WHEN TO GROUND, CLARIFY, OR ABSTAIN: EVIDENCE-AWARE SELECTION AND GROUNDING IN REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have recently advanced remote sensing visual grounding (RSVG). However, existing RSVG methods typically assume that every query refers to one or more identifiable targets present in the image. This assumption forces models to produce a localization even when the intended target lacks sufficient visual support or the query does not contain enough information to distinguish it from similar objects. Existing work either handles these cases separately or maps both to abstention, rather than selecting between clarification and abstention according to the available evidence. To address this limitation, we formulate RSVG as a three-way selective grounding problem and propose EASG, an evidence-aware selection and grounding framework that requires the model to select Ground when one or more intended targets are sufficiently supported, Clarify when multiple candidates remain plausible, or Abstain when visual evidence is insufficient. Specifically, EASG uses a detector to extract coarse-grained candidate evidence and presents it to the MLLM as non-binding guidance, enabling the model to complete selection and grounding according to the available evidence. To support this task, we construct RS-EviBench, a benchmark comprising 68,328 samples that cover reliable grounding, referential ambiguity, and insufficient visual evidence. On the RS-EviBench test set, EASG achieves 96.89% decision accuracy, 58.83% mIoU, and 55.04% Acc@0.7, improving mIoU and Acc@0.7 by 8.01 and 13.36 percentage points, respectively, over the baseline, while experiments on VRSBench-VG and DIOR-RSVG further demonstrate its generalization to grounding-only RSVG benchmarks.

1 INTRODUCTION

Recently, multimodal large language models (MLLMs) have substantially advanced remote sensing visual grounding (RSVG) by strengthening cross-modal alignment and linguistic reasoning. However, existing RSVG studies have typically focused on improving localization accuracy, often overlooking failures that arise from unreliable inputs rather than solely from limitations in model capability.

Such input unreliability is inherent to real-world RSVG and arises from both visual and linguistic sources. Changes between image acquisition and query formation can make the description inconsistent with the current scene (Fig. 1(a)), while occlusion and imaging degradation can further weaken the observability of the target (Fig. 1(b)). Moreover, the user may not have access to the current scene and therefore may be unable to provide the attributes such as color, counts, or spatial relations needed to distinguish among similar objects; even when the user has access to the scene, such information may be omitted for brevity (Fig. 1(c)). Consequently, multiple candidates may satisfy the literal description, while the intended referent remains underdetermined, leading to competing interpretations.

PostAlign rejects non-existent referents to avoid forced localization (Wu et al., 2026), while GRIL requests clarification when essential information is missing (Qiu et al., 2026). AbstainEQA distinguishes referential underspecification from unavailable visual evidence, but maps both to abstention (Wu et al., 2025). In contrast, we formulate RSVG as a three-way selective grounding

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.

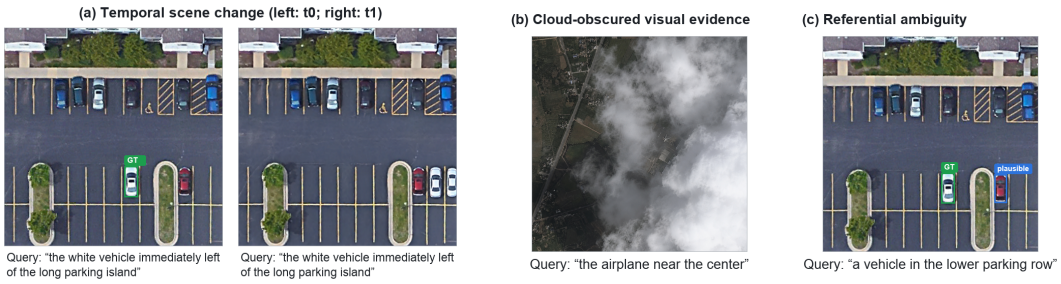


Figure 1: Three sources of unreliable input evidence in real-world RSVG: (a) temporal scene change can invalidate a previously valid referring expression; (b) cloud cover can obscure the visual evidence needed for localization; and (c) an under-specified expression can remain compatible with multiple plausible targets.

problem: *Ground* when one or more intended targets are sufficiently supported, *Clarify* when multiple candidates remain plausible, and *Abstain* when no candidate has sufficient visual support.

To address this problem, we propose Evidence-Aware Selection and Grounding (EASG), a framework that adaptively selects among grounding, clarification, and abstention. EASG integrates a frozen remote-sensing object detector to extract coarse-grained candidate evidence. Given the image, query, and candidate regions, the MLLM compares the query with the candidates and determines whether the available evidence is sufficient for grounding. The candidate regions provide guidance but do not constrain the final prediction, and the model outputs a bounding box only when grounding is justified.

To support model learning for three-way selective grounding, we construct RS-EviBench from the Non-Unique subset of VRSBench-VG (Li et al., 2024c), where multiple same-category objects and complex spatial relations make it possible to weaken a query while keeping it valid for the ground-truth target and at least one other object of the same category. RS-EviBench contains 22,776 triplets, comprising 68,328 samples in total. Each triplet is derived from a single VRSBench-VG source sample and contains three variants corresponding to the three decisions. The full-evidence variant supports *Ground*; weakening the query so that it matches the ground-truth target and at least one other object of the same category yields *Clarify*; and weakening the image by masking the target yields *Abstain*.

Our contributions are summarized as follows:

- We identify a limitation in existing RSVG methods: although prior work supports abstention, it collapses insufficient referring expression and insufficient visual evidence into the same *Abstain* response. As a result, real-world grounding failures cannot be correctly attributed. To address this limitation, we formulate RSVG as a three-way selective grounding problem and introduce EASG, which selects *Ground*, *Clarify*, or *Abstain* according to the available evidence.
- We construct RS-EviBench, comprising 22,776 evidence triplets (68,328 samples). Each triplet keeps one source sample while changing its textual or visual evidence. The constructed data undergoes both MLLM and human review.
- On RS-EViBench, EASG achieves 96.89% accuracy in three-way response selection while also improving grounding precision through detector-assisted grounding. It reaches 58.83% mIoU and 55.04% Acc@0.7, exceeding the baseline by 8.01 and 13.36 percentage points, respectively, with the larger gain at the high IoU threshold highlighting more precise localization. Moreover, results on VRSBench-VG and DIOR-RSVG demonstrate that the localization improvements brought by EASG generalizes across benchmarks.

2 Related Work

RSVG and detector-assisted grounding. Remote-sensing visual grounding was introduced with GeoVG and the DIOR-RSVG benchmark (Sun et al., 2022; Zhan et al., 2023). Subsequent methods

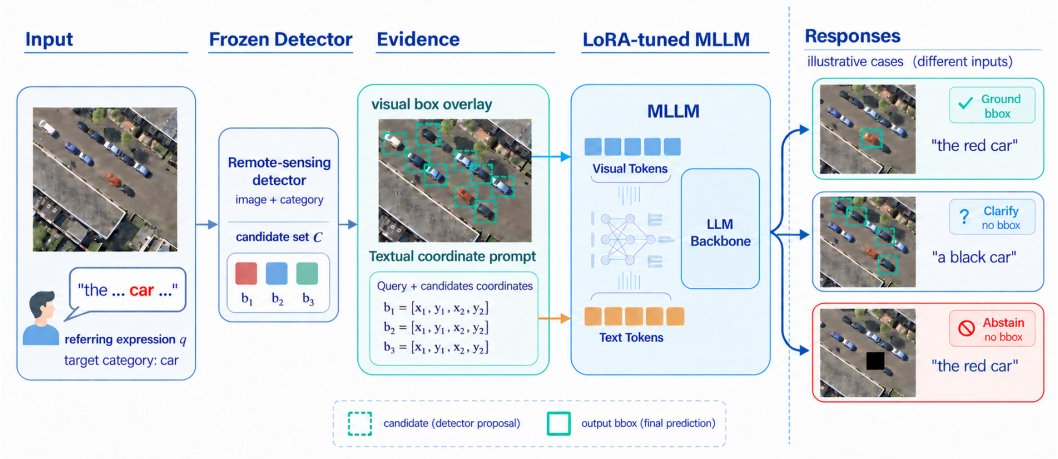


Figure 2: Overview of EASG. The left and middle panels show detector-assisted visual and coordinate prompting; the right panel illustrates the three responses for different inputs.

improve multi-scale visual-text alignment, regional reasoning, and MLLM-based localization (Lan et al., 2024; Li et al., 2024b;a; Ding et al., 2025; Kuckreja et al., 2023; Zhang et al., 2025). Detector-assisted methods use candidate regions to guide grounding, either by scoring proposals (Yu et al., 2018; Shen et al., 2023) or by exposing marked regions and visual tokens to an MLLM (Yang et al., 2023; Ma et al., 2024). These methods mainly optimize localization or choose among candidate regions; they assume that the query has one or more identifiable targets and do not jointly decide whether the available text and image support grounding.

Rejection and clarification. Inference-time and training-based methods improve visual faithfulness by verifying outputs, gathering extra evidence, abstaining, or using preference supervision (Mishra et al., 2026; Xu et al., 2026; Liu et al., 2024a; Zhao et al., 2024; Yu et al., 2024). Other work asks for clarification when a query is ambiguous or key information is missing (Shridhar & Hsu, 2018; Zhang et al., 2023; Qiu et al., 2026), while PostAlign rejects unsupported referents (Wu et al., 2026). AbstainEQA distinguishes several causes of insufficient evidence in its taxonomy, but maps them to one abstention response (Wu et al., 2025). Thus, prior work does not provide one grounding model that selects *Ground* when one or more intended targets are sufficiently supported, *Clarify* when multiple candidates remain plausible, and *Abstain* when visual support is insufficient. EASG addresses this gap with detector-provided evidence and three-way response supervision.

3 Method

Figure 2 summarizes the workflow of EASG.

3.1 Evidence-Conditioned Response Space

Multi-modal input and structured output. Given an image $\mathcal{I}$ and a referring expression q, EASG outputs a response action and a possibly empty list of predicted boxes:

$$(d, \hat{\mathcal{B}}) = \mathrm{EASG}(\mathcal{I}, q), \tag{1}$$

where $d \in \{\mathrm{Ground}, \mathrm{Clarify}, \mathrm{Abstain}\}$ specifies the response behavior. For Ground, $\hat{\mathcal{B}} = [\hat{b}_1, \ldots, \hat{b}_M]$ is a non-empty list of predicted boxes, whereas for Clarify and Abstain, $\hat{\mathcal{B}} = \emptyset$. The output format permits multiple boxes for potential extension to multi-target grounding. However, RS-EViBench is derived from VRSBench-VG and inherits its single-target grounding protocol. Consequently, every Ground sample considered in this work has $M = 1$.

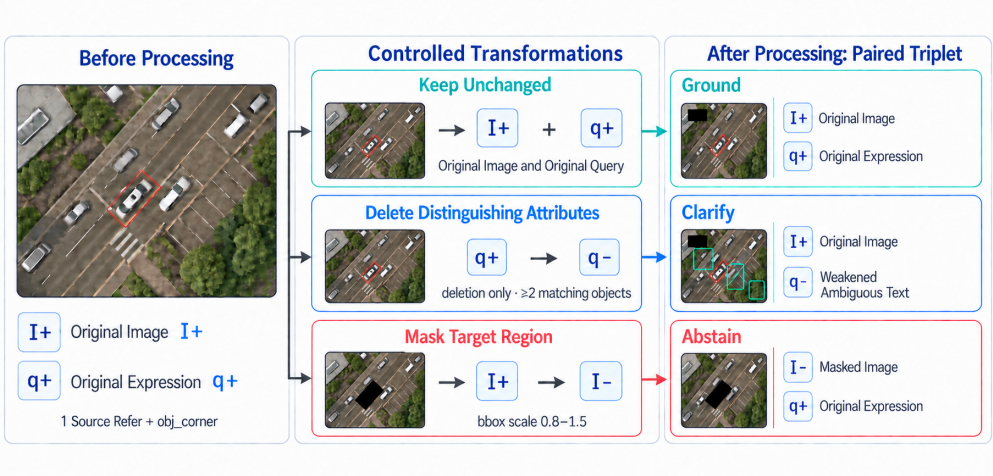


Figure 3: Construction of evidence triplets in RS-EviBench.

3.2 Candidate-First, Non-Binding Evidence Prompting

For each sample, we extract the object category in the referring expression by matching it against a list of category names. If no category is found, we use the full expression as the detector input. We then condition the frozen LAE-DINO detector D (Pan et al., 2025) on the resulting input c:

$$\mathcal{C} = D(\mathcal{I}, c), \qquad c = g(q). \tag{2}$$

Specifically, we first extract the target object category $c = g(q)$ from the referring expression q, then feed the image $\mathcal{I}$ and category c into the detector D to obtain the candidate box set $\mathcal{C} = D(\mathcal{I}, c)$. The detector returns a candidate set $\mathcal{C} = \{b_i\}_{i=1}^{N}$ with normalized box coordinates and $N \geq 0$. We present $\mathcal{C}$ as both an image overlay and a textual coordinate prompt. When $N = 0$, both contain no box entries and the prompt includes a `NONE` option. The overlay image and textual prompt are jointly encoded as multimodal inputs to fine-tune the MLLM. Importantly, $\mathcal{C}$ is not a closed answer set: the model may generate a supported box outside the proposals, request clarification, or abstain.

3.3 Detector-Integrated Structured Fine-Tuning

The frozen detector and evidence construction are identical during training and inference, followed by one MLLM generation pass. Ground samples supervise one target box, while Clarify and Abstain samples supervise an empty box list.

3.4 Training Objective

We fine-tune via LoRA (Hu et al., 2021). Let $\mathbf{s}^* = (s_1^*, \ldots, s_T^*)$ denote the serialized target response, which contains the action label and includes box coordinates only for Ground samples. Let z_d be the logit of the first-token label for action d, with $\mathcal{D} = \{\text{Ground}, \text{Clarify}, \text{Abstain}\}$. The training objective has two terms: (1) an autoregressive generation loss over the complete serialized response, including box coordinates for Ground samples; and (2) a weighted cross-entropy loss for the first-token action decision, which explicitly supervises the Ground, Clarify, and Abstain labels.

$$\mathcal{L} = -\sum_{t=1}^{T} \log p_\theta\big(s_t^* \mid \mathcal{I}_{\text{overlay}}, \mathcal{T}_{\text{input}}, s_{<t}^*\big) - \lambda_{\text{dec}} \log \frac{\exp z_{d^*}}{\sum_{d \in \mathcal{D}} \exp z_d}. \tag{3}$$

4 RS-EviBench

4.1 Evidence Triplet Construction

Source selection and triplet construction. We select Non-Unique samples from VRSBench-VG, where each image contains multiple objects from the same category as the target. This makes it

possible to weaken the referring expression while keeping it valid for the ground-truth target and at least one other object of the same category. As shown in Figure 3, each original VRSBench-VG sample is used to construct the following three RS-EviBench samples:

$$\begin{aligned} \text{Ground} &: (\mathcal{I}^+, q^+), \\ \text{Clarify} &: (\mathcal{I}^+, q^-), \\ \text{Abstain} &: (\mathcal{I}^-, q^+). \end{aligned} \tag{4}$$

Here, $\mathcal{I}^+$ and q^+ are the original image and expression. For Ground, they provide sufficient evidence to localize the target. For Clarify, Qwen3.7-Flash weakens q^+ by removing target-distinguishing attributes such as color, size, position, comparison, or reference relations, while preserving the target category name. We retain q^- only when it matches the ground-truth target and at least one other same-category object, so multiple targets remain plausible. For Abstain, we keep q^+ unchanged and mask the target with a safety margin in $\mathcal{I}^-$ using a target-adaptive rectangle, removing sufficient visual evidence for localization.

Mask-shortcut control. To prevent the mask itself from becoming a trivial shortcut for the response condition, we apply an auxiliary mask to Ground and Clarify samples. Guided by a strong MLLM (Qwen3.7-Flash), the auxiliary mask is constrained to protect the target, candidate regions, other labeled objects of the same category, spatial comparisons, and any reference object or region required to interpret the query. We deliberately omit the auxiliary mask for Ground queries specified solely by non-spatial attributes (e.g., color or size), because an unobserved object beneath the mask could still satisfy such a query. The final validated artifact applies an auxiliary mask to 19,574 of 22,776 Ground-Clarify pairs (85.94% overall and 86.63% on the test dataset).

4.2 Quality Control

We use Qwen3.7-Flash to review the q^- expressions and auxiliary masks in the constructed data. For Clarify, it checks whether q^- matches the ground-truth target and at least one other object of the same category. For the auxiliary masks applied to Ground and Clarify, it checks that the mask does not cover the ground-truth target, required reference objects, or candidate regions, and does not affect the attribute comparisons needed to understand the query. Each review is labeled Accept, Uncertain, or Reject. We manually review a random 20% of accepted cases and all uncertain and rejected cases, correcting the labels when needed.

5 Experiments

5.1 Datasets

VRSBench-VG (Li et al., 2024c) provides 52,472 object references over 29,614 images and a 16,159-sample test set with Unique and Non-Unique subsets. RS-EviBench contains 22,776 triplets (68,328 samples), split into 41,535 train, 6,840 validation, and 19,953 test samples. DIOR-RSVG contains 38,320 image-expression pairs from 17,402 images across 20 categories (Zhan et al., 2023); we use its 7,500-sample test split.

5.2 Metrics

Classification. Decision accuracy measures three-way action selection; per-class F1 balances precision and recall, and macro-F1 weights the three classes equally.

Localization. Let G and $\hat{G}$ be ground-truth and predicted Ground sets, and let C_τ contain samples in both sets with IoU at least τ. We report

$$\text{Acc@}\tau = \frac{|C_\tau|}{|G|}, \qquad \text{Prec@}\tau = \frac{|C_\tau|}{|\hat{G}|}, \qquad \tau \in \{0.5, 0.7\}. \tag{5}$$

Table 1: Results on the RS-EviBench test dataset (19,953 samples; %). The best result in each metric column is bold and the second best is underlined. Dec. Acc denotes decision accuracy; mIoU, mean intersection over union; Acc@τ, localization accuracy at IoU threshold τ; Pr@τ, localization precision at threshold τ; Valid, valid-output rate; T-EM, triplet exact match; and FLR, forced localization rate.

(a) Classification and reliability

Model	Params	Classification	Reliability		
		Dec. Acc	Valid	T-EM	FLR
General-purpose MLLMs (direct evaluation)					
MiniGPT-v2 (Chen et al., 2023)$^+$	7B	10.07	0.00	0.00	22.21
LLaVA-1.5 (Liu et al., 2024b)	7B	29.14	87.63	0.00	84.92
Qwen2.5-VL (Bai et al., 2025)	3B	37.44	99.77	0.57	75.62
Remote-sensing MLLMs (direct evaluation)					
GeoChat$^+$	7B	23.47	60.25	0.00	54.54
VHM (Pang et al., 2024)$^+$	7B	11.18	26.13	0.00	23.94
Task-specific models (RS-EviBench SFT)					
Qwen2.5-VL SFT	3B	<u>96.44</u>	**99.99**	<u>89.58</u>	<u>2.49</u>
EASG	3B	**96.89**	**99.99**	**90.93**	**2.26**

(b) Localization

Model	Params	Localization				
		mIoU	Acc@.5	Acc@.7	Pr@.5	Pr@.7
General-purpose MLLMs (direct evaluation)						
MiniGPT-v2 (Chen et al., 2023)$^+$	7B	0.49	0.02	0.00	0.02	0.00
LLaVA-1.5 (Liu et al., 2024b)	7B	6.12	1.23	0.15	0.48	0.06
Qwen2.5-VL (Bai et al., 2025)	3B	30.46	33.78	21.55	14.44	9.21
Remote-sensing MLLMs (direct evaluation)						
GeoChat$^+$	7B	8.09	3.98	0.62	2.25	0.35
VHM (Pang et al., 2024)$^+$	7B	2.60	0.81	0.06	1.00	0.07
Task-specific models (RS-EviBench SFT)						
Qwen2.5-VL SFT	3B	<u>50.82</u>	<u>59.92</u>	<u>41.68</u>	<u>60.22</u>	<u>41.89</u>
EASG	3B	**58.83**	**70.32**	**55.04**	**70.37**	**55.09**

$^+$MiniGPT-v2, GeoChat, and VHM have low strict format-valid rates; their metrics other than Valid are therefore computed using weak-format parsing of semantically recoverable responses. Valid always reports exact-format compliance. All other models are evaluated under the strict three-way output protocol.

Reliability. Triplet exact match requires all three decisions in the same triplet to be correct. For non-Ground samples U, forced localization rate is

$$\mathrm{FLR} = \frac{|\{i \in U : \hat{d}_i = \mathrm{Ground} \ \vee \ \hat{b}_i \neq \emptyset\}|}{|U|}. \tag{6}$$

5.3 Implementation Details

We use Qwen2.5-VL-3B-Instruct (Bai et al., 2025) with a frozen LAE-DINO detector (Pan et al., 2025). LoRA uses rank 16, scaling 32, and dropout 0.05. We discard detector predictions below 0.15 confidence, apply NMS at IoU 0.5, retain the top 20 boxes, and normalize coordinates to $[0, 1]$. We train one epoch on eight GPUs in bf16 with an effective batch of 32 triplets, using learning rates 2×10^{-4} for LoRA and 2×10^{-5} for the visual merger; the validation checkpoint is reported.

5.4 Main Results

Finding 1: Weak response classification without task-specific fine-tuning. As detailed in Table 1, five evaluated MLLMs without task-specific fine-tuning achieve limited decision accuracies of 11.18–37.44%. Notably, GeoChat and VHM fail to generate valid structured outputs on a large portion of samples (60.25% and 26.13% validity, respectively). Even among valid outputs, they exhibit an overwhelming "Ground" bias (97.76% and 99.98%), yielding subset decision accuracies of only 38.55% and 41.18%. This reveals a severe deficiency in transferring general remote-sensing knowledge to evidence-based decision-making.

Finding 2: SFT substantially improves response selection, while Ground-Abstain confusion remains. As shown in Table 1 and Appendix Table 8, supervised fine-tuning (SFT) dramatically raises decision accuracy from 37.44% to 96.44%, while substantially improving all three per-class F1 scores.

Clarify exhibits the most dramatic recovery (F1 surging from 11.99% to 98.81%), compared with *Ground* (49.66% to 94.75%) and *Abstain* (30.80% to 95.78%). However, among the 710 residual errors post-SFT, 77.9% (553 samples) stem from confusion between *Ground* and *Abstain*, highlighting that determining visual evidence sufficiency remains the core bottleneck.

The 99.99% valid-output rate reflects rare formatting errors; our box-list interface also supports future multi-target grounding, where multiple boxes remain a valid *Ground* response when the query identifies a target set, rather than being classified as *Clarify* simply because it refers to multiple targets.

Finding 3: Candidate-level evidence improves localization in the full EASG system. Compared with standard SFT, EASG improves mIoU by 8.01 points, Acc@0.5 by 10.40 points, and Acc@0.7 by 13.36 points. Decision accuracy increases by only 0.45 points, while Appendix Table 8 shows positive F1 changes for Ground, Clarify, and Abstain. Candidate evidence therefore strengthens localization without sacrificing response selection, helping the model judge whether the available visual evidence is sufficient.

5.5 Generalization to Conventional RSVG

To evaluate whether the localization ability learned by EASG transfers to conventional remote-sensing visual grounding, we test on VRSBench-VG and DIOR-RSVG. Because RS-EviBench is constructed from VRSBench-VG, VRSBench-VG instead evaluates transfer from the Ground-Clarify-Abstain selective-grounding protocol back to conventional forced localization. DIOR-RSVG is distinct from RS-EviBench and evaluates cross-dataset generalization.

VRSBench-VG. As shown in Table 2, EASG reaches 62.50% Acc@0.5 and 45.89% Acc@0.7 on Unique samples. This strong result without Unique-specific fine-tuning indicates transfer beyond the ambiguous cases emphasized by RS-EviBench. On Non-Unique samples, EASG reaches 54.44% Acc@0.5 and 37.81% Acc@0.7, exceeding the strongest reported baseline by 8.14 and 14.91 points, respectively.

On the All split, EASG obtains 57.80% Acc@0.5 and 41.18% Acc@0.7. GeoViS remains stronger at both thresholds, while RSGround-R1 is higher at Acc@0.5 but lower at Acc@0.7 (38.32% versus 41.18%). Unlike these methods, which specifically optimize grounding through full-parameter visual-search tuning or a localization-specific positional reward, EASG uses LoRA to learn selective Ground-Clarify-Abstain behavior. Its competitive strict-threshold accuracy shows that selective response learning preserves localization quality.

DIOR-RSVG. On DIOR-RSVG, EASG reaches 57.12% Acc@0.5 and 46.63% Acc@0.7. It outperforms the directly evaluated generic baselines, while remaining below the in-domain GeoViS and RSGround-R1 specialist references at both thresholds. Because the standard DIOR-RSVG annotations do not support three-way decision supervision, EASG is evaluated without DIOR-RSVG adaptation. The results therefore support cross-benchmark transfer, while the in-domain specialist scores serve only as reference upper bounds.

Table 2: Task transfer and cross-dataset generalization on conventional remote-sensing grounding benchmarks (%). Both benchmarks report localization accuracy at IoU thresholds 0.5 and 0.7. Within each panel, the best result in each metric column is bold and the second best is underlined.

(a) VRSBench-VG: task transfer

Model	Params	Unique		Non-Unique		All	
		Acc@.5	Acc@.7	Acc@.5	Acc@.7	Acc@.5	Acc@.7
LoRA fine-tuned on VRSBench-VG							
MiniGPT-v2	7B	40.7	18.9	32.4	15.2	35.8	16.8
LLaVA-1.5	7B	51.1	16.4	34.8	11.5	41.6	13.6
GeoChat	7B	57.4	22.6	44.5	18.0	49.8	19.9
Mini-Gemini (Li et al., 2024d)	7B	41.1	9.6	22.3	4.9	30.1	6.8
VHM (Pang et al., 2024)	7B	**68.9**	**46.7**	<u>46.3</u>	<u>22.9</u>	55.7	32.8
In-domain specialist references							
GeoViS*	3B	–	–	–	–	**68.5**	**45.7**
RSGround-R1*	3B	–	–	–	–	<u>63.72</u>	38.32
Qwen2.5-VL SFT	3B	42.07	27.15	39.66	23.82	40.66	25.21
EASG	3B	<u>62.50</u>	<u>45.89</u>	**54.44**	**37.81**	57.80	<u>41.18</u>

(b) DIOR-RSVG: cross-dataset generalization

Model	Params	Acc@.5	Acc@.7
Direct evaluation without DIOR-RSVG adaptation			
MiniGPT-v2	7B	29.4	10.2
LLaVA-1.5	7B	8.47	1.45
GeoChat	7B	24.83	7.57
Qwen2.5-VL	3B	36.23	25.83
In-domain specialist references			
GeoViS*	3B	**79.8**	**70.1**
RSGround-R1*	3B	<u>71.81</u>	<u>58.71</u>
EASG	3B	57.12	46.63

In the VRSBench-VG panel, Qwen2.5-VL SFT and EASG are fine-tuned on RS-EviBench, while the reported-baselines group contains published VRSBench-VG LoRA-fine-tuned baselines (Li et al., 2024c). In the DIOR-RSVG panel, EASG is evaluated without adaptation because the benchmark lacks three-way labels. * denotes results from the original protocols. Dashes indicate unreported metrics; Acc@τ and Pr@τ are equivalent here (Zhan et al., 2023).

Table 3: Core-component ablation on the RS-EviBench test dataset (%). The best result in each metric column is bold and the second best is underlined; lower is better for FLR.

Setting	Cand. evidence	Dec. loss	Dec. Acc	Macro-F1	mIoU	Acc@.7	Pr@.7	Triplet EM	FLR
Qwen2.5-VL SFT	×	×	96.44	96.45	50.82	41.68	41.89	89.58	2.49
Qwen2.5-VL SFT w/ dec. loss	×	✓	<u>96.57</u>	<u>96.57</u>	50.11	41.42	41.59	<u>90.00</u>	2.43
EASG w/o dec. loss	✓	×	94.96	94.93	<u>54.16</u>	<u>49.83</u>	<u>53.05</u>	85.48	**2.17**
EASG	✓	✓	**96.89**	**96.89**	**58.83**	**55.04**	**55.09**	**90.93**	<u>2.26</u>

5.6 Ablation Study

We next analyze component contributions and decision-loss sensitivity, followed by diagnostics for shortcut behavior and proposal dependence.

Component contributions. Table 3 shows complementary effects. Candidate evidence alone raises mIoU by 3.34 points, Acc@0.7 by 8.15 points, and Pr@0.7 by 11.16 points over Qwen2.5-VL SFT, but lowers decision accuracy by 1.48 points and triplet exact match by 4.10 points. With candidate

evidence fixed, decision loss improves decision accuracy by 1.93 points, mIoU by 4.67 points, and triplet exact match by 5.45 points; without candidate evidence, its accuracy gain is only 0.13 points and localization slightly decreases. The interaction gains of 5.38 mIoU and 5.47 Acc@0.7 points confirm that decision supervision converts candidate cues into coordinated selection and localization. The full model is best on all core metrics.

Decision-loss sensitivity. We evaluate the sensitivity to the decision-loss weight by setting λ_{dec} to 0, 0.1, and 0.2. The results in Appendix Table 5 show that the default $\lambda_{\text{dec}} = 0.1$ performs best: relative to zero weight, it increases decision accuracy, mIoU, Acc@0.7, and triplet exact match by 1.93, 4.67, 5.21, and 5.45 points, respectively. Increasing the weight further reduces both decision and localization performance.

Mask-shortcut control. RS-EviBench places non-target masks in Ground and Clarify samples to prevent the model from associating the mere presence of a visible mask with Abstain. Appendix Table 6 tests whether EASG nevertheless learns this shortcut by reporting decision accuracy and False Abstain across different mask areas, locations, and aspect ratios. False Abstain measures the fraction of true Ground and Clarify samples incorrectly classified as Abstain.

EASG remains robust across mask positions, sizes, and orientations: decision accuracy stays high and False Abstain remains low in every slice. Center masks are the most challenging, while edge and corner masks are easier. These differences may also reflect variation in scene content and sample difficulty, so they do not indicate a mask-position shortcut.

Oracle-proposal comparison. On the same Ground samples, an oracle selects the proposal with the highest ground-truth IoU and is unavailable at inference. As shown in Appendix Table 7, EASG exceeds this oracle by 1.87 mIoU, 0.35 Acc@0.5, and 2.57 Acc@0.7 points. Appendix Table 9 shows that most outputs do not copy a proposal and often improve on the best one, confirming that candidate boxes are evidence rather than a closed answer set.

6 Limitations

Our evaluation follows the single-target VRSBench-VG protocol and does not yet measure multi-target grounding or multi-turn clarification. The candidate evidence experiments use a remote-sensing detector and do not test transfer to natural-image detectors; the benchmark controls only selected forms of query ambiguity and visual degradation and covers a limited range of benchmarks and model scales.

7 Conclusion

We introduced evidence-aware selective grounding for remote sensing. EASG uses detector candidates as non-binding visual and coordinate evidence and selects Ground, Clarify, or Abstain according to the available evidence. RS-EviBench provides three-sample triplets that separate referential ambiguity from insufficient visual evidence.

On RS-EviBench, EASG achieves 96.89% decision accuracy, 58.83% mIoU, 55.04% Acc@0.7, 90.93% triplet exact match, and a 2.26% forced-localization rate. Candidate evidence improves mIoU and Acc@0.7 by 8.72 and 13.62 points over the matched baseline, while transfer results show that selective response learning preserves conventional grounding ability.

The same principle can extend to GUI and robotic agents. For example, after a user asks a GUI agent to "delete the report I just downloaded," the agent may open a file page containing several matching reports; *Clarify* should identify the intended file before deletion. Likewise, when a robot is asked to "bring me the red cup," clarification can prevent an ambiguous instruction from becoming an unsafe physical action.

AI Use Statement

Generative AI tools were used to assist with language editing and LaTeX formatting. All AI-assisted text, claims, and artifacts were reviewed by the authors, who take responsibility for the final content.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023. URL https://arxiv.org/abs/2310.09478.

Yueli Ding, Di Wang, Ke Li, Xiaohong Zhao, and Yifeng Wang. Visual grounding of remote sensing images with multi-dimensional semantic-guidance. *Pattern Recogn. Lett.*, 189(C):85–91, March 2025. ISSN 0167-8655. doi: 10.1016/j.patrec.2025.01.013. URL https://doi.org/10.1016/j.patrec.2025.01.013.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing, 2023. URL https://arxiv.org/abs/2311.15826.

Meng Lan, Fu Rong, Hongzan Jiao, Zhi Gao, and Lefei Zhang. Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. URL https://api.semanticscholar.org/CorpusID:270152736.

Chongyang Li, Wenkai Zhang, Hanbo Bi, Jihao Li, Shuoke Li, Haichen Yu, Xian Sun, and Hongqi Wang. Injecting linguistic into visual backbone: Query-aware multimodal fusion network for remote sensing visual grounding. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024a. doi: 10.1109/TGRS.2024.3450303.

Ke Li, Di Wang, Haojie Xu, Haodi Zhong, and Cong Wang. Language-guided progressive attention for visual grounding in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024b. doi: 10.1109/TGRS.2024.3423663.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding, 2024c. URL https://arxiv.org/abs/2406.12384.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models, 2024d. URL https://arxiv.org/abs/2403.18814.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2024a. URL https://arxiv.org/abs/2306.14565.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024b. URL https://arxiv.org/abs/2310.03744.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models, 2024. URL https://arxiv.org/abs/2404.13013.

Ashish Mishra, Tarun Kumar, Arpit Shah, Suparna Bhattacharya, and Martin Foltin. Cebc: Conformal evidence-bounded control for low-hallucination vision-language generation. In *Annual Meeting of the Association for Computational Linguistics*, 2026. URL https://api.semanticscholar.org/CorpusID:289773494.

Jiancheng Pan, Yanxing Liu, Yuqian Fu, Muyuan Ma, Jiahao Li, Danda Pani Paudel, Luc Van Gool, and Xiaomeng Huang. Locate anything on earth: Advancing open-vocabulary object detection for remote sensing community, 2025. URL https://arxiv.org/abs/2408.09110.

Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, and Conghui He. Vhm: Versatile and honest vision language model for remote sensing image analysis, 2024. URL https://arxiv.org/abs/2403.20213.

Yiwen Qiu, Linjuan Wu, Yizhou Liu, Yuchen Yan, Jin Ma, Xu Tan, Yao Hu, Daoxin Zhang, Wenqi Zhang, Weiming Lu, Jun Xiao, and Yongliang Shen. Pause or fabricate? training language models for grounded reasoning, 2026. URL https://arxiv.org/abs/2604.19656.

Haozhan Shen, Tiancheng Zhao, Mingwei Zhu, and Jianwei Yin. Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection, 2023. URL https://arxiv.org/abs/2312.15043.

Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction, 2018. URL https://arxiv.org/abs/1806.03831.

Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 404–412, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548316. URL https://doi.org/10.1145/3503161.3548316.

Tao Wu, Chuhao Zhou, Guangyu Zhao, Haozhi Cao, Yewen Pu, and Jianfei Yang. When robots should say "i don't know": Benchmarking abstention in embodied question answering, 2025. URL https://arxiv.org/abs/2512.04597.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. Postalign: Multimodal grounding as a corrective lens for MLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zJnPyb2xrp.

Jian Xu, Yanning Wu, Delu Zeng, John Paisley, and Qibin Zhao. Look again before you abstain:budgeted conformal evidence acquisition for reliable vision-language model, 2026. URL https://arxiv.org/abs/2606.16667.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023. URL https://arxiv.org/abs/2310.11441.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension, 2018. URL https://arxiv.org/abs/1801.08186.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, 2024. URL https://arxiv.org/abs/2312.00849.

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. ISSN 1558-0644. doi: 10.1109/tgrs.2023.3250471. URL http://dx.doi.org/10.1109/TGRS.2023.3250471.

Hanbo Zhang, Jie Xu, Yuchen Mo, and Tao Kong. Invig: Benchmarking interactive visual grounding with 500k human-robot interactions, 2023. URL https://arxiv.org/abs/2310.12147.

Peirong Zhang, Yidan Zhang, Luxiao Xu, Jinliang Lin, Zonghao Guo, Fengxiang Wang, Xue Yang, Kaiwen Wei, and Lei Wang. Geovis: Geospatially rewarded visual search for remote sensing visual grounding, 2025. URL https://arxiv.org/abs/2512.02715.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2024. URL https://arxiv.org/abs/2311.16839.

A Additional Analysis

Random-seed replication. Table 4 reports results from three training seeds. Performance on decision accuracy and FLR is stable. Localization varies more, but the weakest run still exceeds the Qwen2.5-VL SFT with decision loss baseline under the same setup by 4.27 points in mIoU, 7.96 points in Acc@0.7, and 9.66 points in Pr@0.7. The localization gains are therefore not an artifact of the strongest seed.

Table 4: EASG results across three training seeds on the RS-EviBench test dataset (19,953 samples; %). Rows are independent seeds rather than competing methods, so no best/second-best ranking is assigned.

Seed	Dec. Acc	Macro-F1	mIoU	Acc@.5	Acc@.7	Pr@.5	Pr@.7	Valid	Triplet EM	FLR
42	96.89	96.89	58.83	70.32	55.04	70.37	55.09	99.99	90.93	2.26
43	95.07	95.05	54.38	66.38	49.38	68.90	51.25	99.99	85.99	2.71
44	96.03	96.02	55.48	67.93	50.17	70.04	51.73	99.99	88.50	2.17

Table 5: Decision-loss sensitivity on the RS-EviBench test dataset (%) with candidate evidence fixed. Best and second-best results are bold and underlined; lower is better for FLR.

λ_{dec}	Dec. Acc	Macro-F1	mIoU	Acc@.7	Triplet EM	FLR
0	<u>94.96</u>	<u>94.93</u>	<u>54.16</u>	<u>49.83</u>	<u>85.48</u>	**2.17**
0.1	**96.89**	**96.89**	**58.83**	**55.04**	**90.93**	2.26
0.2	95.39	95.36	54.59	49.65	86.65	<u>2.24</u>

Table 6: Mask-shortcut diagnostic for EASG on the RS-EviBench test dataset (%). False Abstain is the fraction of true Ground or Clarify samples predicted as Abstain. Rows are diagnostic slices rather than competing methods, so no best/second-best ranking is assigned.

Factor	Bin	N	Dec. Acc ↑	False Abs. ↓
Overall	All	19,953	96.89	2.15
Area	Small ($< 5\%$)	13,806	96.43	2.71
	Medium (5–10%)	2,682	97.95	1.24
	Large ($> 10\%$)	1,687	97.81	1.20
Location	Center	2,671	94.68	4.75
	Edge	8,763	96.70	2.22
	Corner	6,741	97.72	1.63
Aspect	Horizontal	8,573	96.75	2.41
	Vertical	9,602	96.81	2.37

Table 7: Comparison with the oracle best proposal on the 6,651 Ground samples in the test dataset (%). The best result in each metric column is bold and the second best is underlined. A missing EASG box is counted as incorrect. The oracle uses the ground-truth box and is not an inference-time method.

Method	Output	mIoU	Acc@.5	Acc@.7
Oracle best proposal	Proposal box	<u>56.96</u>	<u>69.97</u>	<u>52.47</u>
EASG	Free-form box	**58.83**	**70.32**	**55.04**

Table 8: Per-class decision F1 on the RS-EviBench test dataset (%). The best result in each class is bold and the second best is underlined.

Model	Ground	Clarify	Abstain
Qwen2.5-VL	49.66	11.99	30.80
Qwen2.5-VL SFT	<u>94.75</u>	<u>98.81</u>	<u>95.78</u>
EASG	**95.45**	**98.84**	**96.39**

Table 9: Proposal coverage and output-dependence diagnostic on true Ground samples in the test dataset (%). Match requires an output-proposal IoU of at least 0.95; Exceed requires the output IoU with the target to be strictly higher than the best proposal IoU. Both use all samples in each row as the denominator, with a missing output counted as neither. Rows are diagnostic slices rather than competing methods, so no best/second-best ranking is assigned.

Best-proposal IoU	N	Recall@.5	Recall@.7	EASG mIoU	Match	Exceed
Overall	6,651	69.97	52.47	58.83	7.61	58.35
< 0.5	1,997	0.00	0.00	30.69	1.55	53.13
$[0.5, 0.7)$	1,164	100.00	0.00	59.50	5.07	69.33
≥ 0.7	3,490	100.00	100.00	74.71	11.92	57.68