, Luc Van Gool, Begüm Demir, Nicu Sebe, and Paolo Rota. Earthmind: Leveraging cross-sensor data for advanced earth observation interpretation with a unified multimodal llm. *arXiv preprint arXiv:2506.01667*, 2025.

[43] Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5158–5171, 2025.

[44] Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in vision-language models. In *International Conference on Learning Representations*, volume 2026, pages 133928–133965, 2026.

[45] Chengcheng Wang, Jianyuan Guo, Hongguang Li, Yuchuan Tian, Ying Nie, Chang Xu, and Kai Han. Circle-rope: Cone-like decoupled rotary positional embedding for large vision-language models, 2026. URL `https://arxiv.org/abs/2505.16416`.

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

[47] Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

[48] Xiaolong Wang, Lixiang Ru, Ziyuan Huang, Kaixiang Ji, Dandan Zheng, Jingdong Chen, and Jun Zhou. Argenseg: Image segmentation with autoregressive image generation model. *arXiv preprint arXiv:2510.20803*, 2025.

[49] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 36:21429–21453, 2023.

[50] XuDong Wang, Shaolun Zhang, Shufan Li, Konstantinos Kallidromitis, Kehan Li, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Segllm: Multi-round reasoning segmentation. *arXiv preprint arXiv:2410.18923*, 2024.

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[52] Yana Wei, Liang Zhao, Kangheng Lin, En Yu, Yuang Peng, Runpei Dong, Jianjian Sun, Haoran Wei, Zheng Ge, Xiangyu Zhang, et al. Perception in reflection. *arXiv preprint arXiv:2504.07165*, 2025.

[53] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024.

[54] Zepeng Xin, Kaiyu Li, Luodi Chen, Wanchen Li, Yuchen Xiao, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. Segearth-r2: Towards comprehensive language-guided segmentation for remote sensing images. *arXiv preprint arXiv:2512.20013*, 2025.

[55] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098, 2025.

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

[57] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[58] 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.

[59] Liang Yao, Fan Liu, Delong Chen, Chuanyi Zhang, Yijun Wang, Ziyun Chen, Wei Xu, Shimin Di, and Yuhui Zheng. Remotesam: Towards segment anything for earth observation. In *Proceedings of the 33rd ACM international conference on multimedia*, pages 3027–3036, 2025.

[60] Liang Yao, Fan Liu, Hongbo Lu, Chuanyi Zhang, Rui Min, Shengxiang Xu, Shimin Di, and Pai Peng. Remotereasoner: Towards unifying geospatial reasoning workflow. *arXiv preprint arXiv:2507.19280*, 2025.

[61] Kaining Ying, Henghui Ding, Guangquan Jie, and Yu-Gang Jiang. Towards omnimodal expressions and reasoning in referring audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22575–22585, 2025.

[62] Zuyao You and Zuxuan Wu. Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. *arXiv preprint arXiv:2506.22624*, 2025.

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

[64] Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. Rrsis: Referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

[65] Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023.

[66] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024.

[67] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv e-prints*, pages arXiv–2505, 2025.

[68] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

[69] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[70] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

[71] Junyi Zhou, Yufei Wang, Qiyuan Zhang, Chuhan Wu, Zhengxi Li, Wentao Zhao, Jie Zhou, Le Tian, Xiao Zhou, and Chen Ma. Workborne: Discovering agentic tasks from real-world artifact transitions, 2026. URL `https://arxiv.org/abs/2610.14915`.

[72] Yue Zhou, Mengcheng Lan, Xiang Li, Litong Feng, Yiping Ke, Xue Jiang, Qingyun Li, Xue Yang, and Wayne Zhang. Geoground: A unified large vision-language model for remote sensing visual grounding. *arXiv preprint arXiv:2411.11904*, 2024.

[73] Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3686–3696, 2025.

A Implementation Details

In this part, we provide additional details regarding the experimental setup, including data composition, hyperparameter configurations, and the prompt templates used for training and inference.

A.1 Data Composition and Preprocessing

To endow GeoReason with foundational geographical cognition and robust understanding of remote sensing (RS) semantic attributes, we incorporate auxiliary data from **GeoChatInstruct** [16] and **VRSBench** [23] during the training phase. These datasets are large-scale RS benchmarks covering diverse tasks. To maintain a focus on general RS perception, we exclusively extract the Visual Question Answering (VQA) and Captioning subsets.

Data Leakage Prevention. A significant portion of images in GeoChatInstruct and VRSBench originate from the DIOR dataset [22], which is also the source for the RRSIS-D [64] benchmark. To ensure the integrity of our evaluation and prevent data leakage, we implemented a strict data cleansing process. We identified and removed all images (and their corresponding crops) from the auxiliary sets that overlap with the validation or test sets of RRSIS-D. After filtering, we obtained **21.2k** samples from GeoChatInstruct and **34.7k** samples from VRSBench. During training, these samples are interleaved with the segmentation datasets. Notably, these auxiliary tasks do not involve the `[SEG]` token or mask generation.

A.2 Training Configurations

Optimization. We employ the AdamW [31] optimizer with an initial learning rate of 4×10^{-5} and a weight decay of 0.05. A linear warmup strategy is applied for the first 5% of iterations. To ensure training stability, we use gradient clipping with a maximum norm of 1.0.

Model Fine-tuning. The vision encoder of the MLLM remains frozen throughout the training process. The MLP adapter and the SAM2 mask decoder are fully updated. For the Large Language Model (LLM), we utilize LoRA with a rank $r = 128$, $\alpha = 256$, and a dropout rate of 0.05. The maximum sequence length is set to 8192. Regarding the multi-task optimization, the balancing coefficients λ_{text}, λ_{mask}, and λ_{pta} are all set to 1.0 in our implementation.

Hardware and Efficiency. The model is trained for a single epoch with a total effective batch size of 64. We utilize $4\times$ NVIDIA RTX Pro 6000 GPUs, employing a per-GPU batch size of 2 combined with 8-step gradient accumulation. The training is conducted using BF16 mixed-precision and takes approximately 9 hours to complete. Benchmarked on the EarthReason validation set, GeoReason increases latency (1223.6 ms vs. 774.1 ms) and token count (73.3 vs. 45.1) compared to the baseline (Qwen3VL-4B + SAM2 without our proposed methods). This overhead stems primarily from the autoregressive generation of reasoning and bounding box tokens rather than model complexity, representing a highly favorable trade-off for the significant performance gains.

A.3 Prompt Templates and Formats

We define specific prompt templates for Reasoning Segmentation (RS) and Referring Expression Segmentation (RES). The bounding box coordinates used in the `<think>` tags are derived from the ground-truth masks by calculating the minimum axis-aligned box that covers all target pixels.

Prompt Templates for GeoReason Training and Inference

Task A: Reasoning Segmentation

Query Template:

```
<image> [Question] When responding, please output your detailed reasoning
and the bbox of the target in JSON format in <think> </think> tags, and then
provide your answer in <answer> </answer> tags, i.e., <think> reasoning and
bbox here </think> <answer> answer with [SEG] here </answer>.
```

Model Response Format:

```
<think> [Detailed Reasoning Content] The bbox of the target: {"bbox_2d":
[xmin, ymin, xmax, ymax]} </think> \n <answer> [Answer Content] </answer>.
```

Task B: Referring Expression Segmentation

Query Template:

```
<image> [Question] When responding, please output the bbox of the target in
JSON format in <think> </think> tags, and then provide your answer in <answer>
</answer> tags, i.e., <think> bbox here </think> <answer> answer with [SEG]
here </answer>.
```

Model Response Format:

```
<think> The bbox of the target: {"bbox_2d": [xmin, ymin, xmax, ymax]}
</think> \n <answer> [Answer Content] </answer>.
```

Note: During training, both the `[SEG]` token in the query and the `[SEG]` token(s) in the `<answer>` tag are jointly supervised using $\mathcal{L}_{bce}$, $\mathcal{L}_{dice}$, and $\mathcal{L}_{pta}$.

Supervision and Inference. During training, our segmentation-related loss is applied to both the prompt-side `[SEG]` token and any `[SEG]` tokens generated within the `<answer>` tags. During inference, we extract the latent embeddings of the prompt-side `[SEG]` token and all `[SEG]` tokens present in the generated answer. Each token produces a candidate mask through the SAM2 decoder, and the final segmentation result is obtained by taking the union of all predicted masks.

B Visual-Evidence Dilution in Long Reasoning

B.1 Analysis of Visual-Evidence Dilution

During visually grounded tasks, extended textual reasoning can weaken multimodal models' reliance on visual evidence in subsequent answers. Several studies have examined this phenomenon. Sun et al. [43] analyze long-CoT *visual forgetting*: removing the image midway through reasoning changes MathVista-Hard accuracy only from 43.1% to 40.9%, while attention to visual tokens declines as reasoning proceeds. Tian et al. [44] use early decisions to examine reasoning progress: MMStar and HallusionBench accuracy drops by over two percentage points after its peak; among Vision-R1's full-reasoning errors, 55.23% are perception errors, and 32.35% of these can be corrected at an earlier decision point. MME-CoT [14] reports negative stability when switching from direct answers to CoT on perception tasks for most models, with a maximum drop of 6.8 percentage points. OPERA [13] observes that longer generation can cause later tokens to over-trust nearby summary tokens and neglect earlier image tokens.

In the embedding-as-mask setting, longer reasoning may dilute visual evidence through two mechanisms: **(i) visual-token competition**, where additional reasoning tokens compete with image tokens for attention from the answer-side `[SEG]`, and **(ii) mask-query drift**, where the evolving reasoning context shifts its hidden representation away from mask-relevant visual evidence. We analyze these mechanisms through image-attention mass and segmentation-query representation drift.

Let $\mathcal{R}_L$ be a reasoning prefix of L tokens, $\mathcal{I}_{img}$ the image-token indices, and $\mathcal{C}_L$ the context visible to the final answer-side `[SEG]`. For one attention head at layer ℓ, its image-attention mass is

$$A_{img}^{(\ell)}(L) = \sum_{i \in \mathcal{I}_{img}} \frac{\exp\Big((\mathbf{q}_{seg}^{(\ell)}(L))^\top \mathbf{k}_i^{(\ell)} / \sqrt{d}\Big)}{\sum_{j \in \mathcal{C}_L} \exp\Big((\mathbf{q}_{seg}^{(\ell)}(L))^\top \mathbf{k}_j^{(\ell)} / \sqrt{d}\Big)}, \tag{8}$$

where d is the head dimension. Additional reasoning tokens enlarge the competing linguistic context, while the segmentation query and attention logits can also change.

Using the notation of Sec. 3.1, the answer-side mask after this prefix is

$$\mathcal{M}_{pred}^{ans}(L) = \mathcal{D}\big(\mathcal{F}_{img}, f_\phi\big(\mathbf{h}_{seg}^{ans}(L)\big)\big)\,. \tag{9}$$

We measure the representation change relative to the zero-length prefix as

$$D_h(L) = 1 - \cos\big(\mathbf{h}_{seg}^{ans}(0), \mathbf{h}_{seg}^{ans}(L)\big)\,. \tag{10}$$

Within each model, the image features and decoder remain fixed across the intervention. Changes in $\mathbf{h}_{seg}^{ans}(L)$ can therefore alter the decoder's latent condition without retraining or replacing the segmentation architecture.

B.2 Controlled Prefix-Length Experiments

We construct the experiment from the EarthReason test set, using one canonical trajectory saved from GeoReason's naturally generated response for each instance. For a trajectory containing T tokenizer tokens, we retain its first $\lfloor rT \rfloor$ tokens for $r \in \{0, 0.25, 0.5, 0.75, 1\}$ and concatenate this prefix with the same final answer. We compare GeoReason with the Qwen3-VL+SAM2 baseline by feeding the resulting sequence into each model. To examine these mechanisms, we measure image attention and hidden-state drift of the answer-side `[SEG]` across prefix lengths.

Table 6: Effect of reasoning-prefix length on the answer-side `[SEG]` on EarthReason test. Image-attention mass is summed over image tokens and averaged across heads, the last four layers, and instances. Hidden-state drift is the mean cosine distance between the last-layer `[SEG]` representations at the retained and zero-length prefixes.

Prefix (%)	Image-attention mass		Hidden-state drift	
	Baseline	GeoReason	Baseline	GeoReason
0	0.2389	0.2593	0.0000	0.0000
25	0.2206	0.2502	0.1955	0.0233
50	0.2135	0.2491	0.1896	0.0248
75	0.2199	0.2481	0.2097	0.0280
100	0.2003	0.2446	0.2076	0.0495

From the zero to the full prefix, the baseline's image attention decreases from 0.2389 to 0.2003 (16.15% relative attenuation), whereas GeoReason changes from 0.2593 to 0.2446 (5.69%). GeoReason reduces the absolute endpoint attenuation by 61.7%. Its full-prefix hidden-state drift is 0.0495, compared with 0.2076 for the baseline, a 76.2% reduction. The baseline's reduced image attention and increased hidden-state drift are consistent with visual-token competition and mask-query drift, respectively. GeoReason mitigates both effects, preserving visual access and stabilizing the segmentation query under longer reasoning prefixes.

C Extended Ablation Studies on PTA Loss

We conduct extensive ablation experiments to validate the design of the *Latent Patch-to-Token Alignment* (PTA) loss, focusing on the PTA head architecture, spatial supervision granularity, pooling strategy, and loss formulations.

Architecture of the PTA Head. To distinguish from the mask-decoder adapter (f_ϕ), we denote the auxiliary alignment module as the **PTA Head (g_θ)**. This head projects the `[SEG]` token into a dedicated alignment space during training. As shown in Table 7, removing the PTA head (*w/o PTA Head*) leads to a significant performance drop, especially on EarthReason (-4.67% in test cIoU). This observation supports our hypothesis that while the `[SEG]` token and image patch tokens share the same dimensionality, they reside in distinct semantic feature spaces. Directly forcing alignment without a cross-modal projection head (like our GLU-based g_θ) compromises the token's representational capacity, causing it to lose critical information required for subsequent mask generation. Furthermore, our GLU-based PTA head outperforms a simple MLP, effectively modeling complex latent interactions for better alignment.

Granularity of Spatial Supervision. We compare our mask-based alignment with a coarser bounding box-based strategy. While the BBox approach aligns the `[SEG]` token with all patches within the rectangular area, our default strategy strictly maps only positive pixels to their corresponding tokens. Table 7 shows that this shape-aware supervision is superior, as it filters out background noise within the BBox and provides more precise geometric guidance.

Pooling Strategy. We compare our default *Average Pooling*, which uniformly aggregates target patches, with *Attention Pooling* that dynamically weights patches using the `[SEG]` token as a query. Interestingly, Attention Pooling leads to performance degradation. This is likely because the attention mechanism over-focuses on local salient features while neglecting the overall footprint and boundaries, which are essential for maintaining the structural integrity of masks in complex remote sensing scenes.

Loss Formulation. Finally, we evaluate different loss functions for PTA. While MSE and MAE provide competitive results, Cosine Similarity achieves the most balanced performance across both benchmarks, demonstrating its robustness in aligning high-dimensional latent vectors.

Table 7: Ablation studies on PTA Loss configurations. Δ denotes the performance gap relative to our final GeoReason model (red for decrease, blue for increase). Results are reported in (%).

Method / Variant	RRSIS-D (Referring)				EarthReason (Reasoning)			
	Val gIoU	Val cIoU	Test gIoU	Test cIoU	Val gIoU	Val cIoU	Test gIoU	Test cIoU
GeoReason	**68.36**	**79.87**	**67.32**	**79.39**	**77.61**	**77.63**	**78.61**	**80.60**
w/o PTA Head	68.26 -0.10	78.26 -1.61	67.24 -0.08	78.38 -1.01	77.20 -0.41	75.92 -1.71	77.31 -1.30	75.93 -4.67
Simple PTA Head (MLP)	68.39 +0.03	78.85 -1.02	67.46 +0.14	78.80 -0.59	77.51 -0.10	78.60 +0.97	78.34 -0.27	79.78 -0.82
BBox-based Alignment	68.56 +0.20	79.26 -0.61	67.28 -0.04	78.65 -0.74	77.26 -0.35	77.71 +0.08	78.64 +0.03	79.47 -1.13
PTA Pooling: Attention	67.76 -0.60	79.05 -0.82	66.69 -0.63	78.11 -1.28	77.47 -0.14	77.47 -0.16	78.10 -0.51	78.65 -1.95
PTA Loss: MSE	68.69 +0.33	79.45 -0.42	67.12 -0.20	78.30 -1.09	77.40 -0.21	77.06 -0.57	78.49 -0.12	79.02 -1.58
PTA Loss: MAE	68.27 -0.09	79.60 -0.27	67.23 -0.09	78.61 -0.78	77.54 -0.07	77.22 -0.41	78.66 +0.05	78.40 -2.20

D Additional Benchmark Results on GeoSeg-Bench

GeoSeg-Bench [33] is a geospatial segmentation benchmark encompassing Interactive, Referring, and Reasoning tasks across diverse scenes and categories. To further demonstrate GeoReason's generalizability across diverse and challenging geospatial contexts, we extend our evaluation to this benchmark. We evaluate the GeoReason checkpoint trained in the main experiments on the Referring and Reasoning subsets of GeoSeg-Bench, without fine-tuning on GeoSeg-1M, the training dataset introduced in UniGeoSeg [33]. Table 8 compares GeoReason with eleven methods evaluated without GeoSeg-1M fine-tuning, as well as fine-tuned Geopixel and Earthmind, using results reported in UniGeoSeg [33].

As shown in Table 8, GeoReason achieves the highest gIoU on both tasks among the compared methods. Compared with Earthmind, the next-best method without GeoSeg-1M fine-tuning, GeoReason improves Referring and Reasoning gIoU by 10.16 and 3.66 percentage points, respectively. Referring cIoU improves by 5.97 points, while Reasoning cIoU is nearly tied (30.99 versus 31.01, a difference of -0.02 points). Despite using no GeoSeg-1M training data, GeoReason also surpasses the fine-tuned baselines on multiple metrics. These results further demonstrate its effectiveness beyond EarthReason and RRSIS-D.

E Qualitative Visualizations and Analysis

In this section, we provide detailed qualitative evidence to further elucidate how GeoReason addresses the challenges in remote sensing (RS) reasoning segmentation.

E.1 Case Study

We further illustrate GeoReason with a representative case study against the baseline (Qwen3VL-4B + SAM2 without our proposed components). As shown in Fig. 8, the baseline attends to an incorrect structure and fails to isolate the intended target, whereas GeoReason correctly identifies the target

Table 8: Results on GeoSeg-Bench. FT indicates fine-tuning on the GeoSeg-1M training dataset [33]. Baseline scores are reproduced from UniGeoSeg [33]. Best results are in **bold**; second-best are underlined.

Method	FT	Referring		Reasoning	
		gIoU	cIoU	gIoU	cIoU
LISA [17]	✗	4.56	3.53	5.77	7.09
PixelLM [39]	✗	5.80	6.04	6.30	6.36
PSALM [69]	✗	18.91	31.77	9.27	11.88
HIPIE [49]	✗	39.14	29.25	11.76	7.90
SegLLM [50]	✗	35.27	15.90	16.41	11.06
Geopixel [41]	✗	40.14	37.34	26.71	27.36
Geopix [34]	✗	28.52	28.81	18.25	20.80
RemoteSAM [59]	✗	29.31	12.06	9.01	8.09
Earthmind [42]	✗	46.48	44.53	28.80	31.01
LISAT [36]	✗	40.14	37.87	20.09	22.06
SegEarth-R1 [21]	✗	9.39	14.65	11.97	12.72
Geopixel [41]	✓	45.50	42.70	28.51	30.25
Earthmind [42]	✓	49.24	48.09	25.71	**36.84**
GeoReason	✗	**56.64**	**50.50**	**32.46**	30.99

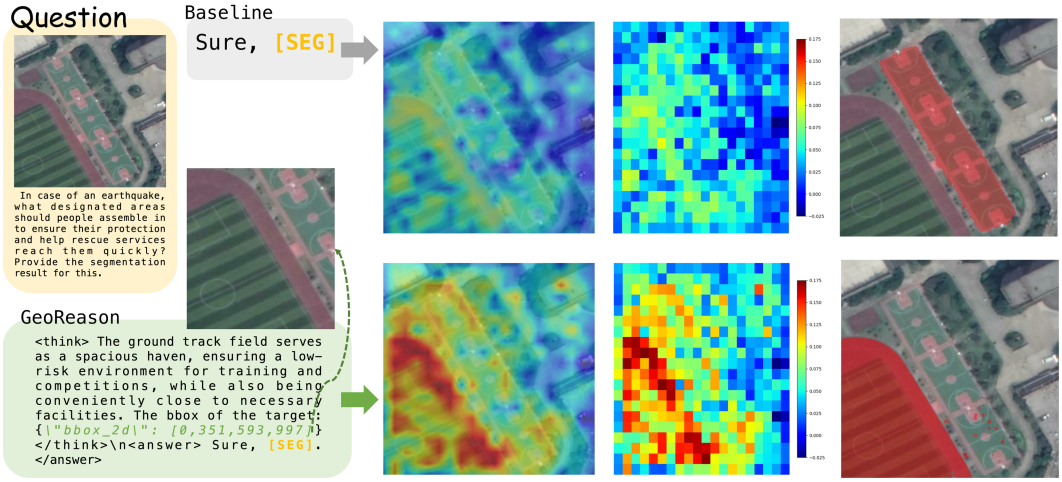


Figure 8: Qualitative case study comparing the baseline and GeoReason. GeoReason first grounds the correct region with an intermediate bbox in CoT, then exhibits stronger target-focused [SEG] and produces a more accurate final mask.

region. A key difference is that GeoReason first localizes the relevant area through the intermediate bbox generated in CoT, which provides an explicit coarse spatial anchor before mask decoding.

We then visualize the cosine similarity between the [SEG] hidden state and image patch tokens. Compared with the baseline, GeoReason produces a much stronger and more concentrated response on the true target region, while suppressing unrelated areas. This sharper contrast is consistent with the final segmentation result: GeoReason yields the correct mask, whereas the baseline remains distracted by competing regions and produces an inferior prediction.

This case supports the complementary roles of our components. The Location-aware CoT improves coarse target localization through bbox reasoning, while the Anticipatory Prior and PTA loss make the [SEG] representation more target-sensitive, leading to more reliable final segmentation.

E.2 Analysis of Typical Failure Modes

We further provide additional qualitative results to illustrate how GeoReason addresses common failure modes in remote sensing reasoning segmentation (Fig. 9):

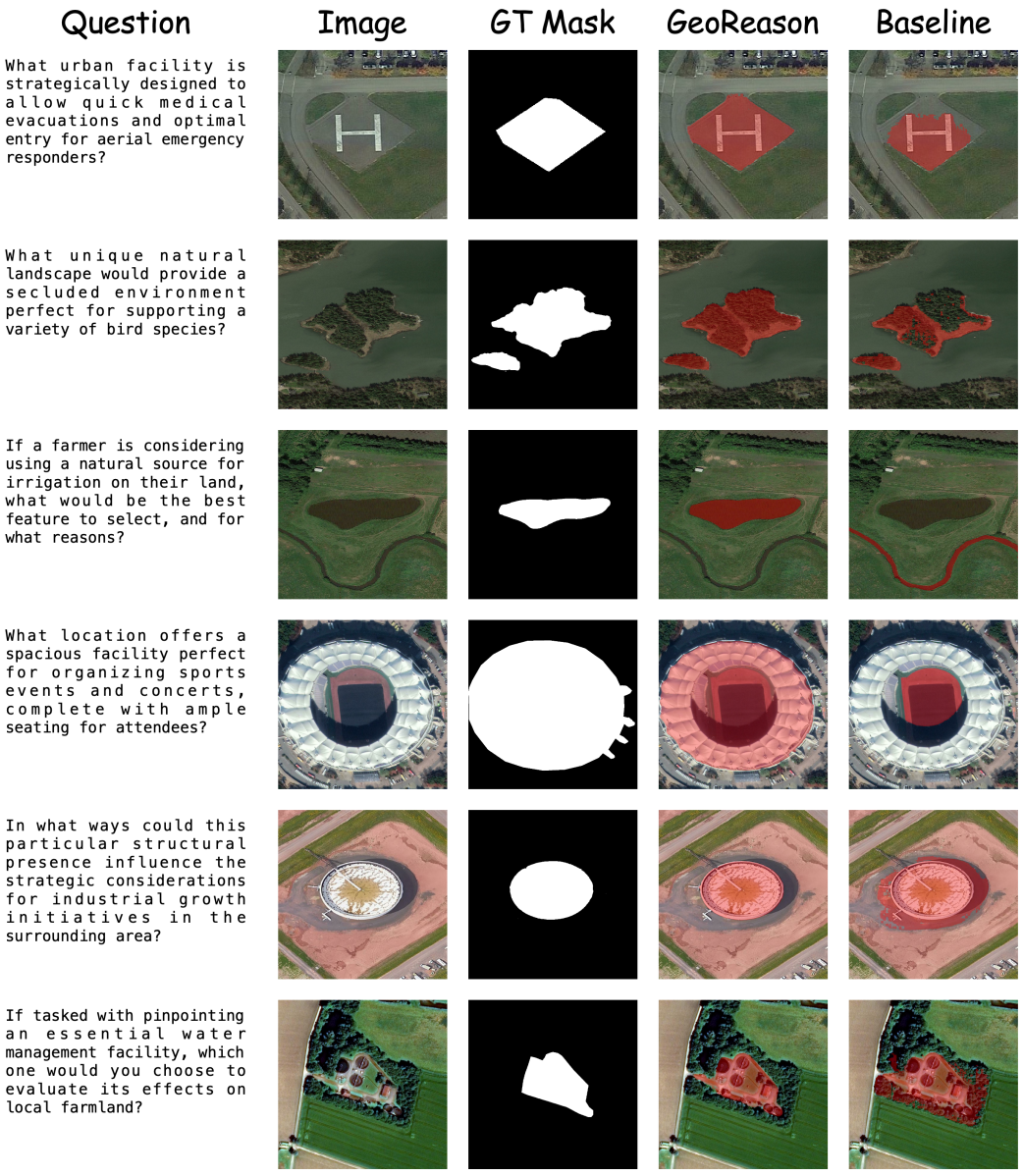


Figure 9: Mitigation of typical failure modes in RS reasoning segmentation.

1. **Spatial Fidelity Degradation:** Due to information decay, baseline models often generate noisy or fragmented masks with blurry boundaries. GeoReason establishes a spatial anchor on the prompt-side via the *Anticipatory Prior*, anchoring visual evidence before autoregressive generation to maintain mask integrity.
2. **Target Misidentification:** Weak visual reasoning often leads standard models to segment incorrect objects or partial targets. Our *Location-aware CoT* enforces an explicit coarse localization (bounding box) before generating the `[SEG]` token, ensuring the model focuses on the correct location.
3. **Boundary Redundancy:** Insufficient understanding of RS textures causes baseline masks to exhibit redundant fragments at the boundaries. By applying the *PTA Loss*, we align the `[SEG]` token with ground-truth image patches, enhancing its representational quality and eliminating blurry mask redundancies.

F Limitations

Our study has several limitations. First, the main experiments use Qwen-VL and SAM-family models. Although the proposed design is modular, its transferability to other MLLM backbones or segmentation decoders has not yet been systematically evaluated. Second, our benchmarks mainly emphasize single-target segmentation, and more complex settings such as multi-target reasoning, heavy occlusion, or highly compositional queries remain for future study.

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 abstract, introduction, and experiments consistently describe the scope of the paper: improving remote-sensing reasoning segmentation with Anticipatory Prior, Location-aware CoT, and PTA. The empirical claims are aligned with the reported results on EarthReason and RRSIS-D.

   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: We discuss limitations in Appendix F, including the evaluated Qwen-VL and SAM-family models, the scope of architecture transfer, and the current focus on mainly single-target settings.

   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 does not present formal theoretical results requiring theorem statements or proofs; it is primarily a method and empirical study.

   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 paper and appendix specify the architecture, training strategy, datasets, metrics, ablations, prompt formats, and implementation details needed to reproduce the main results. We also describe optimization, hardware, and inference settings in the implementation details section.

   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: Code and instructions for environment setup, data preparation, training, and evaluation are available at `https://github.com/Fanc0126/GeoReason`. The experiments use existing datasets and models, whose usage is described in the paper and appendix.

   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 main paper provides the core experimental setup, and the appendix further specifies the optimizer, learning rate, LoRA configuration, batch size, sequence length, loss weights, data composition, and prompt templates.

   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: The main benchmark results are reported as single-run performance without error bars or repeated-run significance tests. We instead provide multiple ablations, mechanistic analyses, and qualitative studies to support the main claims.

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 appendix reports the hardware setup, effective batch size, precision, approximate training time, and relative inference latency and token cost. This information is provided in "Hardware and Efficiency."

   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: To the best of our knowledge, the research complies with the NeurIPS Code of Ethics. It uses standard model training and evaluation on existing remote-sensing datasets and does not involve human-subject experiments.

   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 paper does not include a dedicated discussion of broader impacts. Potential positive uses include geographic understanding and Earth observation analysis, while potential negative uses may include surveillance or other misuse of remote-sensing systems.

    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 paper does not release a high-risk generative model or a scraped dataset. Our release focuses on research code for segmentation experiments built on existing datasets and backbones.

    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: [No]

    Justification: We cite the original papers for the datasets, models, and baselines used in this work, but the paper does not explicitly enumerate the licenses and terms of use of all external assets.

    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: The implementation and documentation for setup, training, and evaluation are available at `https://github.com/Fanc0126/GeoReason`.

    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 or 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 human-subject research or data collection requiring IRB review.

    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: [Yes]

    Justification: LLMs are a core component of the method rather than a writing aid. The paper explicitly describes the Qwen3VL-based MLLM backbone and analyzes how segmentation tokens behave during multimodal reasoning and mask generation.

    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.