RS-REF: BENCHMARKING REASONING IN REMOTE SENSING REFERRING EXPRESSION TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring Expression Comprehension (REC) bridges natural language descriptions with specific image regions, serving as a fundamental capability for Vision-Language Models (VLMs). However, existing REC benchmarks in Remote Sensing (RS) frequently rely on simplistic expressions and lack sufficient distractors. This allows models to exploit single-attribute shortcuts rather than performing visual-language reasoning. In this paper, we investigate reasoning-intensive visual grounding in RS imagery, focusing on scenarios that demand multi-constraint logical reasoning amidst abundant same-category distractors. To this end, we introduce **RS-Ref**, a new benchmark specifically designed for multi-step remote sensing referring expression grounding. All images and expressions are strictly curated to ensure the presence of multiple same-category distractors and require compositional descriptions involving multiple constraints. Furthermore, we propose **GeoREC**, a two-stage visual chain-of-thought framework for referring expression. GeoREC introduces a grounded chain-of-thought mechanism to enable multi-step spatial reasoning, and subsequently employs GRPO-based reinforcement learning to optimize these reasoning trajectories. Extensive evaluations of 20 representative VLMs on RS-Ref, alongside validation on seven existing RS visual grounding benchmarks, demonstrate both the rigorous difficulty of RS-Ref and the state-of-the-art performance of our proposed framework.

1 INTRODUCTION

Vision-Language Models (VLMs) Qwen Team (2026); Team et al. (2025a; 2026a); Yang et al. (2025) have significantly advanced remote sensing (RS) interpretation Kuckreja et al. (2024); Liu et al. (2026b;a); Pang et al. (2025). By modeling and fusing heterogeneous modalities, they enable GeoAI systems to better connect multi-source observations with human-centric queries. Within this context, Referring Expression Comprehension (REC) is a core task requiring models to localize target regions based on natural language descriptions, thereby facilitating fine-grained visual grounding in complex scenes. REC supports critical applications such as disaster monitoring Lenton et al. (2024); Misra et al. (2025), urban infrastructure retrieval Xiao et al. (2024); Yuan et al. (2025), and object search Uijlings et al. (2013); Yu & Shin (2026) in large-scale imagery. Reflecting its importance, several recent benchmarks for RS-VLMs, including CHOICE An et al. (2025), GeoBench-VLM Danish et al. (2025), and OmniEarth Fu et al. (2026), have incorporated language-based localization tasks to assess model capabilities. Ref-Adv Dong et al. (2026) evaluates multi-constraint grounding in natural images, while its applicability to remote sensing remains underexplored.

Despite this progress, existing remote sensing REC evaluations are not always sufficient for diagnosing compositional grounding ability. As illustrated in Fig. 1, we identify three recurring issues that weaken their diagnostic value. **(1) Single-attribute shortcuts.** Many queries are highly templated and contain only the target category or a simple attribute, e.g., "plane" or "large ship". In such cases, the task can degenerate into category-level object detection rather than language-conditioned grounding. **(2) Insufficient same-category distractors.** A referred object is often the only instance of its category in the image. When no visually plausible same-category alternatives exist, additional modifiers become weakly discriminative, because the target can be recovered from category priors alone. **(3) Limited multi-constraint reasoning.** Many expressions contain coarse location hints but do not require combining multiple spatial or contextual constraints to identify a unique instance.

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

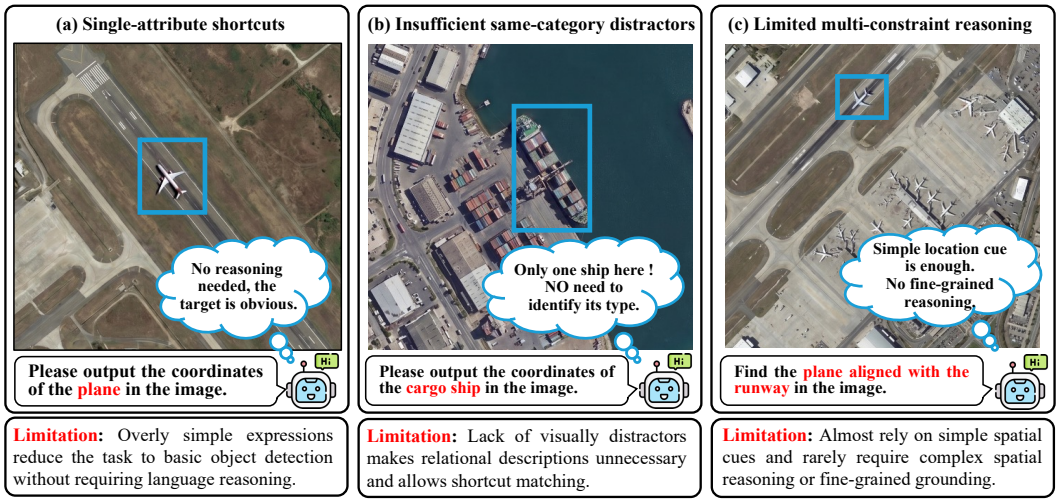


Figure 1: Limitations of Existing Remote Sensing Referring Expression Benchmarks.

This makes it difficult to determine whether a model has truly aligned the expression with the visual scene or has instead relied on shallow category and position biases.

To probe these issues quantitatively, we conduct two controlled perturbation experiments on existing language-based localization tasks. **Category replacement.** On CHOICE, we replace the target category with the generic token "object" while keeping the remaining expression unchanged, e.g., replacing *ship* with *object*. This perturbation removes the explicit category cue and tests whether models can still localize targets from residual cues and dataset biases. As shown in Table 1, InternVL3.5-8B and Qwen3.5-9B still retain Acc@0.5 scores of 34.3% and 41.3%, respectively. Although performance drops compared with the original expressions, the retained accuracy suggests that explicit category semantics are not always necessary for successful localization on this benchmark. **Relation removal.** On GeoBench-VLM, we remove spatial relation phrases while preserving the target category, e.g., deleting phrases such as "at the top of ...". This perturbation tests whether relational descriptions are essential for disambiguating the referred instance. As shown in Table 2, removing relation phrases reduces the Acc@0.5 of InternVL3.5-8B from 23.6% to 12.8%, but unexpectedly improves Qwen3.5-9B from 38.6% to 41.3%. This mixed behavior indicates that relation phrases in existing benchmarks are not consistently decisive for target identification; for some models, they may even introduce noise rather than provide necessary discriminative information.

Table 1: Effect of Replacing Target Categories with a Generic Token in CHOICE.

Model	Setting	Acc@0.5	Acc@0.75	Acc@0.9
InternVL3.5-8B	Original	51.7	13.3	1.3
	Object	34.3	7.0	0.7
Qwen3.5-9B	Original	71.0	34.5	3.2
	Object	41.3	20.0	2.2

Table 2: Effect of Removing Relational Descriptions in GeoBench-VLM.

Model	Setting	Acc@0.5	Acc@0.75	Acc@0.9
InternVL3.5-8B	Original	23.6	4.9	0.5
	w/o Relations	12.8	4.0	0.2
Qwen3.5-9B	Original	38.6	15.2	2.6
	w/o Relations	41.3	20.0	2.2

These observations motivate the need for a benchmark that makes shortcut solutions less effective. To this end, we construct **RS-Ref**, a new benchmark for reasoning-intensive remote sensing referring expression grounding. RS-Ref is designed around three principles. First, each expression is semantically rich and requires compositional interpretation rather than single-category detection. Second, each image contains multiple same-category candidates, creating a challenging distractor environment in which category priors alone are insufficient. Third, each expression includes at least two spatial or contextual constraints, requiring the model to combine multiple pieces of evidence to identify the target. By construction, RS-Ref evaluates whether a model can perform instance-level discrimination under dense visual ambiguity. Table 3 compares RS-Ref with existing datasets, and Table 4 provides qualitative examples illustrating the increased grounding difficulty.

Beyond benchmarking, we further study how to improve REC models under dense same-category distractors. In RS imagery, objects often share similar appearance, scale, and orientation, making direct final-box prediction prone to localization confusion. We therefore argue that models should

explicitly reason over candidate instances and progressively resolve ambiguity according to the referring expression. We propose **GeoREC**, a two-stage grounded visual reasoning framework for remote sensing REC. GeoREC first learns reasoning trajectories that interleave textual deductions with intermediate bounding-box predictions, allowing the model to decompose expressions, verify constraints, and refine localization. It then optimizes these trajectories with Group Relative Policy Optimization (GRPO), encouraging accurate localization and step-wise spatial consistency. Experiments show that GeoREC performs robustly in dense scenes and transfers to existing RS visual grounding benchmarks.

Our contributions are threefold. First, we introduce **RS-Ref**, a distractor-controlled benchmark with 1,560 curated image-expression pairs, same-category distractors, and multi-constraint descriptions. Second, we propose **GeoREC**, a two-stage framework combining grounded reasoning trajectories with GRPO-based optimization. Third, we evaluate 20 representative VLMs on RS-Ref and validate GeoREC on seven existing RS visual grounding datasets, demonstrating both the challenge of RS-Ref and the generalizability of GeoREC.

Table 3: Statistical comparison of RS-Ref benchmarks. Avg Dis denotes the average number of same-category distractors per image, Avg O/I Ratio denotes the average object-to-image area ratio.

Dataset	#Img	#Expr	#Cat	Avg Len	Vocab	#Avg Dis	Avg Img Res	Avg O/I Ratio	Expr Type
DIOR-RSVG Zhan et al. (2023)	922	2180	20	7.6	76	1.3	800×800	5.61%	Free
CHOICE An et al. (2025)	600	600	12	1.2	19	0.0	800×800	6.38%	Template
RSHRBench Dang et al. (2025)	145	200	31	2.0	53	0.2	7552×6870	0.29%	Free
GeoBench-VLM Danish et al. (2025)	439	1313	33	9.2	233	0.6	807×807	2.01%	Free
VRSbench Li et al. (2024)	9318	16159	26	13.9	1718	0.8	1024×1024	7.35%	Free
RS-Ref (Ours)	1560	1560	48	17.8	1098	5.1	812 ×774	1.98%	Free

2 Related Works

Remote Sensing Large Vision–Language Models. The rapid development of vision–language models has advanced multimodal understanding in RS. Building on general multimodal models such as InternVL-3.5 Wang et al. (2025b), GLM-4.1V Team et al. (2026b), and Qwen3-VL Bai et al. (2025), a series of remote sensing vision–language models (RS-VLMs) have been proposed, including GeoChat Kuckreja et al. (2024), EarthGPT Zhang et al. (2024), EarthDial Soni et al. (2025), and SkyMoE Liu et al. (2026a). Meanwhile, works further enhance RS-VLMs with chain-of-thought reasoning Liu et al. (2026b); Fiaz et al. (2025); Wang et al. (2026); Zhang et al. (2025) and large-image perception strategies such as region selection, visual pruning, and progressive zooming Wang et al.; Liu et al. (2025a); Li et al. (2026). Among these capabilities, visual referring expression comprehension is a fundamental task for RS-VLMs and an important indicator of their ability to perform fine-grained target localization and reasoning. Although qualitative results in remote sensing have been encouraging, rigorous quantitative evaluation remains essential for systematically assessing and comparing the reasoning and localization capabilities of different models.

Remote Sensing Referring Expression Benchmarks. In remote sensing, referring expression grounding is evaluated through both dedicated datasets and broader RS-VLM benchmarks. For instance, DIOR-RSVG Zhan et al. (2023); Li et al. (2020) serves as a foundational dataset specifically designed for object localization from natural language. Beyond dedicated datasets, comprehensive benchmarks, such as VRSBench Li et al. (2024), GeoBench-VLM Danish et al. (2025), CHOICE An et al. (2025), and OmniEarth Fu et al. (2026), integrate visual grounding as a key sub-task to assess multi-level perception and reasoning. Recently, RSHR-Bench Dang et al. (2025) further extended this evaluation to super-high-resolution large-scene settings. However, as detailed in Table 3, these prior datasets predominantly feature simplistic expressions, a lack of same-category distractors, and low semantic complexity. These limitations significantly undermine their effectiveness in rigorously evaluating the multi-step reasoning capabilities of VLMs in complex grounding scenarios.

3 The RS-Ref Benchmark

3.1 Data Source

We construct RS-Ref from publicly available RS object detection datasets, including DIOR Li et al. (2020), DOTA Xia et al. (2018), HRSC2016 Liu et al. (2025b), FAIR1M Sun et al. (2022), and

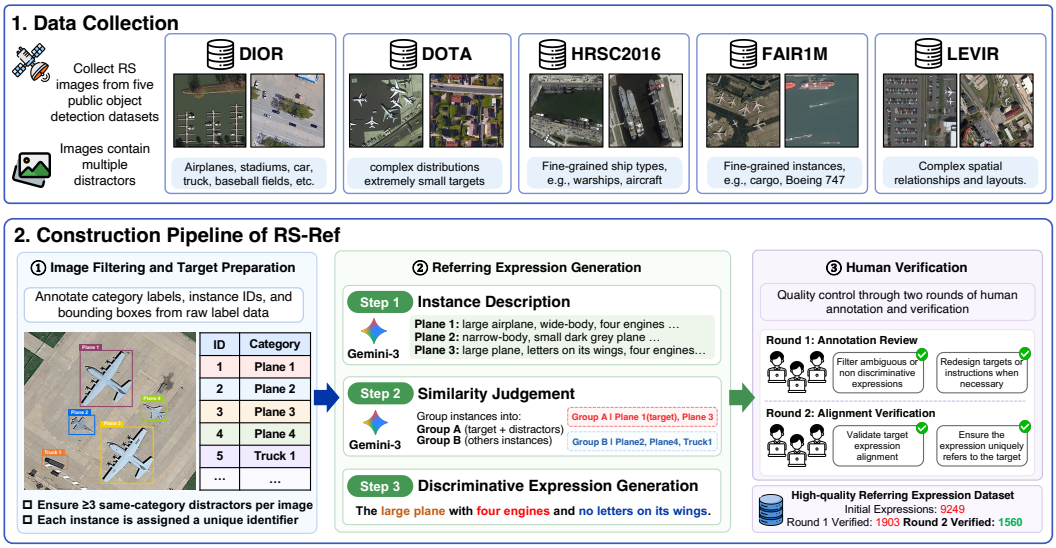


Figure 2: Overview of the RS-Ref construction pipeline, including data collection, target preparation, expression generation, and human verification.

LEVIR Zou & Shi (2017). Given an image with instance-level bounding-box annotations and category labels, our pipeline generates target-expression pairs for referring expression grounding. To avoid overlap with existing RS referring-expression benchmarks and reduce template-induced bias, we exclude images from datasets such as DIOR-RSVG. We also preserve fine-grained category labels when available, such as motorboats, cargo ships, and liquid cargo ships, which increases category ambiguity and reduces reliance on coarse category cues.

3.2 Construction Pipeline of RS-Ref

Image Filtering and Target Preparation. We first filter images according to their instance annotations. To avoid category-only grounding, we retain images containing at least three instances from the same category, which serve as candidate targets and same-category distractors. We also remove overly crowded images whose object counts exceed a preset threshold, since such scenes often lead to ambiguous or unnatural expressions. For each retained image, we visualize bounding boxes and category labels, and assign temporary instance identifiers, e.g., plane 1, plane 2, and plane 3, to distinguish candidates during construction. These identifiers are used only during annotation and are removed from the final expressions.

Annotation Rules. We define annotation rules to ensure that each expression is clear, discriminative, and uniquely grounded. Each expression must: (1) refer to exactly one target instance; (2) distinguish the target from same-category candidates using multiple visual or spatial cues; (3) contain 2–3 descriptive constraints, such as category, appearance, relative location, spatial relation, or nearby context; and (4) avoid shortcut descriptions based on a single salient attribute, such as the largest object, the leftmost building, or the only red object. These rules encourage expressions that progressively narrow the candidate set until only the target instance remains.

Referring Expression Generation. We use Gemini-3 DeepMind (2025) to generate candidate referring expressions through a three-step pipeline. First, given an image with annotated bounding boxes and category labels, the model describes each candidate instance and extracts preliminary visual cues, including category, appearance, shape, orientation, relative location, and local context. Second, we group visually similar same-category instances as hard candidates: one instance is selected as the target, while the remaining similar instances serve as hard distractors; other annotated objects are used as contextual references when they provide useful spatial or semantic cues. Third, the model identifies discriminative attributes at two levels: separating the target from contextual references and distinguishing it from hard same-category distractors. Based on these attributes and the annotation rules, Gemini-3 drafts multi-constraint expressions that are expected to uniquely localize the target instance. This LLM-assisted stage only produces candidate target-expression pairs, whose final validity is determined by human verification.

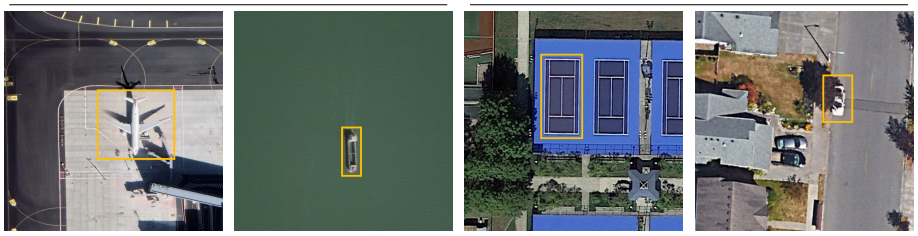
	RS-Ref (Ours)			
Image				
Expression	The airplane on the left side of the terminal that is not connected to the jet bridge.	The overpass crossing the railway on the right side of the image.	The medium-sized yacht moving in the opposite direction to the cargo ship.	The merchant ship with a reddish deck that is not floating on the water.
	CHOICE An et al. (2025)		**VRSBench Li et al. (2024)**	
Image				
Expression	Please output the coordinates of the airplane in the image.	Please output the coordinates of the ship in the image.	The tennis court with a larger area located on the left side of the image.	The vehicle located near the top of the image.

Table 4: Examples from different referring expression datasets. The top row shows examples from our proposed **RS-Ref**. The bottom rows present examples from **CHOICE** and **VRSBench**.

Human Verification. To ensure annotation quality, we conduct two rounds of human verification with 10 trained graduate annotators and additional reviewers with remote sensing experience. The two rounds are performed independently by different annotators to reduce subjective bias. During verification, annotators perform three types of interventions: (1) directly removing expressions with low descriptive quality or incorrect grounding; (2) revising expressions whose target is correct but whose wording is ambiguous or insufficiently discriminative; and (3) manually redesigning the target-expression pair when the selected target is suboptimal and another instance in the image supports a clearer description.

Starting from 9,249 LLM-generated candidates, the first round retains 1,903 expressions after filtering and revision. In the second round, two independent annotators further assess each remaining candidate according to grounding accuracy, uniqueness, and target disambiguation. Among these candidates, 1,560 expressions are consistently judged as valid, 273 are judged as invalid, and 70 receive inconsistent labels and are excluded. The second-round annotations yield a Cohen's Kappa of 0.86. The final RS-Ref benchmark contains 1,560 verified expressions, corresponding to a final keep rate of 16.9%.

4 Baseline Framework

Accurate target localization in RS imagery is challenging because the target is often surrounded by visually similar distractors, while the referring expression may require fine-grained spatial and contextual reasoning. To address this challenge, we propose GeoREC, a two-stage framework for RS referring expression comprehension. GeoREC introduces Grounded Chain-of-Thought (G-CoT), which represents localization as an ordered trajectory of interleaved textual reasoning and bounding-box predictions. The first stage uses supervised fine-tuning to align reasoning steps with visual grounding evidence. The second stage applies Group Relative Policy Optimization (GRPO) to encourage accurate final localization and step-wise spatial consistency, improving grounding in complex RS scenes.

4.1 GROUNDED CHAIN-OF-THOUGHT LEARNING

We formulate G-CoT as an ordered trajectory $\mathcal{Z} = ((r_t, B_t))_{t=1}^{T}$, where r_t denotes the textual reasoning state and B_t denotes the corresponding grounding prediction at step t. Unlike conventional text-only CoT, G-CoT explicitly couples each reasoning step with a spatial prediction, enabling coarse-to-fine localization toward the final target B_T. Following this formulation, we construct a training set, **G-CoT-Instruct**, comprising 35,967 trajectories with images at a resolution of 512×512. Construction details are in the Appendix B.

To train the model with this structured reasoning format, we perform SFT on trajectory-formatted data. Each sample contains an image I, a query Q, and a target sequence Y, where Y is obtained by linearizing the trajectory $\mathcal{Z}$ into:

$$Y = \langle\text{think}\rangle(r_1, B_1), \ldots, (r_T, B_T)\langle/\text{think}\rangle\langle\text{answer}\rangle B_T\langle/\text{answer}\rangle.$$

The sequence is tokenized as $Y = (y_1, \ldots, y_{|Y|})$, and the model is optimized with the standard maximum likelihood objective:

$$\mathcal{L}_{SFT}(\theta) = -\mathbb{E}_{(I,Q,Y)\sim\mathcal{D}_{SFT}} \sum_{i=1}^{|Y|} \log P_\theta(y_i \mid I, Q, y_{<i}). \tag{1}$$

This stage provides a structured initialization that teaches the model to associate intermediate reasoning steps with corresponding spatial predictions.

4.2 TRAJECTORY REFINEMENT VIA GROUP RELATIVE POLICY OPTIMIZATION

Although SFT provides structured supervision, it optimizes only the likelihood of reference trajectories and does not explicitly compare alternative reasoning paths. To further improve grounding quality, we apply GRPO with on-policy sampling. For each input (I, Q), the model samples a group of candidate trajectories, and trajectories with better grounding accuracy and smoother spatial refinement receive higher rewards. We define an outcome reward based on the final prediction: $R_{out} = \text{IoU}(B_T, B^*)$, where B^* is the ground-truth box. To encourage coarse-to-fine refinement, we also introduce a trajectory reward. Let $\Delta_t = \text{IoU}(B_t, B^*) - \text{IoU}(B_{t-1}, B^*)$. Then,

$$R_{traj} = \frac{1}{T-1} \sum_{t=2}^{T} \left[\max(0, \Delta_t) - \lambda \max(0, -\Delta_t)\right], \tag{2}$$

which rewards step-wise localization improvement and penalizes degraded intermediate predictions. The overall reward is

$$R = R_{out} + \alpha R_{traj}.$$

The GRPO training set $\mathcal{D}_{GRPO}$ is constructed by incorporating DIOR-RSVG with G-CoT-Instruct to improve data diversity, resulting in 70,711 samples. During this stage, supervision is provided solely through bounding box annotations, without any intermediate reasoning. We ensure that no RS-Ref images, expressions, annotations, or filtered candidate samples are used during training. For each input (I, Q), we sample G trajectories $\{Y_1, \ldots, Y_G\}$ from the old policy $\pi_{\theta_{old}}$ and compute their rewards. The rewards are normalized within each group to obtain group-relative advantages. The policy is optimized with the clipped GRPO objective:

$$\begin{aligned}
\mathcal{L}_{GRPO}(\theta) = -\mathbb{E}_{\left[(I,Q)\sim\mathcal{D}_{GRPO},\{Y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|I,Q)\right]} \\
\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|Y_i|}\sum_{t=1}^{|Y_i|}\min\left(r_{t,i}(\theta)\hat{A}_i, \text{clip}(r_{t,i}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\right) - \beta D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}), \\
r_{t,i}(\theta) = \frac{\pi_\theta(y_{i,t}|I,Q,y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|I,Q,y_{i,<t})}, \hat{A}_i = \frac{\mathcal{R}_i - \text{mean}(\mathcal{R})}{\text{std}(\mathcal{R})},
\end{aligned} \tag{3}$$

where π_{ref} is the reference policy initialized from the SFT checkpoint. This stage encourages the model to prefer trajectories with more accurate final boxes and more consistent intermediate refinements, leading to more reliable grounding under dense distractors.

Table 5: Performance comparison on RS-Ref. We report Acc@0.5, Acc@0.75, Acc@0.9, and mAcc. We also group samples by the number of same-category distractors. Δ denotes the difference between the group-wise Acc@0.5 and the overall Acc@0.5 of the same model.

Model	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	3–4	Δ	5–6	Δ	≥7	Δ
Qwen3-VL-4B-Instruct Yang et al. (2025)	7.6	1.7	0.4	3.2	9.0	+1.4	6.7	-0.9	5.7	-1.9
Qwen3-VL-8B-Instruct Yang et al. (2025)	8.6	2.1	0.3	3.7	10.3	+1.7	7.5	-1.1	7.0	-1.5
Qwen3-VL-32B-Instruct Yang et al. (2025)	9.1	2.3	0.4	3.9	10.8	+1.7	8.2	-0.9	7.0	-2.1
Qwen3.5-4B Qwen Team (2026)	12.9	3.5	0.6	5.7	17.2	+4.2	9.9	-3.0	11.5	-1.5
Qwen3.5-27B Qwen Team (2026)	15.6	4.7	1.3	7.2	18.3	+2.7	14.3	-1.3	12.8	-2.9
Qwen3.5-35B-A3B Qwen Team (2026)	15.1	4.5	1.1	6.9	18.0	+2.9	13.2	-2.0	13.7	-1.5
GLM-4.6V-Flash-10B Team et al. (2026b)	10.1	2.9	0.6	4.5	11.3	+1.2	10.2	+0.2	7.0	-3.0
Gemma-4-E4B-it Google (2026)	1.1	0.1	0.0	0.4	1.6	+0.5	0.9	-0.2	0.0	-1.1
Gemma-4-26B-A4B-it Google (2026)	2.7	0.4	0.0	1.0	3.9	+1.2	1.9	-0.8	1.3	-1.4
Gemma-4-31B-it Google (2026)	3.6	0.5	0.0	1.4	5.4	+1.8	2.6	-1.0	1.8	-1.8
Kimi-VL-A3B-Instruct Team et al. (2025b)	19.4	8.1	1.3	9.6	19.6	+0.2	19.6	+0.2	18.5	-0.9
InternVL3-8B Zhu et al. (2025)	4.9	1.8	0.6	2.4	5.1	+0.2	4.5	-0.3	5.7	+0.9
InternVL3.5-8B Wang et al. (2025a)	28.0	10.0	1.7	13.2	28.0	0.0	<u>29.1</u>	+1.1	26.0	-2.0
VHM-7B Pang et al. (2025)	1.0	0.2	0.0	0.4	2.1	+1.1	0.4	-0.6	0.0	-1.0
GeoChat Kuckreja et al. (2024)	4.3	0.3	0.1	1.6	5.7	+1.4	3.4	-0.9	1.8	-2.5
SkySenseGPT-7B Luo et al. (2024)	4.0	0.4	0.0	1.5	4.6	+0.6	3.9	0.0	1.3	-2.7
EarthDial Soni et al. (2025)	14.6	3.3	0.3	6.1	19.0	+4.4	11.7	-2.9	10.1	-4.4
RSThinker Liu et al. (2026b)	<u>29.4</u>	<u>22.5</u>	<u>10.3</u>	<u>20.7</u>	<u>29.8</u>	+0.4	28.5	-0.9	**27.3**	-2.0
GeoSolver Sun et al. (2026)	26.4	19.8	8.8	18.3	28.2	+1.7	24.0	-2.4	<u>26.9</u>	+0.5
GeoREC (Ours)	**37.0**	**28.8**	**11.8**	**25.9**	**38.6**	+1.6	**37.7**	+0.7	<u>26.9</u>	-10.1

5 Evaluation Results

5.1 Experimental Setup and Metrics

We evaluate 20 representative vision–language models on RS-Ref, including open-source general-domain models, and RS-VLMs. In addition, we evaluate our proposed baseline method on seven widely used RS visual grounding benchmarks. All models are evaluated in a zero-shot setting using their default generation configurations. To accommodate different output conventions for visual grounding, we design model-specific prompts following official guidelines. Since models output bounding boxes in varying coordinate formats (e.g., $[0, 1000]$ for GLM-4.6V vs. $[0, 1]$ for Kimi-VL), we parse and convert all predictions into a unified coordinate system. This ensures fair comparison and precise metric calculation across all baselines. Experiments are run on two NVIDIA H200 GPU.

We follow the standard evaluation protocol used in referring expression comprehension Liu et al. (2026b;a; 2024). A grounding prediction is considered correct if the IoU between the predicted bounding box and the ground-truth box exceeds a predefined threshold. The commonly used metric Acc@0.5 reports the proportion of predictions with IoU greater than 0.5. To further assess localization precision under stricter criteria, we additionally report Acc@0.75 and Acc@0.9. Moreover, we compute the mean accuracy (mAcc) by averaging accuracy across multiple IoU thresholds.

5.2 Main Results

We first evaluate 20 representative VLMs on the proposed RS-Ref. All models are evaluated using their officially recommended configurations and pretrained weights. We also analyze model performance under different scene complexities by varying the number of distractor instances in the image. Furthermore, we evaluate the proposed baseline on additional RS referring expression datasets to further verify its effectiveness. It is worth noting that the proposed GeoREC baseline does not use any raw data involved in the RS-Ref construction process during training, including the preliminary generated samples that were filtered out during dataset curation, in order to avoid potential data leakage.

Results on RS-Ref. As shown in Table 5, reasoning-based methods such as RSThinker and GeoSolver achieve competitive performance, with mAcc of 20.7 and 18.3, respectively, suggesting that explicit reasoning benefits the understanding of complex referring expressions. GeoREC achieves the best performance, reaching 37.0 Acc@0.5, 28.8 Acc@0.75, 11.8 Acc@0.9, and 25.9 mAcc. Nevertheless, RS-Ref remains highly challenging for current VLMs, as even the best-performing model attains only 11.8 Acc@0.9 under strict criteria, indicating that precise multi-constraint grounding under same-category distractors settings remains an open challenge.

Table 6: Comparison with RS-VLMs on external visual grounding benchmarks. We report Acc@0.5. Bold and underline denote the best and second-best results among the evaluated methods, respectively.

Method	VRSBench	DIOR-RSVG	RRSIS-D	RSVG	CHOICE	GEOBench	OmniEarth
VHM-7B Pang et al. (2025)	33.9	55.9	64.0	2.5	27.5	1.5	51.1
GeoChat Danish et al. (2025)	56.3	31.4	10.0	5.5	26.0	7.5	31.1
SkySenseGPT-7B Luo et al. (2024)	63.5	60.8	69.0	2.0	10.5	6.0	17.2
EarthDial Soni et al. (2025)	14.4	46.1	72.5	2.0	<u>32.0</u>	5.0	48.9
RSThinker Liu et al. (2026b)	<u>75.7</u>	73.4	**87.5**	<u>6.0</u>	19.5	<u>8.5</u>	**81.6**
GeoSolver Sun et al. (2026)	74.0	<u>73.8</u>	82.5	<u>6.0</u>	16.0	8.0	79.4
GeoREC (Ours)	**76.3**	**87.8**	<u>85.0</u>	**22.5**	**63.0**	**13.5**	<u>81.0</u>

Among general-domain VLMs, InternVL3.5-8B and Kimi-VL-A3B-Instruct achieve relatively strong performance, with 28.0 and 19.4 Acc@0.5, respectively. This suggests that recent general-domain VLMs can provide competitive grounding ability even without remote-sensing-specific training. However, their performance still lags behind GeoREC, particularly under strict IoU thresholds. We also observe that scaling improves performance within some model families, such as Qwen3-VL from 4B to 32B, but the trend is not universal across all families, indicating that grounding ability depends on both model scale and visual-language alignment.

Results under Different Distractors. We further analyze performance under different levels of same-category ambiguity. As shown in Table 5, many models show lower accuracy when the number of same-category distractors increases, confirming that dense sets of visually similar objects make REC substantially harder. GeoREC performs best in the 3–4 and 5–6 distractor groups, achieving 38.6 and 37.7 Acc@0.5, respectively. However, its performance drops to 26.9 when the number of distractors reaches ≥ 7. This indicates that although GeoREC improves grounding under moderate distractor density, resolving very dense same-category ambiguity remains challenging. These results also support the diagnostic value of RS-Ref, since performance is strongly affected by the degree of instance-level ambiguity.

Results on Additional REC Benchmarks. To evaluate generalization beyond RS-Ref, we compare GeoREC with existing RS-VLMs on seven external RS visual grounding benchmarks(Table 6). GeoREC achieves the best performance on five datasets: VRSBench, DIOR-RSVG, RSVG, CHOICE, and GeoBench-VLM, with particularly large gains of 14.0, 16.5, and 31.0 points over previous best results on DIOR-RSVG, RSVG, and CHOICE, respectively. On RRSIS-D and OmniEarth, GeoREC ranks second, with only a small gap to the best method. These results suggest that the proposed trajectory-based grounding framework can generalize effectively across diverse benchmarks. Nevertheless, the relatively low absolute accuracy on RSVG and GeoBench-VLM indicates that complex RS grounding remains challenging and leaves substantial room for improvement.

5.3 Benchmark Robustness Analysis

To examine whether RS-Ref reduces shortcut-based grounding, we conduct three diagnostic perturbation tests. We use the same probing model for all compared datasets and measure the relative change in Acc@0.5 before and after perturbation. Since different perturbations may be applicable to different subsets of samples, we report the relative change within each test. **(1) Bias Test.** We replace the original expression with a category-agnostic prompt, such as "the object", while keeping the image unchanged. This tests whether a model can localize the target using dataset bias, object saliency, or positional priors without relying on category semantics. RS-Ref shows the largest performance drop, from 28.0 to 11.9, corresponding to a 57.5% relative decrease. In comparison, CHOICE and GeoBench-VLM show smaller drops of 33.7% and 45.8%, respectively. **(2) Text-Structure Sensitivity Test.** We shuffle the word order of each expression while preserving the same words. This perturbation weakens compositional structure while keeping individual keywords unchanged. RS-Ref again exhibits a much larger decline, from 28.0 to 13.7, corresponding to a 51.1% decrease. By contrast, GeoBench-VLM and DIOR-VG drop by only 16.5% and 17.6%, respectively. **(3) Descriptor-Deletion Test.** We remove non-category descriptors, including visual attributes, spatial relations, and contextual cues, while preserving the target category. This tests whether additional constraints are necessary for target disambiguation. RS-Ref drops from 28.0 to 21.6, corresponding to a 22.9% decrease, which is larger than the drops observed on GeoBench-VLM and DIOR-VG. Overall, these perturbation results suggest that RS-Ref is less solvable by shortcut cues. Successful

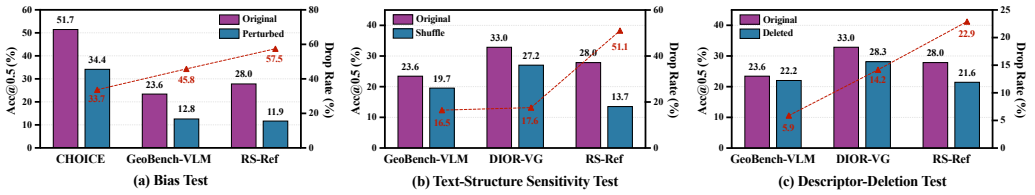


Figure 3: Acc@0.5 degradation under diagnostic perturbations. Larger drops indicate that the benchmark depends more on complete referring expressions and is less solvable by shortcut cues.

localization on RS-Ref depends more strongly on the complete referring expression, including word order, contextual descriptors, and spatial constraints.

5.4 Ablation Study

To analyze the contribution of each component in GeoREC, we conduct an ablation study on RS-Ref. As shown in Table 7, the base model achieves 33.6 Acc@0.5, 16.9 Acc@0.75, and 4.7 Acc@0.9. After supervised fine-tuning with grounded reasoning trajectories, performance improves to 34.8 Acc@0.5, 26.2 Acc@0.75, and 10.1 Acc@0.9. The larger gains under stricter IoU thresholds suggest that trajectory supervision mainly improves localization precision rather than only coarse target selection. Adding GRPO without the trajectory reward further improves Acc@0.5 to 36.0 and Acc@0.9 to 11.0, but leaves Acc@0.75 unchanged. The full GeoREC model, which includes the trajectory reward, achieves the best results across all thresholds, reaching 37.0 Acc@0.5, 28.8 Acc@0.75, and 11.8 Acc@0.9. Compared with the base model, GeoREC improves the three metrics by 3.4, 11.9, and 7.1 percentage points, respectively. These results indicate that both stages are useful: supervised trajectory learning provides a strong initialization for grounded reasoning, while trajectory-aware GRPO further improves spatial consistency and final localization accuracy.

Table 7: Ablation of training stages and trajectory reward on RS-Ref.

Model	Acc@0.5	Acc@0.75	Acc@0.9
Base (GLM-4.1V-9B-Base)	33.6	16.9	4.7
+ SFT	34.8	<u>26.2</u>	10.1
+ SFT + GRPO (w/o R_{traj})	<u>36.0</u>	<u>26.2</u>	<u>11.0</u>
GeoREC (Ours)	**37.0**	**28.8**	**11.8**

6 Conclusion

In this work, we study reasoning-intensive referring expression comprehension in remote sensing imagery. We introduce RS-Ref, a distractor-controlled benchmark designed to evaluate multi-constraint grounding among same-category instances, revealing that current VLMs still struggle with fine-grained instance-level disambiguation. We further propose GeoREC, a two-stage grounded visual reasoning framework that combines trajectory-based supervision with GRPO-based optimization. Experiments on RS-Ref and seven external remote sensing visual grounding benchmarks show that GeoREC achieves strong performance among the evaluated methods and generalizes beyond the proposed benchmark. We hope RS-Ref and GeoREC can facilitate future research on more reliable and interpretable visual grounding in complex remote sensing scenes.

7 Broader Impacts

RS-Ref provides a distractor-controlled benchmark for reasoning-intensive referring expression comprehension in remote sensing imagery, facilitating research on fine-grained instance-level grounding and interpretable visual reasoning. The benchmark may benefit applications such as disaster response, environmental monitoring, and urban analysis, while also exposing challenges in robustness and generalization under complex scenes. As advances in remote sensing grounding may introduce risks in surveillance-related applications, we encourage responsible use and careful consideration of ethical implications in future research and deployment.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

RS-Ref is constructed solely from five publicly available remote sensing datasets without introducing private data, real-time observations, operational information, or additional sensitive metadata. All samples undergo systematic filtering and two rounds of independent human verification to ensure grounding accuracy, uniqueness, and discriminability. Military or military-adjacent objects and potentially sensitive infrastructure, if present, are inherited solely from the original source datasets rather than specifically curated for such purposes. The release will provide transparent documentation of data provenance, licenses, and intended use. RS-Ref is designed exclusively for academic evaluation and research purposes; applications involving surveillance, military targeting, individual tracking, or safety-critical decision-making are explicitly outside its intended scope. Users are required to comply with the original dataset licenses and all applicable regulations.

REPRODUCIBILITY STATEMENT

To facilitate the reproducibility, we refer the reader to the Appendix for comprehensive details regarding the training configurations and the complete data processing pipeline.

REFERENCES

Xiao An, Jiaxing Sun, Zihan Gui, and Wei He. CHOICE: Benchmarking the Remote Sensing Capabilities of Large Vision-Language Models, 2025. URL https://arxiv.org/abs/2411.18145.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, , et al. Qwen3-VL Technical Report, 2025. URL https://arxiv.org/abs/2511.21631.

Yunkai Dang, Meiyi Zhu, Donghao Wang, Yizhuo Zhang, Jiacheng Yang, Qi Fan, Yuekun Yang, Wenbin Li, Feng Miao, and Yang Gao. A benchmark for ultra-high-resolution remote sensing mllms, 2025. URL https://arxiv.org/abs/2512.17319.

Muhammad Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. GEOBench-VLM: Benchmarking Vision-Language Models for Geospatial Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7132–7142, 2025.

DeepMind. Gemini-3. https://deepmind.google/models/gemini/, 2025.

Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. Ref-adv: Exploring mllm visual reasoning in referring expression tasks. *arXiv preprint arXiv:2602.23898*, 2026.

Mustansar Fiaz, Hiyam Debary, Paolo Fraccaro, Danda Paudel, Luc Van Gool, Fahad Khan, and Salman Khan. GeoVLM-R1: Reinforcement Fine-Tuning for Improved Remote Sensing Reasoning, 2025. URL https://arxiv.org/abs/2509.25026.

Ronghao Fu, Haoran Liu, Weijie Zhang, Zhiwen Lin, Xiao Yang, Peng Zhang, and Bo Yang. OmniEarth: A Benchmark for Evaluating Vision-Language Models in Geospatial Tasks, 2026. URL https://arxiv.org/abs/2603.09471.

Google. Gemma 4. https://modelscope.cn/collections/google/Gemma-4, 2026. Accessed: 2026-04-29.

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*, pp. 27831–27840, 2024.

Timothy M Lenton, Jesse F Abrams, Annett Bartsch, Sebastian Bathiany, Chris A Boulton, Joshua E Buxton, Alessandra Conversi, Andrew M Cunliffe, Sophie Hebden, Thomas Lavergne, et al. Remotely sensing potential climate change tipping points across scales. *nature communications*, 15 (1):343, 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.

Qingyun Li, Shuran Ma, Junwei Luo, Yi Yu, Yue Zhou, Fengxiang Wang, Xudong Lu, Xiaoxing Wang, Xin He, Yushi Chen, et al. Co-Training Vision-Language Models for Remote Sensing Multi-Task Learning. *Remote Sensing*, 18(2):222, 2026.

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.

Jiaqi Liu, Ronghao Fu, Lang Sun, Haoran Liu, Xiao Yang, Weipeng Zhang, Xu Na, Zhuoran Duan, and Bo Yang. SkyMoE: A vision-language foundation model for enhancing geospatial interpretation with mixture of experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7168–7178, 2026a.

Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. Towards Faithful Reasoning in Remote Sensing: A Perceptually-Grounded GeoSpatial Chain-of-Thought for Vision-Language Models. In *Proceedings of the International Conference on Learning Representations*, 2026b. URL https://openreview.net/attachment?id=lJ7zecny2e&name=pdf.

Ruixun Liu, Bowen Fu, Jiayi Song, Kaiyu Li, Wanchen Li, Lanxuan Xue, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. ZoomEarth: Active Perception for Ultra-High-Resolution Geospatial Vision-Language Tasks, 2025a. URL https://arxiv.org/abs/2511.12267.

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*, pp. 26658–26668, 2024.

Zikun Liu, Hongzhen Wang, Lubin Weng, and Yiping Yang. Hrsc2016, 2025b. URL https://dx.doi.org/10.21227/km4k-0385.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, and Yansheng Li. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding, 2024. URL https://arxiv.org/abs/2406.10100.

Amit Misra, Kevin White, Simone Fobi Nsutezo, William Straka III, and Juan Lavista. Mapping global floods with 10 years of satellite radar data. *Nature Communications*, 16(1):5762, 2025.

Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, et al. VHM: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6381–6388, 2025.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14303–14313, 2025.

Lang Sun, Ronghao Fu, Zhuoran Duan, Haoran Liu, Xueyan Liu, and Bo Yang. Geosolver: Scaling test-time reasoning in remote sensing with fine-grained process supervision, 2026. URL https://arxiv.org/abs/2603.09551.

Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, et al. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models, 2025a. URL https://arxiv.org/abs/2508.06471.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, et al. Kimi-VL technical report, 2025b. URL https://arxiv.org/abs/2504.07491.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, et al. Kimi K2.5: Visual Agentic Intelligence, 2026a. URL https://arxiv.org/abs/2602.02276.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, , et al. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning, 2026b. URL https://arxiv.org/abs/2507.01006.

Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

Di Wang, Shunyu Liu, Wentao Jiang, Fengxiang Wang, Yi Liu, Xiaolei Qin, Zhiming Luo, Chaoyang Zhou, Haonan Guo, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. GeoZero: Incentivizing Reasoning from Scratch on Geospatial Scenes, 2026. URL https://arxiv.org/abs/2511.22645.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Shan Boqi, Long Lan, Yulin Wang, et al. GeoLLaVA-8K: Scaling Remote-Sensing Multimodal Large Language Models to 8K Resolution. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025a. URL https://arxiv.org/abs/2508.18265.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

Congxi Xiao, Jingbo Zhou, Yixiong Xiao, Jizhou Huang, and Hui Xiong. Refound: Crafting a foundation model for urban region understanding upon language and visual foundations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3527–3538, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Chushi Yu and Yoan Shin. A cross-modality feature adaptive interaction approach for rgb-infrared object detection in aerial imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Zhensheng Yuan, Haozhi Huang, Zhen Xiong, Di Wang, and Guanghua Yang. Robust and efficient 3d gaussian splatting for urban scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26209–26219, 2025.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE transactions on geoscience and remote sensing*, 61:1–13, 2023.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. doi: 10.1109/TGRS.2024.3409624.

Zilun Zhang, Zian Guan, Tiancheng Zhao, Haozhan Shen, Tianyu Li, Yuxiang Cai, Zhonggen Su, Zhaojun Liu, Jianwei Yin, and Xiang Li. Geo-R1: Improving Few-Shot Geospatial Referring Expression Understanding with Reinforcement Fine-Tuning, 2025. URL https://arxiv.org/abs/2509.21976.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

Zhengxia Zou and Zhenwei Shi. Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. *IEEE Transactions on Image Processing*, 27(3): 1100–1111, 2017.

A Qualitative Analysis

In this section, we evaluate InternVL3.5, GLM-4.6V, Kimi-VL, VHM, EarthDial, RSThinker, and our proposed GeoREC on the RS-Ref benchmark. As illustrated in Figures 4 through 7, the visualization results clearly demonstrate that our framework consistently outperforms existing state-of-the-art methods. Specifically, GeoREC is able to accurately localize target objects in complex remote sensing scenes, effectively handling cases with multiple distractor objects, semantically similar non-target entities, and fine-grained spatial relationships.

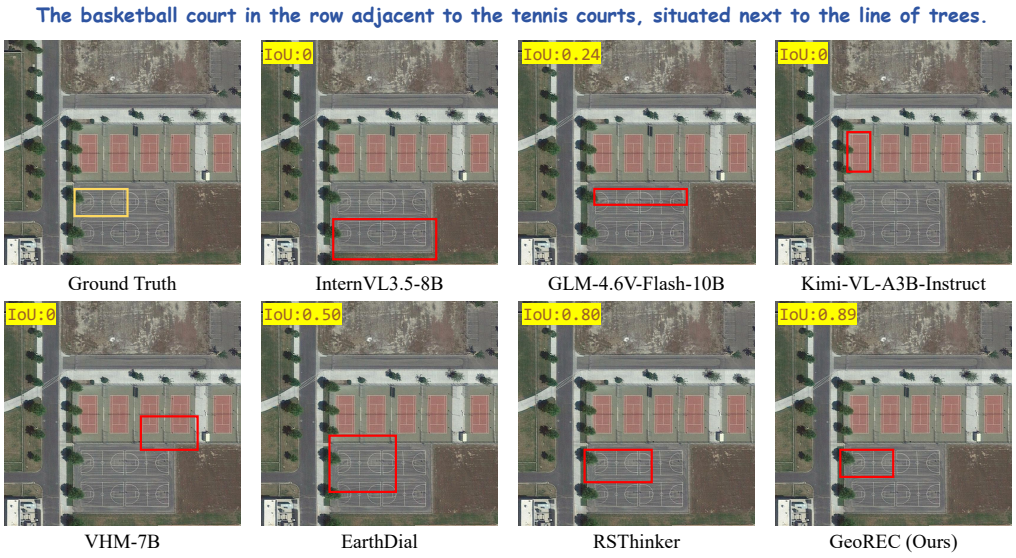


Figure 4: Qualitative examples on RS-Ref.

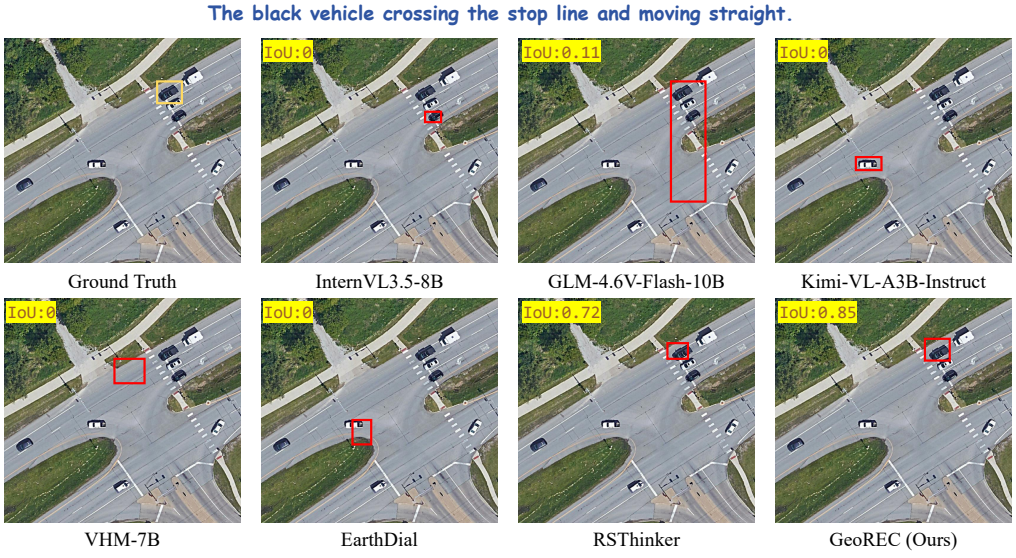


Figure 5: Qualitative examples on RS-Ref.

B G-CoT-Instruct Construction Details

In this section, we describe the construction of G-CoT-Instruct. Built upon the visual grounding subset of VRSBench Li et al. (2024), we first extract the original JSON annotations, which include image metadata such as referring expressions, object IDs, and object positions. These data are then

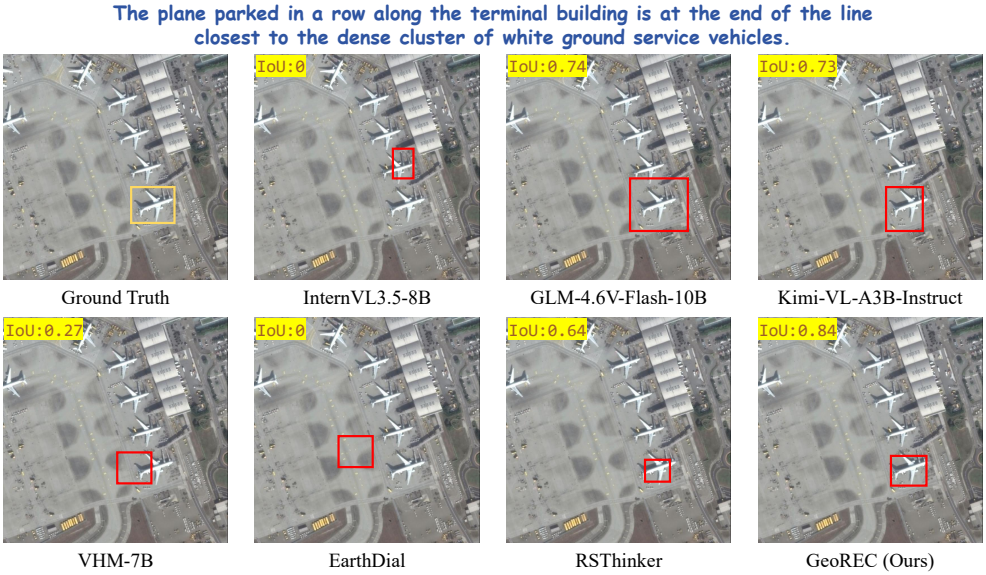


Figure 6: Qualitative examples on RS-Ref.

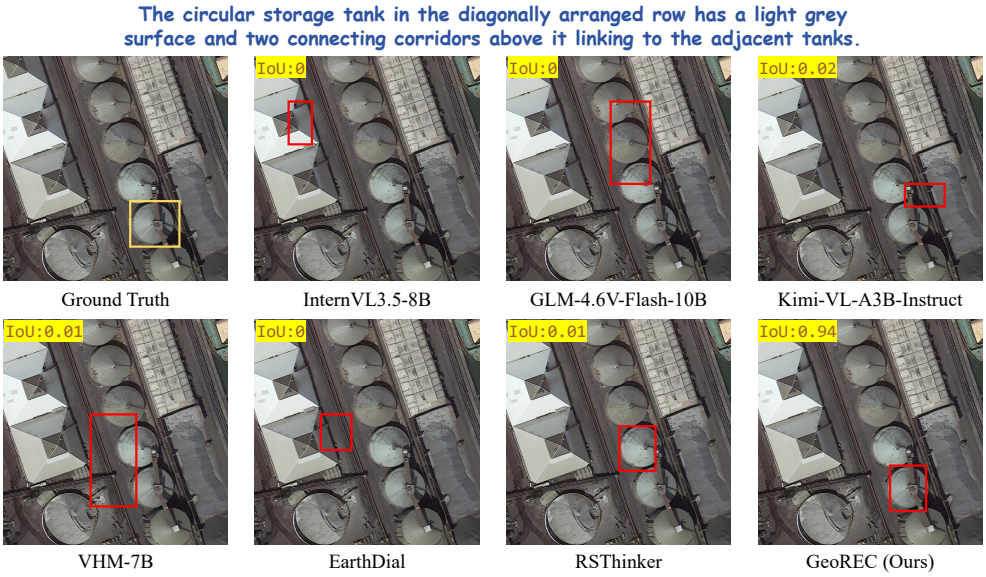


Figure 7: Qualitative examples on RS-Ref.

input into GPT-4o to generate initial grounded chain-of-thought trajectories. To further enhance the reliability of the reasoning and grounding process, we employ an additional GPT-4o to ensure that the textual reasoning states are consistent with the corresponding grounding predictions at each step, while also enforcing a coarse-to-fine reasoning progression. Trajectories that meet both logical consistency and accurate grounding are retained, whereas those failing these criteria are discarded and regenerated. This procedure results in a final set of 35,967 trajectories with images at a resolution of 512×512. We provide the prompt templates below.

C Implementation Details

GeoREC is trained in two stages. In Stage I, we perform full-parameter supervised fine-tuning using LLaMA-Factory, initialized from GLM-4.1V-9B-Base and trained on G-CoT-Instruct. We use a learning rate of 1.0×10^{-6}, train for 5 epochs, and set the batch size to 16. In Stage II, we conduct

GRPO training using EasyR1. The GRPO training set combines DIOR-RSVG and G-CoT-Instruct to improve data diversity. We use a global batch size of 16, optimize with the reward functions R_{traj} and R_{out}, and train for 530 steps. We fix $\alpha = 1$ to assign equal weights to R_{traj} and R_{out} without additional tuning. Additionally, we set $\lambda = 0$, which encourages positive refinement steps while avoiding explicit penalties on temporary regressions.

To ensure fair evaluation, we follow each model's officially recommended prompting strategy and output format. Different models may use different coordinate conventions; for example, GLM-4.6V and Qwen3-VL output relative coordinates in [0,1000], while Kimi-VL uses normalized coordinates in [0,1]. We parse the predicted bounding boxes from each response and transform them into the original image coordinate. All models are then evaluated using the same IoU calculation and accuracy criteria. For responses with invalid or unparseable bounding boxes, we regenerate the output using the same prompt and inference configuration, with a maximum of five attempts. Importantly, regeneration is triggered only by formatting failures and never depends on localization performance, ensuring that the evaluation does not introduce additional bias.

D BROADER IMPACTS

RS-Ref provides a distractor-controlled benchmark for reasoning-intensive referring expression comprehension in remote sensing imagery, facilitating research on fine-grained instance-level grounding and interpretable visual reasoning. The benchmark may benefit applications such as disaster response, environmental monitoring, and urban analysis, while also exposing challenges in robustness and generalization under complex scenes. As advances in remote sensing grounding may introduce risks in surveillance-related applications, we encourage responsible use and careful consideration of ethical implications in future research and deployment.