BEYOND ZOOMING: LEARNING MULTI-TOOL VISUAL REASONING FOR ULTRA-HIGH-RESOLUTION REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultra-high-resolution (UHR) remote-sensing (RS) imagery provides fine-grained Earth-observation evidence over city-scale scenes, but poses a fundamental challenge for multimodal large language models (MLLMs): task-relevant evidence is often sparse, local, and spatially dispersed across extremely large visual contexts. A natural solution is to equip MLLMs with zoom-in tools for active local inspection. However, through a pilot study on XLRS-Bench, we find that zoom-in is only partially effective: it resolves easy and medium-level tasks with locally recoverable evidence, but saturates on hard cases requiring global search, multi-region comparison, path planning, or dispersed-evidence reasoning. Motivated by this finding, we move beyond single-tool zoom-in and introduce **GeoMTVR**, a large-scale Geospatial Multi-Tool Visual Reasoning dataset built from wide-area satellite imagery. GeoMTVR contains 13K UHR VQA samples with interleaved reasoning trajectories, diverse visual tool calls, and returned visual observations, enabling models to learn question decomposition, tool selection, regional inspection, object-level grounding, auxiliary visual reasoning, and cross-tool evidence integration. Beyond supervised fine-tuning, we propose a tool-attention-focused reinforcement learning algorithm that concentrates optimization on critical tool-use decisions, including when to invoke tools, which tool to select, where to apply it, and how to interpret tool outputs. By combining SFT on GeoMTVR with our RL algorithm, we develop **GeoLens**, a multi-tool visual reasoning MLLM for UHR RS. Experiments show that GeoLens achieves competitive or leading performance across three UHR RS benchmarks. Our results suggest that UHR RS MLLMs should evolve from passive image understanding and single-operation exploration toward active, task-adaptive, multi-tool visual reasoning.

1 INTRODUCTION

Ultra-high-resolution (UHR) remote-sensing (RS) imagery greatly increases the observable detail of the Earth's surface. Unlike conventional natural images, a single UHR satellite image often covers city-scale or regional-scale areas while preserving small objects, local structures, and long-range geographic relations. This makes UHR RS imagery an important yet challenging domain for multimodal large language models (MLLMs) (Wang et al., 2025b;c; Dang et al., 2025).

Recent advances in MLLMs have substantially improved visual understanding and reasoning in both general and remote-sensing domains (Kuckreja et al., 2024; Zhang et al., 2024c; Hu et al., 2025; Zhan et al., 2025). However, UHR RS understanding remains far from solved. The core bottleneck is visual evidence acquisition: the model must identify task-relevant evidence from extremely large visual contexts, where useful regions are often sparse, local, and spatially dispersed. Directly compressing the entire image into a global representation may discard fine-grained details, while exhaustive local inspection is computationally inefficient and difficult to learn (Luo et al., 2025; Wang et al., 2025b).

A natural solution is to equip MLLMs with visual tools, especially crop-and-zoom, so that they can actively inspect high-resolution local regions during reasoning. Given an image and a question, the model interleaves textual reasoning with visual actions and decides whether to continue reasoning or invoke a zoom-in tool to acquire finer local evidence (Zheng et al., 2025; Hong et al., 2025; Shen et al.,

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

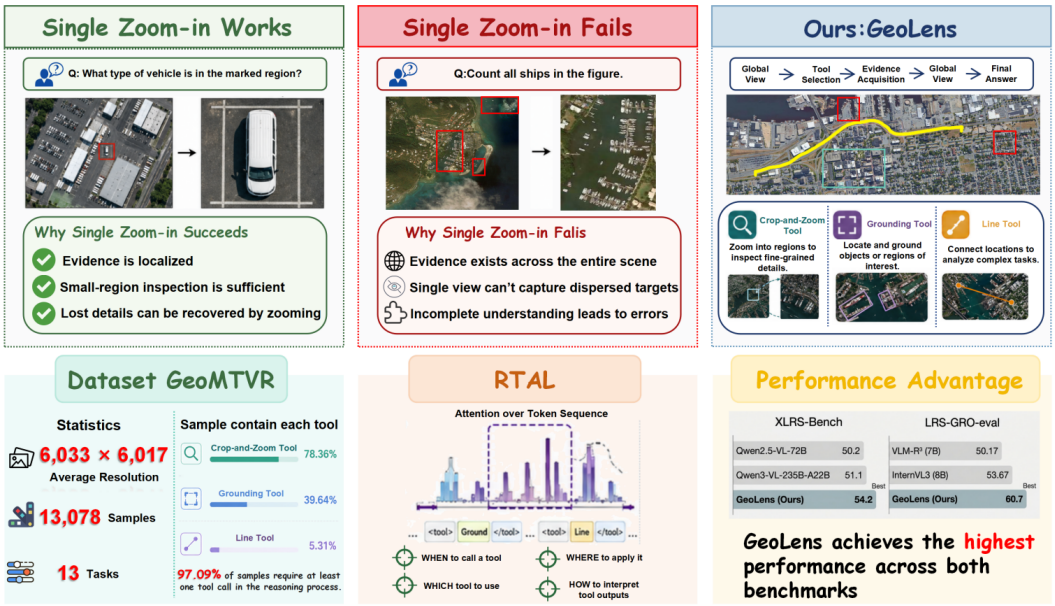

Figure 1: **Overview of GeoLens.** GeoLens enables multi-tool visual reasoning for ultra-high-resolution remote-sensing imagery.

2025). This paradigm has recently attracted attention in both general-domain and RS MLLMs (Wang et al., 2025a; Liu et al., 2025; Wang et al., 2026a). However, recent general-domain analyses suggest that zoom-in is not a universally effective operation: its benefit depends heavily on whether the task can be solved through localized evidence recovery, and indiscriminate zoom-in may introduce noisy or incomplete evidence (Hou et al., 2025; Wei et al., 2026; Ma et al., 2026b). In contrast, recent UHR RS studies emphasize zoom-in as a key mechanism for resolving fine-grained details in large satellite images (Liu et al., 2025; Wang et al., 2026a). This discrepancy raises a central question:

Is single-tool zoom-in sufficient for UHR remote-sensing reasoning, or do UHR RS MLLMs require more diverse visual tools for active evidence acquisition?

To answer this question, we first conduct a pilot study on XLRS-Bench using a zoom-in-enabled reinforcement learning framework (Li et al., 2026). Our analysis shows that zoom-in is useful, but its benefit is highly task-dependent. It mainly improves easy and medium-level tasks whose missing evidence is locally recoverable, such as object recognition, local counting, and attribute identification. However, it saturates on hard cases that require global search, multi-region comparison, long-range spatial reasoning, or heterogeneous perceptual operations, such as global counting, path planning, and dispersed-evidence reasoning. These results suggest that the limitation is not simply an insufficient number of zoom-in calls. Rather, single-tool zoom-in provides only one form of evidence acquisition, while UHR RS reasoning often requires models to decide which visual operation to perform, where to apply it, and how to integrate evidence returned by different tools.

Motivated by this finding, we move beyond single-tool zoom-in and study multi-tool visual reasoning for UHR RS. Specifically, we consider three complementary visual tools: `crop_and_zoomin` for progressive local inspection, `grounding` for precise object-level localization, and `line` for auxiliary-line drawing in spatial relation and path-oriented reasoning (Hu et al., 2024; Liu et al., 2023; Ma et al., 2026a). To support this setting, we construct **GeoMTVR** (**Geo**spatial **M**ulti-**T**ool **V**isual **R**easoning), a large-scale supervised fine-tuning dataset from wide-area satellite imagery. Each sample in GeoMTVR provides an interleaved trajectory of textual reasoning, visual tool calls, and returned visual observations. Different from existing high-resolution RS VQA datasets, which are mostly static question-answer datasets or zoom-in-only tool-interaction datasets (Wang et al., 2025b; Luo et al., 2025; Liu et al., 2025; Wang et al., 2026a), GeoMTVR enables models to learn question decomposition, tool selection, regional inspection, object-level grounding, auxiliary visual reasoning, and cross-tool evidence integration.

However, supervised learning alone may imitate tool-use traces without fully optimizing tool-use decisions. In UHR RS reasoning, the key challenge is not only generating correct answers, but also

allocating model attention to critical tool-related decisions, including when to call a tool, which tool to select, where to apply it, and how to interpret the returned observation. To this end, we propose Reinforced Tool Attention Learning (RTAL), a tool-attention-focused reinforcement learning algorithm inspired by attention-level post-training objectives (Li et al., 2026). RTAL places stronger optimization emphasis on tokens and attention patterns associated with tool-use decisions. By concentrating policy updates around tool invocation and tool-output interpretation, RTAL encourages more adaptive and effective evidence acquisition, rather than uniformly reinforcing all generated tokens.

Based on GeoMTVR and RTAL, we develop **GeoLens**, a multi-tool visual reasoning MLLM for UHR remote sensing. GeoLens is first trained via supervised fine-tuning (SFT) on multi-tool reasoning trajectories and then further optimized with RTAL. As a result, the model learns to actively select visual tools, inspect relevant regions, and integrate evidence from heterogeneous observations. Experiments on UHR RS reasoning benchmarks show that GeoLens consistently outperforms direct reasoning and single-tool zoom-in baselines, achieving stronger performance, better evidence grounding, and more effective tool-use trajectories.

In summary, our main contributions are as follows:

(1) We systematically analyze the effectiveness and limitations of single-tool zoom-in in UHR RS reasoning. Our pilot study shows that zoom-in mainly solves localized easy and medium-level tasks, such as object recognition, local counting, and attribute identification, but saturates on hard cases requiring global coverage, multi-region comparison, or structured spatial reasoning.

(2) We construct **GeoMTVR**, a large-scale UHR RS multi-tool visual reasoning dataset with interleaved textual reasoning, three types of visual tool calls, and returned visual observations. To our knowledge, it is the first dataset to support multi-tool visual reasoning in UHR RS, covering crop-and-zoom, grounding, and auxiliary-line drawing.

(3) We propose RTAL and develop **GeoLens**, a multi-tool visual reasoning MLLM for UHR remote sensing. By combining SFT on GeoMTVR with RTAL, GeoLens learns to actively select visual tools, inspect relevant regions, and integrate heterogeneous visual evidence, achieving consistent improvements over other methods.

2 Related Work

Remote Sensing Large Multimodal Models. Recent advances in large multimodal models (MLLMs) have promoted remote sensing (RS) understanding across multi-task reasoning (Hu et al., 2025; Zhan et al., 2025), captioning and retrieval (Silva et al., 2024), interactive dialogue (Zhang et al., 2024c), unified instruction following (Kuckreja et al., 2024; Muhtar et al., 2024; Li et al., 2024c), multi-sensor analysis (Shu et al., 2025; Soni et al., 2025; Wang et al., 2026b), grounded understanding (Wang et al., 2025d; Zhang et al., 2024b), and mixture-of-experts modeling (Liu & Lian, 2026). As RS imagery moves toward ultra-high-resolution (UHR) inputs, the prohibitive number of visual tokens becomes a central bottleneck. Existing solutions mainly follow two directions. One line reduces visual tokens through pruning or evidence selection, using adaptive clustering, attention-guided cropping, or reference-guided retrieval to remove redundant regions (Wang et al., 2025b; Shalev-Arkushin et al., 2025). These methods improve efficiency but often depend on hand-crafted rules or static selection strategies, which may not generalize to diverse VQA tasks. Another line equips RS MLLMs with visual tools for iterative perception. Representative systems such as ZoomEarth (Liu et al., 2025), VICoT-Agent (Wang et al., 2025a), and GeoEyes (Wang et al., 2026a) progressively refine visual evidence, typically through zoom-in operations over regions of interest. However, current tool-based RS models still rely mainly on zoom-in as the dominant interaction form, leaving complex UHR scenarios underexplored, where evidence may be distributed across different scales, locations, and semantic contexts.

Remote Sensing Vision-Language Datasets. The RS community has developed many vision-language datasets to support multimodal learning, including large-scale image-text resources for geo-foundation models (Yuan et al., 2024). Early datasets such as RSVQA (Lobry et al., 2020), RS-IVQA (Zheng et al., 2021), and EarthVQA (Wang et al., 2024a) construct VQA samples from geographic annotations or automatically generated questions, while RSSA (Pang et al., 2024),

SkySenseGPT (Luo et al., 2024), VHM (Pang et al., 2025), and VRSBench (Li et al., 2024a) expand instruction-following and task coverage. However, these datasets are generally limited in spatial resolution. Recent UHR datasets, including SuperRS-VQA (Wang et al., 2025b), HighRS-VQA (Wang et al., 2025b), LRS-VQA (Luo et al., 2025), XLRS-Bench (Wang et al., 2025c), and RSHR-Bench (Dang et al., 2025), push RS evaluation toward the 8K scale. Tool-augmented datasets such as GeoEyes UHR-CoZ (Wang et al., 2026a) and LRS-GRO (Liu et al., 2025) further introduce external visual operations into RS reasoning. Nevertheless, existing datasets still provide limited support for multi-step and multi-tool reasoning in UHR settings. Most UHR datasets focus on single-step perception or evaluation, while current tool-augmented datasets mainly rely on shallow zoom-in interactions. General-domain multi-tool instruction data is also not tailored to RS imagery or UHR inputs, creating a gap for training MLLMs to perform complex thinking with visual tools in real-world RS scenarios.

3 DATASET

Recent advances in remote sensing multimodal large models have led to substantial VQA datasets, including SuperRS-VQA (Wang et al., 2025b), HighRS-VQA (Wang et al., 2025b), RSVQA-HR (Lobry et al., 2020), UHR-CoZ (Wang et al., 2026a), and LRS-GRO (Liu et al., 2025). However, existing high-resolution RS VQA datasets mainly fall into two categories: static question–answer datasets and tool-interactive datasets centered on zoom-in operations. They therefore provide limited support for diverse visual tool use, which is essential in ultra-high-resolution (UHR) scenarios where reasoning often requires coordinated inspection, localization, and spatial comparison across multiple regions. Table 1 compares our dataset with existing datasets.

Table 1: **Comparison with RS UHR VQA datasets.**

Dataset	Avg. Resolution	Tool Types	VQA Volume
RSVQA (Lobry et al., 2020)	512×512	None	111,134
RSVQA-HR (Lobry et al., 2020)	1,024×1,024	None	1,066,316
LRS-GRO (Liu et al., 2025)	5,000×5,000	Zoom-in	13,245
UHR-CoZ (Wang et al., 2026a)	2,178×2,051	Zoom-in	25,467
GeoMTVR	**6,033×6,017**	**Zoom-in, Grounding, Line**	**13,078**

Table 2: **Pilot Experiments of Dataset.** We only modify the SFT data. Accuracy is reported based on results from XLRS-Bench (Wang et al., 2025c).

Model	SFT Data					Acc.
	Data Source	Volume	Average Resolution	Domain	Tools	
Qwen-2.5VL-7B	SuperRSVQA (Wang et al., 2025b)	13k	8,376×8,376	Remote Sensing	None	48.9
Qwen-2.5VL-7B	SuperRSVQA (Wang et al., 2025b)	13k	8,376×8,376	Remote Sensing	None	42.7↓
	vqa_cold_start (Wu et al., 2025)	10k	478×390	General Domain	Object Mapper	
	maze_cold_start (Wu et al., 2025)	7.3k	588×588	General Domain	Object Mapper	
	GPT4Scene_cold_start (Wu et al., 2025)	6.1k	504×364	General Domain	Object Mapper	
	SR_91k_cold_start (Wu et al., 2025)	11.4k	504×364	General Domain	Object Mapper	

To fill this gap, we follow recent paradigms such as ViLaSR (Wu et al., 2025), which introduce tool invocation and visual-process annotation, and extend them to UHR satellite imagery. We design a staged, semi-automatic, process-driven annotation framework that progressively converts existing UHR RS VQA samples into interpretable tool-use reasoning trajectories. Through multi-stage collaboration, the framework produces high-quality annotations involving diverse visual tools, including crop-and-zoom, grounding, and auxiliary-line operations. Using this pipeline, we construct approximately 13K high-resolution VQA samples with an average resolution of about 6K $\times$ 6K, each enriched with detailed visual tool-use reasoning traces.

Pilot Experiments of Dataset. General-domain datasets have recently provided supervised traces for visual tool use, making it tempting to combine them with high-resolution RS VQA data for training tool-augmented RS MLLMs. We test this strategy by only changing the mixture data in the SFT stage. As shown in Table 2, Qwen2.5-VL-7B (Bai et al., 2025c) trained with SuperRSVQA (Wang et al., 2025b) alone achieves 48.9% accuracy on XLRS-Bench (Wang et al., 2025c), whereas adding several general-domain tool-interaction datasets reduces the accuracy to 42.7%. This result suggests that simply increasing SFT data with visual tool-use trajectories does not improve UHR RS reasoning. Most added datasets are built on low-resolution natural or synthetic images, whose visual scale, object distribution, and evidence layout differ substantially from UHR RS imagery. Consequently, their tool-use traces may provide weak or even misleading supervision for RS scenarios, where task-relevant evidence is sparse, spatially dispersed, and often requires coordinated exploration across

multiple regions. These findings reveal a clear domain mismatch and motivate UHR RS-specific multi-tool interaction data.

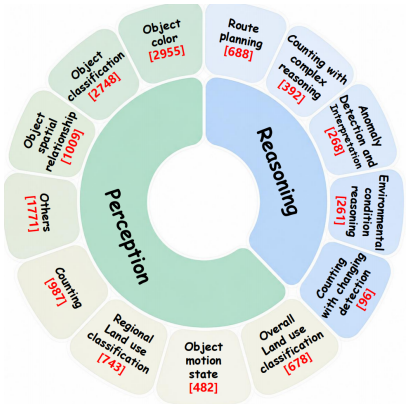

Figure 2: **Our dataset covers perception and reasoning dimensions across 13 sub-tasks.**

Table 3: **Main statistics of the GeoMTVR dataset.**

Metric	Value
Total samples	13,078
Average image resolution	6,033 × 6,017
Average question length	60.72 words
Average reasoning length	126.92 words
Tool-used samples	97.09%
Average number of tool calls per sample	1.53
Average number of tool types per sample	1.23
Samples containing zoomin	78.36%
Samples containing grounding	39.64%
Samples containing line	5.31%

Data source selection. Our dataset is built on existing high resolution remote sensing VQA resources. We integrate multi-source imagery and apply unified re-annotation to improve data diversity and training utility. Specifically, we use high-resolution VQA sources including SuperRSVQA (Wang et al., 2025b) and LRS-GRO-train (Liu et al., 2025). To ensure data quality and fair evaluation, we perform strict image-level deduplication and remove samples overlapping with existing benchmarks such as XLRS-Bench (Wang et al., 2025c) and LRS-GRO-eval.

Annotation. Directly using current MLLMs, such as GPT-5.4 (OpenAI, 2026) and Claude 4.6 (Anthropic, 2026), to annotate tool-use behaviors in UHR satellite imagery often leads to unreliable results. We observe pronounced hallucination and instability, especially in the initial question-region cropping stage, where localization shifts or target misidentification can cascade into incorrect bounding-box and auxiliary-line annotations. To reduce such errors, we adopt the staged annotation framework shown in Figure 3. Given a question, image, and answer, we first use Qwen3-VL-235B-A22B (Bai et al., 2025b) to coarsely crop the question-relevant region. We then incorporate VQA semantics and use GPT-5.2 (OpenAI, 2025) to generate tool-augmented reasoning traces, following the general idea of constructing detailed visual reasoning processes through chained visual manipulations (Qi et al., 2024). Finally, GroundingDINO (Liu et al., 2023) is applied to precisely localize the objects involved in the question, enabling the progressive construction of structured visual reasoning annotations.

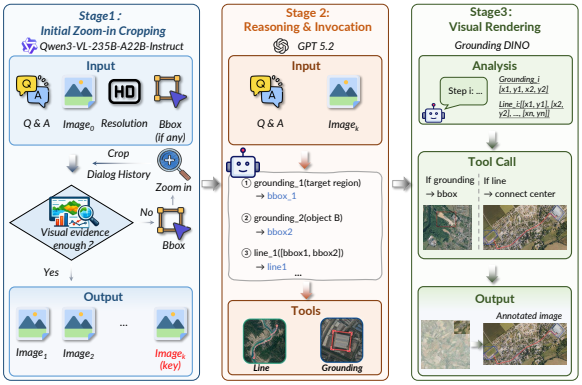

Figure 3: **Data Construction Pipeline.**

Statistics. Fig. 2 visualizes the capability dimensional coverage of both datasets. It can be seen that our datasets span a wider range of task dimensions, showing strong alignment with real-world scenarios. What's more, we summarize the key statistics of our datasets in Tab. 3, including average question/answer lengths, the diversity and quantity of annotated objects, average resolution, and other relevant attributes.

4 METHOD

4.1 PILOT STUDY: LIMITATIONS OF SINGLE ZOOM-IN TOOL IN MLLM.

Existing studies have incorporated tool invocation into high-resolution image VQA by interleaving textual reasoning with visual operations such as zoom-in (Zheng et al., 2025; Hong et al., 2025; Shen et al., 2025; Su et al., 2025; Wang et al., 2025a). Given an image and a question, the model decides

whether to answer directly or acquire higher-resolution local evidence through visual actions. Recent general-domain studies have questioned the broad effectiveness of zoom-in, arguing that its gains may mainly come from reinforcement learning on effective data rather than from the tool itself (Hou et al., 2025; Wei et al., 2026). In contrast, RS UHR studies such as GeoEyes and ZoomEarth report clear benefits of zoom-in for ultra-high-resolution remote-sensing understanding. **This discrepancy raises a key question: whether zoom-in is genuinely effective in RS UHR scenarios, where its limitations emerge, and which types of problems it can solve.**

To examine this question, we conduct a pilot study on XLRS-Bench using Qwen3-VL-8B-Instruct as the base model. The model is trained with GRPO and equipped with only a zoom-in tool (DeepSeek-AI, 2024; 2025). The training data include approximately 7K samples from the original MED training set (Ma et al., 2026b) and 6K samples from SuperRSVQA, forming a mixed corpus for tool-use RL in UHR remote-sensing scenarios. We use a batch size of 256, set the rollout number to 8, and evaluate the model every 10 steps for 220 training steps. Following MED, we decompose tool-induced performance changes into **Call Gain, Schema Gain, Call Harm, and Schema Harm**, and further factorize Call Gain into **Mass, Policy, and Quality**. We focus on the Quality term, namely the probability that a tool call leads to a correct answer on failed samples.

We consider three failure sets: the dynamic failure set $\mathcal{D}_{\text{fail}}(t)$, containing samples answered incorrectly without tools at checkpoint t; the fixed failure set $\mathcal{D}_{\text{fail}}(0)$, defined by the base model before RL; and their intersection $\mathcal{D}_{\text{fail}}(0) \cap \mathcal{D}_{\text{fail}}(t)$, which captures persistently difficult samples. As shown in Fig. 4, the fixed Quality increases steadily from 0.00 to 0.25, indicating that RL learns useful zoom-in behaviors on the original failure set. However, the dynamic Quality first increases and then declines, peaking at around 0.15 and ending at 0.13. This divergence reveals a moving failure set effect: as easier samples are solved and leave $\mathcal{D}_{\text{fail}}(t)$, the remaining failures become harder, and a single zoom-in operation can no longer provide sufficient evidence. By the final checkpoint, the fixed Quality is nearly twice the dynamic Quality, suggesting that evaluation on the initial failure set substantially overestimates tool effectiveness on current hard cases.

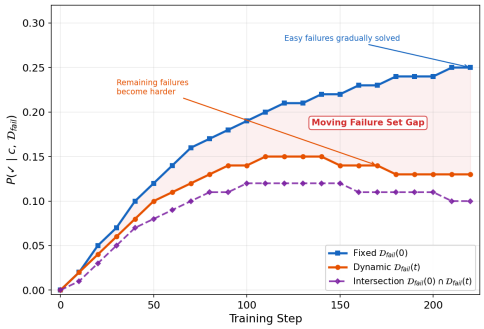

Figure 4: **Quality factor under three failure-set definitions on XLRS-Bench.**

Takeaway

Zoom-in tool in MLLM mainly solves easy UHR RS tasks with localized evidence, such as object recognition, local counting, and attribute identification, but remains insufficient for hard tasks requiring global coverage, multi-region comparison, or long-range spatial reasoning.

A qualitative analysis (see Appendix, Table 8) further explains this effect. Solved samples usually depend on evidence from one localized region, such as recognizing a vehicle type, counting objects in a specified port area, identifying a roof color, or determining the relative position of nearby landmarks. In these cases, the main challenge is to locate the relevant region, and one zoom-in operation often provides enough detail. In contrast, persistently failed samples involve global counting, path planning, land-use enumeration, change counting across image pairs, or reasoning over spatially dispersed clues. These tasks require systematic spatial coverage, multi-region comparison, or heterogeneous perceptual capabilities beyond zoom-in. Although the framework allows up to five zoom-in calls and retains the global image in context, limited coverage and the absence of structured exploration still prevent the model from collecting sufficient evidence.

4.2 GeoLens

GeoLens is trained in two stages. Supervised fine-tuning (SFT) on GeoMTVR initializes multi-tool visual reasoning through interleaved trajectories of reasoning, tool invocation, and visual observation.

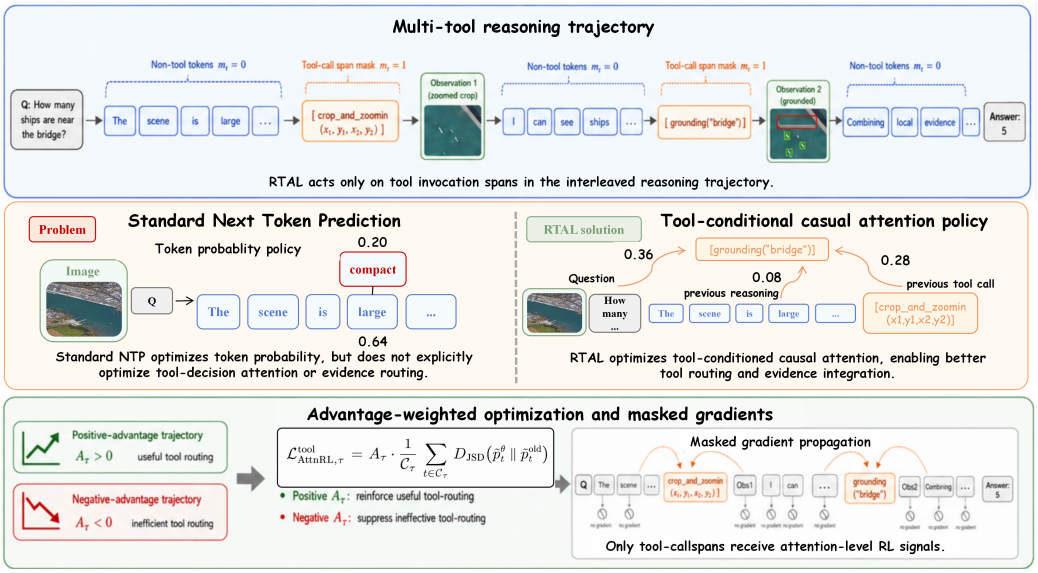

Figure 5: **Overview of Reinforced Tool Attention Learning.** RTAL focuses reinforcement signals on tool-call spans, encouraging the model to route attention toward task-relevant context and tool observations during multi-tool visual reasoning.

Reinforced Tool Attention Learning (RTAL) then refines the attention patterns underlying tool-use decisions.

In UHR remote-sensing reasoning, tool selection and argument generation determine which visual evidence becomes available. Standard output-level RL does not explicitly target the attention patterns at these decision points. RTAL therefore adds an advantage-weighted attention objective at tool-call positions, encouraging tool decisions to incorporate relevant context and previous observations (Fig. 5).

Tool-Conditioned Causal Attention Policy. Let $S = (x_1, \ldots, x_T)$ contain the prompt $x_{1:P}$ and an interleaved continuation. We define the tool-call positions as

$$\mathcal{C} = \{t \mid P < t \le T,\ x_t \in \texttt{ToolsCall}\}, \tag{1}$$

including function names, argument key-value pairs, and structural markers. For a final-layer attention head, let $\alpha_{t,i}$ denote attention from position t to a preceding position i. Under current model parameters θ, normalization over the causal context gives

$$\tilde{p}_t^\theta(i) = \frac{\alpha_{t,i}}{\sum_{j=1}^{t-1} \alpha_{t,j}}, \quad i \in \{1, \ldots, t-1\},\ t \in \mathcal{C}. \tag{2}$$

The context includes the prompt, earlier reasoning, tool calls, and returned observations. The head index is omitted for clarity; attention losses are computed per head and then averaged.

Advantage-Weighted Divergence over Tool Tokens. For a sampled trajectory τ, let $\mathcal{C}_\tau$ denote its tool-call positions and A_τ its GRPO advantage, obtained by normalizing the final outcome reward by the mean and standard deviation within the same prompt's rollout group. With the frozen pre-update attention policy $\tilde{p}_t^{\text{old}}$, the per-head objective is

$$L^{\text{tool}}_{\text{AttnRL},\tau} = \frac{A_\tau}{|\mathcal{C}_\tau|} \sum_{t \in \mathcal{C}_\tau} D_{\text{JSD}}\big(\tilde{p}_t^\theta \,\|\, \tilde{p}_t^{\text{old}}\big), \tag{3}$$

where D_{JSD} is the Jensen–Shannon divergence. Minimizing this loss penalizes departures from the old attention policy when $A_\tau > 0$ and favors departures when $A_\tau < 0$. The auxiliary loss and gradient are defined as zero for trajectories without tool calls. Averaging over sampled trajectories and final-layer heads gives $L^{\text{tool}}_{\text{AttnRL}}$, which is combined with the standard output-level RL loss:

$$L_{\text{total}} = L_{\text{RL}} + \lambda_{\text{attn}}\, L^{\text{tool}}_{\text{AttnRL}}, \tag{4}$$

Table 4: **Experimental results on the perception and reasoning dimensions on XLRS-Bench, with models ranked by average performance.** We mark the highest score in red.

Method	Perception								Reasoning					
Sub-tasks (L-3 Capability)	**OC**	**RC**	**OLUC**	**RLUC**	**OCC**	**OCL**	**OMS**	**OSR**	**AD**	**ECR**	**RP**	**RCCD**	**CCR**	**Avg.**
Remote Sensing MLLMs														
GeoChat Kuckreja et al. (2024)	16.7	29.0	2.0	23.0	21.1	16.8	35.0	24.2	33.0	43.0	10.0	-	21.0	22.9
ZoomEarth Liu et al. (2025)	30.0	38.0	8.0	63.0	35.1	33.9	65.0	31.4	66.0	73.0	22.0	28.3	40.0	41.1
GeoLLaVA-8K Wang et al. (2025b)	26.7	38.0	49.0	69.0	41.6	31.6	65.0	35.0	67.0	78.0	66.0	50.0	52.0	51.5
GeoEyes Wang et al. (2026a)	38.3	40.0	24.0	73.5	59.5	66.1	68.3	32.2	77.0	80.0	56.0	40.0	50.0	54.2
Closed-source MLLMs														
Claude 3.7 Sonnet Anthropic (2025)	27.6	22.7	17.4	68.4	30.5	29.9	63.6	27.6	64.8	78.4	34.5	27.8	32.6	40.5
Gemini 2.5 Pro Google DeepMind (2025)	–	–	–	–	–	–	–	–	–	–	–	–	–	45.2
GPT-5.2 OpenAI (2025)	30.0	37.0	17.0	70.5	43.0	41.4	68.3	34.0	74.0	76.0	52.0	36.7	38.0	47.5
Open-source MLLMs														
ViLaSR Wu et al. (2025)	28.3	38.0	17.0	72.0	31.1	33.1	65.0	39.4	70.0	76.0	27.0	41.7	47.0	45.1
InternVL3-8B Zhu et al. (2025)	40.0	39.0	10.0	71.5	44.5	30.8	65.0	25.2	77.0	82.0	36.0	21.7	50.0	45.6
Qwen2-VL-7B Wang et al. (2024b)	26.7	40.0	11.0	73.0	35.9	34.6	61.7	31.8	70.0	81.0	35.0	46.7	48.0	45.8
InternVL2.5-8B Chen et al. (2024)	38.3	37.0	10.0	77.0	33.4	35.5	65.0	21.6	73.0	83.0	34.0	50.0	43.0	46.2
Qwen2.5-VL-7B Bai et al. (2025c)	33.3	40.0	31.0	77.0	40.6	40.5	66.7	36.2	68.0	72.0	27.0	38.3	45.0	47.4
InternVL3-78B Zhu et al. (2025)	23.3	49.0	33.0	74.0	42.5	37.4	66.7	30.0	76.0	81.0	40.0	45.0	42.0	49.2
Qwen3-VL-8B Bai et al. (2025b)	21.7	50.0	26.0	81.5	46.6	43.1	66.7	30.4	74.0	79.0	37.0	43.3	51.0	50.0
Qwen2.5-VL-72B Bai et al. (2025c)	33.3	47.0	39.0	80.0	45.3	42.1	65.0	34.0	71.0	74.0	37.0	43.3	42.0	50.2
Qwen3-VL-235B-A22B Bai et al. (2025b)	61.7	82.5	38.6	36.7	44.0	39.0	38.9	49.0	73.0	82.0	37.8	33.3	48.0	51.1
Intern-S1-mini Bai et al. (2025a)	–	–	–	–	–	–	–	–	–	–	–	–	–	51.6
GeoLens (OURS)	35.0	40.0	37.0	76.5	46.5	45.6	65.0	36.8	72.0	82.0	62.0	51.7	55.0	54.2

where λ_{attn} controls the contribution of attention-level optimization.

Masked Gradient Propagation for Tool-Specific Attention. For each tool-call position, define

$$J_t = D_{\text{JSD}}\big(\tilde{p}_t^{\theta} \,\|\, \tilde{p}_t^{\text{old}}\big)\,. \tag{5}$$

Let $e_{t,i}$ denote the attention logit. At a fixed t, write $\tilde{p}_i = \tilde{p}_t^{\theta}(i)$ and $e_j = e_{t,j}$. The softmax Jacobian is

$$\frac{\partial \tilde{p}_i}{\partial e_j} = \tilde{p}_i(\delta_{ij} - \tilde{p}_j), \tag{6}$$

where δ_{ij} is the Kronecker delta. The chain rule then gives

$$\nabla_{e_{t,i}} J_t = \tilde{p}_t^{\theta}(i)\left(\nabla_{\tilde{p}_t^{\theta}(i)} J_t - \sum_{j=1}^{t-1} \tilde{p}_t^{\theta}(j)\nabla_{\tilde{p}_t^{\theta}(j)} J_t\right). \tag{7}$$

Holding sampled trajectories, advantages, and the old policy fixed, the tool-masked parameter gradient is

$$\nabla_\theta L_{\text{AttnRL}}^{\text{tool}} = \mathbb{E}_{\tau,h}\left[\frac{A_\tau}{|\mathcal{C}_\tau|}\sum_{t\in\mathcal{C}_\tau}\sum_{i=1}^{t-1}\nabla_{e_{t,i}} J_t \cdot \frac{\partial e_{t,i}}{\partial \theta}\right], \tag{8}$$

where h indexes final-layer heads. The mask selects tool-call query positions for the auxiliary loss, without restricting their causal context or changing the forward graph. Standard output-level RL continues to optimize generated response tokens.

5 Experiments

Experimental Setup We first conduct supervised training with Qwen2.5-VL-7B on our newly annotated GeoMTVR dataset, using interleaved image-text supervision and explicit tool-call annotations. Specifically, we model three explicit tool types: `crop_and_zoomin`, `grounding`, and `line`. In the reinforcement stage, based on the our RTAL method, we perform reinforcement learning on cleaned SuperRS-VQA and LRS-GRO-train data. For evaluation, we test on XLRS-Bench, RSHR-Bench, and LRS-GRO-eval. **Additional details, extended ablations are provided in the Appendix.**

Main Results. GeoLens achieves the best or tied-best overall performance among the compared models across all three benchmarks. On XLRS-Bench (Table 4), it reaches 54.2%, matching GeoEyes and exceeding Qwen3-VL-235B-A22B and Qwen2.5-VL-72B by 3.1 and 4.0 percentage points, respectively, with a 7B backbone. On RSHR-Bench (Table 5), GeoLens obtains the highest overall score of 48.8, surpassing GPT-4o by 4.8 points. On LRS-GRO-eval (Table 6), it achieves 60.7%, improving over GeoEyes by 4.65 percentage points and obtaining the highest scores in all three categories: Global, Region, and Object.

Table 5: **Detailed results on RSHR-Bench, with models ranked by average performance within each category.** We mark the highest score in red.

Model	Perception										Reasoning						Advanced			Overall
Metric	COL	SHP	DET	OC	REL	OGD	RG	OCN	RCN	Avg.	AD	FP	MRJC	MRJCS	OSJ	Avg.	MAD	MTFP	MOSJ	Avg.
Remote Sensing VLMs																				
EarthDial Soni et al. (2025)	41.0	22.0	21.0	30.0	32.5	30.5	27.1	18.0	31.0	28.1	42.0	30.0	29.5	32.0	52.0	37.1	56.7	60.0	73.5	37.0
GeoChat Kuckreja et al. (2024)	32.5	22.0	24.0	29.5	40.0	25.0	22.9	22.5	29.0	25.9	30.0	24.0	25.5	30.0	32.0	28.3	48.3	46.0	62.9	32.1
VHM Pang et al. (2025)	25.5	25.0	26.0	26.5	55.0	25.0	22.9	25.0	25.0	25.7	26.0	24.0	26.5	34.0	28.0	27.7	45.0	53.3	46.2	31.7
GeoLLaVA-8K Wang et al. (2025b)	25.0	24.0	25.0	25.0	25.0	25.0	21.4	25.0	25.0	24.5	24.0	0.0	0.0	34.0	22.0	16.0	25.0	24.7	47.7	23.4
GeoEyes Wang et al. (2026a)	56.5	25.0	34.0	36.0	55.0	19.5	34.3	36.5	33.0	36.6	64.0	42.0	45.0	34.0	60.0	49.0	47.6	47.3	70.2	43.5
Open-source VLMs																				
VILA-HD Shi et al. (2025)	40.0	22.0	22.0	37.0	35.5	26.0	21.4	24.5	24.0	28.0	58.0	30.0	55.0	32.0	58.0	46.6	65.0	57.3	57.6	39.1
MiniCPM2 6 Yao et al. (2024)	21.5	28.0	30.0	24.0	19.5	29.5	34.3	22.0	29.0	27.4	26.0	30.0	35.0	32.0	30.0	30.6	26.7	23.3	31.1	27.8
InternVL 3.5 8B Wang et al. (2025e)	21.5	28.0	18.0	21.5	29.0	28.5	30.0	26.5	25.0	25.3	20.0	16.0	29.0	34.0	26.0	25.0	30.0	22.7	40.2	26.2
Phi-3.5-Vision Microsoft GenAI Team (2024)	25.0	24.0	25.0	25.0	23.5	25.0	22.9	25.0	25.0	24.5	24.0	22.0	23.5	30.0	22.0	24.3	28.3	24.7	47.0	26.0
Deepseek-VL Lu et al. (2024)	22.5	22.0	21.0	25.0	20.5	26.0	28.6	20.5	22.0	23.1	22.0	28.0	50.0	32.0	20.0	30.4	20.0	23.3	33.3	25.7
InternVL2.5-8B Chen et al. (2024)	25.5	22.0	26.0	26.0	22.5	24.5	30.0	22.5	20.0	24.3	26.0	20.0	22.5	34.0	20.0	24.5	25.0	28.7	35.6	25.3
Qwen2.5-VL-7B Bai et al. (2025c)	29.5	25.0	22.0	28.0	25.0	24.5	24.3	26.5	22.0	25.2	26.0	28.0	25.0	10.0	20.0	21.8	21.7	24.0	10.6	23.1
Closed-source VLMs																				
GPT-4o OpenAI (2024a)	49.5	23.0	15.0	35.5	30.5	28.0	27.1	22.5	41.0	30.2	68.0	56.0	30.5	32.0	64.0	50.1	70.0	72.0	84.1	44.0
GPT5 OpenAI (2026)	29.0	10.0	23.0	23.0	37.0	24.5	31.4	20.0	23.0	24.5	74.0	58.0	35.0	34.0	66.0	53.4	78.3	73.3	86.4	42.7
GPT-4o-mini OpenAI (2024b)	41.5	16.0	29.0	31.5	31.5	32.0	28.6	19.5	32.0	29.1	54.0	54.0	31.5	48.0	54.0	48.3	78.3	68.0	75.0	42.6
Gemini-2.5-pro Google DeepMind (2025)	55.0	18.0	31.0	40.0	41.5	32.5	45.7	25.0	25.0	34.9	66.0	32.0	41.5	38.0	50.0	45.5	56.7	60.0	57.6	42.1
GeoLens (OURS)	58.5	22.7	27.0	35.0	57.5	33.0	38.6	34.5	39.0	38.4	68.0	46.0	50.0	42.0	62.0	53.6	70.0	61.1	85.4	48.8

Table 6: **Results on the LRS-GRO-eval benchmark.**

Leaderboard	MaxSize	Global	Region	Object	Avg
LLaVA-OV-1.5 An et al. (2025)	2304×2304	62.43	36.18	39.92	45.33
IXC-2.5 Zhang et al. (2024a)	4096×4096	62.43	40.00	47.32	50.00
InternVL3 Zhu et al. (2025)	3200×3200	71.60	44.58	47.80	53.67
GeoChat Kuckreja et al. (2024)	504×504	62.43	36.79	40.72	45.88
Qwen2.5-VL (7B) Bai et al. (2025c)	1024×1024	69.39	38.47	43.58	49.62
Qwen2.5-VL (3B) Bai et al. (2025c)	1024×1024	59.01	31.91	37.46	42.25
VLM-R^3 (w/ tools) Jiang et al. (2025)	512×512	69.72	44.83	37.40	50.17
GeoChat* Kuckreja et al. (2024)	504×504	58.78	37.25	42.83	46.09
ZoomEarth Liu et al. (2025)	512×512	63.09	46.11	51.80	53.76
GeoEyes Wang et al. (2026a)	1024×1024	74.87	45.03	48.24	56.05
GeoLens (OURS)	1024×1024	79.0	48.9	57.6	60.7

Although GeoLens matches GeoEyes in overall XLRS-Bench performance, their task-level profiles differ. GeoLens improves RP, RCCD, and CCR from 56.0%, 40.0%, and 50.0% to 62.0%, 51.7%, and 55.0%, respectively, with the latter two being the highest scores among the compared models. These task-specific gains align with our focus on complex spatial and multi-step reasoning beyond local zoom-in.

Table 7: **Contributions of SFT and RTAL.** XLRS-Bench accuracy and tool-use statistics across training stages. Accuracy gains are relative to the base model; the invalid-call rate reduction is relative to SFT + GRPO. All differences in accuracy or rates are measured in percentage points.

Stage	XLRS Avg.	Invalid calls	Avg. calls	Valid calls	Avg. types
Qwen2.5-VL-7B	47.4	-	-	-	-
SFT	51.7 (+4.3)	23.6%	1.81	76.4%	1.09
SFT + GRPO	52.4 (+5.0)	19.8%	1.89	80.2%	1.13
SFT + RTAL (GeoLens)	54.2 (+6.8)	15.7% (-4.1)	2.40	84.3%	1.26

Further Analysis. Table 7 summarizes the contributions of the two training stages. SFT on GeoMTVR raises the XLRS-Bench average from 47.4% to 51.7%. Starting from the SFT checkpoint, RTAL further improves the score to 54.2%, outperforming SFT alone and SFT + GRPO by 2.5 and 1.8 percentage points, respectively. Compared with GRPO, RTAL increases the average number of calls from 1.89 to 2.40 and tool types from 1.13 to 1.26 per question, while reducing the invalid-call rate from 19.8% to 15.7%. The detailed results in the Appendix further show a 5.0-percentage-point improvement in reasoning average over GRPO (64.5% vs. 59.5%). Together, these results show that RTAL improves reasoning performance alongside more frequent and varied tool use with a lower proportion of invalid calls.

6 Conclusion

Our pilot study on XLRS-Bench showed that single-tool zoom-in helps recover local evidence but struggles with global search, multi-region comparison, and structured spatial reasoning in ultra-high-resolution (UHR) remote sensing. We introduced GeoMTVR, a multi-tool visual reasoning dataset with interleaved reasoning, tool calls, and returned visual observations covering crop-and-zoom, grounding, and auxiliary-line operations. We developed GeoLens, a 7B multi-tool MLLM, through supervised fine-tuning on GeoMTVR followed by our proposed Reinforced Tool Attention Learning (RTAL), which augments output-level RL with an advantage-weighted attention objective at tool-call positions. Among the compared models, GeoLens ties the highest average accuracy on XLRS-Bench (54.2%) and achieves the highest scores on LRS-GRO-eval (60.7%) and RSHR-Bench (48.8 overall). These results highlight the value of domain-specific interaction data and targeted tool-use optimization for UHR remote-sensing reasoning beyond single-tool zoom-in.

AI USE STATEMENT

We used generative AI models to assist the construction of the GeoMTVR dataset and to polish the writing of this paper. As described in Section 3, we employed Qwen3-VL-235B-A22B to coarsely crop question-relevant regions, GPT-5.2 to generate tool-augmented reasoning trajectories covering crop-and-zoom, grounding, and auxiliary-line operations, and GroundingDINO to precisely localize the objects involved in each question, forming a staged, semi-automatic annotation pipeline. We did not use generative AI for code generation or idea generation.

The authors manually reviewed all AI-assisted content and outputs, including the generated tool-use reasoning trajectories, visual tool calls, returned observations, and quality-screening results. Human annotators validated and filtered the annotations produced by the semi-automatic pipeline, and the dataset further underwent strict image-level deduplication and removal of samples overlapping with existing benchmarks as part of quality control. The authors take full responsibility for the manuscript, dataset, methodology, reported results, and conclusions, including all content produced with assistance from generative AI.

REPRODUCIBILITY STATEMENT

The data sources, the staged semi-automatic annotation pipeline, and the quality-control procedures (including image-level deduplication and removal of samples overlapping with existing benchmarks) for the GeoMTVR dataset are described in Section 3. The GeoLens model architecture, the supervised fine-tuning and RTAL training stages, the three tool interfaces (crop-and-zoom, grounding, and auxiliary-line drawing), and the corresponding training data and hyperparameters are specified in Section 5 and Appendix A.2. The evaluation benchmarks, evaluation protocol, and scoring rules are described in Section 5 and Appendix A.2; the ablation studies and the task-level tool-usage analysis are reported in Appendix A.3 and A.5. We will release the GeoMTVR dataset, the code, and the model checkpoints.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Anthropic. Claude 3.7 sonnet. https://www.anthropic.com, 2025. Accessed: 2026-05.

Anthropic. The claude 4.6 model family. Technical report, Anthropic, 2026. URL https://www.anthropic.com/research.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025c.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Yunkai Dang, Meiyi Zhu, Donghao Wang, Yizhuo Zhang, Jiacheng Yang, Qi Fan, Yuekun Yang, Wenbin Li, Feng Miao, and Yang Gao. A benchmark for ultra-high-resolution remote sensing mllms. *arXiv preprint arXiv:2512.17319*, 2025.

DeepSeek-AI. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

Xinhai Hou, Shaoyuan Xu, Manan Biyani, Mayan Li, Jia Liu, Todd C Hollon, and Bryan Wang. Codev: Code with images for faithful visual reasoning via tool-aware policy optimization. *arXiv preprint arXiv:2511.19661*, 2025.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:272–286, 2025.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139348–139379, 2024.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*, 2025.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

Bangzheng Li, Jianmo Ni, Chen Qu, Ian Miao, Liu Yang, Xingyu Fu, Muhao Chen, and Derek Zhiyuan Cheng. Reinforced attention learning. *arXiv preprint arXiv:2602.04884*, 2026.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12384*, 2024a.

Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo, Qi Wang, Youming Deng, Wenbin Wang, Xian Sun, Haifeng Li, et al. Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(5):6, 2024b.

Zhenshi Li, Dilxat Muhtar, Feng Gu, Xueliang Zhang, Pengfeng Xiao, Guangjun He, and Xiaoxiang Zhu. Lhrs-bot-nova: Improved multimodal large language model for remote sensing vision-language interpretation. *arXiv preprint arXiv:2411.09301*, 2024c.

Ruixun Liu, Bowen Fu, Jiayi Song, Kaiyu Li, Wanchen Li, Lanxuan Xue, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. Zoomearth: Active perception for ultra-high-resolution geospatial vision-language tasks. *arXiv preprint arXiv:2511.12267*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Xu Liu and Zhouhui Lian. Rsunivlm: A unified vision-language model for remote sensing via granularity-oriented moe. *Pattern Recognition*, 179:113717, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2026.113717. URL https://www.sciencedirect.com/science/article/pii/S0031320326006825.

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

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

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

Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. *arXiv preprint arXiv:2503.07588*, 2025.

W. Ma, S. Sun, T. Yu, R. Wang, Tat-Seng Chua, and J. Bian. Thinking with blueprints: Assisting vision-language models in spatial reasoning via structured object representation, 2026a.

Yan Ma, Weiyu Zhang, Tianle Li, Linge Du, Xuyang Shen, and Pengfei Liu. What does vision tool-use reinforcement learning really learn? disentangling tool-induced and intrinsic effects for crop-and-zoom. *arXiv preprint arXiv:2602.01334*, 2026b.

Microsoft GenAI Team. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

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

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o, 2024a.

OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence, 2024b.

OpenAI. Gpt-5.2: Advancing science and math. https://openai.com/zh-Hans-CN/index/introducing-gpt-5-2/, 2025. Accessed: 2026-05-06.

OpenAI. Introducing gpt-5.4. https://openai.com/zh-Hans-CN/index/introducing-gpt-5-4/, 2026. Accessed: 2026-05-06.

Chao Pang, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Xingxing Weng, Shuai Wang, Litong Feng, Gui-Song Xia, et al. H2rsvlm: Towards helpful and honest remote sensing large vision language model. *arXiv preprint arXiv:2403.20213*, 2024.

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

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236*, 2024.

Rotem Shalev-Arkushin, Rinon Gal, Amit H. Bermano, and Ohad Fried. Imagerag: Dynamic image retrieval for reference-guided image generation, 2025. URL https://arxiv.org/abs/2502.09411.

H. Shen, K. Zhao, T. Zhao, R. Xu, Z. Zhang, M. Zhu, and J. Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6613–6629, 2025.

Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, and Hongxu Yin. Scaling vision pre-training to 4k resolution, 2025. URL https://arxiv.org/abs/2503.19903.

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

J. D. Silva, J. Magalhaes, D. Tuia, and B. Martins. Large language models for captioning and retrieving remote sensing images, 2024. URL https://arxiv.org/abs/2402.06475.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danisho, Paolo Fraccaro, Campbell Watson, Levente J. Klein, Salman Khan, and Fahad Khan. Earthdial: Turning multi-sensory earth observations to interactive dialogues. *ArXiv*, 2025. URL https://arxiv.org/abs/2412.15190.

Z. Su, P. Xia, H. Guo, Z. Liu, Y. Ma, X. Qu, J. Liu, Y. Li, K. Zeng, Z. Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

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

C. Wang, Z. Luo, R. Liu, C. Ran, S. Fan, X. Chen, and C. He. Vicot-agent: A vision-interleaved chain-of-thought framework for interpretable multimodal reasoning and scalable remote sensing analysis. *arXiv preprint arXiv:2511.20085*, 2025a.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Boqi Shan, Long Lan, Yulin Wang, Hongzhen Wang, Wenjing Yang, Bo Du, and Jing Zhang. Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. *arXiv preprint arXiv:2505.21375*, 2025b.

Fengxiang Wang, Hongzhen Wang, Mingshuo Chen, Di Wang, Yulin Wang, Zonghao Guo, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? *arXiv preprint arXiv:2503.23771*, 2025c.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Yajie Yang, Yifan Zhang, Long Lan, Xue Yang, Hongda Sun, Yulin Wang, Di Wang, et al. Geoeyes: On-demand visual focusing for evidence-grounded understanding of ultra-high-resolution remote sensing imagery. *arXiv preprint arXiv:2602.14201*, 2026a.

J. Wang, Z. Zheng, Z. Chen, A. Ma, and Y. Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5481–5489, 2024a.

J. Wang, Y. Zhong, Z. Chen, Z. Zheng, A. Ma, and L. Zhang. Earthvl: A progressive earth vision-language understanding and generation framework, 2026b. URL https://arxiv.org/abs/2601.02783.

P. Wang, H. Hu, B. Tong, Z. Zhang, F. Yao, Y. Feng, Z. Zhu, H. Chang, W. Diao, Q. Ye, and X. Sun. Ringmogpt: A unified remote sensing foundation model for vision, language, and grounded tasks. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–20, 2025d. doi: 10.1109/TGRS.2024.3510833.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

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

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing, 2025. URL https://arxiv.org/abs/2506.09965.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Zhenghang Yuan, Zhitong Xiong, Lichao Mou, and Xiao Xiang Zhu. Chatearthnet: A global-scale image-text dataset empowering vision-language geo-foundation models. *Earth System Science Data Discussions*, pp. 1–24, 2024.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024a.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, Jun Li, and Xuerui Mao. Earthmarker: A visual prompting multi-modal large language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024b.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 2024c.

Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A APPENDIX

A.1 OVERVIEW OF THE APPENDIX

This appendix supplements the proposed **GeoLens** and our datasets **GeoMTVR** with details excluded from the main paper due to space constraints.

The appendix is organized as follows:

- Sec. A.2: More implement details of GeoLens.
- Sec. A.3: Ablation studies of GeoLens.
- Sec. A.4: Stability, mechanism, and contributions of RTAL.
- Sec. A.5: Task-level tool usage analysis.
- Sec. A.6: Visualizations of samples and challenging cases.
- Sec. A.7: Potential societal impact of GeoLens and GeoMTVR.

A.2 MORE DETAILS OF PILOT EXPERIMENTS AND GEOLENS

Table 8 in Appendix further explains this effect. Solved samples usually depend on evidence from one localized region, such as recognizing a vehicle type, counting objects in a specified port area, identifying a roof color, or determining the relative position of nearby landmarks. In these cases, the main challenge is to locate the relevant region, and one zoom-in operation often provides enough detail. In contrast, persistently failed samples involve global counting, path planning, land-use enumeration, change counting across image pairs, or reasoning over spatially dispersed clues. These tasks require systematic spatial coverage, multi-region comparison, or heterogeneous perceptual capabilities beyond zoom-in. Although the framework allows up to five zoom-in calls and retains the global image in context, limited coverage and the absence of structured exploration still prevent the model from collecting sufficient evidence.

A.2.1 DATASET DETAILS OF LRS-GRO AND XLRS-BENCH

LRS-GRO. Following LRS-VQA (Luo et al., 2025), LRS-GRO (Liu et al., 2025) is curated from high-resolution remote-sensing imagery sourced from FAIR1M-1.0 (Sun et al., 2022), GLH-Bridge, and STAR (Li et al., 2024b). The final benchmark contains 1,224 high-resolution images with an average resolution of approximately 5,000×5,000 pixels, 3,592 bounding boxes, and 13,245 questions. Among them, 1,000 samples are equipped with detailed step-by-step chain-of-thought annotations and are used as the SFT subset. LRS-GRO emphasizes active perception: models are expected to decide whether regional cropping is necessary and to use local visual evidence when the question scope cannot be resolved from the global view.

The benchmark defines 17 major question categories organized into three hierarchical spatial levels. The *Global* level includes counting, season, urban–rural judgment, and scene type. The *Region* level includes counting, existence, status, visual features, function, and category. The *Object* level includes function, material/surface, category, state, relative position, shape/structure, and color/pattern. For region- and object-level tasks, reference bounding boxes are provided to support supervision and evaluation. This hierarchical design enables evaluation of whether a model can adaptively match its visual evidence acquisition strategy to the spatial scope of a question.

XLRS-Bench. XLRS-Bench (Wang et al., 2025c) is used as the main ultra-high-resolution remote-sensing VQA benchmark in our experiments. It evaluates both perception and reasoning abilities over extremely large remote-sensing images. As shown in Table 9, the benchmark is organized into two Level-1 categories, *Perception* and *Reasoning*, and further decomposed into Level-2 and Level-3 sub-tasks. All Level-3 tasks are formulated as multiple-choice VQA questions, covering 3,080 samples in total.

A.2.2 MORE DETAILS OF GEOLENS

Model architecture. GeoLens is based on the Qwen2.5-VL architecture, specifically `Qwen2_5_VLForConditionalGeneration`. The language backbone contains 28 transformer

Table 8: **More details of Pilot Experiments.** Qualitative analysis of samples solved by single-tool zoom-in and those that remain persistently failed. Single zoom-in is effective for localized evidence acquisition, but struggles with global coverage, multi-region comparison, and structured spatial reasoning in UHR remote-sensing images.

Group	Representative tasks	Key observation
Samples solved during RL training		
Object classification	Vehicle-type recognition	A single zoom-in to the target region reveals discriminative visual details, such as the difference between trucks and buses.
Regional counting	Counting ships in a specified port	The relevant region is spatially localized, and objects become clearly separable after zoom-in.
Attribute recognition	Identifying roof color	Zoom-in mitigates the loss of fine-grained appearance caused by global-view downsampling.
Local spatial relation	Reasoning over nearby landmarks	Both targets can be covered by one local crop, making their relative layout directly observable.
Persistently failed samples		
Global counting	Counting vehicles in a city-scale image	Limited zoom-in calls cannot systematically cover the full UHR image, leaving many targets unobserved.
Path planning	Finding a route between distant points	The global view lacks road-level details, while local zoom-in crops are insufficient to form a coherent long-range path.
Land-use enumeration	Identifying all land-use types	The answer depends on multiple heterogeneous regions, but limited local inspections cover only a small subset of them.
Change counting	Comparing object counts across image pairs	Both images require multiple local inspections, making the fixed tool budget insufficient for reliable comparison.
Complex reasoning	Inferring conditions from dispersed clues	The required evidence is distributed across distant regions, and crop-and-zoom lacks a structured exploration mechanism.

Table 9: **Characteristics of XLRS-Bench.** The benchmark is used to evaluate perception and reasoning abilities in ultra-high-resolution remote-sensing VQA.

L1-Task	L2-Task	L3-Task	Abbr.	Annotation Format	Number of Samples	Answer Type
Perception	Counting	Overall Counting	OC	VQA	60	Multiple Choice (A/B/C/D)
		Regional Counting	RC	VQA	100	Multiple Choice (A/B/C/D)
	Scene Classification	Overall Land Use Classification	OLUC	VQA	100	Multiple Choice (A/B/C/D)
		Regional Land Use Classification	RLUC	VQA	200	Multiple Choice (A/B/C/D)
	Object Spatial Relationship	Object Spatial Relationship	OSR	VQA	500	Multiple Choice (A/B/C/D)
	Object Properties	Object Classification	OCC	VQA	800	Multiple Choice (A/B/C/D)
		Object Color	OCL	VQA	800	Multiple Choice (A/B/C/D)
		Object Motion State	OMS	VQA	60	Multiple Choice (A/B for Yes/No)
Reasoning	Route Planning	Route Planning	RP	VQA	100	Multiple Choice (A/B/C/D)
	Anomaly Reasoning	Anomaly Detection and Interpretation	AD	VQA	100	Multiple Choice (A/B/C/D)
	Complex Reasoning	Environmental Condition Reasoning	ECR	VQA	100	Multiple Choice (A/B/C/D)
		Counting with Complex Reasoning	CCR	VQA	100	Multiple Choice (A/B/C/D)
	Spatiotemporal Reasoning	Regional Counting with Change Detection	RCCD	VQA	60	Multiple Choice (A/B/C/D)

layers with hidden size 3584, intermediate size 18944, 28 attention heads, and 4 key-value heads. The vocabulary size is 152,064. The vision encoder contains 32 layers with hidden size 1280, 16 attention heads, patch size 14, and spatial merge size 2. Visual features are projected into the 3584-dimensional language hidden space. Overall, GeoLens follows the standard Qwen2.5-VL multimodal design, coupling a ViT-style visual encoder with an autoregressive language model.

GeoLens is formulated as an agentic remote-sensing vision-language model. Given an ultra-high-resolution remote-sensing image and a question, the model generates an interleaved trajectory consisting of brief reasoning, structured tool calls, visual observations, and the final answer. It can invoke three tools: `crop_and_zoomin`, `grounding`, and `line`. These tools support progressive local inspection, object-level localization, and auxiliary-line-based spatial reasoning. Returned observations are appended to the context so that subsequent reasoning can condition on newly acquired visual evidence.

Supervised fine-tuning stage. The RTAL stage starts from a supervised checkpoint trained from Qwen2.5-VL-7B-Instruct. The SFT stage uses 13,078 cold-start multimodal instruction samples from three sources. These samples provide multi-tool reasoning trajectories and teach the model the basic format of tool invocation, evidence interpretation, and final answer generation.

RTAL stage. GeoLens is then optimized with RTAL from the SFT checkpoint. The implementation is based on `verl`, using 8 GPUs, vLLM rollout, and FSDP actor/reference execution. The RTAL training data contains 9,979 prompts from SuperVQA and LRS-GRO, with an additional validation set of 244 SuperVQA samples. Tables 10 and 11 summarize the SFT and RTAL data, respectively.

Table 10: **SFT data used for GeoLens.** The SFT stage uses 13,078 cold-start multimodal instruction samples.

Source	Samples
SuperVQA cleaned/accepted data	9,211
LRS-GRO zoom-based open-answer data	3,479
route_visual_accepted_388	388
Total	13,078

Table 11: **RTAL training data used for GeoLens.**

Source	Samples
SuperVQA	8,066
LRS-GRO	1,913
Total	9,979

The SFT training is performed with LLaMA-Factory in a full fine-tuning setting for the language model. The vision tower and multimodal projector are frozen, while the language model is updated. The training uses 8 GPUs with DeepSpeed ZeRO-3, bf16 precision, per-device batch size 1, and gradient accumulation 4, resulting in a global batch size of 32. The learning rate is set to 1×10^{-5} with a cosine scheduler and warmup ratio 0.1. The model is trained for 3 epochs with cutoff length 32768 and maximum image pixels 2,359,296. The final SFT run reaches 1224 optimization steps with a training loss of approximately 0.1199. Tables 12 and 13 list the main SFT and RTAL hyperparameters.

Table 12: **SFT training configuration of GeoLens.**

Parameter	Value
Initialization	Qwen2.5-VL-7B-Instruct
Training framework	LLaMA-Factory
Trainable modules	Language model
Frozen modules	Vision tower and multimodal projector
GPUs	8
Distributed strategy	DeepSpeed ZeRO-3
Precision	bf16
Per-device batch size	1
Gradient accumulation	4
Global batch size	32
Learning rate	1×10^{-5}
Scheduler	Cosine
Warmup ratio	0.1
Epochs	3
Cutoff length	32768
Maximum image pixels	2,359,296
Optimization steps	1224
Final training loss	$\approx$0.1199

Table 13: **RTAL training configuration of GeoLens.**

Parameter	Value
Algorithm	RTAL with GRPO-style policy optimization
Implementation	`verl`
Rollout backend	vLLM
Actor/reference execution	FSDP
GPUs	8
Rollout number n	8
Train batch size	32
PPO mini-batch size	32
Actor PPO micro-batch per GPU	1
Max prompt length	8192
Max response length	1024
Actor learning rate	1×10^{-6}
KL loss	Disabled
Entropy coefficient	0
vLLM tensor parallel size	2
vLLM GPU memory utilization	0.6
Max batched tokens	16384
Save frequency	Every 10 steps
Checkpoint keep limit	2

RTAL attention regularization. During RTAL, the attention regularization is applied to tool-action-related spans and valid generated response tokens through attention JSD terms. The tool-action mask is constructed from spans marked by `Action` and the corresponding closing delimiter, targeting the region where the model decides which tool to call and how to parameterize it. The response mask covers valid generated tokens. This design concentrates optimization on high-impact reasoning and tool-use regions while preserving the standard policy-optimization objective for answer correctness.

Reward and advantage. We use outcome-only GRPO: each complete response receives one scalar reward $0.9 \cdot \text{correctness} + 0.1 \cdot \text{format}$, with no handcrafted reward for a tool, tool type, or observation. Correctness is A–E matching for SuperVQA/XLRS and normalized short-answer matching for LRS-

GRO; format requires a valid <answer>...</answer> output. A tool is therefore rewarded only indirectly through a correct, valid final answer.

For the eight rollouts of each prompt, rewards are normalized by their within-group mean and standard deviation. The resulting group-relative advantage is broadcast to valid response tokens, is computed only within that prompt group, and is neither batch-normalized nor separately computed for tool tokens. RTAL keeps this reward, grouping, and advantage estimator unchanged; it adds the attention-JSD term with the same advantage over Action ... } spans plus a valid-response branch. Its gain thus comes from improved tool-related routing, not altered reward shaping.

Prompting setup. The RTAL prompt frames GeoLens as a remote-sensing visual reasoning assistant capable of calling crop_and_zoomin, grounding, and line. The system prompt requires brief reasoning before tool calls, normalized coordinates in $[0, 1000]$, and final answers inside <answer>...</answer>. The user prompt asks the model to solve the remote-sensing question using visual evidence, optionally invoking tools when necessary, while allowing direct answering when the initial evidence is sufficient.

A.3 Ablation Studies of GeoLens

Table 14: **Data ablation on tool-augmented SFT on XLRS-Bench.** 'Avg.' represents the average accuracy across sub-tasks.

SFT Data	OC	RC	OLUC	RLUC	OCC	OCL	OMS	OSR	AD	ECR	RP	RCCD	CCR	Avg.
w/o Tool Data (SuperRSVQA)	31.7	39.0	37.0	76.0	41.2	38.0	63.3	30.0	72.0	80.0	33.0	45.0	50.0	48.9 ↓ +0.7
SuperRSVQA w/ Zoom-in Tool	31.7	36.0	43.0	73.5	42.5	32.0	60.0	30.0	75.0	79.0	52.0	40.0	50.0	49.6 ↓ +0.7
SuperRSVQA w/ Full Tool Data	31.7	40.0	29.0	74.5	40.1	33.1	66.7	42.0	71.0	81.0	57.0	38.3	50.0	50.3

Data Ablation on Tool-Augmented SFT. To systematically analyze the effect of tool-related data during supervised fine-tuning, we construct three progressively enhanced variants based on SuperRSVQA, as shown in Table 14. The first setting uses only the original SuperRSVQA data and is denoted as w/o Tool Data. The second setting further introduces zoom-in local-crop operations into SuperRSVQA, providing single-tool active-perception supervision. The third setting augments SuperRSVQA with the full tool set, including progressive zoom-in, object grounding, and auxiliary-line-based spatial reasoning. Under the same evaluation protocol on XLRS-Bench, performance improves as tool information is gradually introduced. Zoom-in data strengthens local-region perception and benefits fine-grained recognition and regional reasoning, while the complete multi-tool chain further improves multi-step reasoning and complex decision-making. This indicates that tool-augmented data provides not only additional supervision but also structural guidance for learning how to acquire and integrate visual evidence. However, the gain from tool annotations alone remains limited in complex reasoning scenarios, suggesting that an effective learning mechanism is necessary to fully exploit the potential of tool-use data.

Table 15: **Module-level ablation on training stages on XLRS-Bench.** 'Avg.' represents the average accuracy across sub-tasks.

Training Stage	OC	RC	OLUC	RLUC	OCC	OCL	OMS	OSR	AD	ECR	RP	RCCD	CCR	Avg.
Qwen2.5-VL-7B	33.3	40.0	31.0	77.0	40.6	40.5	66.7	36.2	68.0	72.0	27.0	38.3	45.0	47.4 ↓ +4.3
Base Model + SFT	26.7	36.0	45.0	82.5	44.6	39.2	65.0	32.4	71.0	82.0	55.0	43.3	49.0	51.7 ↓ +2.5
Base Model + SFT + RTAL (GeoLens)	35.0	40.0	37.0	76.5	46.5	45.6	65.0	36.8	72.0	82.0	62.0	51.7	55.0	54.2

Module-Level Ablation on Training Stages. To verify the contribution of each training stage, we compare three settings in Table 15: the Qwen2.5-VL-7B base model, the model after supervised fine-tuning, and the final GeoLens model trained with SFT and RTAL. Introducing SFT substantially improves task alignment and basic visual understanding, showing that supervised training provides an effective initialization for remote-sensing tool use. Building on this checkpoint, RTAL further improves performance by optimizing the internal information routing and tool-use decisions during reasoning. This confirms the effectiveness of the progressive training paradigm from base modeling, to supervised alignment, and then to reinforcement-based attention optimization.

Effectiveness of RTAL against Token-Level RL. To further validate RTAL, we compare it with a standard token-level policy-optimization method, GRPO, under the same training data and ini-

Table 16: **Effectiveness of RTAL against token-level RL on XLRS-Bench.** 'Avg.' represents the average accuracy across sub-tasks.

RL Method	OC	RC	OLUC	RLUC	OCC	OCL	OMS	OSR	AD	ECR	RP	RCCD	CCR	Avg.
Base Model + SFT + GRPO	31.7	39.0	40.0	79.0	47.6	45.5	65.0	35.6	70.0	82.0	57.0	43.3	45.0	52.4 ↓ +1.8
Base Model + SFT + RTAL (GeoLens)	35.0	40.0	37.0	76.5	46.5	45.6	65.0	36.8	72.0	82.0	62.0	51.7	55.0	54.2

tialization conditions. As summarized in Table 16, RTAL consistently outperforms token-level RL in overall performance. Qualitative case analysis further shows that RTAL encourages clearer and more structured reasoning paths during tool invocation, leading to more reasonable tool selection and parameter generation. In contrast, token-level optimization is more prone to redundant calls or unstable decisions. These results suggest that optimizing the model's internal attention distribution provides more effective credit assignment than directly optimizing only the output-token distribution, thereby improving performance on complex reasoning tasks.

A.4 Stability, Mechanism, and Contributions of RTAL

A.4.1 Stability of RTAL attention optimization

RTAL does not treat the final-layer attention weights as a complete causal explanation of model decisions. Instead, we view the attention distribution as an interface for optimizing information routing rather than as an explicit interpretability signal. RTAL therefore does not force the model to attend to predefined locations or simply maximize the attention weights of specific tokens. Instead, it constrains tool-related attention distributions through an advantage-weighted Jensen–Shannon divergence, encouraging them to approach effective routing patterns observed in high-reward trajectories.

Regarding the concern that optimizing gradients only on tool-call spans may cause attention collapse, RTAL contains two complementary attention optimization branches. The tool-oriented branch constructs masks over structured `Action` spans—including tool names, parameter fields and values, and structural markers—to optimize critical decisions such as tool selection and parameter generation. However, RTAL does not suppress normal response tokens: it simultaneously retains a valid-response attention regularization branch that applies normalized attention-JSD constraints to valid response tokens. RTAL thus improves tool-oriented decision making while avoiding excessive restriction of general language generation.

Furthermore, although standard GRPO often adopts output-level KL regularization to constrain the deviation between the current and reference policies, RTAL already introduces explicit distribution constraints within its attention optimization. Specifically, the attention-level objective optimizes the JSD between the current and previous attention policies rather than forcing attention toward a fixed pattern. Removing the GRPO KL term therefore does not imply that RTAL attention distributions can drift without constraint.

Although our analysis suggests that KL regularization is not necessary for RTAL stability, we further add two ablation studies: (1) a KL on/off ablation to examine the effect of output policy constraints on RTAL, and (2) a sensitivity analysis with different λ_{attn} values to evaluate RTAL robustness to the attention loss weight. As shown in Tables 17 and 18, KL brings no additional gain under the same configuration, while all tested λ_{attn} values remain close. These results support stable RTAL optimization without dependence on a narrow coefficient or output-level KL in this setting.

Table 17: **Ablation of output-level KL regularization.** RTAL without KL achieves performance comparable to RTAL with KL, indicating that the attention-level constraints are sufficient for stable optimization.

Method	KL	XLRS Avg.	Reasoning Avg.	Invalid calls	Avg. calls
SFT + GRPO	0.01	52.4	59.5	19.80%	1.89
RTAL w/o KL	0	54.2	64.5	15.70%	2.40
RTAL + KL	0.01	54.1	64.3	15.85%	2.39

Table 18: **Sensitivity of RTAL to the attention loss weight λ_{attn}.** All tested values remain close, indicating stable optimization without dependence on a narrow coefficient.

λ_{attn}	XLRS Avg.	Perception Avg.	Reasoning Avg.
0.05	54.0	47.8	63.9
0.10 (default)	54.2	47.8	64.5
0.20	54.1	47.7	64.3

A.4.2 TOOL-CALL QUERY MASKING

The phrase "focusing on tool tokens" refers to where the auxiliary RTAL loss is triggered, not where the model is required to place its attention. More precisely, RTAL treats the model's attention patterns during tool-call sequence generation as actions: it applies the auxiliary tool-specific attention objective at tool-call query tokens, while retaining the causal context as attention keys and values. Consider an interleaved trajectory

$$\text{Question} \rightarrow \text{reasoning} \rightarrow \texttt{Action}(\cdot) \rightarrow \text{observation} \rightarrow \text{reasoning} \rightarrow \texttt{Action}(\cdot) \rightarrow \text{answer}.$$

For a token position inside `Action(tool, arguments)`, RTAL computes its attention distribution over all preceding valid tokens. A later tool call may therefore attend to the original question, image tokens, previous reasoning, earlier tool calls, and returned observations. The mask determines where the auxiliary attention loss is evaluated, not which contextual tokens the model is allowed to use.

RTAL also does not replace ordinary sequence optimization. Standard GRPO continues to optimize generated responses, while the tool-specific JSD branch targets the positions where executable decisions are serialized, including the tool name, argument fields, and argument values. A separate valid-response regularization branch is retained for valid response tokens. The tool-call mask thus provides targeted credit assignment for tool selection and parameter generation without suppressing general reasoning or language generation. The masking behavior is summarized in Table 19.

Table 19: **Summary of masking behavior in RTAL.** The mask determines where the auxiliary attention loss is evaluated, not which contextual tokens the model is allowed to use.

Component	Query positions receiving the loss	Available causal context
Standard GRPO	Generated response tokens	Standard autoregressive context
Tool-specific attention JSD	(`Action...`) positions only	All preceding valid tokens
Valid-response regularization	Valid response positions	All preceding valid tokens

The masks are implemented as follows:

```
tool_query_mask = parse_action_spans(input_ids)
                  # tool names, fields, values, delimiters
response_mask   = parse_responses(input_ids)

jsd = JensenShannon(current_attn, previous_attn)

tool_loss    = lambda_tool * masked_mean(jsd, mask=tool_query_mask)
               # omitting advantage for brevity
response_reg = lambda_resp * masked_mean(jsd, mask=response_mask)
```

A.4.3 CONTRIBUTIONS OF SFT AND RTAL

SFT and RTAL are consecutive contributions: GeoMTVR SFT establishes tool calling, observation use, and evidence integration, while RTAL further improves routing around tool actions and call quality. As shown in Table 20, relative to GRPO, RTAL adds 0.51 calls and 0.13 tool types per

question while increasing valid calls by 4.10 pp; relative to SFT, it reduces invalid calls by 7.90. It therefore yields more diverse and more reliable interaction, consistent with its +5.0 reasoning-average gain (64.5 vs. 59.5).

Table 20: **Contributions of SFT and RTAL.** RTAL improves routing around tool actions and call quality over both SFT and GRPO.

Stage	XLRS Avg.	Invalid calls	Avg. calls	Valid calls	Avg. types
Qwen2.5-VL-7B	47.4	-	-	-	-
SFT	51.7 (+4.3)	23.6%	1.81	76.4%	1.09
SFT + GRPO	52.4 (+5.0)	19.8%	1.89	80.2%	1.13
SFT + RTAL (GeoLens)	54.2 (+6.8)	15.7% (-4.1)	2.40	84.3%	1.26

A.5 Task-Level Tool Usage Analysis

Remote-sensing tasks exhibit significant task heterogeneity: many tasks can be solved with a single or even zero tool calls, while more challenging tasks, such as route planning, require coordinated multi-tool interactions. GeoLens is specifically designed to address the limitation of existing single-tool remote-sensing MLLMs on such complex tasks; it consistently outperforms GeoEyes on challenging tasks such as Route Planning (RP).

To quantify this, we analyze tool usage across all 3,080 XLRS-Bench samples. The preferred tool trajectories contain 7,389 tool calls in total, averaging 2.399 calls per question. The average usage frequencies of the Zoom-in, Grounding, and Line tools are 1.261, 1.059, and 0.079, respectively. Table 21 reports the per-task statistics.

Table 21: **Per-task tool usage statistics on XLRS-Bench.** Tool usage is highly task-dependent; complex tasks such as RC, OSR, and RP require coordinated multi-tool interactions.

Task	Samples	Zoom-in	Grounding	Line	Total calls
OC	60	0.383	0.900	0.050	1.333
RC	100	1.720	1.130	0.000	2.850
OLUC	100	0.430	0.300	0.000	0.730
RLUC	200	1.560	1.040	0.010	2.610
OCC	800	1.627	0.956	0.015	2.599
OCL	800	1.484	1.204	0.004	2.691
OMS	60	1.000	0.533	0.000	1.533
OSR	500	0.842	1.744	0.178	2.764
AD	100	0.390	0.220	0.270	0.880
ECR	100	0.750	0.540	0.130	1.420
RP	100	1.180	0.650	0.830	2.660
RCCD	60	1.350	0.950	0.000	2.300
CCR	100	0.490	0.280	0.120	0.890
Overall	3,080	1.261	1.059	0.079	2.399

These statistics show that tool usage is highly task-dependent rather than uniformly applied. Different tasks require different forms of evidence: simple tasks may rely on a single operation, while complex tasks requiring local details, object localization, and spatial structure reasoning benefit from coordinated tool usage. For example, Route Planning (RP) requires complementary evidence from multiple tools, with an average of 2.660 tool calls per question, indicating that a single zoom-in operation is insufficient for fully supporting such reasoning.

These task-level statistics further support our pilot study (Sec. 4.1), which shows that a single zoom-in tool is insufficient for complex UHR remote-sensing reasoning. Importantly, GeoLens does not enforce multi-tool usage for all tasks. Instead, it adaptively coordinates different tools only when tasks require complementary evidence from local details, object-level localization, and spatial structure reasoning (e.g., RC, OSR, and RP).

A.6 VISUALIZATIONS OF SAMPLES IN GEOMTVR

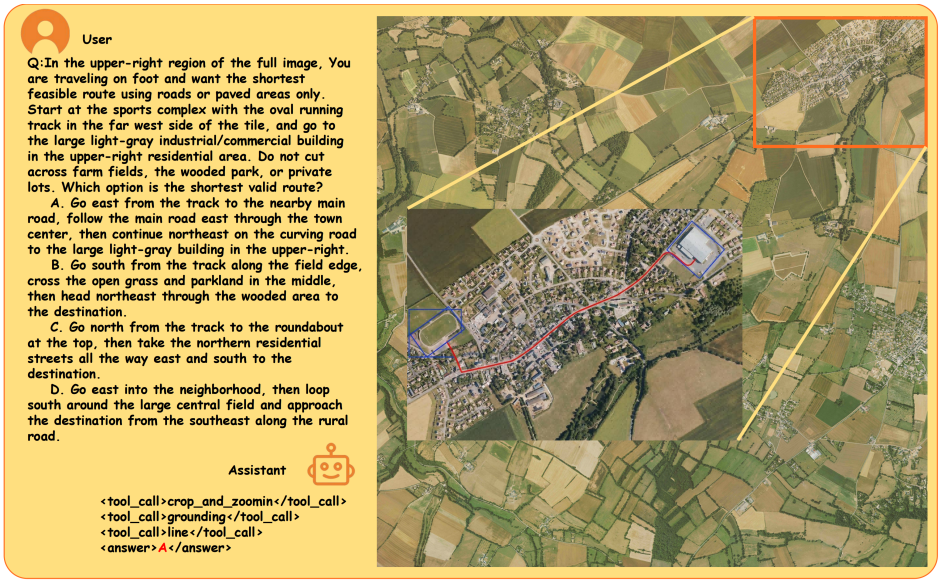

Figure 6: **Visualization of a GeoMTVR sample.** The example illustrates the interleaved reasoning trajectory and visual tool use in GeoMTVR.

A.7 POTENTIAL SOCIETAL IMPACT

Potential positive societal impact. GeoLens and GeoMTVR are designed to improve visual evidence acquisition and reasoning over ultra-high-resolution remote-sensing imagery. Such capabilities can support applications where analysts need to inspect large-area scenes efficiently, such as environmental monitoring, urban planning, land-use analysis, disaster assessment, infrastructure inspection, and ecological observation. Compared with direct global-image reasoning, the multi-tool reasoning paradigm encourages the model to expose intermediate evidence through zoom-in, grounding, and auxiliary-line operations, which may make remote-sensing VQA systems more interpretable and easier to audit. GeoMTVR may also provide a useful research resource for studying active visual reasoning, tool-use behavior, and reliable multimodal decision making under very large image resolutions.

Potential negative societal impact. The proposed model and dataset are intended for research and decision-support scenarios rather than autonomous high-stakes deployment. As with other remote-sensing analysis systems, incorrect predictions may arise from cloud cover, seasonal changes, outdated imagery, low-quality annotations, or distribution shifts across sensors and geographic regions. Therefore, GeoLens should not be treated as a substitute for expert review in operational settings where errors may have material consequences. In addition, improved analysis of large-area imagery could be misapplied if used without appropriate governance or domain constraints. We therefore recommend that future releases include clear documentation of intended use, evaluation scope, known limitations, and responsible-use terms. Practical deployments should keep humans in the loop, verify important outputs with external evidence, and follow applicable regulations and institutional review procedures.