Grounding What Matters: Evidence-Aware Credit Assignment for Remote Sensing Reasoning

Anonymous ACL submission

Abstract

Remote sensing has emerged as an important domain for multimodal large language models (MLLMs), where reliable understanding requires locating small, dense, and spatially subtle visual evidence. However, our diagnostic analysis shows that current MLLMs often fail because their reasoning is not grounded to the correct visual evidence, leading to wrong-region attention or missed task-relevant objects. To address this issue, we introduce RsEvi-8B, an evidence-aware remote sensing MLLM that **reasons with parseable bbox-grounded evidence**. To enable this capability, we construct RsEvi-248K, a large-scale supervised dataset spanning VQA, visual grounding, object detection, object counting, and scene classification. The dataset provides verified bbox-grounded reasoning traces, where each intermediate visual claim is aligned with supporting evidence regions and filtered for both answer correctness and evidence faithfulness. To improve evidence-faithful reasoning, we further propose the evidence-aware GRPO, which uses an evidence verifier to regulate learning on generated evidence spans while preserving task-level outcome optimization. Extensive experiments across five remote-sensing task families show that RsEvi-8B improves the dataset-balanced average score by 17.15 points over Qwen3-VL-8B-Instruct, establishing explicit visual grounding as a key ingredient for reliable remote-sensing reasoning.

1 Introduction

Remote sensing has become an important domain for multimodal large language models (MLLMs), where models must understand complex overhead scenes from image level semantics to object and region level details. Recent remote sensing MLLMs and benchmarks have advanced instruction following, region dialogue, open ended VQA, and multi task evaluation (Kuckreja et al., 2024; Zhang et al., 2024; Luo et al., 2024; Li et al., 2024; Pang et al., 2024; Muhtar et al., 2024). However, the decisive evidence in remote sensing images is often small, dense, and embedded in large scenes (Shabbir et al., 2025). A correct answer may depend on tiny aircraft, ships near a dock, or subtle spatial relations.

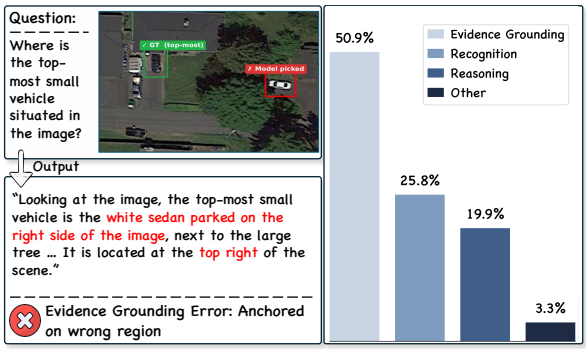


Figure 1: Diagnostic error analysis on 2,000 RSVQA-HR and VRSBench questions. Left: example grounding failure. Right: distribution of incorrect predictions.

To examine this challenge, we conduct a diagnostic error analysis on 2,000 randomly sampled questions from RSVQA-HR (Lobry et al., 2020) and VRSBench (Li et al., 2024). We evaluate Qwen3-VL-8B-Instruct on the corresponding evaluation splits and categorize its 632 incorrect predictions with GPT-5.4 (OpenAI, 2026). The judge is given only the image, question, gold answer, and model response, and assigns one primary error type following the template in Fig. 15. As shown in Fig. 1, evidence grounding failures are the largest category, accounting for 322 of 632 errors. They typically involve attending to the wrong region, missing small task-relevant objects, or assigning attributes and counts to the wrong visual target. This suggests that reliable remote-sensing reasoning requires explicit visual evidence, not answer prediction alone.

To make visual evidence explicit, a natural direction is to move from *unanchored* textual reasoning to *Think with Grounding*, where intermediate visual claims are anchored to supporting image regions.

General vision language research has explored related forms of this idea through intermediate boxes, object centric grounding, interleaved box coordinates, and replayed visual evidence (Shao et al., 2024; Man et al., 2025; Fan et al., 2025; Wang et al., 2025a). In a related but distinct line, remote sensing has developed CoT reasoning works that leverage existing labels, annotations, or rewards (Koksal and Alatan, 2025; Liu et al., 2025; Fiaz et al., 2025; Xu et al., 2025; Yao et al., 2025). These works improve textual or structured CoT, but they do not directly train models to ground each visual claim with a parseable evidence span. This gap comes from a mismatch between existing annotations and the evidence needed for reasoning. Many VQA and scene classification benchmarks provide task level outputs, rather than evidence boxes for reasoning steps (Lobry et al., 2020; Cheng et al., 2017; Xia et al., 2017). Spatial annotations are more common in visual grounding and object detection, where boxes serve as final targets rather than intermediate reasoning evidence. Intermediate visual claims therefore remain hard to verify and optimize independently from the final answer.

This raises a practical question: *Can remote sensing MLLMs be trained to reason with explicit, verifiable grounded visual evidence?* We address this question with two ingredients: bbox-grounded data construction and evidence-aware optimization.

We first construct RsEvi-248K, a large-scale SFT dataset for remote sensing Think with Grounding. Our pipeline unifies diverse tasks across visual question answering (VQA), visual grounding (VG), object detection (OD), object counting (OC), and scene classification (SC). An MLLM extracts task-relevant and visually groundable evidence labels from the question, image, and reference answer. These labels guide SAM3 (Carion et al., 2025) to propose candidate evidence regions, which are then verified to form evidence sets. A teacher model generates grounded reasoning traces from the resulting evidence sets, and rejection sampling keeps only responses with correct answers and faithful evidence.

Using RsEvi-248K, we train RsEvi-8B, a remote sensing MLLM that reasons with explicit grounded visual evidence. The model writes intermediate reasoning inside `<think>` and represents each visual claim as a `<box>` span with an evidence label and a normalized bounding box. We further optimize RsEvi-8B with evidence-aware GRPO. Outcome rewards verify the final answer but not the visual faithfulness of the supporting evidence. When the same response-level advantage is assigned to all tokens, unsupported boxes in correct responses may be reinforced, while faithful boxes in incorrect responses may be penalized. We therefore keep the reward outcome-based but apply evidence-aware advantage masking to evidence spans: an evidence verifier judges whether each box supports its label, and the advantage on that span is adjusted accordingly. Across diverse remote-sensing tasks, RsEvi-8B improves the dataset-balanced average score by 17.15 points over Qwen3-VL-8B-Instruct and reduces Evidence Grounding errors by 57.1% relative to the base model.

Our contributions are summarized as follows:

- We identify evidence selection and localization failures as a key bottleneck in remote sensing MLLMs, and introduce RsEvi-8B, a model that reasons with explicit grounded visual evidence and improves task performance.
- We construct RsEvi-248K, a large-scale SFT dataset with verified bbox-grounded reasoning traces, built by evidence construction and rejection sampling.
- We propose evidence-aware GRPO, which keeps outcome rewards for final answers while modulating evidence-span advantages according to visual faithfulness.

2 SFT Dataset Construction

We construct RsEvi-248K, a multimodal SFT dataset for remote sensing that pairs task-specific answers with reasoning traces grounded in verifiable visual evidence where applicable, providing cold-start data before reinforcement learning. We use Qwen3.5-397B throughout the construction pipeline (Team, 2026). See Appendix B.1 for further details.

Data sources and task normalization. We construct RsEvi-248K from public remote-sensing datasets spanning five task families: visual question answering, visual grounding, object detection, object counting, and scene classification. The sources include RSVQA-HR, VRSBench, VersaD, DIOR-RSVG, HRRSD, NWPU-RESISC45, AID, fMoW, UCMerced, METER-ML, FAIR1M, and DOTA (Lobry et al., 2020; Li et al., 2024; Pang et al., 2024; Sun et al., 2022b; Zhang et al., 2019; Cheng et al., 2017; Xia et al., 2017; Christie et al.,

2018; Yang and Newsam, 2010; Zhu et al., 2022; Sun et al., 2022a; Xia et al., 2018). We normalize all samples into a unified format with task labels and reference answers.

Object-level evidence construction. For each sample, we construct a set of visual evidence boxes that can support the final response. When object or grounding annotations are available, we directly convert them into the shared $[0, 1000]$ coordinate space. Otherwise, we prompt Qwen3.5-397B to select task-relevant and visually localizable evidence labels from the question, image, and reference answer, following the evidence label selection template in Fig. 9. These labels are then used to obtain candidate regions from SAM3 (Carion et al., 2025). We retain only candidates whose boxes support the selected labels, using the Evidence Box Verification in Fig. 10. This yields a verified evidence set $\mathcal{E}_k = \{(\ell_i, b_i)\}_{i=1}^{N_k}$ for sample k, where ℓ_i denotes an object label and $b_i = [x_{i1}, y_{i1}, x_{i2}, y_{i2}]$ denotes its bounding box.

Evidence-guided teacher response generation. Given the image, instruction, task type, reference answer, and verified evidence set, we sample K candidate responses from Qwen3.5-397B as the teacher model. All tasks follow a unified response protocol, with task-specific generation templates provided in Fig. 11. Each response contains a reasoning trace r enclosed by `<think>` and `</think>`, followed by a final answer a in `\boxed{}`. The reasoning trace interleaves natural language with structured evidence references serialized as `<box>` spans, where each span references an evidence item $(\ell_i, b_i) \in \mathcal{E}_k$.

Rejection sampling with answer and evidence verification. We retain only candidates that are both answer-correct and evidence-faithful. We first enforce the response format and require every `<box>` reference to match a verified evidence box, which prevents unsupported spatial claims. We also remove responses that reveal annotation provenance using the evidence provenance check in Fig. 14. We then apply task-specific answer verification: answer matching with semantic equivalence checking for VQA, exact matching for SC and OC, and IoU-based localization matching for VG and OD. Among valid candidates, we use Qwen3.5-397B to score each reasoning trace for evidence usage, factual consistency, clarity, and conciseness, and select the highest-scoring response. The answer equivalence check and Grounded Trace Scoring templates are shown in Figs. 12 and 13, respectively. Samples without valid candidates are discarded. Table 1 reports the task distribution.

Task	# Samples	Ratio
Visual Question Answering (VQA)	100,000	40.40%
Visual Grounding (VG)	59,436	24.01%
Scene Classification (SC)	40,662	16.43%
Object Counting (OC)	22,982	9.28%
Object Detection (OD)	24,448	9.88%
Total	**247,528**	**100.00%**

Table 1: Distribution of the RsEvi-248K SFT dataset.

3 Method

We present RsEvi-8B, a multimodal large language model for remote sensing reasoning that produces grounded visual evidence across diverse tasks. We first review SFT and GRPO in Sec. 3.1. The training procedure then consists of two stages: cold-start supervised fine-tuning in Sec. 3.2, and GRPO with evidence-aware credit assignment in Sec. 3.3.

3.1 Preliminaries

Supervised fine-tuning. Given a supervised dataset $\mathcal{D}_{\text{sft}} = \{(x, y)\}$, supervised fine-tuning optimizes the policy π_θ by minimizing the negative log-likelihood of the target response:

$$\mathcal{L}_{\text{sft}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{sft}}} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}). \tag{1}$$

Group relative policy optimization. For each input x, GRPO samples a group of G responses $\{y_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$, and each response receives a reward R_i. The group-relative advantage is computed by normalizing rewards within the sampled group:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G}) + \epsilon}. \tag{2}$$

The policy is then updated with a clipped objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\Big(\rho_{i,t}\hat{A}_i, \text{clip}\big(\rho_{i,t}, 1-\varepsilon, 1+\varepsilon\big)\hat{A}_i\Big)\Bigg] - \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big). \tag{3}$$

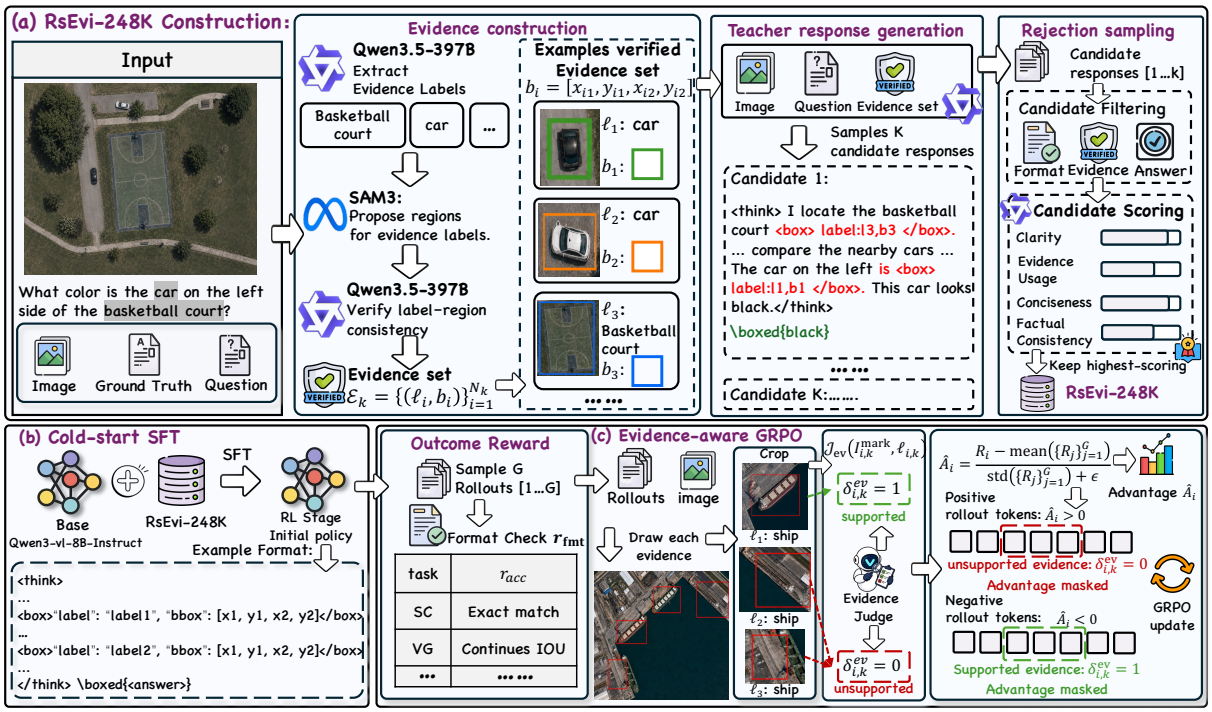


Figure 2: **The RsEvi pipeline.** (a) **RsEvi-248K construction** extracts evidence labels across tasks, verifies candidate regions, and filters teacher traces for answer correctness and evidence faithfulness. (b) **Cold-start SFT** initializes Qwen3-VL-8B-Instruct with grounded traces into a parsable evidence-aware policy. (c) **Evidence-aware GRPO** combines outcome rewards with evidence judgments during policy optimization. Crops show representative judgments, including supported evidence (ℓ_1), label–region mismatch (ℓ_2), and box–target misalignment (ℓ_3).

Here, $\rho_{i,t}$ is the token-level policy ratio between π_θ and $\pi_{\theta_{\text{old}}}$ for token $y_{i,t}$, and π_{ref} denotes the reference policy.

3.2 Cold-Start Initialization

Directly applying RL to a base multimodal policy is challenging for remote sensing reasoning. Before reward signals become informative, the model needs to learn the response format, task-specific output spaces, the normalized coordinate convention, and the placement of visual evidence within the reasoning trace. Early rollouts are therefore frequently unparsable or visually ungrounded, leaving the RL signal sparse and noisy.

We initialize RsEvi-8B with supervised fine-tuning on RsEvi-248K, which spans the same five tasks used in RL: scene classification, object counting, visual grounding, object detection, and VQA. The dataset separates intermediate reasoning from the final boxed answer, with evidence boxes embedded in the trace whenever applicable. This yields a parsable, evidence-aware starting policy, allowing the subsequent RL stage to focus on task accuracy and evidence faithfulness.

3.3 Evidence-Aware GRPO

In the RL stage, we optimize both final answer accuracy and evidence faithfulness. Standard GRPO assigns one response-level advantage to all tokens, which may reinforce unsupported evidence in correct rollouts or penalize valid evidence in incorrect ones. We therefore keep the reward outcome-based and use an external multimodal judge only to adjust evidence span advantages.

Outcome reward. We use a separate outcome reward for each task to score the final answer. For a rollout y with reference answer a and task type τ, the reward is defined as

$$R(y, a, \tau) = (1-\lambda) r_{\text{acc}}(y, a, \tau) + \lambda r_{\text{fmt}}(y). \quad (4)$$

The binary format reward $r_{\text{fmt}}(y)$ checks whether y follows the required response format, which encloses reasoning with `<think>` and `</think>`, places the final answer in a `\boxed{}` expression, and represents each evidence item with a `<box>` tag containing a label ℓ and normalized coordinates $[x_1, y_1, x_2, y_2] \in [0, 1000]^4$. The accuracy term r_{acc} uses normalized exact matching for scene classification, thresholded count matching for object

counting, continuous IoU scoring for visual grounding, and a soft set-level score with best-IoU and count consistency for object detection. For VQA, we use normalized answer matching and call a text-only semantic judge for unresolved cases, following the answer equivalence check in Fig. 12. We set $\lambda = 0.1$ in all experiments. Complete reward definitions are given in Appendix C.2.

Evidence-aware advantage masking. Outcome reward is a coarse credit signal: it verifies the final answer but not the visual faithfulness of the supporting evidence. As a result, correct answers may reinforce unsupported evidence, while supported evidence in incorrect responses may be penalized. To reduce this credit assignment noise, we keep the reward outcome-based and modify only the advantage on judged evidence spans.

For each generated evidence item $e_{i,k} = (b_{i,k}, \ell_{i,k})$ in response y_i, where $b_{i,k}$ is the predicted region and $\ell_{i,k}$ is its text label, we draw $b_{i,k}$ as a red bounding box on the image. We then crop the image around the box and enlarge the crop before judging, which improves the visibility of small objects in remote sensing images. Let $I^{\text{mark}}_{i,k}$ denote the resulting visual input. We use a fixed MLLM evidence judge $\mathcal{J}_{\text{ev}}$ to assign a support label $\delta^{\text{ev}}_{i,k} = \mathcal{J}_{\text{ev}}(I^{\text{mark}}_{i,k}, \ell_{i,k}) \in \{0, 1\}$, using the verification prompt in Fig. 10. Here, $\delta^{\text{ev}}_{i,k} = 1$ indicates that the highlighted region supports label $\ell_{i,k}$. Let $\mathcal{T}^{\text{ev}}_{i,k}$ denote the token span of evidence item $e_{i,k}$. We decompose the response advantage into $[\hat{A}_i]_+ = \max(\hat{A}_i, 0)$ and $[\hat{A}_i]_- = \min(\hat{A}_i, 0)$.

For tokens in a judged evidence span, we keep only the advantage component whose sign is consistent with the evidence judgment. In the expression below, k denotes the unique evidence item whose span contains token t:

$$\widetilde{A}_{i,t} = \begin{cases} \delta^{\text{ev}}_{i,k}[\hat{A}_i]_+ + (1 - \delta^{\text{ev}}_{i,k})[\hat{A}_i]_-, & t \in \mathcal{T}^{\text{ev}}_{i,k}, \\ \hat{A}_i, & t \notin \bigcup_k \mathcal{T}^{\text{ev}}_{i,k}. \end{cases} \tag{5}$$

Consequently, supported evidence under negative rollout advantage and unsupported evidence under positive rollout advantage receive zero evidence-span advantage, rather than being updated in the direction implied by the sequence-level outcome.

Policy update. The policy is optimized with the clipped GRPO objective in Eq. 3, with the per-token surrogate using $\widetilde{A}_{i,t}$ in place of $\hat{A}_i$. At each iteration, we sample G rollouts, compute the outcome reward R_i (Eq. 4) and the group relative advantage $\hat{A}_i$ (Eq. 2), query the judge on every generated evidence item to compute $\delta^{\text{ev}}_{i,k}$, and form the masked advantage $\widetilde{A}_{i,t}$ (Eq. 5). Since the judge only modulates evidence-span advantages, the reward remains outcome-based while encouraging visually faithful evidence generation.

4 Experiments

4.1 Experimental Setup

Tasks and benchmarks. We evaluate RsEvi-8B on a unified remote-sensing multimodal suite covering VG, OD, OC, SC, and VQA. The benchmarks include DIOR-RSVG (Zhan et al., 2023), RSVG (Sun et al., 2022b), VRSBench-VG (Li et al., 2024), and RRSIS-D (Liu et al., 2024) for VG; HRRSD (Zhang et al., 2019), RSOD (Long et al., 2017), and NWPU VHR-10 (Cheng et al., 2014) for OD; DOTA (Xia et al., 2018), RSVQA-HR (Lobry et al., 2020), HRRSD, and NWPU VHR-10 for OC; AID (Xia et al., 2017), SIRI-WHU (Zhao et al., 2015), WHU-RS19 (Dai and Yang, 2010), and UC-Merced (Yang and Newsam, 2010) for SC; and RSVQA-HR and VRSBench for VQA. See Appendix D.1 for more details.

Baselines. We compare RsEvi-8B with three groups of models: general open-source MLLMs, closed-source commercial MLLMs, and remote-sensing MLLMs. The general open-source baselines include GLM-4.6V-Flash (Hong et al., 2025), Gemma-4-31B-it (DeepMind, 2026), InternVL3.5-14B (Wang et al., 2025b), Qwen3.5-397B-A17B (Team, 2026), and Qwen3-VL-8B-Instruct (Bai et al., 2025). The closed-source baselines include Claude-Haiku-4.5 (Anthropic, 2025), Gemini-3-Flash (Google DeepMind, 2025), and GPT-5.4-mini (OpenAI, 2026). The open-source remote-sensing baselines include GeoChat (Kuckreja et al., 2024), SkySenseGPT (Luo et al., 2024), VHM (Pang et al., 2024), and EarthDial (Soni et al., 2025). All models are evaluated with temperature set to 0 using greedy decoding.

Training Details. RsEvi-8B is initialized from Qwen3-VL-8B-Instruct (Bai et al., 2025) and trained with SFT on RsEvi-248K followed by evidence-aware GRPO. The RL stage uses the multitask outcome reward with evidence-span advantages masked by box–label support. See Appendix C for more training details.

Model	DIOR-RSVG				RSVG				VRSBench-VG				RRSIS-D			
	@0.25	@0.5	@0.75	mIoU	@0.25	@0.5	@0.75	mIoU	@0.25	@0.5	@0.75	mIoU	@0.25	@0.5	@0.75	mIoU
Open-source MLLMs																
GLM-4.6V-Flash	63.68	49.83	32.40	47.42	34.47	21.92	9.37	22.67	68.96	54.40	29.12	48.59	63.37	49.56	31.83	46.99
Gemma-4-31B-it	61.80	44.93	23.51	42.57	34.15	13.77	2.53	17.95	69.46	50.11	23.93	46.34	61.32	43.87	22.47	42.00
InternVL3.5-14B	16.93	13.87	9.43	12.57	11.41	7.01	1.87	7.99	20.89	16.22	9.33	14.63	17.71	14.69	9.97	13.26
Qwen3.5-397B-A17B	64.31	53.71	37.51	49.92	49.31	**37.00**	**18.42**	**33.23**	73.22	61.46	34.67	53.27	64.87	53.54	38.37	50.16
Qwen3-VL-8B-Instruct	61.95	49.80	33.51	46.48	40.83	28.93	14.59	27.13	70.93	58.35	32.08	50.91	62.55	50.49	34.13	46.87
Close-source Commercial MLLMs																
Claude-Haiku-4.5	21.97	6.13	0.68	12.89	1.71	0.08	0.00	1.37	21.06	6.48	0.96	12.64	22.45	6.75	0.98	13.20
Gemini-3-Flash	26.63	12.88	5.28	16.91	19.40	12.22	6.03	12.57	22.59	11.27	3.86	14.13	26.58	13.90	5.80	17.23
GPT-5.4-mini	34.36	14.24	1.97	20.34	3.10	0.49	0.08	3.00	38.87	16.60	2.93	22.42	34.66	14.08	1.90	20.17
Open-source Remote Sensing MLLMs																
GeoChat	57.96	29.27	4.09	32.22	13.20	2.36	0.16	7.41	43.31	17.01	2.30	23.24	58.03	28.67	3.99	31.99
SkySenseGPT	41.80	17.31	2.08	23.20	5.46	0.49	0.08	3.48	31.07	9.24	0.66	17.12	41.05	17.05	1.77	22.67
VHM	66.75	51.15	21.11	44.10	3.26	0.57	0.08	2.21	47.39	23.83	4.11	26.89	65.63	50.21	21.23	43.52
Earthdial	59.93	40.05	15.68	37.67	19.15	6.11	0.57	10.30	47.73	22.93	6.18	27.55	59.12	39.82	15.57	37.41
RsEvi-8B	**84.24**	**78.69**	**62.81**	**69.56**	**51.02**	31.78	11.90	31.01	**85.80**	**78.35**	**56.81**	**68.30**	**85.41**	**80.41**	**64.31**	**70.93**

Table 2: Experimental results on Visual Grounding tasks. The best and second-best results are highlighted in **bold** and underlined, respectively.

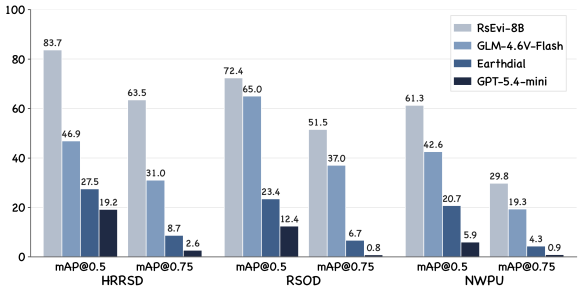


Figure 3: Performance Comparison of RsEvi-8B with State-of-the-Art Closed-Source, Open-Source, and Remote-Sensing MLLMs on Object Detection Tasks.

4.2 Main Results

Overall performance. Tables 2, 3, and 4, together with Figure 3, show that RsEvi-8B achieves strong results across the evaluation suite. It leads on most grounding, detection, counting, scene classification, and VQA metrics, while remaining competitive with much larger general MLLMs and outperforming remote-sensing MLLMs.

Visual grounding and object detection. We first examine whether RsEvi-8B can produce faithful evidence boxes. In visual grounding, Table 2 shows that RsEvi-8B ranks first on 13 out of 16 columns across the four benchmarks. Averaged over them, it improves Acc@0.75 by 16.72 points and mIoU by 13.31 points over the strongest average baseline, Qwen3.5-397B-A17B. The high-threshold gain is important because evidence boxes must be accurate enough to support the corresponding visual claim. In object detection, Figure 3 shows that RsEvi-8B achieves the best mAP@0.5 and mAP@0.75 on all three benchmarks among the baselines shown. Its average mAP@0.75 is 19.17 points higher than the average of the best baseline values shown in the figure. Together, these results show that RsEvi-8B grounds task-relevant objects in evidence boxes that are precise enough to be verified.

Scene classification and VQA. For scene classification, Table 3 shows that RsEvi-8B achieves the best or second-best result on all four benchmarks, with an average accuracy of 90.54 and a 3.21-point gain over the strongest average baseline, VHM. This suggests that grounding local visual claims can support scene semantics without weakening global image recognition. The advantage is larger on VQA, where RsEvi-8B ranks first on all seven reported columns, reaching an average accuracy of 82.09 and outperforming the strongest average baseline, Gemini-3-Flash, by 9.36 points. The stronger gains on presence, position, and category questions are consistent with our diagnostic finding that many VQA errors stem from grounding the answer in the wrong visual entity or relation.

Object counting. Table 4 shows that RsEvi-8B obtains the best accuracy on all four counting benchmarks: DOTA, RSVQA-HR, HRRSD, and NWPU VHR-10. Its average counting accuracy reaches 52.46, which is 7.17 points higher than the strongest average baseline, Qwen3.5-397B-A17B. The gain is especially clear on HRRSD, improving

Model	Scene Classification				HR VQA		VRSBench				
	AID	SIRI	WHU	UCM	Comp	Pre	Existence	Position	Category	Scene	Shape
Open-source MLLMs											
GLM-4.6V-Flash	71.90	66.08	86.90	74.48	68.50	63.90	86.38	62.27	63.73	66.56	67.96
Gemma-4-31B-it	76.43	66.29	92.20	85.57	74.40	70.00	83.55	61.87	55.10	66.50	65.57
InternVL3.5-14B	77.40	65.50	93.50	87.57	83.00	63.70	86.79	66.58	62.33	55.68	45.61
Qwen3.5-397B-A17B	76.60	67.00	91.50	82.19	72.60	67.70	85.77	65.36	58.04	60.87	70.41
Qwen3-VL-8B-Instruct	66.83	68.08	88.30	79.76	78.00	70.60	83.03	58.84	58.78	64.69	67.04
Close-source Commercial MLLMs											
Claude-Haiku-4.5	65.80	62.42	86.80	63.81	74.20	59.10	80.25	50.48	48.88	60.28	66.90
Gemini-3-Flash	81.07	74.12	94.40	92.10	77.30	71.40	86.96	65.99	65.90	71.10	70.48
GPT-5.4-mini	73.10	65.54	89.80	78.14	70.50	65.40	83.35	59.71	57.03	65.50	66.69
Open-source Remote Sensing MLLMs											
GeoChat	73.13	53.00	84.80	70.43	80.40	56.90	85.84	65.03	70.21	51.99	44.48
SkySenseGPT	87.83	67.58	95.50	90.19	84.00	65.20	85.38	63.85	72.84	53.99	47.43
VHM	92.00	70.50	96.40	90.43	82.60	67.80	86.90	51.97	67.96	62.78	58.89
Earthdial	87.73	60.50	95.90	**93.67**	81.80	59.10	82.08	58.25	64.96	51.17	42.59
RsEvi-8B	**94.93**	**75.79**	**98.00**	93.43	**87.50**	**91.80**	**91.27**	**72.87**	**82.55**	**73.51**	**75.12**

Table 3: Experimental results on Scene Classification and VQA tasks. The best and second-best results are highlighted in **bold** and underlined, respectively.

Model	DOTA		RSVQA-HR		HRRSD		NWPU	
	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE
Open-source MLLMs								
GLM-4.6V-Flash	27.50	6.8250	28.10	2.4194	65.20	0.8231	42.21	2.5066
Gemma-4-31B-it	34.17	3.7227	34.80	2.2649	61.06	0.8961	37.12	2.0029
InternVL3.5-14B	35.00	4.1583	28.70	2.5780	66.24	0.7012	40.32	2.2897
Qwen3.5-397B-A17B	36.67	4.1500	42.30	2.2530	65.21	0.6617	36.97	2.0218
Qwen3-VL-8B-Instruct	32.50	5.1917	42.10	2.3920	60.50	0.7517	40.61	1.8661
Close-source Commercial MLLMs								
Claude-Haiku-4.5	20.83	6.8750	19.40	2.8047	56.18	0.9743	36.54	2.4920
Gemini-3-Flash	38.33	**3.0333**	32.80	2.0201	67.33	1.6665	42.50	1.6943
GPT-5.4-mini	29.17	4.1917	28.90	2.5510	48.27	1.4786	33.92	2.1718
Open-source Remote Sensing MLLMs								
GeoChat	11.67	7.2167	33.30	4.1310	46.39	1.4993	31.15	3.3377
SkySenseGPT	21.67	4.4583	26.30	1.9444	57.16	0.8562	36.83	2.5240
VHM	27.50	6.5417	13.10	2.0280	62.68	0.7842	39.59	2.2445
Earthdial	26.67	5.8917	37.20	1.9190	61.93	0.8627	37.70	2.7394
RsEvi-8B	**43.33**	4.4750	**45.50**	**1.8360**	**76.90**	**0.3613**	**44.10**	**1.4076**

Table 4: Experimental results on Object Counting tasks.

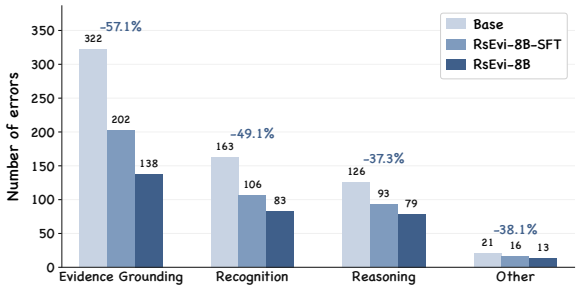


Figure 4: Diagnostic error reduction on 2,000 randomly sampled questions from RSVQA-HR and VRSBench. Percentages indicate the relative error reduction of RsEvi-8B compared with Base.

the best baseline accuracy from 67.33 to 76.90 and the best baseline MAE from 0.6617 to 0.3613. On DOTA, although RsEvi-8B does not achieve the lowest MAE, it still obtains the best exact-count accuracy. Overall, explicit evidence helps the model track visible instances before counting.

Diagnostic error reduction. We analyze errors on 2,000 randomly sampled questions from RSVQA-HR and VRSBench. We compare three models: Base denotes Qwen3-VL-8B-Instruct, RsEvi-8B-SFT denotes the checkpoint after supervised fine tuning on RsEvi-248K, and RsEvi-8B denotes the final model after evidence-aware GRPO. We classify incorrect predictions into the same four categories used in the introduction: Evidence Grounding, Recognition, Reasoning, and Other. Figure 4 reports the error counts, with percentages indicating the relative reduction of RsEvi-8B over Base. RsEvi-8B reduces all error types, with the largest drop in Evidence Grounding errors, from 322 to 138, or 57.1%. The share of Evidence Grounding errors also decreases, suggesting that the model is less likely to ground answers in the wrong visual evidence. Recognition errors show the second largest reduction at 49.1%. These results support our diagnostic finding that explicit grounded visual evidence mitigates a dominant failure mode in remote-sensing question answering.

4.3 Ablation Study

We ablate the two main components of RsEvi: SFT on RsEvi-248K and evidence-aware GRPO. Table 5 reports dataset-balanced task-family averages for VG, VQA, SC, OD, and OC. VG uses Acc@0.5, OD uses mAP@0.75, OC uses exact-count accu-

Model	VG	VQA	SC	OD	OC	Avg.
Qwen3-VL-8B-Instruct	46.9	70.0	75.7	19.8	43.9	51.3
RsEvi-8B-SFT	56.7	78.2	89.3	45.1	46.0	63.1
RsEvi-8B-SFT + GRPO	65.1	82.6	89.6	46.4	49.2	66.6
RsEvi-8B	**67.3**	**83.6**	**90.5**	**48.2**	**52.5**	**68.4**

Table 5: Ablation study with dataset-balanced task-family averages.

racy, and SC/VQA use accuracy. The final column averages the five task-family scores.

Effect of RsEvi-248K. RsEvi-8B-SFT improves the average score from 51.27 to 63.06 over Qwen3-VL-8B, showing that RsEvi-248K provides an effective cold-start initialization. The largest gain appears on OD, followed by SC, VG, and VQA. This shows that verified bbox-grounded traces provide an effective cold-start initialization, helping the model learn the response format, normalized coordinates, and evidence placement before RL.

Effect of evidence-aware GRPO. Starting from RsEvi-8B-SFT, standard GRPO raises the average score from 63.06 to 66.59. Evidence-aware GRPO further improves it to 68.42 by adjusting updates on generated evidence spans according to box–label support. The gain over standard GRPO is consistent across task families, largest on OC, followed by VG and OD. These results support evidence-aware credit assignment as a key component of RsEvi-8B. It improves tasks that depend on grounded visual evidence while benefiting scene understanding and question-conditioned reasoning.

4.4 Case Study

We provide qualitative examples in Figs. 5, 6, 7, and 8. The examples cover object counting, object detection, scene classification, VQA, and visual grounding. They show that RsEvi-8B uses explicit <box> spans to anchor task-relevant visual claims, such as airplane instances for counting, vehicle boxes for detection and comparison, scene cues for resort classification, and the reference vehicle and target basketball court for visual grounding. These cases illustrate how grounded evidence is used before producing the final answer.

5 Related Work

Grounded Reasoning in MLLMs. Grounded reasoning anchors language outputs to visual evidence by referring to image regions (Peng et al., 2024; Chen et al., 2023; You et al., 2023; Shao et al., 2024; Fan et al., 2025). Existing work spans grounded MLLMs that align regions with language and reasoning frameworks that interleave visual evidence with chain-of-thought. The former evolved from coordinate tokens (Peng et al., 2024) to natural-language coordinates (Chen et al., 2023) and free-form region grounding (You et al., 2023). Grounding-augmented reasoning frameworks use localized evidence as intermediate reasoning signals: Visual CoT (Shao et al., 2024) uses bounding boxes to guide later steps, Argus (Man et al., 2025) formalizes object-centric grounding over task-relevant regions, GRIT (Fan et al., 2025) interleaves box coordinates with textual reasoning.

Remote Sensing MLLMs. Remote sensing MLLMs adapt vision-language models to aerial and satellite imagery, addressing fine-grained categories, small objects, and geospatial context (Yao et al., 2025). GeoChat (Kuckreja et al., 2024) introduces region-level RS-grounded dialogue, while EarthGPT (Zhang et al., 2024) unifies multi-task instruction tuning. Subsequent works (Luo et al., 2024; Pang et al., 2024; Muhtar et al., 2024) target fine-grained scene reasoning, hallucination-aware instruction tuning, and scalable VGI-based data construction, respectively. GeoPixel (Shabbir et al., 2025) advances pixel-level grounding for high-resolution imagery with small objects. Reasoning-oriented methods progress from RS CoT with GRPO (Koksal and Alatan, 2025) and Geo-CoT supervision (Liu et al., 2025) to RL-based geospatial reasoning (Fiaz et al., 2025; Xu et al., 2025).

6 Conclusion

Remote sensing reasoning requires models to produce not only plausible answers, but also the visual evidence that supports their intermediate claims. Our diagnostic analysis shows that evidence grounding failures are a major bottleneck for current remote sensing MLLMs, especially in large-scale scenes with small, dense, and spatially subtle targets. To address this issue, we introduce RsEvi-8B, a remote sensing MLLM that performs Think with Grounding through parseable bbox-grounded evidence spans. We further optimize it with evidence-aware GRPO, which mitigates the mismatch between outcome-level rewards and evidence faithfulness by assigning credit at the evidence-span level. Across diverse remote-sensing tasks, explicit evidence grounding improves performance and reasoning faithfulness.

Limitations

We note two limitations of this work. First, the construction of grounded reasoning data relies on external models for evidence extraction, verification, and teacher response generation. Although filtering and rejection sampling reduce low-quality samples, errors from these models may still affect the SFT data. Second, the evidence-aware RL stage uses an external judge to assign binary support labels to generated evidence spans. Although we crop and enlarge highlighted regions before judging, these judgments may still be noisy when the evidence is ambiguous.

Ethical Considerations

All experiments in this work are conducted on publicly available datasets and benchmarks. We do not collect private data or involve human subjects. As with prior work using existing remote sensing datasets, model behavior may be affected by annotation noise, dataset bias, or limited coverage.

References

Anthropic. 2025. System card: Claude haiku 4.5. https://www.anthropic.com/claude-haiku-4-5-system-card. Official system card. Accessed: 2026-05-25.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman R"adle, and 2 others. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883.

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

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. 2018. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180.

Dengxin Dai and Wen Yang. 2010. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geoscience and remote sensing letters*, 8(1):173–176.

Google DeepMind. 2026. Gemma 4. https://deepmind.google/models/gemma/gemma-4/. Official model page. Accessed: 2026-05-25.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. 2025. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*.

Mustansar Fiaz, Hiyam Debary, Paolo Fraccaro, Danda Paudel, Luc Van Gool, Fahad Khan, and Salman Khan. 2025. Geovlm-r1: Reinforcement fine-tuning for improved remote sensing reasoning. *arXiv preprint arXiv:2509.25026*.

Google DeepMind. 2025. Gemini 3 Flash Model Card. https://deepmind.google/models/model-cards/gemini-3-flash/. Official model page. Accessed: 2026-05-25.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Aybora Koksal and A. Aydin Alatan. 2025. Tinyrs-r1: Compact multimodal language model for remote sensing. *arXiv preprint arXiv:2505.12099*.

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

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

Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. 2025. Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. *arXiv preprint arXiv:2509.22221*.

Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. 2024. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26658–26668.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. 2020. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566.

Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. 2017. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, and 1 others. 2024. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. 2024. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *Computer Vision – ECCV 2024*, pages 440–457. Springer.

OpenAI. 2026. Introducing gpt-5.4. `https://openai.com/index/introducing-gpt-5-4/`. Accessed: 2026-05-20.

OpenAI. 2026. Introducing gpt-5.4 mini and nano. `https://openai.com/index/introducing-gpt-5-4-mini-and-nano/`. Official model announcement. Accessed: 2026-05-25.

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

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, volume 2024, pages 51575–51598.

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

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*.

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, and 1 others. 2025. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14303–14313.

Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, and 1 others. 2022a. 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.

Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. 2022b. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International conference on Multimedia*, pages 404–412.

Qwen Team. 2026. Qwen3. 5: Accelerating productivity with native multimodal agents.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and 1 others. 2025a. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

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

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. 2017. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981.

Chenhui Xu, Fuxun Yu, Michael J. Bianco, Jacob Kovarskiy, Raphael Tang, Qi Zhang, Zirui Xu, Will LeVine, Brandon Dubbs, Heming Liao, Cassandra Burgess, Suvam Bag, Jay Patravali, Rupanjali Kukal, Mikael Figueroa, Rishi Madhok, Nikolaos Karianakis, and Jinjun Xiong. 2025. Geo-r1: Unlocking vlm geospatial reasoning with cross-view reinforcement learning. *arXiv preprint arXiv:2510.00072*.

Yi Yang and Shawn Newsam. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279.

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

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. 2023. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE transactions on geoscience and remote sensing*, 61:1–13.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. 2024. 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.

Yuanlin Zhang, Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu. 2019. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5535–5548.

Bei Zhao, Yanfei Zhong, Gui-Song Xia, and Liangpei Zhang. 2015. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 54(4):2108–2123.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. 2025. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pages 400–410.

Bryan Zhu, Nicholas Lui, Jeremy Irvin, Jimmy Le, Sahil Tadwalkar, Chenghao Wang, Zutao Ouyang, Frankie Y Liu, Andrew Y Ng, and Robert B Jackson. 2022. Meter-ml: A multi-sensor earth observation benchmark for automated methane source mapping. *arXiv preprint arXiv:2207.11166*.

A Diagnostic Error Analysis

We perform diagnostic error classification on the incorrect predictions made by Qwen3-VL-8B-Instruct on 2,000 randomly sampled questions from RSVQA-HR and VRSBench. We use GPT-5.4 (OpenAI, 2026) as the multimodal judge with temperature set to 0, and ask it to assign exactly one primary error type following the error attribution template in Fig. 15. The template uses four categories. *Evidence Grounding* denotes failures to ground the answer in the correct visual evidence, such as missing the relevant object, attending to an incorrect or adjacent region, or using only partial evidence. *Recognition* denotes cases where the model attends to the correct or a highly plausible region but misidentifies its category, scene cue, or attribute. *Reasoning* denotes cases where the relevant visual evidence is mostly correct but the model makes an error in counting, comparison, spatial relation, or aggregation. *Other* covers formatting errors, synonym or label mismatches, annotation ambiguity, unclear questions, and cases whose primary cause cannot be reliably determined.

B Dataset Construction

B.1 SFT Dataset Construction Details

Overview. In this section, we provide additional details on the construction of RsEvi-248K. We build the dataset from public remote-sensing resources covering visual question answering, visual grounding, scene classification, object detection, and object counting (Lobry et al., 2020; Li et al., 2024; Pang et al., 2024; Zhan et al., 2023; Zhang et al., 2019; Cheng et al., 2017; Xia et al., 2017; Christie et al., 2018; Yang and Newsam, 2010; Zhu et al., 2022; Sun et al., 2022a; Xia et al., 2018). Qwen3.5-397B-A17B (Team, 2026) is used as the teacher model for label extraction, response generation, answer verification, and reasoning quality assessment. All spatial annotations are normalized into a shared $[0, 1000]$ coordinate space. Table 6 summarizes the source composition of the final SFT set. The training and evaluation sets are non-overlapping.

Data sources and task normalization. We first convert all source datasets into a unified format containing an image, a task label, a user instruction, a reference answer, and optional object-level annotations. VRSBench contributes both VQA and visual grounding samples, while caption-only samples are excluded. RSVQA-HR is used for VQA-style question answering, and DIOR-RSVG provides referring-expression grounding examples. HRRSD provides object detection annotations and also contributes object counting samples. Scene classifi-

cation data is collected from NWPU-RESISC45, AID, fMoW, UCMerced, and METER-ML. VersaD contributes object-centric VQA samples over FAIR1M and DOTA imagery. For all localization tasks, boxes are normalized to $[0, 1000]$ and represented in `xyxy` format.

Object-level evidence construction. For samples with native object or grounding annotations, we directly use the provided boxes as candidate evidence after coordinate normalization. For other samples without object-level evidence, we prompt Qwen3.5-397B-A17B to extract task-relevant and visually localizable object labels from the question, image, and reference answer, following the evidence label selection template in Fig. 9. We use these labels as text prompts for SAM3 (Carion et al., 2025) to obtain candidate regions. We retain SAM3 proposals whose labels match the extracted label set and whose confidence scores are at least 0.55, then normalize their boxes to the shared coordinate system and discard boxes with degenerate geometry. To reduce noisy over-segmentation, labels that produce more than 20 proposals are filtered before teacher verification. The remaining boxes are verified by the teacher model, and only boxes that pass verification are retained, using the evidence box verification template in Fig. 10. This yields a verified evidence set $\mathcal{E}_i = \{(\ell_j, b_j)\}_{j=1}^{N_i}$ for sample i, where ℓ_j is an object label and b_j is its normalized bounding box.

Evidence-guided teacher response generation. Given the image, instruction, task type, reference answer, and verified evidence set, we sample $K = 8$ teacher responses with Qwen3.5-397B-A17B using the Grounded Trace Generation template in Fig. 11, instantiated with task-specific answer requirements and examples. Each response follows a unified protocol: the reasoning trace is enclosed by `<think>` and `</think>`, evidence items are serialized as `<box>` spans, and the final answer is placed in a `\boxed{}` expression. The final answer format is task-dependent: short text for VQA, a class name for scene classification, an integer for object counting, a single box for visual grounding, and a list of boxes for object detection.

Rejection sampling with answer and evidence verification. We retain only responses with correct answers and faithful evidence use. First, we enforce the response format and require each evidence box mentioned in the reasoning trace to match the verified evidence set. We also discard responses that reveal annotation provenance or auxiliary hints. This leakage filter combines rule-based screening with a separate teacher-based leakage check that identifies cases where the response relies on annotation provenance rather than visual evidence, with the evidence provenance check shown in Fig. 14. We then apply task-specific answer checks: normalized semantic matching for VQA, exact matching for scene classification and object counting, and IoU-based matching for visual grounding and object detection with an IoU threshold of 0.75. Among valid candidates, the teacher model scores the reasoning trace along five axes: factual consistency, evidence usage, hallucination avoidance, clarity, and conciseness. We keep the highest-scoring candidate for SFT. The answer equivalence check and Grounded Trace Scoring templates are shown in Figs. 12 and 13, respectively.

Deduplication and resampling. After filtering, we remove exact duplicates from the SFT candidate set. To keep the final SFT set compact while preserving broad task coverage, we cap VQA at 100,000 examples, keep visual grounding, scene classification, and object counting, and upsample OD by a factor of two. In practice, this repeats the HRRSD object detection subset once in the final training set. This produces 247,528 SFT examples in total. The final task distribution is reported in Table 1 in the main text, and the source-level composition is shown in Table 6.

B.2 RL Dataset Construction

The RL training set is built from the same normalized remote-sensing sources and task families as RsEvi-248K. Unlike SFT, it uses verifiable reference answers rather than teacher responses, since supervised reasoning targets are not required for RL. Starting from the same source selection and task weights, we form a more balanced RL mixture by reducing the dominance of visual question answering and increasing the relative coverage of object detection and object counting. The final RL training set contains 153,600 examples, including 132,859 unique image and instruction pairs. Its task distribution is shown in Table 7.

Task	Source	# Samples	Training Source / Split
VQA	RSVQA-HR	47,635	Train split
	VRSBench	34,053	Train split
	VersaD	18,312	Train split
Visual Grounding	VRSBench	34,727	Train split
	DIOR-RSVG	24,709	Train split
Scene Classification	NWPU-RESISC45	25,196	Train split
	AID	6,979	Train split
	fMoW	4,910	Train split
	UCMerced	2,097	Train split
	METER-ML	1,480	Train split
Object Detection	HRRSD	24,448	Train/val split
Object Counting	VRSBench	13,419	Train split
	HRRSD	6,366	Train/val split
	FAIR1M	2,138	Train split
	DOTA	1,059	Train split
Total	–	**247,528**	–

Table 6: Source composition of RsEvi-248K. The training and evaluation sets are non-overlapping.

Task	# Samples	Ratio
Visual Question Answering (VQA)	28,125	18.31%
Visual Grounding (VG)	36,337	23.66%
Scene Classification (SC)	32,933	21.44%
Object Counting (OC)	28,247	18.39%
Object Detection (OD)	27,958	18.20%
Total	**153,600**	**100.00%**

Table 7: Distribution of the RL training dataset.

C Training Details

C.1 Training setup.

In this section, we describe the training details of RsEvi-8B. All training is conducted on 4 NVIDIA H200 GPUs. The model is trained in two stages: supervised fine tuning (SFT) followed by reinforcement learning (RL) with evidence-aware GRPO. Both stages adopt Qwen3-VL-8B-Instruct (Bai et al., 2025) as the backbone. The SFT and RL stages are implemented based on LLaMA-Factory (Zheng et al., 2024) and EasyR1 (Zheng et al., 2025), respectively.

Supervised fine tuning. For the SFT stage, we perform full parameter fine tuning on the language model and the multimodal projector while keeping the vision tower frozen. The SFT corpus comprises 247,528 samples spanning five task families: visual question answering, visual grounding, scene classification, object detection, and object counting. We train for one epoch and use the resulting checkpoint to initialize the RL policy. Key SFT hyperparameters are reported in Table 8.

Reinforcement learning. For the RL stage, we extend the GRPO implementation in EasyR1 with evidence-aware supervision. Following DAPO (Yu et al., 2026), we use token level policy gradient loss averaging and clip-higher, with clipping factors $\epsilon_l = 0.2$ and $\epsilon_h = 0.28$, and disable the reference KL term ($\beta = 0$). The scalar reward is the multitask outcome reward defined in Appendix C.2, with the format weight set to $\lambda = 0.1$. We employ Qwen-3-VL-30B-A3B (Bai et al., 2025) as the evidence judge. It uses the Evidence Box Verification prompt in Fig. 10 to verify whether each generated evidence box supports its text label in the rendered image region. To focus the judge on local visual cues, we render a crop around each candidate box with a minimum padding of 64 pixels and a padding ratio of 0.5. The judge uses the Evidence Box Verification in Fig. 10. Other RL hyperparameters are summarized in Table 9. More implementation details can be found in the EasyR1 codebase.

Training overhead. We measure the additional cost of evidence-aware GRPO on 4 NVIDIA H200 GPUs. Both runs use the same SFT checkpoint, RL data, rollout batch size of 384, and actor batch size of 128. We average `perf.time_per_step` over RL training logs, excluding SFT time and the validation record at step 0. As shown in Table 10, evidence-aware GRPO takes 209.7 seconds per step, compared with 182.7 seconds for standard GRPO. The 14.8% increase indicates that evidence verification does not add substantial training cost.

C.2 Reward Details

Notation. For a rollout y, let $\hat{a}$ denote the model's final answer extracted from y. Given a reference answer a and task type $\tau \in \{\mathrm{SC}, \mathrm{OC}, \mathrm{VG}, \mathrm{VQA}, \mathrm{OD}\}$, the outcome reward combines a task specific accuracy term with a binary format term:

$$R(y, a, \tau) = (1 - \lambda)\, r^{\tau}_{\mathrm{acc}}(y, a) + \lambda\, r_{\mathrm{fmt}}(y). \quad (6)$$

The accuracy term r^{τ}_{acc} takes a task specific form, which we define in the remainder of this section.

Format reward. The format reward enforces compliance with the prescribed reasoning and an-

Stage	lr	Epochs	Max Len.	Batch	Opt.	Sched.	Warmup	WD	Freeze VT
SFT	1e-5	1	4096	128	AdamW	Cosine	0.03	0.01	True

Table 8: Hyperparameters for supervised fine tuning. Max Len. denotes the maximum sequence length, Batch denotes the training batch size, Opt. denotes the optimizer, Sched. denotes the learning rate scheduler, WD denotes weight decay, and Freeze VT indicates whether the vision tower is frozen.

Algorithm	lr	Max Len.	Steps	Opt.	Rollout BS	n	Actor BS	Temp.	top-p/top-k
GRPO	1e-6	4096 / 2048	400	AdamW	384	8	128	1	1.0 / -1

Table 9: Hyperparameters for reinforcement learning. Max Len. denotes the maximum prompt/response lengths, Opt. denotes the optimizer, Rollout BS and Actor BS denote the rollout and actor batch sizes, respectively, and n denotes the number of rollouts per prompt.

RL setting	Hardware	Time / step
Standard GRPO	4 H200	182.7 s
Evidence-aware GRPO	4 H200	209.7 s (↑14.8%)

Table 10: Training overhead of evidence-aware GRPO on 4 NVIDIA H200 GPUs.

swer template:

$$r_{\text{fmt}}(y) = \mathbf{1}[y \in \mathcal{T}_{\text{fmt}} \ \wedge \ \text{ValidBox}(y)], \quad (7)$$

where $\mathcal{T}_{\text{fmt}}$ denotes the required `<think>` reasoning template terminated by a `\boxed{}` answer, and $\text{ValidBox}(y)$ verifies that all evidence boxes are well formed.

Scene classification. We adopt normalized exact matching:

$$r_{\text{acc}}^{\text{SC}}(y, a) = \mathbf{1}[\nu(\hat{a}) = \nu(a)], \quad (8)$$

where $\nu(\cdot)$ lowercases the text and collapses consecutive whitespace.

Object counting. Let $\hat{c}$ and $c^\star$ denote the predicted and reference counts. For small reference counts ($c^\star \leq 2$), we require an exact match, $r_{\text{acc}}^{\text{OC}}(y, a) = \mathbf{1}[\hat{c} = c^\star]$. For larger counts, we use a clipped relative score, $r_{\text{acc}}^{\text{OC}}(y, a) = \max(0,\, 1 - |\hat{c} - c^\star| / \max(|\hat{c}|, |c^\star|))$.

Visual grounding. Let $\hat{\mathcal{B}}$ and $\mathcal{B}$ denote the predicted and reference grounding boxes. We use a continuous IoU reward:

$$r_{\text{acc}}^{\text{VG}}(y, a) = \max_{\hat{b} \in \hat{\mathcal{B}},\, b \in \mathcal{B}} \text{IoU}(\hat{b}, b). \quad (9)$$

Visual question answering. For VQA, we first attempt deterministic normalized matching and invoke a semantic judge only when the rule based comparison is inconclusive. Let $\mu(\cdot)$ denote the VQA text normalizer and $\mathcal{Y} = \{\texttt{yes}, \texttt{no}\}$ the yes/no answer set. The reward $r_{\text{acc}}^{\text{VQA}}(y, a, q)$ equals 1 if $\mu(\hat{a}) = \mu(a)$. It equals 0 if $\mu(\hat{a})$ and $\mu(a)$ both lie in $\mathcal{Y}$ but disagree. Otherwise, it falls back to $J(q, a, \hat{a}) \in \{0, 1\}$, the output of a semantic equivalence judge conditioned on the question q.

Object detection. Let $\hat{\mathcal{B}} = \{\hat{b}_i\}_{i=1}^{n}$ and $\mathcal{B} = \{b_j\}_{j=1}^{m}$ denote the predicted and reference box sets, respectively. We use a soft set level reward that linearly combines, with equal weight $w_{\text{od}} = 0.5$, a per reference best IoU term $s_{\text{iou}} = \frac{1}{m} \sum_{j=1}^{m} \max_{1 \leq i \leq n} \text{IoU}(\hat{b}_i, b_j)$ that rewards localization accuracy, and a count consistency term $s_{\text{cnt}} = \min(n, m) / \max(n, m)$ that penalizes both overprediction and underprediction:

$$r_{\text{acc}}^{\text{OD}}(y, a) = w_{\text{od}}\, s_{\text{iou}} + (1 - w_{\text{od}})\, s_{\text{cnt}}. \quad (10)$$

D Evaluation Details and Additional Results

D.1 Evaluation Details

We evaluate RsEvi-8B and all baselines on a suite of remote-sensing benchmarks that cover grounded evidence extraction, scene-level understanding, object counting, and visual question answering. Table 11 summarizes the benchmark setting, including task types, sample numbers, and evaluation metrics.

Evaluation Datasets. The benchmarks jointly test object-level localization, countable evidence, scene semantics, and question-conditioned remote-sensing understanding.

- **NWPU VHR-10** (Cheng et al., 2014) is a ten-class very-high-resolution object detec-

Benchmark	Task	Samples	Metric
NWPU VHR-10 (Cheng et al., 2014)	OD OC	1,601 687	mAP@0.5, mAP@0.75 Acc, MAE
RSOD (Long et al., 2017)	OD	936	mAP@0.5, mAP@0.75
DOTA (Xia et al., 2018)	OC	120	Acc, MAE
HRRSD (Zhang et al., 2019)	OD OC	14,049 14,049	mAP@0.5, mAP@0.75 Acc, MAE
RSVG (Sun et al., 2022b)	VG	1,227	Acc@0.25/0.5/0.75, mIoU
DIOR-RSVG (Zhan et al., 2023)	VG	7,500	Acc@0.25/0.5/0.75, mIoU
RRSIS-D (Liu et al., 2024)	VG	7,720	Acc@0.25/0.5/0.75, mIoU
UC-Merced (Yang and Newsam, 2010)	SC	2,100	Acc
WHU-RS19 (Dai and Yang, 2010)	SC	1,000	Acc
SIRI-WHU (Zhao et al., 2015)	SC	2,400	Acc
AID (Xia et al., 2017)	SC	3,000	Acc
RSVQA-HR (Lobry et al., 2020)	OC VQA	1,000 2,000	Acc, MAE Acc
VRSBench (Li et al., 2024)	VG VQA	16,159 37,409	Acc@0.25/0.5/0.75, mIoU Acc

Table 11: Evaluation benchmarks, task types, sample numbers, and metrics used in our experiments. VG, OD, OC, SC, and VQA denote visual grounding, object detection, object counting, scene classification, and visual question answering, respectively.

tion dataset. We use it for object detection and object counting, and report additional detection splits by instance number and object scale.

- **RSOD** (Long et al., 2017) contains representative aerial object categories such as aircraft, oil tanks, overpasses, and playgrounds. We use it to evaluate object detection on compact but canonical remote-sensing targets.
- **DOTA** (Xia et al., 2018) is a large-scale aerial object detection dataset with dense objects and strong scale and orientation variation. We use a DOTA-derived split for object counting.
- **HRRSD** (Zhang et al., 2019) is a high-resolution remote-sensing object detection dataset with large-scale annotations. We use it for both object detection and object counting.
- **RSVG** (Sun et al., 2022b) is a remote-sensing visual grounding benchmark built around geospatial referring expressions. We use it to evaluate whether language descriptions can be localized to image regions.
- **DIOR-RSVG** (Zhan et al., 2023) extends DIOR imagery with expression-box annotations for referring object localization. We use it for visual grounding.
- **RRSIS-D** (Liu et al., 2024) provides referring expressions for remote-sensing objects with diverse scales and orientations. We convert its annotations into box-level targets for visual grounding.
- **UC-Merced** (Yang and Newsam, 2010) is a 21-class land-use scene classification dataset. We use it to evaluate image-level remote-sensing scene recognition.
- **WHU-RS19** (Dai and Yang, 2010) is a 19-class high-resolution satellite scene classification benchmark. We use it for scene-level category recognition.
- **SIRI-WHU** (Zhao et al., 2015) contains 2,400 remote-sensing scene images across 12 land-use categories. We use it for scene classification.
- **AID** (Xia et al., 2017) is a large-scale aerial scene classification benchmark with 30 categories and diverse imaging conditions. We use it for scene classification.
- **RSVQA-HR** (Lobry et al., 2020) evaluates visual question answering over high-resolution remote-sensing images. We use its comparison, presence, and counting questions for visual question answering and object counting.
- **VRSBench** (Li et al., 2024) is a versatile remote-sensing vision-language benchmark covering captions, object references, and question-answer pairs. We use its visual grounding and visual question answering splits.

Metric Definitions. For visual grounding, Acc@t denotes the percentage of predictions whose intersection-over-union with the ground-truth box is at least threshold t, and mIoU is the mean intersection-over-union over all evaluated samples. For object detection, mAP@t is the mean average precision computed at IoU threshold t over object categories. For scene classification and visual question answering, Acc is the normalized answer accuracy. For object counting, Acc measures exact numerical correctness after answer normalization, while MAE measures the mean absolute difference between the predicted and ground-truth counts. All metrics are reported as percentages except MAE.

D.2 Additional Results

Table 12 provides additional object detection results on the five NWPU VHR-10 splits. Single and

Model	Single		Multiple		Small		Medium		Large	
	@0.5	@0.75	@0.5	@0.75	@0.5	@0.75	@0.5	@0.75	@0.5	@0.75
Open-source MLLMs										
GLM-4.6V-Flash	46.74	23.17	35.66	13.01	36.27	12.67	43.06	18.59	49.44	30.63
Gemma-4-31B-it	45.74	15.25	23.30	3.80	30.45	6.90	36.63	10.57	46.27	17.80
InternVL3.5-14B	30.69	11.03	39.81	13.16	25.12	4.49	33.52	11.38	45.09	20.14
Qwen3.5-397B-A17B	52.96	25.98	43.94	14.83	38.71	10.85	50.64	21.95	58.75	34.88
Qwen3-VL-8B-Instruct	46.64	22.67	16.96	5.40	29.44	6.56	36.94	16.08	38.94	26.00
Close-source Commercial MLLMs										
Claude-Haiku-4.5	1.50	0.00	0.67	0.00	0.94	0.00	0.93	0.00	2.15	0.00
Gemini-3-Flash	9.23	3.11	9.55	2.71	7.62	2.08	8.82	2.83	12.55	4.96
GPT-5.4-mini	8.43	1.40	1.83	0.09	2.47	0.04	5.24	0.23	7.88	2.39
Open-source Remote Sensing MLLMs										
GeoChat	4.01	0.00	2.09	0.11	1.56	0.01	3.14	0.07	5.63	0.00
SkySenseGPT	3.61	0.20	1.71	0.06	0.72	0.00	2.39	0.15	6.68	0.30
VHM	13.04	1.20	2.90	0.20	7.55	1.06	7.99	0.65	14.36	1.00
Earthdial	28.18	6.62	8.34	0.51	15.93	2.75	19.40	4.41	29.52	5.83
RsEvi-8B	**65.70**	**36.01**	**54.08**	**19.43**	**52.30**	**16.52**	**62.33**	**29.31**	**68.79**	**46.87**

Table 12: Results on the five NWPU object detection splits. The best and second-best results are highlighted in **bold** and underlined, respectively.

Model	Overall	Quant	Q-MAE	Color	Image	Size	Reason	Direct	Rural
Open-source MLLMs									
GLM-4.6V-Flash	68.48	47.29	0.9938	71.10	82.55	52.72	72.28	60.17	83.39
Gemma-4-31B-it	65.79	44.24	1.1074	78.79	83.17	47.48	71.18	58.49	84.75
InternVL3.5-14B	66.25	53.59	0.7733	64.31	73.60	45.10	68.40	46.33	82.03
Qwen3.5-397B-A17B	67.31	46.97	0.9491	77.24	81.75	53.51	73.28	57.44	82.71
Qwen3-VL-8B-Instruct	65.61	44.24	1.1541	78.20	83.17	52.82	70.18	51.57	84.41
Close-source Commercial MLLMs									
Claude-Haiku-4.5	61.18	39.10	1.1915	72.14	81.93	49.46	64.30	50.10	86.44
Gemini-3-Flash	70.98	51.95	1.2198	78.28	86.54	54.20	73.28	**61.84**	84.41
GPT-5.4-mini	65.47	41.97	1.1964	**80.51**	81.58	54.10	69.18	57.65	**86.78**
Open-source Remote Sensing MLLMs									
GeoChat	59.69	25.76	3.1311	44.45	61.82	43.42	62.08	42.14	80.00
SkySenseGPT	62.57	39.58	1.2954	51.49	62.18	45.20	64.08	44.65	79.32
VHM	65.89	47.46	1.0051	71.46	87.42	49.85	65.85	46.54	83.73
Earthdial	60.94	41.32	1.4245	62.85	68.47	46.49	54.66	31.45	72.20
RsEvi-8B	**77.49**	**59.30**	**0.6030**	78.70	**92.83**	**59.64**	**75.61**	60.17	**86.78**

Table 13: Additional results on VRSBench tasks omitted from Table 3. Quant, Q-MAE, Reason, and Direct denote Quantity, Quantity-MAE, Reasoning, and Direction, respectively. The best and second-best results are highlighted in **bold** and underlined, respectively.

Multiple separate images according to whether the queried category appears as one instance or multiple instances, while Small, Medium, and Large group samples by target scale. These splits complement the aggregate object detection results by making scale sensitivity and crowded-instance localization explicit.

Table 13 reports the remaining VRSBench tasks omitted from Table 3. These tasks include the overall VRSBench score as well as quantity, quantity MAE, color, image-level recognition, size, reasoning, direction, and rural-or-urban recognition. Together with the main VRSBench columns in Table 3, the appendix table gives a more complete view of model behavior across both perceptual and reasoning-oriented remote-sensing question types.

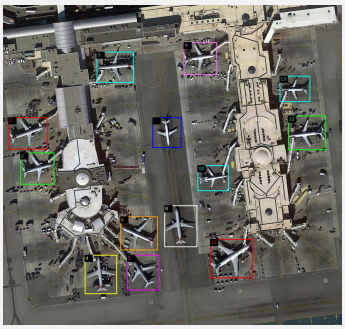
E Case Study

We provide qualitative examples in Figs. 5, 6, 7, and 8. These cases cover object counting, object detection, scene classification, visual question answering, and visual grounding. They illustrate how RsEvi-8B anchors intermediate visual claims to explicit <box> evidence before producing the final answer.

F Prompt Templates

We provide the templates used throughout the diagnostic, data construction, filtering, and optimization pipeline. Figs. 9 and 10 define evidence construction; Fig. 11 defines grounded trace generation; Figs. 12, 13, and 14 define response verification and filtering; and Fig. 15 defines the VQA error attribution procedure used in the diagnostic analysis.

Case Study 1:



Problem: How many airplanes are there in the image? Provide your answer as a single integer.

Output.

<think>

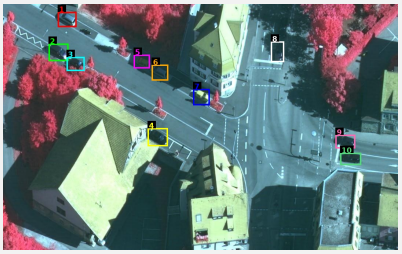
I scan the airport apron systematically to count all visible airplanes, using the beige terminal buildings, jet bridges, and gray taxi lanes as visual landmarks. Starting from the left side, I find one airplane parked near the terminal gate on the left apron: <box>"label":"airplane", [16, 351, 131, 451]</box>. Moving right along the same row, another airplane is visible beside the concourse and service vehicles: <box>"label":"airplane", [51, 461, 151, 561]</box>. Further up near the top left, next to a rectangular terminal block, I spot a third airplane: <box>"label":"airplane", [276, 146, 386, 243]</box>. Below that, in the lower left area near the round terminal section, there is a fourth airplane: <box>"label":"airplane", [241, 781, 336, 903]</box>. To its right, a fifth airplane is parked on the lower apron: <box>"label":"airplane", [366, 781, 461, 891]</box>. Above that one, a sixth airplane is visible near the lower middle-left gate lane: <box>"label":"airplane", [346, 661, 456, 766]</box>. In the center, along the broad vertical runway/taxiway strip, I see a seventh airplane: <box>"label":"airplane", [441, 351, 526, 446]</box>. Below the center, an eighth airplane is parked with its tail toward the bottom of the image: <box>"label":"airplane", [476, 626, 573, 756]</box>. Moving to the upper right section, close to the right-side terminal finger, I find a ninth airplane: <box>"label":"airplane", [533, 116, 631, 218]</box>. Below that, a tenth airplane is visible on the same terminal-side apron: <box>"label":"airplane", [574, 501, 666, 581]</box>. Further right and lower, an eleventh airplane is parked near service roads and gate equipment: <box>"label":"airplane", [611, 731, 736, 853]</box>. Above that, a twelfth airplane is visible on the far right apron: <box>"label":"airplane", [841, 346, 953, 453]</box>. To its left, a thirteenth airplane is parked with wings spanning across the dark pavement: <box>"label":"airplane", [816, 221, 906, 303]</box>. This completes the count of all airplanes in the image.

</think>

$\boxed{13}$ ← Correct

Figure 5: Qualitative example of RsEvi-8B on remote-sensing object counting. The model localizes airplane evidence across the airport scene, and then uses this grounded visual evidence to derive the final count.

Case Study 2:



Problem: {Object Detection} Detect all vehicle in the image and return their bounding boxes. Output ONLY a JSON bbox list: `[[xmin,ymin,xmax,ymax],...]`.

Output.

<think>
I scan the overhead street scene to detect all vehicles, focusing on the paved roads between the green-roofed buildings and bright pink tree canopies while avoiding shadows and road markings. Starting from the upper left, near the curb and the edge of the intersection, I find a dark vehicle on the road: <box>"label":"vehicle", [138, 31, 186, 93]</box>. Moving down along the same diagonal road, partly beside the pink trees, I spot another dark vehicle: <box>"label":"vehicle", [115, 161, 165, 231]</box>. Further along that same road segment, there is another dark vehicle slightly to the lower right: <box>"label":"vehicle", [161, 214, 206, 271]</box>. In the lower central-left area near the corner of the large green-roofed building, I detect a dark vehicle: <box>"label":"vehicle", [366, 503, 417, 577]</box>. Further right along the upper road, I see a yellow vehicle: <box>"label":"vehicle", [331, 206, 371, 261]</box>, followed by another dark vehicle near the road bend: <box>"label":"vehicle", [376, 246, 418, 311]</box>, and a yellow vehicle closer to the central intersection: <box>"label":"vehicle", [481, 345, 523, 411]</box>. Near the top right intersection, on the vertical lane beside the building shadow, there is a dark vehicle: <box>"label":"vehicle", [678, 151, 711, 236]</box>. Moving to the lower right road, where the street runs horizontally past the building edge, I find two more dark vehicles: <box>"label":"vehicle", [841, 531, 891, 586]</box> and <box>"label":"vehicle", [853, 604, 906, 653]</box>.
</think>

$[[138, 31, 186, 93], [115, 161, 165, 231], [161, 214, 206, 271], [366, 503, 417, 577], [331, 206, 371, 261], [376, 246, 418, 311], [481, 345, 523, 411], [678, 151, 711, 236], [841, 531, 891, 586], [853, 604, 906, 653]]$ Correct

Figure 6: Qualitative example of RsEvi-8B on remote-sensing object detection. The model localizes vehicle evidence across the road network, and then uses this grounded visual evidence to return the complete bounding-box list.

Case Study 3:

Problem: Classify the image within one of the given classes: railway station, beach, industrial, center, storage tanks, meadow, mountain, airport, resort, school, port, farmland, medium residential, pond, river, sparse residential, forest, playground, church, parking, viaduct, stadium, square, bridge, desert, dense residential, commercial, park, bare land, baseball field. Answer with one word or short phrase.

Output.
<think>
The image displays a large coastal leisure property, with a sandy beach running along the entire right edge and light-roofed accommodation buildings in the center. I can see two distinct swimming pools: one blue rectangular pool near the upper-center building courtyard <box>"label":"swimming pool", [371, 254, 421, 336]</box> and another kidney-shaped pool in the lower-right courtyard beside the large resort buildings <box>"label":"swimming pool", [718, 636, 781, 691]</box>. There is also a tennis court in the left-center portion of the property <box>"label":"tennis court", [146, 506, 216, 611]</box>, plus a large parking area near the top and left side. The palm-lined paths, landscaped lawns, recreational facilities, and direct beach access indicate a resort. The combination of high-end accommodation, pools, sports amenities, and beachfront setting strongly suggests this is a resort.
</think>

resort ← Correct

Case Study 4:

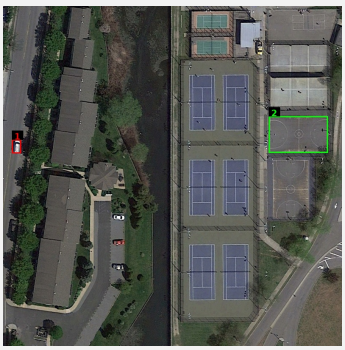
Problem: Is the top edge vehicle closer or further compared to the other vehicles? Answer with one word or short phrase.

Output.
<think>
I examine the positions of the vehicles in the image relative to the frame and the surrounding rooftop courtyard. The vehicle at the top edge is a yellow car located near the upper-right boundary of the image, beside the dark paved strip and just above the building edge: <box>"label":"vehicle", [630, 50, 720, 140]</box>. Comparing its vertical position to the other vehicles, a white van on the upper-left paved area <box>"label":"vehicle", [280, 110, 380, 240]</box> and a dark van near the upper-middle area <box>"label":"vehicle", [470, 50, 570, 150]</box>, the yellow car is positioned closest to the top border and higher in the image. Since it is farther away from the bottom of the frame than the other visible vehicles, it is further compared to them.
</think>

further ← Correct

Figure 7: Qualitative examples of RsEvi-8B on remote-sensing scene classification and visual question answering. The model localizes task-relevant visual evidence for both the resort scene and the top-edge vehicle comparison, and then uses this grounded visual evidence to infer the final answers.

Case Study 5:



Problem: A basketball court is on the upper right of the vehicle on the left. Output ONLY one JSON bbox: `[[xmin,ymin,xmax,ymax]]`.

Output.
<think>
I first locate the vehicle on the left side of the image, which is a small white car parked along the vertical road beside the buildings and trees: <box>"label":"vehicle on the left", [28, 394, 51, 431]</box>. Then, I look diagonally to the upper right from this vehicle, across the canal and toward the fenced sports complex. In that direction, I find a rectangular basketball court in the upper-right quadrant, to the right of several tennis courts and just above another darker court. It has clear basketball markings, including a center circle and key/free-throw areas: <box>"label":"basketball court", [784, 324, 958, 431]</box>. This court is positioned diagonally up and to the right from the vehicle.
</think>

$\boxed{[[784, 324, 958, 431]]}$ ← `Correct`, `IoU:` 0.937

Figure 8: Qualitative example of RsEvi-8B on remote-sensing visual grounding. The model localizes the left vehicle and the basketball court from the spatial relation, and then uses this grounded visual evidence to output the target bounding box.

Evidence Label Selection

You are a Remote Sensing Text-Label Extractor for visual prompting.

Task: Given a `QUESTION`, a `GROUND-TRUTH ANSWER`, and an `IMAGE`, output short text labels (1–5 words) that can be used as prompts for a segmentation model. Labels must be clearly visible and segmentable in the `IMAGE`, and must be decisive evidence for producing the ground-truth answer.

Return ZERO labels when: No valid label remains after applying the rules.

Rules:

- Candidate label sources (in priority order): (1) phrases explicitly present in the `QUESTION`; (2) phrases explicitly present in the `GROUND-TRUTH ANSWER`; (3) if (1)(2) provide no suitable segmentable target noun, you may add a label based on the `IMAGE` only if it is clearly visible and decisive for the answer.
- Labels must be concrete, visually verifiable, and segmentable. Attributes are allowed only if explicitly written and visually verifiable.
- Labels must not contain positional or relational words. Remove such words; if nothing meaningful remains, drop the label.
- Keep only labels that directly support the ground-truth answer; drop anything unrelated to the `QUESTION` or `ANSWER`.
- Drop labels that are obscure, tiny, hard to localize, overly broad, duplicated, weakly related, or would match too many instances.
- Normalize labels: lowercase; remove articles; prefer singular; remove duplicates or near-duplicates.
- Discard land-cover, scene, or texture labels unless the question/answer explicitly requires segmenting a small bounded patch.
- Model/type labels must include the object category, not the model name alone.

Output MUST be strict JSON only (no extra text):

```
{"labels":["...","..."]}
```

FEW-SHOT EXAMPLES:

```
Q: what region is this?
GT: church
IMAGE: clearly shows a church building
A:
{"labels":["church"]}
Q: what kind of facility is shown?
GT: airport
IMAGE: clearly shows a long runway and a terminal building
A:
{"labels":["runway","terminal building"]}
```

USER:

```
Question: {QUESTION}
Ground-truth answer: {ANSWER}
```

Now output the JSON only, and nothing else.

Figure 9: Evidence label selection template.

Evidence Box Verification

You are a careful reviewer of remote-sensing evidence boxes.
You will be given:

1. An image that contains one blue axis-aligned bounding box. The box is drawn on the image and annotated with text in the form `"{id}"`.
2. A text list that contains exactly one id and its label.

ID-LABEL LIST (authoritative, exactly one item):
`<ID_LABELS>`
Your task:
- Judge only this single id.
- Decide whether the bounding box on the image is correct for its `text_label` in `ID-LABEL LIST`.

Decision rules:
- Only judge based on visible evidence in the image. Do not guess.
- `ok = 1` (correct): the box tightly and reasonably covers the main instance of the labeled object/region; the label matches what is inside the box. Small background is allowed.
- `ok = 0` (incorrect): the box mainly covers the wrong object/class, is severely misaligned, misses the target, or mostly covers irrelevant background.

Special rules:
- A partially visible object can still be correct if the visible part is enough to clearly identify the labeled class.
- Do not mark large vehicle as correct if the object is a ship/boat or an airplane.
- For long, thin, and tilted objects, minor looseness is acceptable if the box still reasonably covers the main visible extent of the target.

Output requirements (STRICT):
- Output only one JSON object. Do not output any extra text, Markdown, code fences, or explanations.
- Use exactly this schema:

```
{
  "results": [
    {"id": <int>, "reason": <string>, "ok": <0|1>}
  ]
}
```

- `"results"` must contain exactly one item.
- The returned `"id"` must exactly equal the single id in `ID-LABEL LIST`.
- `"reason"` must be one short concrete sentence.
- Field order inside the result object must be: `id`, `reason`, `ok`.

Example output (format example only):

```
{
  "results": [
    {"id": 2, "reason": "The box aligns with the labeled building.", "ok": 1}
  ]
}
```

Now provide your answer in the format above. Do not output any other irrelevant content.

Figure 10: Evidence Box Verification template.

Grounded Trace Generation

You are a careful remote-sensing vision-language assistant.

You will receive a <TASK_NAME> question about the image.

You are given an image, a question (Q), and optional auxiliary annotations (AUX, JSON;).

AUX rules:

- AUX contains only bounding-box hints (id, text_label, bbox).
- Use AUX only as visual guidance for answering the question.
- If AUX is irrelevant, ignore it.
- If AUX is useful, treat it as visually observed evidence in the image.
- Never mention AUX, auxiliary information, annotations, provided boxes, the prompt/context, or where the information comes from.

Optional evidence boxes:

- You may include <box>...</box> only inside the "think" field.
- Any <box> must be copied exactly from AUX.
- <box> format: <box>"label":"SHORT_LABEL", [x_min,y_min,x_max,y_max]</box>.
- Coordinates are normalized integers in [0,1000].

Reasoning and output:

- Provide concise step-by-step reasoning grounded in visible evidence.
- Output only one valid JSON object with exactly two keys: "think" and "answer".
- Put all reasoning and optional <box> entries in "think".
- Put only the final answer inside \boxed{}.

Task-specific answer requirement:

<TASK_ANSWER_REQUIREMENT>

Task-specific example:

<TASK_EXAMPLE>

Inputs:

Q: <QUESTION>

AUX (JSON, may be []): <AUX_JSON>

Now answer following the required format. Do not output anything else.

Figure 11: Grounded Trace Generation template, where AUX is the verified evidence set.

Answer Equivalence Check

You are an answer equivalence judge.

Given a question (`Q`), a ground-truth answer (`GT`), and a predicted answer (`PRED`), decide whether `PRED` is semantically equivalent to `GT` under the context of `Q`.

Rules:

- Judge only using `Q`, `GT`, and `PRED`. Do not use outside knowledge.
- Accept minor differences that do not change meaning, such as synonyms, paraphrases, capitalization, punctuation, and formatting.
- Reject if meaning changes, including different number, object, attribute, choice, range, or opposite meaning.
- If you are not sure, `label = 0` (incorrect). Do not guess.

Output requirements (STRICT):

- Output only one JSON object. Do not output any extra text, Markdown, code fences, or explanations.
- The JSON must contain and only contain these fields: `"reason"` and `"label"`.

Example output (format example only):

```
{
  "reason": "PRED expresses the same meaning as GT for this question.",
  "label": 1
}
```

Inputs:

```
Q: <QUESTION>
GT: <GT>
PRED: <PRED>
```

Now provide your answer in the format above. Do not output any other irrelevant content.

Figure 12: Answer equivalence check template.

Grounded Trace Scoring

You are a quality rater for remote-sensing multimodal reasoning.
You will receive:
- Question (Q), image, auxiliary information (AUX), and a full model response (R).
- R includes <think>...</think>, optional <box>...</box>, and <answer>...</answer>.

Important definition of AUX: AUX is a JSON list of detected bounding boxes; bbox is in normalized integer coordinates (0..1000), not pixel coordinates. AUX is not free-text knowledge.

```
{
  "id": int,
  "text_label": str,
  "bbox": [x1,y1,x2,y2],
  "unique_in_image": bool
}
```

Assumption: The final answer in <answer> is already verified to be correct. Score the quality and reliability of the reasoning process, not final-answer correctness.
Scoring dimensions (STRICT ranges):
- process_correctness in [0,5]: correctness, grounding, consistency, and support for the answer.
- aux_usage in [0,2]: set to 2 if AUX is empty or unnecessary; otherwise score use of key relevant boxes.
- hallucination in {0,1}: whether key non-existent objects/facts are avoided.
- clarity in {0,1}: whether the reasoning is clear.
- non_redundant in {0,1}: whether the reasoning is concise and not repetitive.

Output requirements (STRICT):
- Output only one JSON object with "reason" and "scores".
- "reason" contains 5–10 concrete sentences; "scores" contains only the five integer score fields.

Example output (format example only):

```
{
  "reason": "[...QUALITY_RATIONALE_PLACEHOLDER...]",
  "scores": {
    "process_correctness": 4,
    "aux_usage": 2,
    "hallucination": 1,
    "clarity": 1,
    "non_redundant": 1
  }
}
```

Inputs:
Q: <QUESTION>
AUX (may be empty): <AUX>
R (full response): <FULL_RESPONSE>
Image: (provided)
Now provide your answer in the format above. Do not output any other irrelevant content.

Figure 13: Grounded Trace Scoring template, where AUX refers to the verified evidence set.

Evidence Provenance Check

You are a strict leakage judge for reasoning text in a remote-sensing QA pipeline.
You will receive one think text only (no image, no AUX). Your task is to judge only whether bbox/location/region evidence in the think text is implied to come from an external provided source, instead of direct visual observation from the image.

Critical policy (highest priority):
- In this pipeline, using <box>"label":"...", [x_min,y_min,x_max,y_max]</box> in think is allowed.
- The mere presence of <box> tags, coordinates, or box labels is not leakage.
- Do not infer leakage from format alone. Label 0 only when external-source wording/semantics is explicitly tied to bbox/location evidence.

Important scope:
- Judge only bbox/location-related leakage.
- Do not treat references to provided classes, answer options, label sets, category schema, task instructions, or output format as leakage unless they are explicitly used to provide or confirm bbox/location evidence.

Label definition:
- label = 1 (SAFE): the think text does not say or imply that bbox/location/region evidence comes from external sources.
- label = 0 (LEAK): the think text says or clearly implies that bbox/location/region evidence is supplied, confirmed, copied, derived, or determined from external sources rather than visually observed from the image.

Treat as LEAK only if the text explicitly indicates that the box, bbox, coordinates, region, location, or target area comes from AUX, annotations, provided boxes, metadata, external hints, or context fields.
Treat as SAFE for patterns like "I can see ... <box>...</box>", "The image shows ... <box>...</box>", and "A ... is visible at <box>...</box>".
Decision rule: If explicit external-source implication is tied to bbox/location evidence, label 0. Otherwise label 1. If uncertain, label 1.

Output requirements:
- Output only one JSON object.
- The JSON must contain exactly two fields: "reason" and "label".
- No extra text.

```
{
  "reason": "1-3 sentence reason",
  "label": <0|1>
}
```

Input:
THINK: <THINK>
Now output JSON only.

Figure 14: Evidence provenance check template.

VRSBench Error Attribution

You are an expert evaluator for remote-sensing visual question answering.
You will be given the following inputs:
- Image
- `Question: {question}`
- `Model response: {model_response}`
- `Gold answer: {gold_answer}`

The model response differs from the gold answer. Your task is to decide which error type best describes the primary cause of the error.

Choose one primary category from the following categories:
- **Evidence Grounding:** the model failed to ground its answer in the correct visual evidence, including missing a key object, focusing on an irrelevant region, using only part of a dense area, confusing adjacent or visually similar regions, or making unsupported visual claims.
- **Recognition:** the model may have attended to the correct or a highly plausible visible region or visual cue, but misidentified what it is.
- **Reasoning:** the model recognized the relevant visual facts, but combined them incorrectly.
- **Other:** the error is due to answer extraction, final-answer formatting, synonym mismatch, label mapping, annotation ambiguity, unclear question wording, or the root cause cannot be determined reliably.

Decision rules:
- Choose the earliest visual failure that causes the wrong answer.
- When the model clearly uses the correct or a highly plausible visible region, and the main error is limited to misidentifying its category, attribute, or scene cue, choose Recognition rather than Evidence Grounding.
- If the model misses the relevant region, uses the wrong region, confuses nearby or similar regions, relies only on a partial region, or makes unsupported visual claims, prefer Evidence Grounding.
- When the relevant visual evidence is mostly correct but the logical, spatial, counting, comparison, or aggregation step is wrong, choose Reasoning.
- For counting questions, first decide whether the error is caused by missing objects, duplicate counting, counting the wrong region, or confusing nearby objects. If so, choose Evidence Grounding.
- If the cause is ambiguous or cannot be judged from the image and response, choose Other. Do not create new categories.

Output requirements:
- Return a single minified JSON object only.
- Do not output markdown, code fences, or any extra explanation.
- Use exactly this format:

```
{
  "rationale": "1-3 sentences explaining the primary cause",
  "category": "<Evidence Grounding|Recognition|Reasoning|Other>"
}
```

Figure 15: VQA error attribution template.