GEOSEG: TRAINING-FREE REASONING-DRIVEN SEGMENTATION IN REMOTE SENSING IMAGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in multimodal large language models (MLLMs) have reframed segmentation from fixed-category prediction toward open-ended, instruction-grounded localization. While reasoning-based segmentation has progressed rapidly in natural scenes, extending these advances to remote sensing remains an open challenge, owing to the scarcity and high annotation cost of reasoning-oriented training data, as well as domain-specific obstacles such as overhead viewpoints and extreme scale variation. We present **GeoSeg**, a training-free framework for reasoning-driven remote sensing segmentation that circumvents the supervision bottleneck by composing pretrained vision–language models for zero-shot inference. GeoSeg bridges MLLM reasoning and pixel-level localization through two key components: (i) **bias-aware coordinate refinement**, which corrects the systematic spatial offset introduced when MLLMs process overhead imagery, and (ii) **dual-route prompting**, which fuses visual keypoint cues with semantic text prompts for robust mask prediction. We also introduce **GeoSeg-Bench**, a diagnostic benchmark comprising 810 annotated image–query pairs across four scene domains and three hierarchical difficulty levels, designed to support standardized, zero-shot evaluation of reasoning-driven segmentation methods in remote sensing.

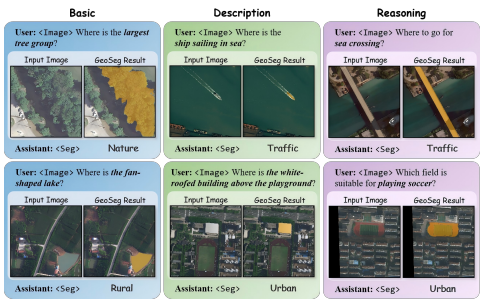


Figure 1: Selected predictions across the three query levels of GeoSeg-R810. The examples illustrate attribute, relation, and intent grounding; aggregate performance is reported separately.

1 INTRODUCTION

Remote sensing segmentation has evolved alongside how we specify *what* to segment. Early closed-set models Chen et al. (2018); Ma et al. (2021); Diakogiannis et al. (2020); Chen et al. (2021) relied on dense pixel supervision over fixed taxonomies. Subsequently, open-vocabulary methods Liu et al. (2024a); Xu et al. (2023); Cao et al. (2024); Liang et al. (2023); Han et al. (2025) generalized beyond training classes via vision–language alignment, while generalist promptable segmenters Kirillov et al. (2023); Ravi et al. (2024); Carion et al. (2025); Wang et al. (2023) further decoupled segmentation from categories by using point or box prompts to "segment anything" Kirillov et al. (2023). This shift facilitates adaptation across diverse Earth observation scenarios.

Concurrently, reasoning-driven segmentation Lai et al. (2024); Yan et al. (2024); Wei et al. (2024); Zhang et al. (2024a); Xia et al. (2024) enables models to interpret complex natural language requests involving attributes, relations, or implicit intents (e.g., *"where can I seek medical help?"*) and output precise pixel-level masks. While successful in natural scenes, transferring these advances to remote sensing is hindered by a domain gap in viewpoint, scale, and scene organization. Grounding errors under overhead views can exclude the intended referent before segmentation begins, bottlenecking open-ended analysis.

Remote sensing further poses unique challenges for reasoning segmentation: overhead perspectives remove typical cues, targets exhibit extreme scale variations and high density, and objects often require spatial layout or functional context rather than texture for disambiguation. Reasoning-oriented remote sensing datasets are also scarce, making heavy, task-specific fine-tuning impractical. Con-

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

sequently, despite the maturation of generalist segmenters, a critical gap remains: a frozen-model pipeline must connect semantic reasoning to reliable spatial support, while evaluation must distinguish attributes, relations, and intent-oriented requests.

To circumvent this supervision bottleneck, we introduce GeoSeg, a training-free framework for reasoning-driven segmentation in remote sensing imagery. As detailed in Sec. 3, GeoSeg decomposes inference into three stages linking MLLM reasoning to pixel-level masks. **Stage 1 (reasoning-driven grounding):** An MLLM $\mathcal{L}$ processes the image–query pair to generate a grounding tuple (b, p)—a coarse bounding box and a referential phrase—bridging high-level intent with a spatial search region. **Stage 2 (bias-aware coordinate refinement):** To counter grounding drift under overhead views, we asymmetrically expand b using margins (α, β) fixed using the reported held-out calibration analysis, yielding a refined crop $I_{b'}$ to improve potential target coverage. **Stage 3 (dual-route segmentation and fusion):** Within $I_{b'}$, *Route A* extracts visual keypoints via CLIP Surgery Li et al. (2023) to prompt SAM3 Carion et al. (2025), parallel to *Route B*, which prompts SAM3 using text p. These crop-level masks are merged via consensus-driven fusion, prioritizing their *intersection* to suppress clutter-induced false positives with conservative fallbacks when only one route provides a valid mask or their outputs do not overlap.

To support systematic evaluation, we introduce GeoSeg-Bench (denoted GeoSeg-R810), a dedicated benchmark comprising 810 image–query pairs spanning diverse scenarios, object categories, spatial contexts, and query types with hierarchical difficulty levels. The name distinguishes this resource from UniGeoSeg's GeoSeg-Bench (Ni et al., 2026). It supports comparisons across general segmenters, reasoning models, and MLLMs with fixed model weights; geometric margins are calibrated separately. Experiments show that GeoSeg achieves the best overall performance among the evaluated methods on GeoSeg-R810 across standard pixel-level metrics, and is additionally evaluated on an external LaSeRS subset. Figure 1 shows selected predictions; detailed ablations characterize the contribution of each core component. In summary, our contributions are three-fold:

- **Task & problem setting.** We study the problem setting of *instruction-grounded, reasoning-driven* segmentation in remote sensing, and identify the key challenges that differentiate it from natural-image benchmarks.
- **Methodological innovation.** We propose GeoSeg, a training-free framework that integrates bias-aware coordinate refinement and dual-route prompting of visual and semantic cues to enable accurate instruction-grounded pixel-level localization.
- **Benchmark & evaluation.** We establish GeoSeg-R810, a benchmark featuring 810 manually annotated image–query pairs across four scenarios and hierarchical difficulty levels, and provide a standardized zero-shot evaluation protocol that supports structured comparison across reasoning-based segmentation methods.

2 Related Work

2.1 Remote Sensing Segmentation Evolution

Traditional closed-set models (Chen et al., 2018; Ma et al., 2021; Diakogiannis et al., 2020; Chen et al., 2021; 2024a) predict pixel masks over predefined taxonomies, constrained by costly supervision. Open-vocabulary methods (Xu et al., 2023; Liang et al., 2023; Han et al., 2025; Lüddecke & Ecker, 2022; Xu et al., 2022; Rao et al., 2022) exploit vision–language pretraining to generalize beyond training categories; several works have adapted these to remote sensing (Liu et al., 2024a; Cao et al., 2024; Zhang et al., 2024b). Despite this progress, most open-vocabulary methods still operate on explicit class names and struggle with instruction-grounded queries involving spatial relations or implicit intent—challenges compounded by overhead viewpoints and extreme scale variation.

2.2 Generalist Promptable Models

Foundation promptable segmenters (Kirillov et al., 2023; Ravi et al., 2024; Carion et al., 2025; Wang et al., 2023) decouple segmentation from category inventories via point, box, or mask prompts. Grounding models such as GLIP (Li et al., 2022) and Grounding DINO (Liu et al., 2024b) enable language-driven spatial localization, catalyzing remote sensing adaptations includ-

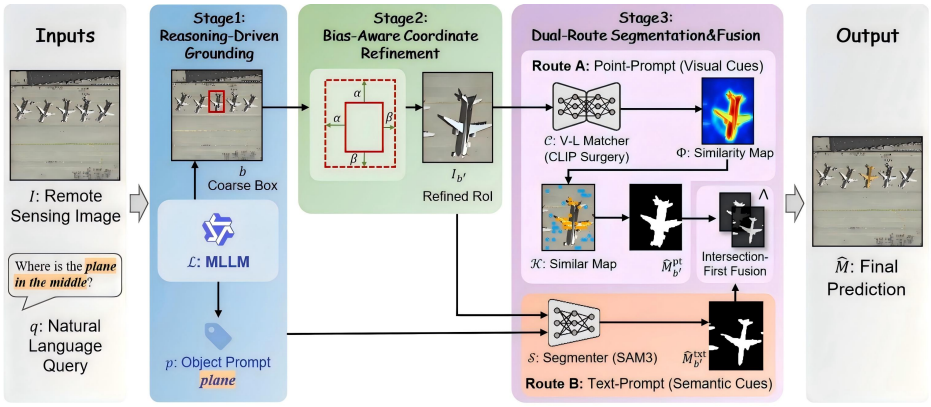


Figure 2: **The GeoSeg pipeline.** Given image I and query q: (1) Reasoning-driven grounding: MLLM $\mathcal{L}$ generates coarse box b and prompt p. (2) Bias-aware coordinate refinement: asymmetric expansion (α, β) yields refined RoI $I_{b'}$. (3) Dual-route segmentation: *Route A* (CLIP Surgery visual cues) and *Route B* (SAM3 with p) are fused via consensus fusion to produce $\hat{M}$.

ing RSPrompter (Chen et al., 2024b) and SAMRS (Wang et al., 2023). A common recipe pairs a grounding module with a SAM-like segmenter. While effective with accurate prompts, these pipelines remain sensitive to grounding biases under overhead views, leading to cascading failures. SAM4MLLM also connects language-model localization and point cues to SAM (Chen et al., 2024c). GeoSeg specializes this interface to frozen components, fixed crop expansion, CLIP-derived points, and a text-prompted parallel route.

2.3 Reasoning-Driven and Geospatial Segmentation

LISA++, PixelLM, GLAMM, GSVA, and Seg-Zero connect language reasoning with mask prediction (Yang et al., 2023; Ren et al., 2024b; Rasheed et al., 2024; Xia et al., 2024; Liu et al., 2025). VisionReasoner integrates reasoning and perception through reinforcement learning (Liu et al., 2026). In remote sensing, LISAt introduces GRES and specialized training, UniGeoSeg provides multi-task geospatial instruction data and GeoSeg-Bench, and PixDLM introduces the UAV-focused DRSeg resource and dual-path representation (Quenum et al., 2025; Ni et al., 2026; Ke et al., 2026). GeoSeg focuses on composing frozen grounding and segmentation components through calibrated crop expansion and complementary point/text prompting. Its intersection-first fusion rule connects these cues to a common output mask. Appendix I summarizes the distinct task, data, and adaptation settings of the related resources.

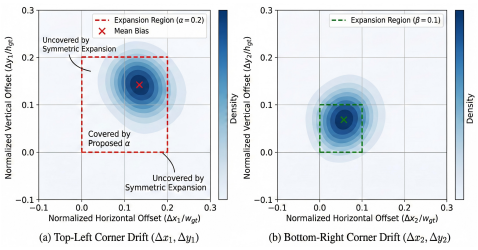
3 Method

Figure 2 organizes inference into three connected stages. Section 3.1 defines the image–query input and mask output. Section 3.2 converts the query into a referential phrase and a calibrated search region. Section 3.3 derives masks from point and text prompts in that region and defines how consensus and fallback rules convert the two predictions into a single output mask.

3.1 Problem Definition and Framework Overview

Let $I \in \mathbb{R}^{H\times W\times 3}$ be a remote sensing image and q be an open-ended natural language query that may involve attributes, spatial relations, or implicit intent (e.g., *"the residential buildings arranged in rows next to the park"*). The goal is to predict a binary mask $M \in \{0,1\}^{H\times W}$, where $M_{u,v} = 1$ indicates pixels belonging to the target region implied by q. A plural query can denote a union of disconnected regions. One enclosing RoI must contain every intended component; the pipeline cannot recover targets excluded at grounding.

Interfaces and spatial support. The phrase p and region b' serve different purposes. The phrase specifies which visible content is relevant; the region bounds where the downstream segmenter can



(a) Top-Left Corner Drift ($\Delta x_1, \Delta y_1$) (b) Bottom-Right Corner Drift ($\Delta x_2, \Delta y_2$)

Figure 3: **Quantification of domain-specific grounding drift.** We analyze coordinate offsets on a held-out calibration set comprising 1,000 images randomly sampled from LoveDA, NWPU-VHR-10, and DIOR datasets. The KDE illustrates Qwen3-VL-32B coordinate errors on these sources and motivates the empirical margins $\alpha = .2, \beta = .1$. It does not establish a universal MLLM bias or a unique margin-selection criterion.

All model weights remain fixed: Qwen3-VL-32B (Yang et al., 2025), CLIP Surgery (Li et al., 2023), and SAM3. Geometric calibration uses reference localization annotations; its source-overlap implications are discussed in Appendix G.

The plot concerns box localization before mask generation. GeoSeg uses these offsets to motivate fixed search-region expansion; the segmenter then resolves foreground support within the crop.

Expansion can retain target pixels excluded by a tight box but also admit same-class distractors. The phrase and prompting routes distinguish the requested pixels within the enlarged region. Geometry is therefore evaluated together with mask prediction.

find it. Neither alone supplies a pixel mask. The two routes therefore share $(I_{b'}, p)$ but expose different controls to SAM3: selected coordinates in Route A and the phrase itself in Route B. Keeping the crop common prevents a route comparison from silently changing the available spatial context.

The final mask is placed on the original $H \times W$ canvas, with zero support outside the crop. For plural referents, the enclosing box may contain intervening distractors; the mask routes separate them from the requested foreground.

3.2 Reasoning-Driven Grounding and Bias-Aware Coordinate Refinement

Reasoning-driven grounding. The reasoning module $\mathcal{L}$ serves as the initial interpreter. We employ a composite prompt to guide $\mathcal{L}$ in decomposing the complex query q into structured spatial and semantic outputs. Given (I, q), $\mathcal{L}$ generates a response from which we parse a grounding tuple:

$$(b, p) = \text{Parse}(\mathcal{L}(I, q)), \tag{1}$$

where $b = [x_1, y_1, x_2, y_2]$ denotes the coarse bounding box in absolute pixel coordinates, and p is a concise referential phrase inferred from the image and query. For implicit intent, its target noun need not occur literally in q.

Bias-aware coordinate refinement. The coordinate analysis uses 1,000 images from LoveDA, NWPU-VHR-10, and DIOR with Qwen3-VL-32B (Figure 3). The observed residual distribution motivates enlarging the search region more strongly toward the upper-left. We use fixed empirical margins calibrated for this model and source collection. This correction increases the spatial support available to both segmentation routes while preserving a simple, shared crop interface; the calibration protocol is described in Appendix G.

We first clamp b to image bounds, then expand it with bias-aware margins:

$$\begin{aligned} w &= x_2 - x_1, \quad h = y_2 - y_1, \\ x_1' &= \text{clip}(x_1 - \alpha w,\, 0,\, W), \\ y_1' &= \text{clip}(y_1 - \alpha h,\, 0,\, H), \\ x_2' &= \text{clip}(x_2 + \beta w,\, 0,\, W), \\ y_2' &= \text{clip}(y_2 + \beta h,\, 0,\, H). \end{aligned} \tag{2}$$

where $\alpha = 0.2$ and $\beta = 0.1$ are the fixed settings used in the reported experiments. This yields the refined crop $I_{b\prime}$; the geometric effect is given below.

Geometric effect of refinement. Before image-boundary clipping,

$$w' = (1 + \alpha + \beta)w = 1.3w, \quad h' = 1.3h, \quad A(b') = 1.69A(b), \tag{3}$$

and the center displacement is $((\beta-\alpha)w/2, (\beta-\alpha)h/2) = (-.05w, -.05h)$. Thus the rule increases area by 69% in the interior. Clipping makes this increase smaller near image boundaries. It improves potential target coverage at the expense of a larger distractor region, so the benefit depends on the location of the intended referent and the surrounding visual clutter.

Expansion also raises the absolute area required by the later validity threshold: for an interior box, $\gamma A(b') = .0169A(b)$ when $\gamma = .01$. Crop refinement and mask validity therefore interact, particularly for small targets. The numerical transformation and its calibration scope are analyzed in Appendix G.

3.3 Dual-Route Segmentation and Fusion

The refined crop supplies complementary point-prompt and text-prompt routes.

Route A: point-prompt segmentation (visual keypoints). This route mines visual cues to guide the segmenter. We employ a vision–language matcher $\mathcal{C}$ (specifically CLIP Surgery (Li et al., 2023)) to compute a similarity map $\Phi = \mathcal{C}(I_{b'}, p)$ between the crop and the prompt p. We select at most k spatially distinct heatmap maxima by NMS:

$$\mathcal{K} = \text{TopK-NMS}(\Phi;\, k, \tau), \tag{4}$$

where τ filters low-confidence responses and k caps the number of points. Unless stated otherwise, we use a fixed setting of $k = 5$ and $\tau = 0.3$ for all experiments. We then feed $\mathcal{K}$ to the segmenter $\mathcal{S}$ to obtain the point-prompted mask:

$$\hat{M}^{\text{pt}}_{b'} = \mathcal{S}(I_{b'}, \mathcal{K}). \tag{5}$$

If $\mathcal{K} = \emptyset$ (an edge case empirically observed in $< 5\%$ of our cases), we treat Route A as invalid in the fusion stage. This route focuses on salient object parts, with extracted keypoints determining which local regions supply the segmentation prompts.

Route B: text-prompt segmentation. In parallel, this route utilizes the semantic capability of $\mathcal{S}$ directly. We feed the phrase p as a text prompt:

$$\hat{M}^{\text{txt}}_{b'} = \mathcal{S}(I_{b'}, p). \tag{6}$$

This route captures global context, but a crop containing nearby objects can also lead the segmenter to include visually similar instances outside the intended referent.

Consensus-driven fusion. Both crop-level masks ($\hat{M}^{\text{pt}}_{b'}, \hat{M}^{\text{txt}}_{b'}$) are mapped back to the original image coordinates to obtain $\hat{M}^{\text{pt}}$ and $\hat{M}^{\text{txt}}$. Concretely, we resize crop-level binary masks to the RoI resolution (nearest-neighbor) and paste them back to the original canvas using the RoI offsets.

We prefer a valid intersection. Intersection can remove false positives and true pixels alike; invalid agreement invokes the fallback rule below.

Let the validity indicator $\mathcal{V}(M)$ be defined as:

$$\mathcal{V}(M) = \mathbb{I}\left(\frac{A(M)}{A(b')} \geq \gamma\right), \tag{7}$$

where $A(\cdot)$ denotes the pixel area and $A(b')$ is the RoI area. We set $\gamma = 0.01$ as a small area-ratio threshold to filter degenerate masks while tolerating small objects. For the point-prompt route, we additionally set $\mathcal{V}(\hat{M}^{\text{pt}}) = 0$ when $\mathcal{K} = \emptyset$ (i.e., no valid keypoints are found).

For compact notation, let $v_{\text{pt}} = \mathcal{V}(\hat{M}^{\text{pt}})$, $v_{\text{txt}} = \mathcal{V}(\hat{M}^{\text{txt}})$, and $v_{\cap} = \mathcal{V}(\hat{M}^{\text{pt}} \cap \hat{M}^{\text{txt}})$. The final prediction $\hat{M}$ is derived as:

$$\hat{M} = \begin{cases} \hat{M}^{\text{pt}} \cap \hat{M}^{\text{txt}}, & v_{\text{pt}} \wedge v_{\text{txt}} \wedge v_{\cap}, \\ \hat{M}^{\text{pt}}, & v_{\text{pt}} \wedge (\neg v_{\text{txt}} \vee \neg v_{\cap}), \\ \hat{M}^{\text{txt}}, & \neg v_{\text{pt}} \wedge v_{\text{txt}}, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{8}$$

The first case prioritizes precision through intersection; the next two cases fall back to Route A or Route B, respectively. The rule makes fallback precedence explicit: when both routes are valid but the intersection is invalid, Route A is selected. The area test is a non-degeneracy heuristic, not a confidence estimate; small true targets below 1% of the crop can be rejected.

Image Sources and Diversity. To reflect real-world diversity, we aggregate imagery from public datasets (LoveDA (Wang et al., 2021), Potsdam (Rottensteiner et al., 2012), NWPU-VHR-10 (Cheng et al., 2014), DIOR (Li et al., 2020)) and supplementary internet sources. These sources vary in scene layout and imaging conditions; combining them broadens source coverage without making the resulting collection a representative sample of all remote-sensing scenes.

Manual Annotation and Quality Assurance. We curate 810 image–query–mask triplets (I, q, M). Queries q are open-ended instructions that avoid fixed taxonomies; while common nouns may appear, queries require interpreting attributes, relations, or intents beyond simple class lookups. Ground-truth masks M supply pixel-level annotations for evaluating referent selection and segmentation boundaries.

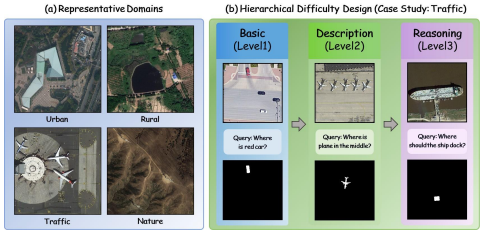


Figure 4: **Overview of GeoSeg-R810.** (a) Representative Domains: We showcase samples from the four domains defined in our scenario taxonomy: *Urban*, *Rural*, *Traffic*, and *Nature*. (b) Hierarchical Difficulty Design: Using the *Traffic* domain, we illustrate the progression across three levels: *Basic* (Level 1), *Description* (Level 2), and *Reasoning* (Level 3), requiring deeper reasoning.

Coordinate and agreement conventions. Both route masks are resized to the RoI with nearest-neighbor interpolation and placed at its original-canvas offset, so that their intersection concerns the same pixel locations. The shared phrase, crop, and segmenter make agreement a mask-selection rule. Intersection emphasizes common foreground support, while the ordered fallbacks retain a valid route mask when the overlap is too small. The coordinate convention keeps both operations consistent with the full-image evaluation.

The crop sets an upper limit on recoverable support. Let G denote a nonempty reference foreground and let B' denote the pixel set inside the refined region. Because both routes are restricted to this crop and pasted into an otherwise empty canvas, the final prediction satisfies $\hat{M} \subseteq B'$. It follows directly that

$$\mathrm{Recall}(\hat{M}, G) = \frac{|\hat{M} \cap G|}{|G|} \leq \frac{|B' \cap G|}{|G|}. \tag{9}$$

This support constraint identifies the role of refinement: enlarging the crop can retain more target pixels, while the routes determine which available pixels are selected. Grounding, geometry, and fusion therefore address referent selection, accessible support, and foreground selection, respectively. The component ablations evaluate how these complementary responsibilities affect the final segmentation.

4 GeoSeg-R810: Benchmark Construction

To rigorously evaluate *reasoning-driven segmentation* in remote sensing, we introduce **GeoSeg-R810**, pairing overhead imagery with open-ended natural language queries and pixel-accurate masks. As a diagnostic testbed, it assesses grounding capabilities from explicit visual descriptions to implicit functional intents. Figures 4 and 5 illustrate query diversity and difficulty distributions; comprehensive metadata formats and reasoning taxonomies are further detailed in Appendix A.

4.1 Data Collection and Annotation Protocol

Figure 4 connects the image, query, and target mask for representative tasks. Annotation must resolve the referent implied by the query, not merely select every visible instance of a category. This distinction is central when spatial relations or intended functions determine the particular region that the prediction must recover from the image.

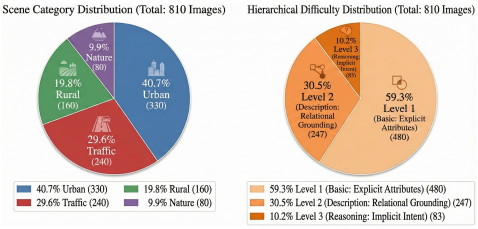


Figure 5: **Distribution statistics of GeoSeg-R810.** Left: Proportional breakdown of four scenarios (Urban, Traffic, Rural, Nature). Right: Observed composition of three difficulty levels (Basic, Description, Reasoning).

GeoSeg-R810 categorizes scenes into four domains (Fig. 5, Left), reflecting diverse remote sensing challenges: **Urban (330 pairs):** Dense built environments (commercial, residential, industrial) with high object density and complex shadows. **Rural (160 pairs):** Agricultural scenes (farmland, irrigation, greenhouses) presenting texture homogeneity and seasonal variations. **Traffic (240 pairs):** Transportation networks (highways, bridges, airports) featuring extreme aspect ratios and connectivity reasoning. **Nature (80 pairs):** Landscapes (water bodies, forests, terrain) with irregular boundaries and large-scale variations.

Table 1: **Quantitative comparison on GeoSeg-R810 and the SegEarth-R2 training dataset.** We benchmark GeoSeg against representative methods from three categories: (1) *Generalist segmentation*, (2) *Reasoning segmentation*, and (3) *Open-source MLLMs*. We report seven pixel-level metrics (%) and inference speed (FPS). Best and second-best results are highlighted in bold and underlined, respectively. Methods that fail to segment any target (IoU=0) are excluded from Accuracy and Specificity rankings. Inference speed is evaluated on GeoSeg-R810 with a single A100 GPU.

Method	GeoSeg-R810 (↑)								SegEarth-R2 training dataset (↑)						
	IoU	Dice	Acc.	Prec.	Rec.	Spec.	BF	FPS	IoU	Dice	Acc.	Prec.	Rec.	Spec.	BF
Generalist segmentation															
SAM3 Carion et al. (2025)	19.9	26.8	89.3	21.7	58.6	91.8	12.9	-	10.5	14.3	77.0	11.2	47.3	78.0	10.9
Grounded SAM Ren et al. (2024a)	18.6	25.1	64.2	20.0	73.1	63.9	10.9	-	11.3	15.1	61.9	13.1	55.2	61.8	11.1
CLIP Surgery + SAM3 Li et al. (2023)	24.7	31.4	73.9	26.8	69.8	74.2	10.4	-	14.5	19.0	69.0	15.5	55.7	69.3	10.1
Reasoning segmentation															
GLAMM Rasheed et al. (2024)	0.0	0.0	94.6	0.0	0.0	100.0	0.0	-	0.0	0.0	95.4	0.0	0.0	100.0	0.0
GSVA Xia et al. (2024)	9.0	12.5	78.3	9.9	31.8	81.2	2.9	-	7.5	10.9	79.9	8.9	32.8	82.3	3.7
LISA++ Yang et al. (2023)	12.4	14.4	95.2	14.0	16.8	**99.2**	5.1	-	3.7	4.4	95.3	4.3	6.1	99.4	2.65
PixelLM Ren et al. (2024b)	0.0	0.0	94.6	0.0	0.0	100.0	0.0	0.175	2.4	3.1	**95.4**	3.5	4.1	**99.8**	1.3
Open-source MLLMs															
Qwen-Edit Wu et al. (2025a)	8.7	13.9	49.6	10.0	68.7	48.4	2.4	0.011	5.8	9.7	57.3	7.2	58.1	57.2	3.9
OmniGen2 Wu et al. (2025b)	6.8	10.8	43.3	7.2	67.4	42.1	2.7	0.011	5.8	9.3	34.8	6.5	74.2	33.4	3.6
BAGEL Deng et al. (2025)	10.7	16.2	57.2	11.3	74.7	56.3	3.5	0.001	8.8	13.4	43.8	9.2	85.4	42.0	5.4
UniWorld Lin et al. (2025)	8.8	14.0	44.0	9.2	78.3	41.6	2.4	0.015	6.4	10.3	47.2	7.1	70.0	45.7	3.8
Hunyuan Cao et al. (2025)	5.6	9.7	10.8	5.6	**99.2**	5.7	1.0	0.002	4.3	7.6	12.0	4.3	**98.1**	8.2	1.7
Ours															
GeoSeg	**56.4**	**64.2**	**96.8**	**69.0**	65.7	98.3	**26.6**	**0.221**	**17.4**	**22.0**	94.7	**29.2**	25.2	97.0	**13.2**

4.2 Scenario Taxonomy and Dataset Composition

The domain and difficulty counts in Figure 5 are separate marginal summaries. A traffic image can support a direct reference, spatial disambiguation, or functional inference. These axes support targeted diagnostics; the aggregate evaluation uses the supplied mixture.

4.3 Hierarchical Difficulty Design

The benchmark contains 480 L1, 247 L2, and 83 L3 pairs (59.3%, 30.5%, and 10.2%; Figure 5). The three levels organize queries by their reasoning requirements. Aggregate results reflect this composition, with explicit-attribute queries contributing the largest share; level labels also retain the distinction between recognition, relational description, and implicit intent.

Level 1 (Basic: Explicit Attributes). Focuses on direct visual recognition using explicit appearance cues (color, texture, shape), sometimes paired with common nouns (e.g., *"where is the blue lake?"*). Evaluates foundational vision-language alignment.

Level 2 (Description: Spatial & Relational). Emphasizes spatial relations and layout-aware reasoning to disambiguate targets within dense scenes. Uses directional predicates like *next to* or *between* (e.g., *"buildings arranged in rows next to the park"*), testing geometric context understanding.

Level 3 (Reasoning: Implicit Intent & Semantics). The hardest tier targets objects identified through implicit semantic associations. It requires functional reasoning (e.g., *"Where can I seek medical help?"* → hospitals), state inference (e.g., *"crops ready for harvest"* → mature fields), or causal logic (e.g., *"Where does the river flow into?"* → river mouth). This level tests whether a model can connect visual evidence with external knowledge and geographic context.

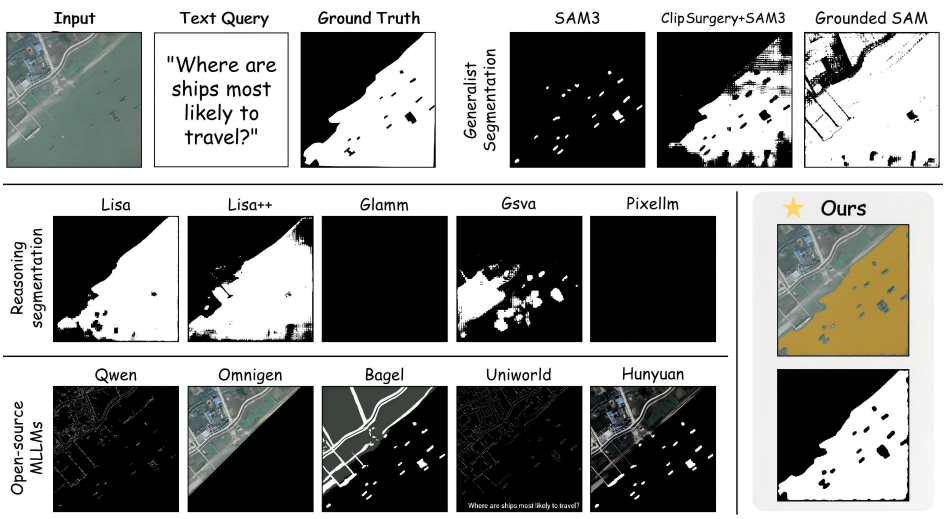


Figure 6: **Qualitative comparison with baselines.** We compare GeoSeg with baselines from generalist segmentation, reasoning segmentation, and open-source MLLMs. The examples show that baselines either miss the queried target or include substantial background noise, whereas GeoSeg produces masks that better match the intended regions. More examples are in Appendix C.

Table 2: **Model and human evaluation on GeoSeg-R810 and SegEarth-R2.** We assess segmentation quality via MLLM judges (Qwen3-VL-8B/32B, used as complementary automated judges) and human evaluators (User). Predictions are rated 1–5 across *Faithfulness* (F.; constraint compliance), *Localization* (L.; boundary precision), and *Robustness* (R.; evaluator-specific rubric). Best and second-best are bold and underlined.

Method	GeoSeg-R810 (↑)									SegEarth-R2 training dataset (↑)								
	Qwen-8B			Qwen-32B			User			Qwen-8B			Qwen-32B			User		
	F.	L.	R.	F.	L.	R.	F.	L.	R.	F.	L.	R.	F.	L.	R.	F.	L.	R.
Generalist segmentation																		
SAM3 Carion et al. (2025)	1.78	1.72	1.81	2.03	1.74	2.00	2.75	2.30	2.68	1.39	1.36	1.40	1.50	1.37	1.46	2.00	1.87	1.99
Grounded SAM Ren et al. (2024a)	1.71	1.65	1.75	1.92	1.64	1.86	2.63	2.17	2.50	1.42	1.39	1.44	1.55	1.41	1.50	2.05	1.88	2.03
CLIP Surgery + SAM3 Li et al. (2023)	1.98	1.91	2.02	2.15	1.88	2.05	2.86	2.51	2.78	1.57	1.52	1.61	1.74	1.54	1.66	2.35	2.05	2.25
Reasoning segmentation																		
GLAMM Rasheed et al. (2024)	1.00	1.00	1.00	1.00	1.00	1.00	1.38	1.30	1.38	1.00	1.00	1.00	1.01	1.01	1.01	1.38	1.35	1.33
GSVA Xia et al. (2024)	1.32	1.30	1.35	1.42	1.30	1.39	1.96	1.74	1.84	1.25	1.23	1.28	1.37	1.25	1.33	1.81	1.72	1.81
LISA++ Yang et al. (2023)	1.57	1.54	1.58	1.62	1.49	1.53	2.22	2.03	2.07	1.16	1.15	1.16	1.18	1.14	1.16	1.64	1.53	1.57
PixelLM Ren et al. (2024b)	1.04	1.04	1.04	1.05	1.03	1.04	1.45	1.40	1.44	1.10	1.09	1.11	1.13	1.09	1.11	1.53	1.49	1.45
Open-source MLLMs																		
Qwen-Edit Wu et al. (2025a)	1.21	1.20	1.22	1.32	1.21	1.26	1.75	1.61	1.66	1.12	1.11	1.13	1.13	1.11	1.14	1.50	1.46	1.52
OmniGen2 Wu et al. (2025b)	1.15	1.14	1.17	1.21	1.15	1.22	1.65	1.54	1.63	1.14	1.13	1.16	1.22	1.14	1.19	1.62	1.52	1.65
BAGEL Deng et al. (2025)	1.31	1.29	1.34	1.43	1.31	1.42	2.20	1.95	2.15	1.26	1.23	1.28	1.40	1.25	1.33	2.10	1.85	2.00
UniWorld Lin et al. (2025)	1.22	1.21	1.24	1.35	1.23	1.32	1.84	1.67	1.75	1.15	1.14	1.17	1.26	1.16	1.21	1.72	1.53	1.62
Hunyuan Cao et al. (2025)	1.06	1.06	1.06	1.10	1.07	1.11	1.53	1.46	1.50	1.04	1.04	1.05	1.10	1.05	1.08	1.50	1.45	1.49
Ours																		
GeoSeg	**3.64**	**3.48**	**3.66**	**3.72**	**3.28**	**3.28**	**4.35**	**4.12**	**4.20**	**1.78**	**1.71**	**1.81**	**1.94**	**1.69**	**1.82**	**2.65**	**2.80**	**2.50**

5 Experiment

5.1 Experimental Settings

We evaluate GeoSeg-R810 and 10,000 LaSeRS training cases used by SegEarth-R2 (Appendix F), with fixed weights for GeoSeg and 12 baselines. The off-benchmark geometric calibration is described in Appendix G. We report IoU, Dice, accuracy, precision, recall, specificity, and Boundary F-score; metric definitions and interpretation are in Appendix H. Qwen3-VL-8B/32B and human raters provide complementary 1–5 faithfulness, localization, and robustness ratings (Appendix J). Component controls remove box refinement or either mask route. Results average seeds 0, 42, and 100 (Appendix G).

5.2 Main Results

Segmentation and ratings. GeoSeg achieves 56.4% IoU and 64.2% Dice on GeoSeg-R810, compared with 24.7% IoU for CLIP Surgery+SAM3 and 12.4% for LISA++ (Table 1). On the external subset it reaches 17.4% IoU, a 2.9-point advantage over CLIP Surgery+SAM3, with substantial remaining error. On GeoSeg-R810, GeoSeg reaches 69.0% precision and 26.6% Boundary F-score,

Table 3: **Ablation of components on GeoSeg-R810 and SegEarth-R2.** Evaluating Box Refinement and the Dual-Route parallel paths (Route A: Point-Prompt; Route B: Text-Prompt). The complete configuration has the highest reported IoU; other metrics show trade-offs across component removals. Best and second-best are **bold** and underlined.

Components			GeoSeg-R810 (↑)							SegEarth-R2 training dataset (↑)						
Box Refine	Route A	Route B	IoU	Dice	Acc.	Prec.	Rec.	Spec.	BF	IoU	Dice	Acc.	Prec.	Rec.	Spec.	BF
✗	✓	✓	51.1	59.9	96.5	62.7	64.3	97.8	21.7	13.4	17.7	93.9	22.9	21.1	96.2	9.7
✓	✗	✓	52.9	61.4	95.8	59.6	**71.5**	96.6	24.3	16.0	20.6	93.1	23.6	**27.9**	95.0	11.6
✓	✓	✗	43.2	49.0	96.5	52.4	48.9	**99.0**	21.2	15.2	19.1	**95.1**	22.7	22.6	**97.6**	11.4
✓	✓	✓	**56.4**	**64.2**	**96.8**	**69.0**	65.7	98.3	**26.6**	**17.4**	**22.0**	94.7	**29.2**	25.2	97.0	**13.2**

Removing box refinement reduces GeoSeg-R810 IoU by 5.3 points (56.4 to 51.1); removing Route A reduces it by 3.5 points (to 52.9); removing Route B reduces it by 13.2 points (to 43.2). Adding Route B increases recall from 48.9 to 65.7 with a smaller decrease in specificity from 99.0 to 98.3, showing its contribution to foreground coverage. On LaSeRS, removing refinement, Route A, or Route B lowers IoU by 4.0, 1.4, and 2.2 points, respectively. Refinement has the largest external removal effect, whereas Route B has the largest effect on GeoSeg-R810. Together, the controls characterize the complementary roles of crop coverage and prompting within the full pipeline (Appendix I).

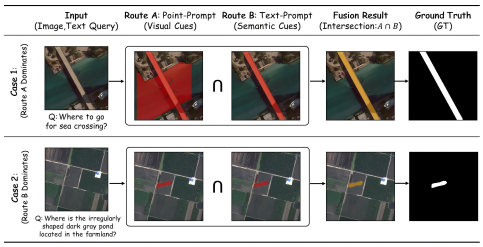


Figure 7: **Selected component ablations.** Columns remove refinement or point/text routes. Further examples appear in Appendix E.

compared with 26.8% and 10.4% for CLIP Surgery+SAM3. Its recall is lower, at 65.7% versus 69.8%, indicating a more selective foreground prediction. On LaSeRS, precision remains higher at 29.2% versus 15.5%, but recall falls to 25.2%. Thus stronger target selectivity accompanies limited external coverage.

Model and human assessment. Table 2 supplies the complete model and human ratings: curated-set human scores are 4.35/4.12/4.20, while all external-set GeoSeg ratings remain below 3. Human localization rises from the strongest baseline's 2.51 to 4.12 on GeoSeg-R810 and from 2.05 to 2.80 externally. These ratings complement pixel overlap by assessing query-conditioned masks; the evaluator-specific robustness rubrics remain separate. Figure 6 illustrates target omissions and excess foreground in selected baseline outputs. The external results delimit transfer.

5.3 Component Ablations

The complete configuration has the highest IoU in Table 3. Figure 7 visualizes foreground changes after removing refinement or either prompting route.

6 Conclusion

GeoSeg connects reasoning-driven target selection to remote-sensing masks through language grounding, calibrated region expansion, and complementary point/text segmentation routes. The framework separates referent selection, spatial coverage, and pixel support, using an intersection-first rule with explicit fallbacks while keeping pretrained weights fixed. GeoSeg-R810 complements the method with 810 image–query–mask triplets across four scene domains and three query levels. The full configuration attains 56.4% IoU on this resource and 17.4% on the external LaSeRS subset, with component ablations characterizing the roles of geometric refinement and dual-route prompting. These results support the proposed composition and provide a structured resource for reasoning-oriented remote-sensing evaluation.

Limitations and future work. Calibration is model-specific, with lower mask accuracy on the external subset. Future work will broaden calibration and improve coverage of small or disconnected targets.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and consistency checking. The authors take responsibility for the final manuscript, including its mathematical derivations, implementation descriptions, experimental values, citations, and claims.

ETHICS STATEMENT

Remote-sensing segmentation can support mapping and environmental analysis, but localization errors and biased language interpretation can also affect communities or sensitive locations. GeoSeg's masks are model predictions, not verified geographic facts, and should receive independent review before consequential use. Source imagery retains its applicable access and redistribution conditions. Human mask ratings should be reported in aggregate with participant privacy protected. The benchmark assesses the stated imagery and query distribution; it does not establish suitability for surveillance or other sensitive deployments.

REFERENCES

Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary remote sensing image semantic segmentation. *arXiv preprint arXiv:2409.07683*, 2024.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

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.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsmamba: Remote sensing image classification with state space model. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024a.

Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024b.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pp. 323–340, 2024c.

Gong Cheng, Junwei Han, Peicheng Zhou, and Lei Guo. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:119–132, 2014.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

Zongyan Han, Jiale Cao, Shuo Chen, Tong Wang, Jorma Laaksonen, and Rao Muhammad Anwer. Openseg-r: Improving open-vocabulary segmentation via step-by-step visual reasoning. *arXiv preprint arXiv:2505.16974*, 2025.

Shuyan Ke, Yifan Mei, Changli Wu, Yonghan Zheng, Jiayi Ji, Liujuan Cao, and Rongrong Ji. Pixdlm: A dual-path multimodal language model for uav reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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*, pp. 4015–4026, 2023.

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*, pp. 9579–9589, 2024.

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.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022.

Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv e-prints*, pp. arXiv–2304, 2023.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7061–7070, 2023.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

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.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *International Conference on Learning Representations*, 2026.

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

Ailong Ma, Junjue Wang, Yanfei Zhong, and Zhuo Zheng. Factseg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.

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*, pp. 34865–34876, 2026.

Jerome Quenum, Wen-Han Hsieh, Tsung-Han Wu, Ritwik Gupta, Trevor Darrell, and David M. Chan. Lisat: Language-instructed segmentation assistant for satellite imagery. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18082–18091, 2022.

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*, pp. 13009–13018, 2024.

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.

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

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*, pp. 26374–26383, 2024b.

Franz Rottensteiner, Gunho Sohn, Jurgen Jung, Markus Gerke, and Christian Heipke. The ISPRS 2d semantic labeling challenge. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1:293–298, 2012.

Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36:8815–8827, 2023.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. Lasagna: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

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*, pp. 3858–3869, 2024.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18134–18144, 2022.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2955–2966, 2023.

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

An Yang, 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.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023.

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

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2024b. doi: 10.1109/TGRS.2024.3449154.

A DETAILS OF GEOSEG-R810

To facilitate future research and ensure reproducibility, we provide comprehensive details regarding the construction, hierarchical taxonomy, and data format of GeoSeg-R810.

A.1 DATASET DIRECTORY STRUCTURE

The benchmark is organized hierarchically by scenario and difficulty level. The file structure is as follows:

Directory Hierarchy

```
RS_ReasonSeg_Benchmark/
+-- images/                  # Original remote sensing images
|   +-- {Urban, Rural, Traffic, Nature}/
|   |   +-- Level1/          # Explicit Reference
|   |   +-- Level2/          # Relational Description
|   |   +-- Level3/          # Implicit Reasoning
+-- masks/                   # Binary masks; mirrored hierarchy
+-- annotations/             # Meta-information in JSON format
    +-- {Urban, Rural, Traffic, Nature}.json
```

A.2 METADATA FORMAT

Each scenario contains a JSON annotation file. We adopt a standardized schema to ensure ease of parsing for zero-shot evaluation:

JSON Annotation Example

```
{
    "id": "traffic_005",
    "file_name": "Traffic/Level1/traffic_l1_005.jpg",
    "mask_path": "Traffic/Level1/traffic_l1_005.png",
    "height": 1024,
    "width": 1024,
    "category": "Traffic",
    "difficulty": "Level1",
    "question": "The␣red␣car␣parked␣near␣the␣intersection."
}
```

A.3 SCENARIO TAXONOMY AND REASONING DIFFICULTY

Section A.4 details the diverse reasoning challenges embedded in GeoSeg-R810. We define three difficulty levels: **Level 1 (Basic)** tests basic alignment; **Level 2 (Description)** tests spatial context; and **Level 3 (Reasoning)** tests external knowledge and causal logic.

A.4 DOMAIN-SPECIFIC QUERY EXAMPLES

Urban L1 "The large stadium with a white roof.": Color & geometric attribute alignment.

Urban L2 "The residential buildings arranged in rows next to the park.": Spatial proximity and layout-aware disambiguation.

Urban L3 "Where can I seek medical help in an emergency?": Functional reasoning: Mapping intent to "Hospital" via visual cues (e.g., Red Cross).

Rural L1 "The circular irrigation pivot.": Domain-specific terminology and shape recognition.

Rural L2 "Farmland located between the forest and the river.": Complex relative positioning (Between A and B).

Rural L3 "Show me the crops that are ready for harvest.": State inference: Golden color $\rightarrow$ ripeness $\rightarrow$ harvest readiness.

Traffic L1 "The passenger aircraft.": Fine-grained classification (Commercial vs. Cargo).

Traffic L2 "The vehicle right next to the red car.": Spatial reference and distractor suppression.

Traffic L3 "Which road section is likely to cause congestion?": Causal logic: Identifying bottlenecks or construction zones.

Nature L1 "The blue lake.": Basic semantic segmentation of water bodies.

Nature L2 "Bare soil area resulting from deforestation.": Contextual texture discrimination.

Nature L3 "Areas with high risk of landslides.": Knowledge-based reasoning: Evaluating slope and lack of vegetation.

B MLLM-AS-A-JUDGE PROMPT TEMPLATE AND GRADING GUIDELINES

To complement standard pixel-level metrics and evaluate the semantic alignment of our segmentation outputs, we employ an MLLM-as-a-judge protocol (powered by Qwen3-VL). Standard overlap metrics can sometimes penalize predictions that are functionally correct but morphologically varied. To address this, we cast the MLLM as an expert in remote sensing image analysis.

The model is provided with a composite image containing the original background, the ground truth mask (rendered in green), the predicted mask (rendered in red), and the overlapping regions (rendered in yellow). It is then instructed to evaluate the segmentation quality across three distinct dimensions—Faithfulness, Localization, and Robustness—using a 1-to-5 Likert scale.

The exact prompt template and detailed scoring rubrics fed to the model are structured as follows:

System Prompt & Scoring Rubric

You are an expert in remote sensing image analysis. Evaluate whether the predicted mask correctly segments the target region specified by the query in this remote sensing image.
The image shows:

- Original image (background)
- Green overlay: Ground truth mask (correct segmentation)
- Red overlay: Predicted mask (model's segmentation)
- Yellow areas: Overlapping regions (where both masks agree)

Image dimensions: [Width] x [Height] pixels
Query: [Query]
Evaluate the segmentation from THREE aspects and provide integer scores from 1 to 5 for each metric.
Use the following scoring standards:
1. Faithfulness: Does the predicted mask correctly follow the query and identify the intended target region?

- **5 (Excellent):** Perfectly correct identification, no confusion with other classes.
- **4 (Good):** Mostly correct, minor confusion with similar classes.
- **3 (Fair):** Generally correct but some confusion with related classes.
- **2 (Poor):** Significant confusion, partially wrong class identification.
- **1 (Very Poor):** Completely wrong class, major misidentification.

2. Localization: Does the predicted mask precisely follow the complex edges?

- **5 (Excellent):** Boundaries perfectly match ground truth, no rounded corners or overflow.
- **4 (Good):** Boundaries mostly accurate, minor deviations at complex edges.

- **3 (Fair):** Generally follows boundaries but noticeable deviations or slight overflow.
- **2 (Poor):** Significant boundary misalignment, obvious rounded corners or overflow.
- **1 (Very Poor):** Severe boundary errors, completely misaligned edges.

3. Robustness: Can the segmentation resist interference from clouds, shadows, seasonal changes, or similar textures?

- **5 (Excellent):** Highly robust, unaffected by environmental variations or similar textures.
- **4 (Good):** Mostly robust, minor sensitivity to environmental factors.
- **3 (Fair):** Moderate robustness, some sensitivity to clouds/shadows/similar textures.
- **2 (Poor):** Low robustness, easily confused by environmental variations.
- **1 (Very Poor):** Very fragile, severely affected by clouds, shadows, or similar textures.

Respond with ONLY a valid JSON object in this format (use integer scores 1-5):
`{"faithfulness": <1-5>, "localization": <1-5>, "robustness": <1-5>}`

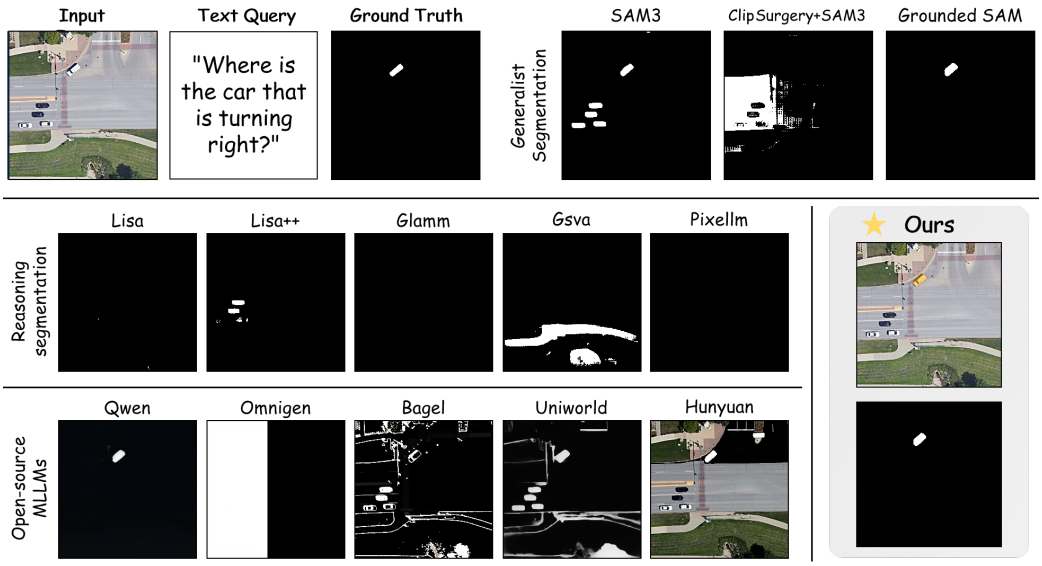
C Additional Qualitative Comparisons with Baselines

In this section, we provide six additional qualitative examples to further analyze the behavior of our proposed approach compared to existing baselines. Consistent with the evaluation in the main text, the comparisons span three major categories: generalist segmentation, reasoning segmentation, and open-source MLLMs.

To comprehensively evaluate the models, these six selected cases encompass varying degrees of query complexity, visual ambiguity, and challenging object scales. As illustrated in Figure 8, the baseline models frequently fail to accurately ground the target objects. They tend to exhibit severe over-segmentation or miss the target entirely due to a limited comprehension of complex, implicit instructions. In contrast, the full pipeline gives more faithful masks in the selected examples; these illustrations do not quantify population-level robustness.

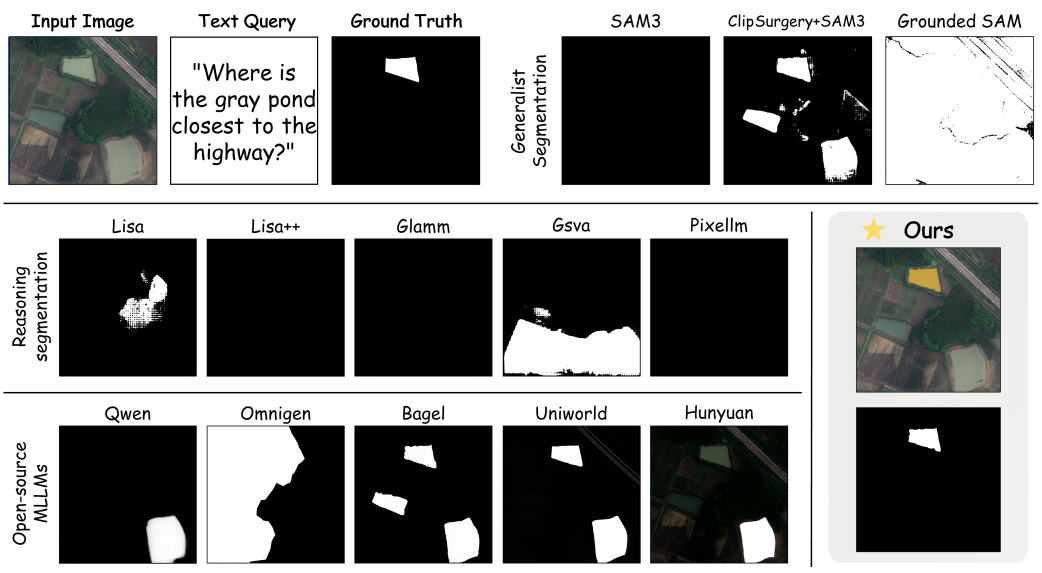


(a) Case 1: Dealing with visual ambiguity.

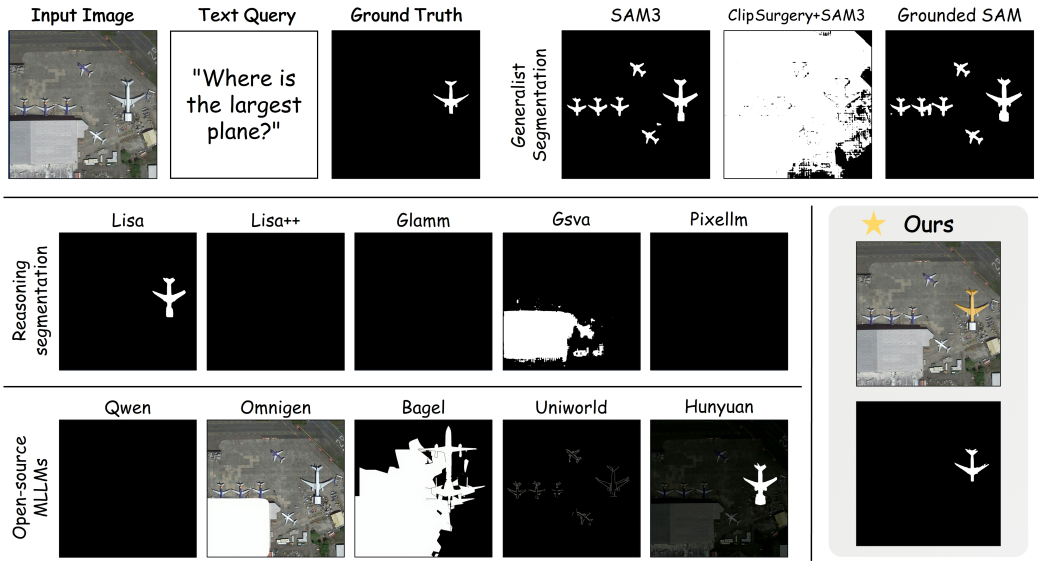


(b) Case 2: Complex spatial reasoning.

Figure 8: **Extended qualitative comparison with multiple baselines.** We present six additional challenging cases. Compared to generalist segmentation, reasoning segmentation, and open-source MLLMs, the full method produces improved masks in these selected examples. (Figure continued on next page.)

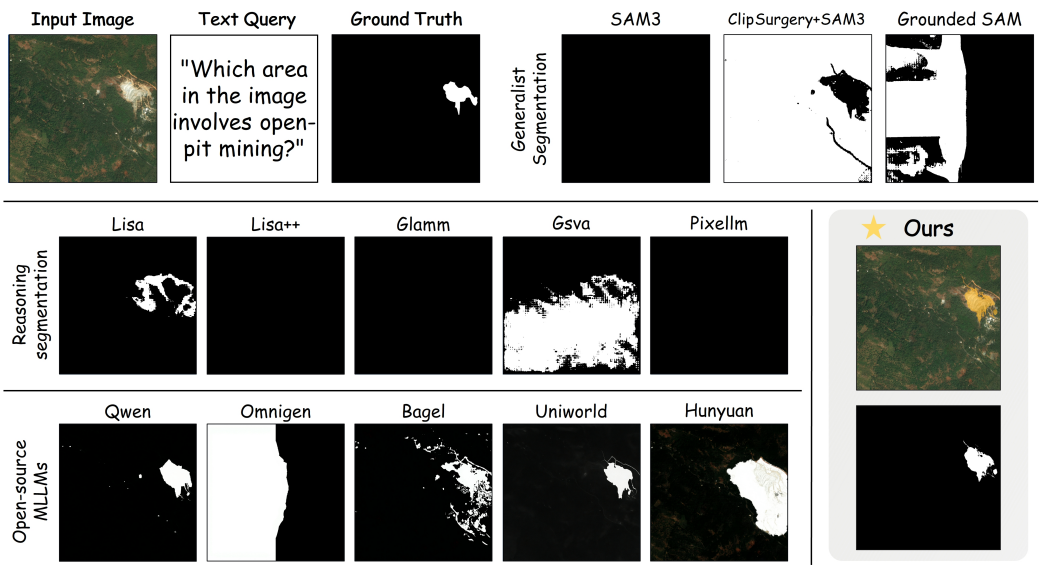


(c) Case 3: Implicit instruction following.

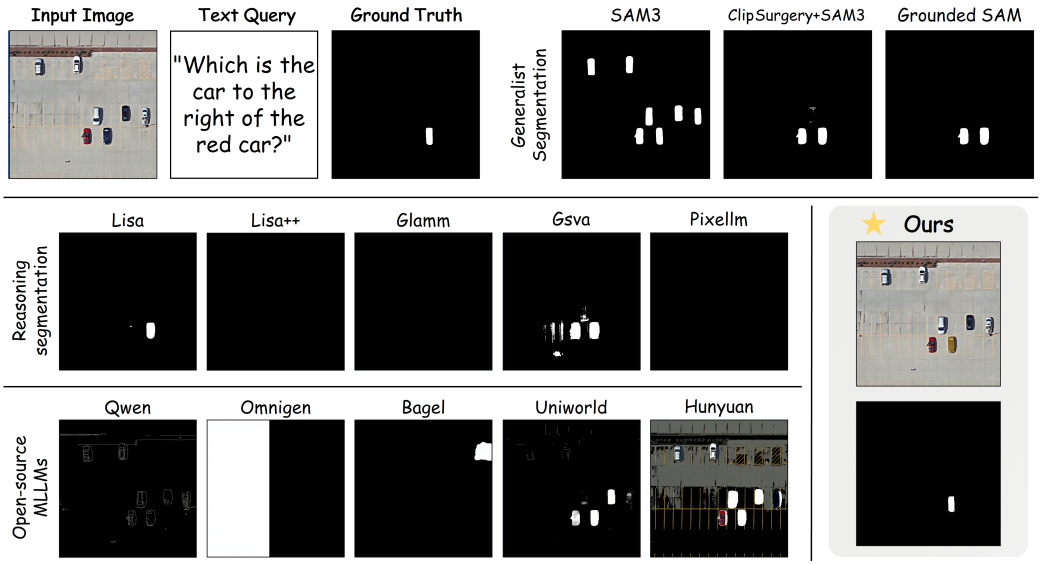


(d) Case 4: Multiple target grounding.

Figure 8: **Extended qualitative comparison with multiple baselines.** (Continued from previous page.)



(e) Case 5: Resolving multi-target spatial relationships.



(f) Case 6: Precise boundary delineation under diverse conditions.

Figure 8: **Extended qualitative comparison with multiple baselines.** (Continued from previous page.)

D Detailed User Study Protocols

To evaluate perceptual quality and alignment with human preferences, we conducted a blind user study with 50 participants, including users with backgrounds in computer vision or remote sensing as well as general users.

Data Preparation and Interface. The study uses 40 selected challenging image–query pairs, including ambiguous queries and same-class distractors. This is a challenge-oriented subset, not an established uniform random sample of the entire benchmark. We developed a custom, web-based blind evaluation interface. For each trial, participants were shown the original image and the text query, followed by the segmentation masks generated by GeoSeg and other baseline models. To reduce presentation bias, the order of the models was strictly anonymized and randomized for every image.

Evaluation Metrics. Participants were instructed to evaluate the predictions using a 1-to-5 Likert scale (where 1 indicates completely unacceptable and 5 indicates perfect prediction) across three carefully defined dimensions:

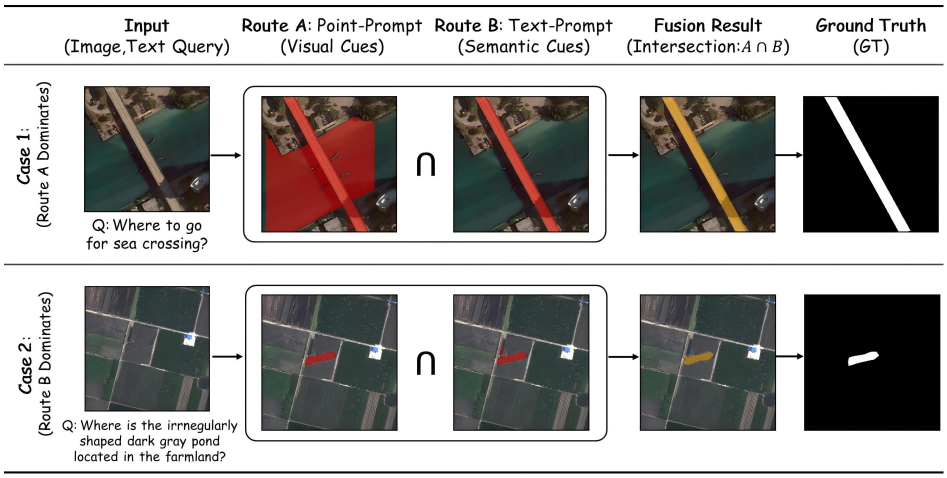
- **Faithfulness:** The degree to which the predicted mask accurately captures the core semantics and intent of the given text query.
- **Localization:** The spatial precision of the segmentation boundaries and how tightly the mask wraps the intended target without bleeding into the background.
- **Robustness:** The model's capability to accurately resolve linguistic ambiguities and strictly ignore surrounding same-class distractors.

Prior to the formal evaluation, all participants completed a brief tutorial with standardized examples to ensure a consistent understanding of the scoring criteria. Qualitative feedback was also collected at the end of the survey; several participants noted that GeoSeg better separated intended targets from same-class distractors in complex scenes.

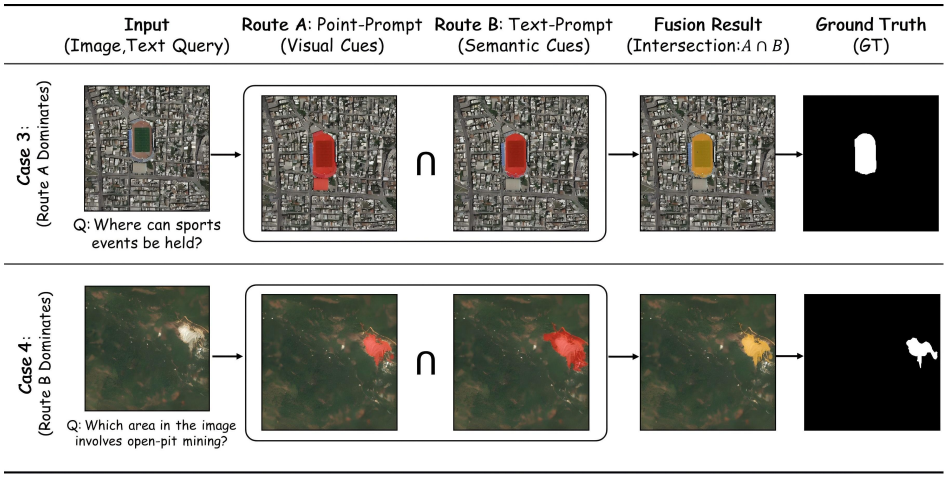
E Additional Qualitative Results for Ablation Study

In this section, we present three additional qualitative examples to further illustrate the critical role of each proposed component, specifically the Bias-Aware Coordinate Refinement (*Box Refine*) and the Dual-Route strategy (*Route A* for Visual Cues and *Route B* for Semantic Cues).

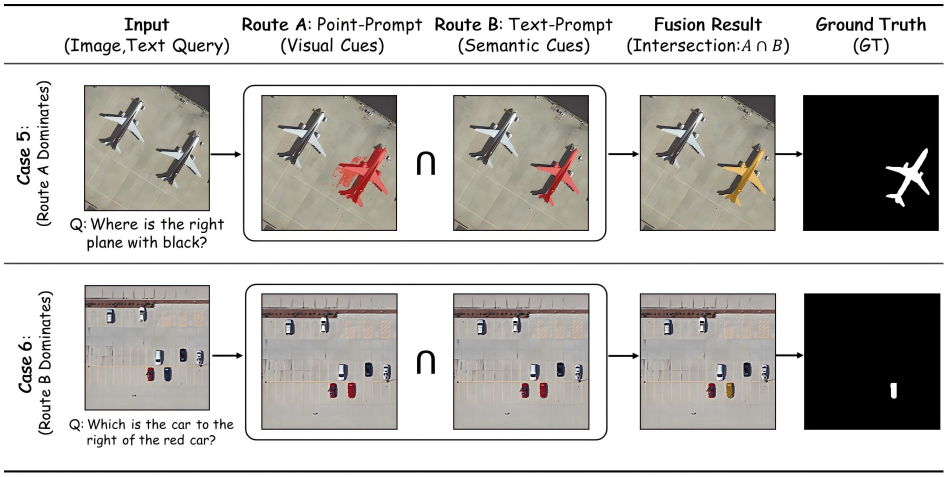
As shown in Figure 9, removing any key module degrades the qualitative results. Without the *Box Refine* module, the predicted masks tend to deviate from the target region. Using only *Route A* or only *Route B* can lead to incomplete masks or incorrect grounding, respectively. The full pipeline combines the two routes with coordinate refinement and produces more accurate masks across the illustrated cases.



(a) Case 1: Impact of missing Visual Cues (*Route A*).



(b) Case 2: Impact of missing Semantic Cues (*Route B*).



(c) Case 3: Inaccurate localization without *Box Refine*.

Figure 9: **Extended qualitative ablation study.** We showcase three additional cases comparing the full pipeline with its variants. Removing *Route A*, *Route B*, or the *Box Refine* module results in noticeable performance degradation, such as semantic misunderstanding or coarse boundaries, illustrating the effects in these selected cases.

F Overview of the Benchmark Dataset

In the evaluation phase of this study, we utilize 10,000 publicly available training instances from the LaSeRS dataset, originally proposed for SegEarth-R2, as an external evaluation subset. This use concerns a public training split, rather than an official held-out test ranking; it does not rule out exposure during upstream pretraining. The data organization, task dimensions, and instance categories of this benchmark are outlined below.

Data Organization. Each instance in this benchmark is structured as a question-answer-mask triple. Alongside precise segmentation masks for pixel-level evaluation, the dataset provides bounding boxes for coarse localization, as well as the corresponding textual instructions and standard textual answers.

Core Task Dimensions. The benchmark systematically covers four critical dimensions of language-guided segmentation in remote sensing imagery:

- **Hierarchical Granularity:** The tasks encompass macroscopic semantic-level queries (e.g., "all tennis courts"), instance-level queries (e.g., "the rightmost tennis court"), and microscopic part-level queries (e.g., "the service area of the tennis court").
- **Target Multiplicity:** The dataset includes both single-target instructions and complex queries demanding the simultaneous grounding of multiple distinct targets.
- **Reasoning Requirements:** Queries range from explicit descriptions of visual attributes to implicit instructions that require the deduction of target regions via geographic commonsense (e.g., inferring a safe refuge during an earthquake).
- **Linguistic Variability:** The textual instructions feature significant linguistic diversity, spanning from concise short queries to highly detailed, long descriptions.

Instance Categories and Scope. Covering a broad spectrum of remote sensing scenarios, the benchmark comprises 122 object categories. The highlighted instances are distributed across three distinct conceptual levels:

- **General Categories:** Common geographical features and facilities, such as airplanes, buildings, bridges, harbors, and various sports fields.
- **Fine-grained Concepts:** Specialized objects differentiated by specific models or functions, including specific passenger jets (e.g., Boeing 737, A350), cargo trucks, container cranes, and engineering ships.
- **Part-level Elements:** Focuses on the precise segmentation of minute details under extreme scale variations, such as airplane engines, the bow or stern of a ship, football nets, and tennis court service boxes.

G Protocol Interpretation and Geometric Analysis

Random seeds and result aggregation. The numerical performance entries in the main-text tables are arithmetic means over runs with seeds 0, 42, and 100, evaluated on the same benchmark pairs. Each seed contributes a complete run-level metric before averaging; dataset sizes and fixed calibration parameters are not averaged. An additional run with seed 1000 is used to check stability and is excluded from the reported mean. Results remain stable across the tested seeds. This repeatability applies to the reported model and calibration settings; cross-domain transfer, calibration sensitivity, and agreement between automated and human evaluation remain distinct properties.

G.1 Training-free scope and calibration

The segmentation and grounding models retain their pretrained weights. The fixed settings are $\alpha = .2$, $\beta = .1$, $k = 5$, activation threshold $\tau = .3$, and validity threshold $\gamma = .01$. Coordinate-error analysis uses reference target annotations to calibrate the geometric correction. Thus, calibration operates on prediction errors while the grounding and segmentation weights remain frozen; the training-free designation refers to this absence of weight updates.

Figure 3 labels horizontal and vertical offsets relative to reference-box width w_{gt} and height h_{gt}, respectively, rather than full-image dimensions. With predicted-minus-reference differences, the normalized errors would be $(x_i - x_i^*)/w_{\mathrm{gt}}$ and $(y_i - y_i^*)/h_{\mathrm{gt}}$. The reference-box normalization in the plot is distinct from the predicted-box widths used by Equation 2. The KDE motivates the chosen margins but does not specify a unique quantile-to-margin estimator. The expansion equation defines the implemented geometric rule.

LoveDA, NWPU-VHR-10, and DIOR contribute to the calibration sources, while the diagnostic set also draws from those datasets. Source disjointness and image disjointness are different. An off-benchmark calibration design must exclude overlapping source images, rescaled copies, crops, and multiple queries on the same asset. The aggregate plot and domain counts cannot certify that exclusion or establish transfer to unseen source domains. Likewise, a mixed collection of public and internet imagery does not establish a unified redistribution license; provenance and rights are properties of each source asset.

G.2 Fusion logic and small-target behavior

The mask-validity test measures area relative to the refined crop. It supplies no semantic guarantee. With valid point and text masks, a valid intersection is used; with an invalid intersection and valid points, the point route wins. If points are unavailable or invalid but text is valid, the text mask is used. With neither valid, the output is empty. This precedence is intentionally asymmetric and should not be described as confidence-based selection.

For a true region occupying less than 1% of an RoI, even a geometrically accurate mask can fail the area test. Expansion may lower that fraction further. Similarly, disconnected targets may be truncated by a single box; consensus cannot recover pixels never presented to either route. These cases identify structural failure modes of the specified algorithm without assigning unmeasured failure frequencies.

H Pixel Metrics and Aggregation

Pixel overlap, foreground recovery, and background rejection are distinct diagnostics. In sparse scenes, high accuracy or specificity can coexist with poor target recovery, so Table 1 retains foreground and boundary columns.

For a prediction P and reference foreground mask G, define $TP = |P \cap G|$, $FP = |P \setminus G|$, $FN = |G \setminus P|$, and TN on their shared image canvas. The usual nonzero-denominator quantities are

$$\mathrm{IoU} = TP/(TP + FP + FN), \qquad \mathrm{Dice} = 2TP/(2TP + FP + FN), \tag{10}$$

$$\mathrm{Precision} = TP/(TP + FP), \qquad \mathrm{Recall} = TP/(TP + FN), \tag{11}$$

$$\mathrm{Specificity} = TN/(TN + FP), \qquad \mathrm{Accuracy} = (TP + TN)/(TP + FP + FN + TN). \tag{12}$$

For a *single* confusion matrix, Dice equals $2\mathrm{IoU}/(1 + \mathrm{IoU})$. The identity need not hold after averaging: applying a nonlinear function to a mean differs from averaging that function. A per-pair macro mean $N^{-1}\sum_i m_i$, a global metric computed from $\sum_i TP_i, \sum_i FP_i, \sum_i FN_i$, and the across-seed mean in Appendix G represent different aggregation levels. IoU and Dice should therefore retain their individual reported values rather than being reconstructed from one another.

Boundary F-score matches predicted and reference boundary pixels within a tolerance and combines boundary precision and recall. Tolerance in absolute pixels differs from tolerance normalized by image size. Since the aggregate table does not specify the matching tolerance, its values do not establish tolerance-invariant boundary quality.

For a nonempty reference and empty prediction, foreground IoU and Dice are zero, while accuracy and specificity may remain high because background dominates. A zero denominator in precision requires an explicit evaluator convention. An empty reference and empty prediction is a separate case; multi-target masks are evaluated as foreground unions only when that matches the annotation semantics. Aggregate scores alone do not identify the empty-mask conventions needed for exact

Table 4: Related resources and the distinct scope of GeoSeg-R810. No comparative performance is inferred.

Resource	Setting	Distinction relevant here
GRES/LISAt	Geospatial language and reasoning segmentation	Supplies task-specific training data and a trained assistant.
GeoSeg-Bench/UniGeoSeg	Multi-task geospatial instruction segmentation	Covers referring, interactive, and reasoning tasks; its benchmark name is distinct from ours.
DRSeg/PixDLM	UAV reasoning segmentation	UAV-focused queries and a learned dual-path representation.
GeoSeg-R810	Frozen-model evaluation on 810 pairs	Compact four-domain, three-level diagnostic set with fixed geometric calibration.

reproduction. In particular, the GLAMM/PixelLM zero-IoU results describe the evaluated outputs and do not distinguish model refusal, parsing failure, empty-mask decoding, or domain mismatch.

I Benchmark Positioning and Evidence Scope

The composition setting. Table 4 distinguishes GeoSeg-R810 from related resources. Its focus is a frozen-model pipeline connecting semantic target selection to pixel-level prediction, with component controls and a diagnostic set organized by scene domain and query difficulty.

The domain counts are Urban 330, Rural 160, Traffic 240, and Nature 80. The difficulty counts are L1 480, L2 247, and L3 83. These sum independently to 810; they are marginal distributions, not a domain-by-level cross-tabulation. Aggregate IoU does not identify any domain or difficulty mean. Selected qualitative successes at L3 therefore cannot establish population-level reasoning robustness.

The 31.7-point curated-set IoU advantage and 2.9-point external advantage are point-estimate differences. The reported component deltas are 5.3, 3.5, and 13.2 points on GeoSeg-R810, and 4.0, 1.4, and 2.2 points on LaSeRS for removing refinement, Route A, and Route B. These differences are computed from Tables 1 and 3; they describe changes within each comparison rather than additive independent component effects.

Attribution of the component gap. The full system exceeds CLIP Surgery+SAM3 by 31.7 IoU points. The no-Route-A and no-refinement variants retain 28.2 and 26.4 points of that gap (89.0% and 83.3%). Thus, the removals reduce the full configuration's score while retaining substantial benefit from the remaining pipeline. Grounding, cropping, phrase generation, and fusion also differ from the baseline; this comparison does not isolate grounding alone. The component controls characterize changes within the pipeline rather than prove a uniquely optimal fusion rule.

J Judge and Human-Study Interpretation

Appendices B and D give the automated rubric and human protocol. The two automated judges share the grounding model family. The 50-participant, 40-case study supplies descriptive evidence without item-level uncertainty estimates. Baseline input/output interfaces and available timings differ across methods; the 0.221 FPS value ranks only configurations with reported timing. The automated rubric uses a static overlay of reference and prediction. Its robustness item concerns visible clouds, shadows, seasonal appearance, and similar textures. A static rating does not measure a controlled perturbation response. The human robustness item instead emphasizes linguistic ambiguity and same-class distractors. The rubric-specific columns are labeled by evaluator; no numerical equivalence or cross-evaluator average is assumed.

Blinding and randomization reduce presentation cues without establishing inter-rater reliability. The participant and case counts do not specify independent ratings per case; reported means cannot recover confidence intervals or agreement coefficients. The shared Qwen family links grounding and judging, so its highest score is not independent evidence of reasoning success.