

## Article

# Crop-Tool-Augmented Active Perception with Reinforcement Learning for High-Resolution Remote Sensing Visual Question Answering

Qian Li, Kailiang Chen, Yitong Han  and Xiangyang Xu \*

School of Automation, Beijing Institute of Technology, Beijing 100081, China

\* Correspondence: xxy1970@bit.edu.cn

## Highlights

### What are the main findings?

- We propose a crop-tool-augmented active perception framework for high-resolution remote sensing visual question answering, enabling the model to actively determine whether local evidence is needed and select question-relevant regions for fine-grained reasoning.
- We introduce a GRPO-based reinforcement learning strategy with a region-aware reward design that combines output-format constraints, reference-region coverage, answer semantic consistency, and cropping penalties to improve cropping decisions and answer generation.

### What are the implications of the main finding?

- The findings indicate that active region selection is an effective strategy for addressing detail loss, background redundancy, and scale variation in ultra-high-resolution remote sensing imagery.
- This study provides an interpretable and trainable framework for developing remote sensing vision–language models with adaptive perception, dynamic tool use, and more reliable question answering capability.

## Abstract

High-resolution remote sensing visual question answering (RS-VQA) requires models to identify question-relevant regions and reason over fine-grained visual evidence. However, existing vision–language models usually rely on fixed global image inputs, which may lose critical local details in ultra-high-resolution imagery and struggle with sparse informative regions, large object-scale variations, and complex spatial layouts. To address these challenges, this paper proposes a crop-tool-augmented active perception framework with reinforcement learning. The framework introduces structured tokens to explicitly organize the reasoning process into question understanding, cropping decision-making, local evidence acquisition and final answer generation. Based on this design, the model can actively determine whether a cropping operation is needed and select task-relevant regions for further inspection. To enable stable tool-use and multi-turn reasoning in a compact vision–language model, we construct teacher-guided cropping reasoning trajectories from high-resolution images, question–answer pairs, and annotated regions in the LRS-GRO dataset, and use them for cold-start supervised fine-tuning of Qwen2.5-VL-3B. Furthermore, we introduce Group Relative Policy Optimization to refine the model’s active perception policy. A region-aware reward function is designed by integrating output-format constraints, reference-region coverage, answer semantic consistency, and cropping penalties, which encourages compact and informative region selection while reducing redundant



Academic Editors: Keshav D. Singh,  
Abdul Bais and Saeid Hodayouni

Received: 19 May 2026

Revised: 23 June 2026

Accepted: 1 July 2026

Published: 8 July 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

tool invocations. Experiments on VRSBench, MME-RealWorld-RS, XLRS-Bench, and LRS-VQA demonstrate that the proposed method achieves competitive overall performance compared with closed-source, open-source, and remote-sensing-specific vision–language models, and obtains the best or comparable results on most benchmarks. Ablation studies further verify the effectiveness of structured supervised fine-tuning, reinforcement learning optimization, and the proposed reward design.

**Keywords:** remote sensing visual question answering; high-resolution remote sensing imagery; vision–language model; Group Relative Policy Optimization

## 1. Introduction

Remote sensing imagery contains rich spectral, spatial, and temporal information, providing essential data support for a wide range of Earth observation applications, such as land-use classification, target recognition, environmental monitoring, disaster assessment, and urban planning [1,2]. In recent years, deep learning-based computer vision methods have achieved substantial progress in predefined remote sensing tasks, including object detection and scene classification [3]. However, these methods are usually designed for fixed task settings and closed output spaces, which limits their ability to respond to open-ended user queries. Such queries often require the joint understanding of visual content, natural language semantics, and spatial relationships. For example, answering the question “How many buildings in this region were damaged after the flood?” requires region localization, fine-grained object recognition, spatial reasoning, and free-form answer generation.

Visual Question Answering (VQA) provides a natural human–computer interaction paradigm by allowing users to query image content through natural language and obtain corresponding answers [4]. In the remote sensing domain, Remote Sensing Visual Question Answering (RS-VQA) is particularly valuable for disaster response, urban governance, environmental assessment, and land resource monitoring [5]. Early RS-VQA methods mainly adopted classification-based formulations, where answers were selected from a predefined candidate set [6,7]. Although these methods demonstrate the feasibility of RS-VQA, their closed-vocabulary design restricts free-form answer generation and makes it difficult to handle questions involving multi-step reasoning or fine-grained spatial understanding [8].

The rapid development of multimodal large language models has provided new opportunities for overcoming these limitations. Vision–language models (VLMs) extend large language models with visual perception modules, enabling image understanding and natural language interaction in a unified architecture [9]. These models have shown strong performance on general-domain VQA benchmarks [10]. With their introduction into remote sensing scenarios, recent studies have begun to explore natural language-based interpretation and analysis of geospatial data [8]. For example, GeoChat [11] and SkyEyeGPT [12] employ instruction tuning on remote sensing datasets to support conversational tasks such as image captioning, visual grounding, and visual question answering. SkySenseGPT [13] adopts a typical vision–language architecture by combining a Contrastive Language–Image Pre-training (CLIP) visual encoder with a large language model, and applies Low-Rank Adaptation (LoRA) fine-tuning on a large-scale instruction dataset covering classification, VQA, and object detection tasks.

Despite these advances, both general-domain VLMs and remote sensing VLMs still face notable challenges in RS-VQA, especially for ultra-high-resolution imagery [14,15]. This limitation mainly stems from two factors. First, remote sensing images often contain

thousands of pixels in height and width. When such images are resized to meet the fixed input resolution of VLMs, many small objects and local visual clues may be compressed or lost [10,16]. Second, remote sensing scenes exhibit substantial scale variation. A single image may contain both large geographic structures and tiny objects, which places high demands on multi-scale visual perception [17,18]. Many existing remote sensing VLMs still follow a passive perception paradigm, where the whole image is processed in a single forward pass. This strategy may overlook question-relevant local details in ultra-high-resolution images, thereby limiting reasoning and answer generation in complex RS-VQA tasks [19].

To enhance reasoning in remote sensing scenarios, recent studies have investigated reasoning-enhanced VLMs. Chain-of-Thought (CoT) prompting encourages models to generate intermediate reasoning steps and has been shown to improve multimodal problem solving [20]. In RS-VQA, RS-EoT [21] proposes a language-driven multi-turn image interaction strategy to reduce premature conclusions. CoT-GeoReason [22] guides models to sequentially perform feature detection, spatial analysis, and answer synthesis, achieving favorable results on the EarthVQA benchmark [23]. These studies indicate that explicit reasoning mechanisms can improve the understanding of complex geospatial questions.

Building upon this direction, tool-use mechanisms have been introduced to enable models to actively acquire local visual evidence rather than relying only on a single global image input. In the active perception paradigm, a model can dynamically invoke cropping or zooming tools according to the question, progressively shifting from global observation to local inspection. This process is better suited to fine-grained evidence acquisition in ultra-high-resolution imagery. In addition to studies on tool use and multimodal reasoning [24–26], recent works have further explored active visual perception for high-resolution or geospatial vision–language tasks. For example, VisionThink [27] enables models to request higher-resolution regions when necessary; DeepEyes [28] studies iterative “thinking with images”; ZoomEarth [29] introduces an adaptive cropping–zooming framework based on GRPO; RSThinker [30] constructs perception-grounded geospatial reasoning trajectories; and GeoEyes [31] analyzes tool-invocation homogenization and evidence-based reward design. Different from ZoomEarth, which mainly focuses on adaptive cropping and zooming, our method treats cropping as an explicit tool-use action for VHR RS-VQA and combines structured prompting, teacher-guided SFT, GRPO-based policy optimization, and region-aware reward design to optimize both crop selection and answer generation. These studies provide useful references for active perception in remote sensing scenarios, but challenges remain in stable crop-tool invocation, fine-grained evidence localization, and efficient reasoning under ultra-high-resolution inputs.

Although RS-VQA has gradually evolved from passive perception to active perception, several challenges remain unresolved. First, question-relevant evidence in ultra-high-resolution remote sensing images is often distributed in local regions, requiring models to identify fine-grained visual clues under global semantic guidance. Second, active perception requires not only accurate answer generation but also stable and executable tool invocation; otherwise, multi-turn perception and reasoning cannot be reliably maintained. Third, although human-annotated crop boxes provide useful supervision, directly training small-scale VLMs to perform multi-turn crop-based reasoning remains difficult. Fourth, during reinforcement learning, traditional Intersection over Union (IoU)-based rewards may suffer from sparsity, while insufficient constraints on tool-use cost can lead to redundant cropping and reduced inference efficiency.

To address these challenges, this paper investigates fine-grained evidence acquisition and active reasoning for ultra-high-resolution RS-VQA. We propose a crop-tool-augmented active perception framework that integrates structured prompting, teacher-guided cold-

start supervised fine-tuning (SFT), and GRPO-based reinforcement learning. Specifically, special tokens are introduced to regulate the model's thinking, cropping, and answering behaviors, allowing the reasoning process to be expressed in a structured and executable form. A teacher model is then used to generate high-quality multi-turn crop-based reasoning trajectories for cold-start SFT of a small-scale VLM. Finally, we design a reward function tailored to the spatial characteristics of remote sensing imagery, further optimizing region selection, tool-use rationality, and final answer accuracy under GRPO-based reinforcement learning.

The main contributions of this paper are summarized as follows:

1. We propose a structured active perception framework for ultra-high-resolution RS-VQA. By introducing special tokens such as <think>, <crop>, and <answer>, the model's reasoning process is organized into explicit stages, including reasoning, crop decision-making, local evidence acquisition, and answer generation. This design improves the model's ability to focus on question-relevant fine-grained regions.
2. We design a teacher-guided cold-start SFT strategy for crop-tool-based reasoning. Based on high-resolution images, question-answer pairs, and human-annotated crop boxes from the LRS-GRO dataset, Seed 1.6 is used to generate multi-turn crop-based reasoning trajectories. These trajectories help Qwen2.5-VL-3B learn structured reasoning patterns, executable tool invocation formats, and local-evidence-based answer generation.
3. We introduce GRPO-based reinforcement learning to optimize the active perception policy and design a reward function adapted to remote sensing spatial characteristics. The reward integrates format constraints, region coverage, answer semantic consistency, and cropping penalties, thereby alleviating IoU reward sparsity and reducing redundant crop-tool invocations while improving region selection quality and final RS-VQA performance.

## 2. Materials and Methods

### 2.1. Training Dataset

This study uses the LRS-GRO dataset as the primary training resource [29]. LRS-GRO is designed for active perception-based RS-VQA on ultra-high-resolution imagery, emphasizing a model's ability to perform semantic understanding, region localization, and fine-grained reasoning over large-scale images. Unlike conventional RS-VQA datasets that focus mainly on global scene comprehension, LRS-GRO encourages models to actively attend to question-relevant regions, making it well-suited for vision-language models with tool-use and region-cropping capabilities.

The dataset aggregates images from multiple public remote sensing sources, including FAIR1M-1.0, GLH-Bridge, and STAR [32–34], covering diverse scenarios such as airports, bridges, ports, industrial sites, urban regions, and rural areas. In total, LRS-GRO contains 1224 high-resolution images with resolutions ranging from 4000 to 5000 pixels, each encompassing extensive geographic coverage and numerous fine-grained objects. The dataset provides 3592 annotated bounding boxes and 13,245 VQA samples. Bounding boxes are organized hierarchically into region-level and object-level annotations. Region-level boxes represent semantically coherent areas composed of multiple objects (e.g., airports or industrial zones), while object-level boxes correspond to individual entities with homogeneous semantics (e.g., airplanes, ships, buildings, bridges). This hierarchy facilitates learning of both region selection and multi-scale visual reasoning.

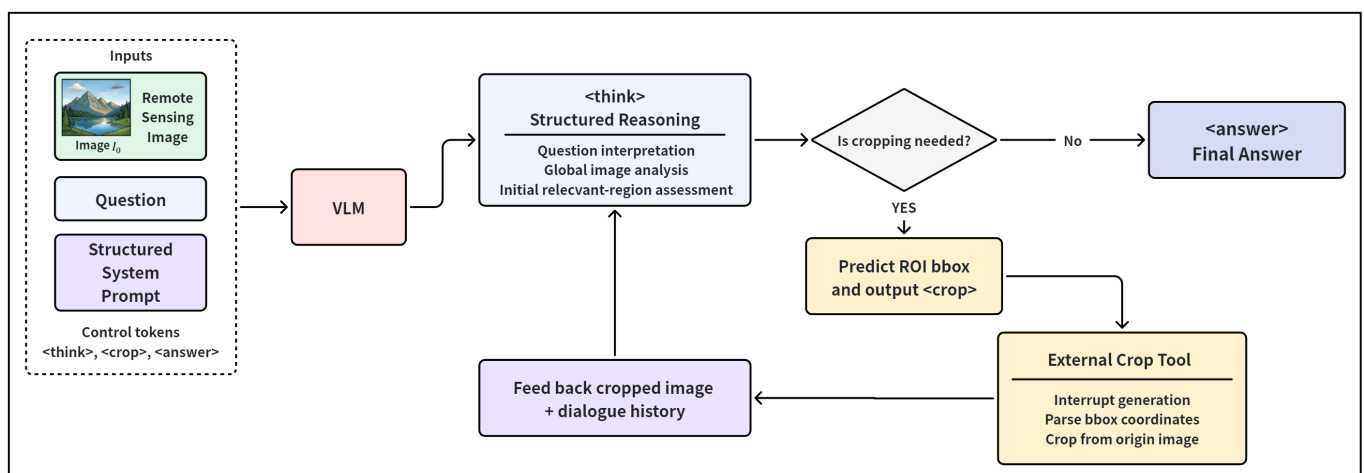
LRS-GRO structures VQA tasks across three spatial levels—global, region, and object—spanning 17 question types. The global level assesses overall scene comprehension, the region level emphasizes localized semantic reasoning, and the object level targets fine-grained

attributes and relationships of individual entities. This multi-level design challenges models to reason across spatial scales and dynamically identify relevant regions, effectively evaluating active perception and fine-grained reasoning capabilities.

## 2.2. Structured Reasoning and Tool-Use Framework

Early studies on tool-augmented large language models mainly relied on predefined triggering rules or externally controlled tool-use strategies. In such settings, the model itself had limited ability to determine when and how external tools should be invoked [24,25]. Recent reasoning-oriented models have shown that explicitly generating intermediate reasoning steps, commonly formulated as Chain-of-Thought (CoT), can improve performance on complex tasks [26,35,36]. This test-time reasoning paradigm suggests that allocating additional computation to structured reasoning during inference can benefit multimodal understanding, particularly in remote sensing scenarios that require spatial analysis and evidence localization.

Motivated by this idea, we design a structured system prompt for RS-VQA that integrates crop-tool invocation into the generation process through specialized control tokens [20,22,29,31]. The prompt constrains the model's output to follow a stepwise procedure of “reasoning–decision–execution–answering”. As shown in Figure 1, the model first analyzes the input image and question, then decides whether local visual evidence is necessary. If the global image already provides sufficient information, the model directly produces the final answer. Otherwise, it predicts a region of interest and invokes the cropping tool to obtain a localized image for further analysis.



**Figure 1.** Structured reasoning and crop-tool-augmented active perception framework for RS-VQA.

Specifically, we introduce three special tokens, namely `<think>`, `<crop>`, and `<answer>`. The `<think>` token elicits structured reasoning, including question intent interpretation, global scene analysis, and preliminary identification of potentially relevant regions [20,26,27]. This stage encourages the model to integrate available information before making a tool-use decision or generating the final answer, thereby reducing the instability of direct response generation.

If the model determines during the `<think>` stage that the current visual input is insufficient, it predicts a region of interest and outputs the corresponding bounding-box coordinates within the `<crop>` token. The `<crop>` token serves as an executable trigger for crop-tool invocation [24,25,29,31]. Once this token is detected, the system pauses generation, parses the enclosed coordinates, and crops the corresponding region from the original remote sensing image using an external Python-based tool. The cropped image is

then returned to the model as an updated visual input together with the dialogue history, forming a multi-turn active perception process.

After receiving the cropped image, the model continues reasoning based on the localized evidence. This iterative procedure enables fine-grained visual analysis and progressively refines the answer. When sufficient evidence has been acquired, the model outputs the final result within the <answer> token. The answer is expressed in concise natural language for subsequent parsing and evaluation. By parsing the sequence of special tokens, the system can record the reasoning trajectory, crop-tool behavior, and final prediction in a unified structured format.

Overall, the proposed prompt design explicitly incorporates active perception into the generative reasoning process [29–31]. It enables the model to reason before deciding whether to invoke the cropping tool, thereby improving local evidence acquisition in complex remote sensing scenes. Moreover, the structured representation of intermediate reasoning steps and region-selection behavior provides interpretable and verifiable signals, which are further used for reinforcement learning optimization.

### 2.3. Cold-Start SFT with Teacher-Generated CoT

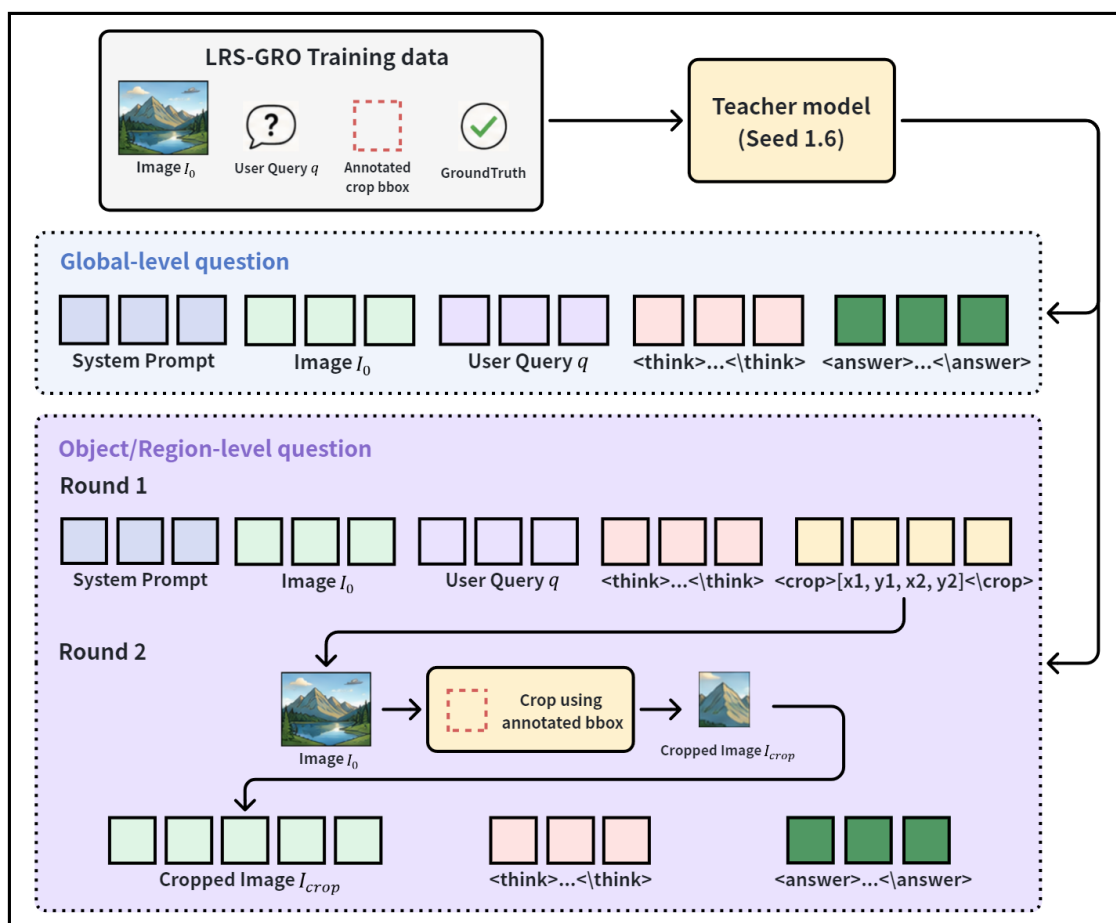
We adopt Qwen2.5-VL-3B as the backbone model for training [37]. Qwen2.5-VL follows a multimodal architecture that integrates a vision encoder with a large language model, enabling unified modeling of visual and textual inputs. Unlike conventional VLMs that rely on fixed-resolution preprocessing, Qwen2.5-VL introduces a dynamic-resolution mechanism that adaptively generates visual tokens according to the original image size. This design helps preserve fine-grained visual details in high-resolution images while maintaining computational efficiency. These properties make Qwen2.5-VL-3B suitable for integration with our crop-tool-augmented active perception framework, where both global scene understanding and local evidence analysis are required.

However, the original Qwen2.5-VL-3B model is not explicitly optimized for structured reasoning or tool invocation (Appendix A). Due to its limited model scale, it may fail to consistently follow system prompts that require multi-step reasoning traces, predefined special tokens, and executable crop-tool instructions. To address this issue, we introduce a teacher-guided cold-start supervised fine-tuning strategy, as shown in Figure 2 [24–26]. Specifically, we use high-resolution images, manually annotated crop boxes, and ground-truth answers from the LRS-GRO training set, and employ Seed 1.6 as the teacher model to generate structured multi-turn reasoning trajectories [29,38]. These trajectories provide explicit demonstrations of reasoning, crop-tool invocation, and local-evidence-based answer generation, enabling the student model to acquire preliminary active perception capabilities during the cold-start stage.

The teacher model is guided by a carefully designed prompt and is required to follow the predefined output format specified by the system prompt. Given an image-question pair, the teacher first provides a concise description of the global image context. It then analyzes the question and determines whether local visual evidence is needed. For region-level and object-level questions, the teacher generates crop-based reasoning trajectories, including target localization, justification for cropping, and coordinate prediction. For global-level questions, the teacher is instructed to complete the reasoning process without invoking the cropping tool and to directly generate the final answer.

For questions requiring local evidence, the teacher outputs the annotated bounding box from the training set in the first reasoning round, enclosed within the special token <crop> [29,37]. Since Qwen2.5-VL uses absolute coordinates to represent spatial regions during pretraining, we retain this coordinate format to avoid additional normalization and denormalization operations. This design reduces the learning difficulty associated with

mapping between tool invocation formats and spatial locations. The original image is then cropped according to the predicted coordinates, and the resulting local region is fed back to the teacher as the second-round visual input. Based on this localized evidence, the teacher continues reasoning and generates the final answer. In this way, the constructed trajectories teach the student model not only how to answer questions, but also when to invoke the cropping tool, which region to select, and how to use local visual evidence for reasoning.



**Figure 2.** Pipeline for constructing multi-turn crop-based reasoning trajectories.

After constructing the cold-start trajectory dataset, we perform supervised fine-tuning on Qwen2.5-VL-3B within the proposed active perception framework. Let the teacher-generated dataset be denoted as

$$\mathcal{D} = \{(I_0, q, \mathcal{T}_i)\}_{i=1}^N, \quad (1)$$

where  $I_0$  represents the original high-resolution image,  $q$  denotes the user query, and

$$\mathcal{T}_i = \{(r_1, v_1), (r_2, v_2), \dots, (r_T, a)\} \quad (2)$$

denotes a multi-turn trajectory with  $T$  interaction rounds. Here,  $r_t$  is the reasoning text at step  $t$ ,  $v_t$  represents the visual input at step  $t$ , and  $a$  is the final answer.

The SFT objective is defined as the negative log-likelihood loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} \left[ \sum_{t=1}^{|\mathbf{y}|} \log P_{\theta}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) \right]. \quad (3)$$

Here,  $x$  consists of the system prompt, visual tokens generated by the dynamic-resolution encoder, and the user query, while  $y$  denotes the target output sequence containing reasoning text, special tokens such as  $\langle \text{think} \rangle$  and  $\langle \text{crop} \rangle$ , and the final answer. For trajectories involving cropping, the second-round visual input corresponds to the cropped local image  $I_{\text{crop}}$ . The model is trained to first generate a valid crop instruction, such as  $\langle \text{crop} \rangle x_1, y_1, x_2, y_2 \langle / \text{crop} \rangle$ , and then continue reasoning based on the returned local evidence.

To improve training efficiency while preserving the general capability of the backbone model, we adopt Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning [39]. Specifically, low-rank decomposition matrices are inserted into the Transformer attention modules and multilayer perceptron layers. For a pretrained weight matrix  $W_0 \in \mathbb{R}^{d \times k}$ , LoRA parameterizes the updated weight as follows:

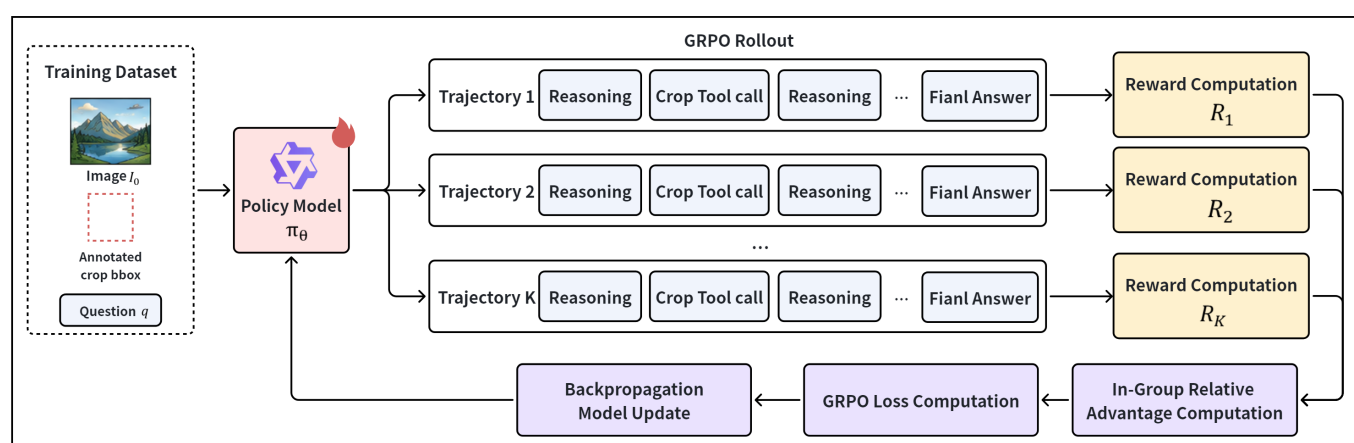
$$W = W_0 + \Delta W = W_0 + BA, \quad (4)$$

where  $B \in \mathbb{R}^{d \times r}$ ,  $A \in \mathbb{R}^{r \times k}$ , and  $r \ll \min(d, k)$ . During fine-tuning, only the low-rank matrices  $A$  and  $B$  are optimized, while the original parameters  $W_0$  remain frozen.

#### 2.4. Group Relative Policy Optimization

After cold-start SFT, the model acquires basic abilities in structured reasoning and crop-tool invocation. However, supervised fine-tuning mainly imitates teacher-generated trajectories and does not directly optimize the model's decision policy according to task-level feedback. In complex ultra-high-resolution RS-VQA scenarios, the model may still select suboptimal regions, invoke the cropping tool unnecessarily, or fail to balance local evidence acquisition with inference efficiency. To further improve its active perception strategy, we introduce Group Relative Policy Optimization (GRPO) for reinforcement learning [35,40].

Compared with conventional supervised learning, reinforcement learning enables the model to optimize its generation policy through reward feedback. Compared with Proximal Policy Optimization (PPO), GRPO estimates advantages based on group-relative rewards and removes the need for an additional value network. This design reduces computational overhead and improves training stability. Figure 3 illustrates the GRPO optimization process in our active perception setting.



**Figure 3.** Overview of the GRPO optimization framework for active perception-based RS-VQA.

Given an input image  $I_0$  and a user query  $q$ , the current policy model  $\pi_\theta$  generates a group of  $K$  candidate trajectories:

$$\{y_1, y_2, \dots, y_K\}. \quad (5)$$

Each trajectory may contain intermediate reasoning steps, crop-tool invocations, and a final answer. The reward function evaluates each trajectory and produces the corresponding reward values:

$$\{R_1, R_2, \dots, R_K\}. \quad (6)$$

The group mean and standard deviation are computed as follows:

$$\mu = \frac{1}{K} \sum_{i=1}^K R_i, \quad \sigma = \sqrt{\frac{1}{K} \sum_{i=1}^K (R_i - \mu)^2}. \quad (7)$$

The normalized advantage for each trajectory is then defined as follows:

$$A_i = \frac{R_i - \mu}{\sigma + \epsilon}, \quad (8)$$

where  $\epsilon$  is a small constant for numerical stability. The GRPO objective with Kullback–Leibler (KL) regularization is formulated as follows:

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{K} \sum_{i=1}^K A_i \log \pi_{\theta}(y_i | I_0, q) + \beta D_{\text{KL}}(\pi_{\theta} \| \pi_{\text{ref}}), \quad (9)$$

where  $\pi_{\text{ref}}$  denotes the reference policy obtained after SFT, and  $\beta$  controls the strength of the KL regularization. The KL term constrains the updated policy from deviating excessively from the SFT model, thereby improving training stability.

### 2.5. Reward Function Design

An effective reward function is critical for learning a robust active perception strategy. Traditional IoU-based rewards evaluate only the geometric overlap between the predicted crop and the reference region [41]. Although suitable for standard localization tasks, they have several limitations for RS-VQA: early in training, poor localization leads to near-zero IoU, providing sparse feedback; IoU penalizes partial coverage even when the region contains sufficient visual evidence; and IoU measures alignment rather than task relevance.

To address this, we define a region reward that combines reference-region coverage and crop compactness:

$$R_{\text{region}} = \frac{|b \cap b^*|}{|b^*| + \epsilon} - \lambda_{\text{area}} \frac{|b|}{|I_0|}, \quad (10)$$

where  $b$  is the predicted crop,  $b^*$  is the reference region, and  $I_0$  is the full image. The coverage term ranges in  $[0, 1]$  and encourages inclusion of question-relevant evidence; the area term penalizes overly large crops, with  $\lambda_{\text{area}} \in [0, 1]$ . In all experiments, we set  $\lambda_{\text{area}} = 0.2$ . If the predicted crop is invalid or non-positive,  $R_{\text{region}}$  is set to 0.

We also define a format reward  $R_{\text{fmt}}$ , which gives a positive score when the output contains valid special tokens (`<think>`, `<crop>`, `<answer>`) and executable crop coordinates; otherwise, it is 0. This ensures structural validity of the reasoning trajectory.

The answer reward  $R_{\text{ans}}$  evaluates semantic correctness using an LLM-as-a-judge strategy [42]:

$$R_{\text{ans}} = \begin{cases} 1, & \text{if the predicted answer is semantically consistent with ground truth} \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

To regulate tool usage, we introduce a multiplicative cropping penalty:

$$P_{\text{crop}} = 1 - \lambda_{\text{crop}} \cdot \mathbb{I}_{\text{crop}}, \quad (12)$$

where  $\mathbb{I}_{\text{crop}} = 1$  if the trajectory contains a crop invocation, and 0 otherwise;  $\lambda_{\text{crop}} \in [0, 1)$  controls the penalty strength. Thus,  $P_{\text{crop}} \in (0, 1]$ . When no crop is invoked,  $P_{\text{crop}} = 1$ ; when cropping occurs,  $P_{\text{crop}} < 1$ , reducing the effective answer reward. In our experiments, the crop penalty coefficient  $\lambda_{\text{crop}}$  is set to 0.1. This value provides a mild penalty for unnecessary crop-tool invocations while still allowing necessary cropping, achieving a reasonable balance between local evidence acquisition and inference efficiency.

The overall reward is then as follows:

$$R = \alpha R_{\text{fmt}} + R_{\text{region}} + R_{\text{ans}} \cdot P_{\text{crop}}, \quad (13)$$

where  $\alpha$  weights the format reward. In our experiments, we set  $\alpha = 0.3$ , and the coefficient of  $R_{\text{region}}$  is fixed to 1.0 in the composite reward. This formulation clarifies that the crop penalty modulates the answer reward without introducing unintended positive reward.

**Intuition and Example.** For a global-level question that does not require cropping, the model should rely on the full image. If it unnecessarily invokes cropping,  $R_{\text{ans}}$  is discounted by  $P_{\text{crop}}$ , reflecting a minor cost. Conversely, for an object-level question requiring local evidence, invoking the crop tool can increase  $R_{\text{region}}$  and help achieve  $R_{\text{ans}} = 1$ , so the model is rewarded for selective cropping.

## 2.6. Evaluation Datasets and Metrics

**Evaluation Datasets.** Table 1 summarizes the four evaluation datasets used in this study. To evaluate the proposed method under diverse remote sensing scenarios, we conduct experiments on VRSBench, MME-RealWorld-RS, XLRB-Bench, and LRS-VQA [10,19,43,44]. These benchmarks cover different image resolutions, scene categories, and question types, allowing us to assess general RS-VQA performance, fine-grained visual perception, spatial reasoning, and ultra-high-resolution image analysis.

**Table 1.** Statistics of the evaluation datasets.

| Dataset               | Avg. Resolution | Images | VQA Pairs |
|-----------------------|-----------------|--------|-----------|
| VRSBench [44]         | 512 × 512       | 29,614 | 123,221   |
| MME-RealWorld-RS [10] | 5602 × 4445     | 1265   | 3738      |
| XLRB-Bench [19]       | 8500 × 8500     | 1400   | 3080      |
| LRS-VQA [43]          | 7099 × 6329     | 1657   | 7333      |

VRSBench evaluates general vision–language understanding in remote sensing scenes and contains a large number of image–question pairs with moderate image resolution. MME-RealWorld-RS focuses on real-world high-resolution remote sensing imagery and is used to examine model robustness in complex visual environments. XLRB-Bench is designed for ultra-high-resolution remote sensing images, where large spatial coverage and small visual targets pose challenges to region localization and multi-step reasoning. LRS-VQA further emphasizes question answering over large-scale remote sensing scenes and is suitable for evaluating fine-grained evidence acquisition in wide-area imagery. Together, these datasets provide a comprehensive testbed for analyzing the effectiveness of the proposed active perception framework across both standard and high-resolution RS-VQA settings.

**Evaluation Metrics.** We adopt answer accuracy (Acc) as the primary evaluation metric. For each test sample, the model generates an answer based on the input image and question. The prediction is then compared with the ground-truth answer to determine whether it is correct. However, RS-VQA answers are often expressed in short natural-language phrases, and many questions allow semantically equivalent answers with different sur-

face forms. Therefore, exact string matching or token-level overlap may underestimate model performance.

To obtain a more reliable evaluation for open-ended answers, we employ an LLM-as-a-judge strategy [42]. Specifically, we use Qwen3-32B as the judging model. For each sample, the question, predicted answer, and ground-truth answer are provided to the judge model. The judge is instructed to determine whether the prediction is semantically consistent with the reference answer and to output a binary decision, i.e., correct or incorrect. The same judging prompt is used for all datasets to ensure consistency across evaluations. The final accuracy is computed as follows:

$$Acc = \frac{N_{correct}}{N_{total}}, \quad (14)$$

where  $N_{correct}$  denotes the number of samples judged as correct, and  $N_{total}$  represents the total number of evaluated samples. Compared with exact string matching, this semantic-consistency-based protocol better reflects answer correctness in open-ended RS-VQA tasks.

### 3. Results

#### 3.1. Comparative Analysis and Discussion

To evaluate the effectiveness of the proposed method, we compare it with representative VLMs on four RS-VQA benchmarks, including VRSBench, MME-RealWorld-RS, XLRS-Bench, and LRS-VQA [10,19,43,44]. For compact presentation, MME-RealWorld-RS is abbreviated as MME-RW-RS in the results' table. The compared models are grouped into three categories: proprietary VLMs, open-source general-purpose VLMs, and remote sensing-specific VLMs.

For proprietary models, we include GPT-4o, GPT-4V, Gemini-2.5-Flash, and Claude 3.5 Sonnet, which have shown strong general visual understanding and reasoning abilities [45–48]. For open-source models, we evaluate InternVL, LLaVA-OV, and several Qwen-VL variants with different model scales, including Qwen3-VL models with enhanced reasoning ability [37,49–52]. These models provide strong baselines for assessing general multimodal understanding. We also compare with remote sensing-specific VLMs, including GeoChat, GeoLLaVA, ZoomEarth, SkySenseGPT, and GeoVLM-R1 [11,13,16,29,53]. These methods are designed or adapted for geospatial imagery through domain-specific instruction tuning, high-resolution input modeling, crop-based interaction, or reasoning-oriented optimization.

As shown in Table 2, the proposed method demonstrates strong performance across the four RS-VQA benchmarks, achieving accuracies of 49.8, 43.3, 40.9, and 34.7 on VRSBench, MME-RW-RS, XLRS-Bench, and LRS-VQA, respectively. These values result in a competitive average accuracy of 42.2 across the evaluated benchmarks. Notably, our approach attains the top performance on VRSBench and LRS-VQA, while remaining competitive on MME-RW-RS and XLRS-Bench.

When compared to proprietary VLMs, such as GPT-4o, GPT-4V, Gemini-2.5-Flash, and Claude 3.5 Sonnet, the proposed method achieves a higher average score. Although these models perform well on general-purpose visual understanding tasks, their performance is less consistent on remote sensing benchmarks, particularly in high-resolution scenarios requiring fine-grained region localization. This indicates that incorporating domain-specific active perception and crop-tool-based reasoning can provide substantial benefits for RS-VQA.

The improvement over open-source general-purpose VLMs is even more pronounced. Using Qwen2.5-VL-3B as the backbone, the proposed method increases the average accuracy from 23.2 to 42.2. This suggests that the performance gains arise not merely from model

scale, but from the integrated effect of structured reasoning, teacher-guided cold-start SFT, and GRPO-based active perception optimization.

**Table 2.** Accuracy (%) comparison with existing vision–language models on RS-VQA benchmarks.

| Model                   | VRSBench    | MME-RW-RS   | XLRS-Bench  | LRS-VQA     | Avg         |
|-------------------------|-------------|-------------|-------------|-------------|-------------|
| GPT-4o [45]             | 42.5        | 28.9        | 39.8        | 27.6        | 34.7        |
| GPT-4V [46]             | 37.2        | 32.2        | 32.1        | 24.5        | 31.5        |
| Gemini-2.5-Flash [47]   | 45.3        | 42.0        | 41.2        | 32.9        | 40.4        |
| Claude 3.5 Sonnet [48]  | 40.2        | 36.5        | 36.5        | 19.7        | 33.2        |
| InternVL2.5-8B [49]     | 34.5        | 32.9        | 32.8        | 25.2        | 31.4        |
| InternVL3-8B [50]       | 36.9        | 38.2        | 34.5        | 26.5        | 34.0        |
| Qwen2-VL [54]           | 32.1        | 26.9        | 26.9        | 25.0        | 27.7        |
| LLaVA-OV [51]           | 31.6        | 26.6        | 26.5        | 26.1        | 27.7        |
| Qwen2.5-VL-3B [37]      | 28.5        | 22.1        | 22.1        | 20.1        | 23.2        |
| Qwen2.5-VL-7B [37]      | 33.8        | 28.4        | 28.3        | 25.8        | 29.1        |
| Qwen2.5-VL-72B [37]     | 38.7        | 36.6        | 35.2        | 27.9        | 34.6        |
| Qwen3-VL-8B [52]        | 35.2        | 35.9        | 32.5        | 25.8        | 32.4        |
| Qwen3-VL-235B-A22B [52] | 41.9        | 38.9        | 38.7        | 29.1        | 37.2        |
| GeoChat [11]            | 40.8        | 28.6        | 22.2        | 25.5        | 29.3        |
| SLiME-8B [10]           | 30.2        | 29.7        | 29.7        | 28.7        | 29.6        |
| GeoLLaVA-8K [16]        | 48.9        | 41.6        | 39.1        | 28.8        | 39.6        |
| ZoomEarth [29]          | 37.5        | 44.1        | 40.2        | 30.7        | 38.1        |
| SkySenseGPT [13]        | 39.8        | 33.5        | 33.5        | –           | 35.6        |
| GeoVLM-R1 [53]          | –           | –           | 36.8        | 26.8        | 31.8        |
| <b>Ours</b>             | <b>49.8</b> | <b>43.3</b> | <b>40.9</b> | <b>34.7</b> | <b>42.2</b> |

Similarly, compared to remote sensing-specific VLMs, the proposed approach delivers competitive results. Models such as GeoLLaVA-8K and ZoomEarth achieve strong performance on high-resolution benchmarks, underscoring the importance of high-resolution input modeling and localized visual evidence. Nevertheless, our method attains the highest overall average and shows a clear advantage on LRS-VQA, where accurate localization of fine-grained evidence across wide-area scenes is crucial. These results indicate that the crop-tool-augmented active perception strategy effectively enhances RS-VQA performance in large-scale, high-resolution settings.

Overall, the experimental outcomes highlight that the proposed method achieves a balanced improvement across general RS-VQA tasks and high-resolution fine-grained reasoning. The substantial gains over the Qwen2.5-VL-3B backbone further validate the effectiveness of our training pipeline, which combines structured reasoning, teacher-generated crop-based trajectories, and GRPO-based optimization.

### 3.2. Ablation Studies

To analyze the contribution of each component, we conduct ablation studies from three perspectives: training strategy, cropping mechanism, and reward design. Table 3 compares different training strategies. Compared with the original Qwen2.5-VL-3B backbone, SFT alone improves the average accuracy from 23.2 to 39.5, yielding an absolute gain of 16.3 points. This result shows that teacher-generated structured trajectories provide an effective cold-start signal, enabling the model to learn basic multi-step reasoning patterns and crop-tool invocation formats.

GRPO alone also improves the backbone model, increasing the average accuracy to 34.8. However, its performance remains lower than that of SFT alone. This suggests that reinforcement learning without sufficient initialization may suffer from unstable exploration, especially when the model needs to simultaneously learn reasoning formats, region selection, and answer generation. When SFT and GRPO are combined, the average

accuracy further increases to 42.2, achieving the best results among all training strategies. This indicates that SFT provides a reliable initialization for structured reasoning and tool use, while GRPO further refines the active perception policy through task-level feedback.

**Table 3.** Ablation study on training strategies. Improvements over Qwen2.5-VL-3B are shown in parentheses.

| Training Strategy  | VRSBench            | MME-RW-RS           | XLRS-Bench          | LRS-VQA             | Avg.                |
|--------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Qwen2.5-VL-3B [37] | 28.5                | 22.1                | 22.1                | 20.1                | 23.2                |
| SFT only           | 47.5 (+19.0)        | 40.1 (+18.0)        | 37.6 (+15.5)        | 32.9 (+12.8)        | 39.5 (+16.3)        |
| GRPO only          | 39.2 (+10.7)        | 38.0 (+15.9)        | 32.5 (+10.4)        | 29.4 (+9.3)         | 34.8 (+11.6)        |
| <b>SFT + GRPO</b>  | <b>49.8 (+21.3)</b> | <b>43.3 (+21.2)</b> | <b>40.9 (+18.8)</b> | <b>34.7 (+14.6)</b> | <b>42.2 (+19.0)</b> |

Table 4 further examines the effect of cropping under different training paradigms. Under the SFT setting, adding cropping only slightly improves the average accuracy from 39.3 to 39.5. The limited gain indicates that supervised learning can teach the model the basic format of crop-tool invocation, but it does not fully optimize when and where to crop. In particular, the performance on MME-RW-RS decreases after adding cropping, suggesting that inaccurate or unnecessary crop decisions may introduce distracting local evidence and affect answer generation.

**Table 4.** Effect of cropping under different training paradigms. Improvements over the baseline are shown in parentheses.

| Settings       | VRSBench    | MME-RW-RS   | XLRS-Bench  | LRS-VQA     | Avg.        |
|----------------|-------------|-------------|-------------|-------------|-------------|
| SFT            | 46.3        | 41.8        | 37.2        | 31.7        | 39.3        |
| SFT + Cropping | 47.5 (+1.2) | 40.1 (−1.7) | 37.6 (+0.4) | 32.9 (+1.2) | 39.5 (+0.2) |
| RL             | 49.1        | 40.9        | 39.4        | 32.6        | 40.5        |
| RL + Cropping  | 49.8 (+0.7) | 43.3 (+2.4) | 40.9 (+1.5) | 34.7 (+2.1) | 42.2 (+1.7) |

In contrast, cropping becomes more beneficial when combined with reinforcement learning. Under the RL setting, the average accuracy increases from 40.5 to 42.2 after introducing cropping, with consistent gains across all four benchmarks. The improvements are especially clear on MME-RW-RS and LRS-VQA, where high-resolution inputs and wide-area scenes require more accurate local evidence acquisition. These results suggest that cropping itself is not always beneficial; its effectiveness depends on whether the model can learn a reasonable crop-selection policy. GRPO helps optimize this policy by rewarding trajectories that select informative regions and produce correct answers.

Overall, the comparison indicates that cropping should be coupled with policy optimization rather than treated as a fixed operation. SFT provides the model with basic executable tool-use behavior, while GRPO improves the timing and spatial precision of crop-tool invocation. This explains why the combination of cropping and reinforcement learning yields the most stable improvement.

We further evaluate the influence of reward design in Table 5. Replacing the conventional IoU reward with the proposed region reward improves the average accuracy from 39.1 to 40.4. This improvement indicates that the coverage-based formulation provides a denser and more task-oriented supervision signal than strict geometric overlap. By measuring whether the predicted crop contains the reference region while penalizing excessive area, the region reward better matches the evidence-seeking nature of active perception in RS-VQA.

**Table 5.** Ablation study on reward design.

| Reward Design |              |            | Performance |             |             |             |             |
|---------------|--------------|------------|-------------|-------------|-------------|-------------|-------------|
| $R_{IoU}$     | $R_{region}$ | $P_{crop}$ | Reasoning   | Object      | Counting    | Scene       | Avg.        |
| ✓             | –            | –          | 39.6        | 42.4        | 14.3        | 52.0        | 39.1        |
| –             | ✓            | –          | 41.5        | 43.7        | 14.6        | 53.6        | 40.4        |
| –             | ✓            | ✓          | <b>42.2</b> | <b>44.3</b> | <b>15.0</b> | <b>53.8</b> | <b>40.9</b> |

The improvement is observed in Reasoning, Object, and Scene tasks, suggesting that better region selection benefits not only localization but also downstream semantic analysis. The gain on Counting tasks is relatively small, which may be because counting often depends on broader scene-level information rather than a single localized crop. This result also suggests that the role of cropping varies across question types.

Adding the cropping penalty further improves the average accuracy from 40.4 to 40.9. The penalty discourages redundant tool invocation and encourages the model to preserve global context when cropping is unnecessary. This effect is particularly important for Scene questions, where excessive local focus may remove useful contextual information. Meanwhile, the gains in Reasoning and Object tasks suggest that the model learns to invoke cropping more selectively and at more appropriate stages. These results verify that the proposed reward design not only guides the model toward informative regions, but also regulates the cost and necessity of crop-tool use.

## 4. Discussion

### 4.1. Method Analysis

High-resolution remote sensing visual question answering presents substantial challenges for conventional vision–language models due to the sparse distribution of informative content and the wide scale variation of key objects. In many scenarios, only a small portion of the image is relevant to the question, while the remaining regions often contain redundant background information that can interfere with reasoning. To address this, we propose an active perception framework that integrates dynamic cropping into the reasoning process of vision–language models. Unlike approaches relying solely on global image representations, the proposed method selectively focuses on task-relevant local regions based on the semantic content of the question, enabling more precise visual understanding during inference.

Experimental results on multiple RS-VQA benchmarks demonstrate consistent performance gains over traditional global-view methods. This improvement indicates that explicitly incorporating region selection into the reasoning pipeline effectively mitigates the impact of irrelevant information in high-resolution imagery. Additionally, the framework enhances the model’s ability to capture fine-grained visual evidence, which is crucial for tasks involving small objects, complex spatial arrangements, or subtle semantic distinctions.

Another key advantage of the proposed framework is its joint modeling of perception and reasoning within a unified decision-making process. Conventional vision–language models typically treat visual encoding as a fixed preprocessing step, while reasoning is performed on the extracted features. In contrast, our method allows reasoning to guide perception, enabling adaptive adjustment of the observation region according to the current question. This active perception mechanism improves reasoning robustness in complex remote sensing scenarios and more closely resembles human visual cognition, where attention is dynamically allocated to informative regions during problem solving.

#### 4.2. Comparison with Existing Methods

Compared with existing remote sensing vision–language models, the proposed method incorporates an explicit and interpretable active perception mechanism at the behavioral level. Methods such as GeoChat primarily enhance semantic understanding through instruction tuning on remote sensing data. Although they perform well on general semantic tasks, their reasoning relies on global image representations and lacks an explicit mechanism to identify question-relevant regions. As a result, when processing high-resolution images with abundant irrelevant content, redundant visual information can negatively affect reasoning performance. GeoVLM-R1 improves answer quality for complex queries by enhancing reasoning capabilities, yet it still relies on globally encoded image features and does not explicitly model spatial information selection. Consequently, despite more sophisticated reasoning, it may struggle to accurately localize key visual evidence in challenging remote sensing scenarios.

In contrast, our framework explicitly integrates cropping operations into the reasoning process, allowing the model to dynamically adjust its perceptual focus based on question semantics and scene complexity. By directly linking perception and reasoning, the model actively identifies informative regions during inference, rather than treating visual encoding and reasoning as separate stages. This unified strategy contributes to more stable performance improvements across datasets and question types.

Compared with ZoomEarth, the proposed method further introduces a cropping penalty during GRPO training, enabling adaptive crop decision-making. Instead of performing cropping indiscriminately, the model learns when and where cropping is necessary, improving both flexibility and efficiency. This is particularly beneficial for questions that can be answered using only global context. Overall, the proposed framework enhances not only region localization but also the model’s ability to autonomously manage its perception behavior.

#### 4.3. Reward Mechanism Analysis

The design of the reward function is central to optimizing crop-tool behavior in the proposed framework. Conventional IoU-based rewards mainly evaluate the geometric overlap between the predicted crop and the annotated region. Although this criterion is suitable for standard localization tasks, it does not fully reflect the task-oriented nature of RS-VQA. In this setting, the selected region should not only overlap with the reference annotation, but also provide visual evidence that is useful for answering the question. Therefore, directly relying on IoU may produce sparse or insufficient learning signals, especially when the predicted crop partially covers the target but fails to obtain a high overlap score.

To alleviate this problem, we introduce a reward mechanism that combines reference-region coverage with an area constraint. The coverage term encourages the model to include question-relevant visual evidence in the selected crop, whereas the area constraint discourages overly large regions that simply cover more content without improving reasoning quality. By balancing these two objectives, the proposed reward guides the model toward compact and informative regions, which is more consistent with the evidence-seeking requirement of active perception in RS-VQA.

In addition, we introduce a cropping penalty to regulate tool invocation. In practical remote sensing scenarios, not every question requires local inspection. Some questions can be answered from the global scene context, and unnecessary cropping may introduce redundant visual evidence, increase inference cost, or even distract the model from the correct reasoning path. The cropping penalty therefore encourages the model to invoke the crop tool more selectively. Instead of treating cropping as a default operation, the model

learns to determine whether local evidence is needed according to the question and visual context. This design improves the adaptability of the perception process across different question types.

The proposed reward mechanism also improves the interpretability of the framework. Since crop decisions are explicitly represented and optimized during training, the resulting reasoning process can be inspected through the generated crop coordinates and intermediate outputs. Compared with implicit attention mechanisms, this explicit crop-tool behavior provides more observable evidence for analyzing how the model acquires and uses visual information. Nevertheless, the current reward design mainly focuses on region coverage, crop compactness, answer correctness, and tool-use cost. More fine-grained reward signals, such as those considering multi-step crop trajectories or the semantic sufficiency of selected regions, remain worth further investigation.

#### 4.4. Limitations and Future Work

Although the proposed framework achieves promising results on multiple RS-VQA benchmarks, several limitations remain that warrant further investigation. First, the current training process relies on manually annotated cropping regions, which increases annotation costs and limits scalability to large-scale unlabeled remote sensing datasets. Despite these annotations providing effective supervision for learning crop-tool behavior, constructing high-quality region labels is time-consuming and may not generalize well to diverse or complex scenarios. Future work could explore reinforcement learning or self-supervised strategies that learn region selection policies directly from question–answer supervision, reducing the dependence on explicit cropping annotations and improving scalability in real-world applications.

Second, the framework currently employs a single-step cropping operation. Although this design simplifies decision-making, a single crop may be insufficient to capture all relevant visual evidence in complex scenes containing multiple spatially distributed regions. In such cases, one cropped region might omit critical information needed for accurate reasoning. Future research could investigate multi-step or hierarchical cropping strategies, enabling the model to progressively refine its perception and integrate information from multiple regions during inference.

Third, cropping is the primary visual operation in the current framework. Although effective at reducing redundant information, additional perception mechanisms could further enhance scene understanding. For example, future extensions could incorporate multi-scale feature exploration, cross-region integration, adaptive zooming, or external knowledge retrieval modules. Integrating such capabilities may yield a more comprehensive active perception system for remote sensing multimodal reasoning tasks.

Overall, this work demonstrates that incorporating active region selection into the reasoning process improves RS-VQA performance and provides a new perspective on modeling active perception in multimodal vision–language systems. Addressing the limitations outlined above will be critical for extending the framework to larger-scale datasets, more complex scenes, and richer perception operations in future studies.

## 5. Conclusions

This paper presents an active perception-based vision–language framework for high-resolution remote sensing visual question answering (RS-VQA), designed to address the challenges posed by sparse informative content, substantial background redundancy, and large-scale variations of key objects in remote sensing images. The framework treats cropping as an explicit visual operation within the reasoning process, enabling the model to selectively focus on question-relevant regions and generate answers based on targeted

visual evidence. By dynamically adjusting its perceptual focus according to the task requirements, the model moves beyond reliance on fixed global image representations.

To further enhance region selection, a reward mechanism is introduced that combines region coverage, area constraints, and a cropping penalty. This design encourages the model to prioritize informative regions containing key evidence, discourages excessive or uninformative crops, and guides adaptive cropping decisions depending on the question context. Consequently, the framework establishes a tighter integration between region localization, cropping behavior, and answer reasoning, facilitating a more flexible and interpretable perception process. Experimental results across multiple RS-VQA benchmarks demonstrate competitive performance and improvements in most evaluated settings, validating the effectiveness of incorporating active region selection into high-resolution multimodal understanding. We also observe that the performance gain varies across datasets, indicating that the effectiveness of crop-tool-augmented active perception is influenced by task composition, question difficulty, and the distribution of local visual evidence.

In summary, this work provides a new modeling perspective for RS-VQA by explicitly embedding visual perception behavior into the reasoning pipeline of vision–language models. Beyond improving task performance, this approach offers insights for the development of vision–language systems with dynamic decision-making, interpretable reasoning, and adaptive perception, and lays the groundwork for future research on more sophisticated active perception mechanisms in multimodal remote sensing applications.

**Author Contributions:** Methodology, Q.L.; software, Q.L.; validation, Q.L. and Y.H.; investigation, Q.L., Y.H. and K.C.; visualization, Q.L. and Y.H.; writing—original draft preparation, Q.L. and K.C.; writing—review and editing, Q.L., K.C. and X.X.; supervision, X.X.; project administration, Q.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Beijing Institute of Technology Science and Technology Innovation Program Project (Grant No. 3060012212522).

**Data Availability Statement:** The evaluation datasets used in this study are all publicly available remote sensing visual question answering or remote sensing vision–language benchmark datasets, including VRSBench, MME-RealWorld-RS, XLRs-Bench, and LRS-VQA. The training data are mainly based on the publicly available LRS-GRO dataset and its associated annotations. The model training data, inference results, and evaluation results generated during this study are available from the corresponding author upon reasonable request. Due to the licenses and terms of use of some original datasets, this study does not redistribute these datasets. Readers may obtain them from the corresponding official dataset pages or original publications.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

|      |                                         |
|------|-----------------------------------------|
| Acc  | Accuracy                                |
| bbox | Bounding box                            |
| CLIP | Contrastive Language–Image Pre-training |
| CoT  | Chain-of-Thought                        |
| GRPO | Group Relative Policy Optimization      |
| IoU  | Intersection over Union                 |
| KL   | Kullback–Leibler divergence             |
| LLM  | Large language model                    |
| LoRA | Low-Rank Adaptation                     |

|        |                                          |
|--------|------------------------------------------|
| MLP    | Multilayer perceptron                    |
| PPO    | Proximal Policy Optimization             |
| RL     | Reinforcement learning                   |
| RS-VQA | Remote sensing visual question answering |
| SFT    | Supervised fine-tuning                   |
| VLM    | Vision–language model                    |
| VQA    | Visual question answering                |

Appendix A

Appendix A.1. Training Details and Hyperparameter Settings

We adopt Qwen2.5-VL-3B as the base model and conduct training in two stages: cold-start SFT and GRPO-based reinforcement learning. In the first stage, we construct the cold-start SFT data from the LRS-GRO training set. Specifically, we use Seed 1.6 to generate trajectory data with high-quality chain-of-thought rationales for the training samples, and then perform LoRA-based fine-tuning on the base model using these generated trajectories. During cold-start SFT, the vision encoder is frozen, and LoRA adapters are inserted only into the attention projection modules of the language model, including q\_proj, k\_proj, v\_proj, and o\_proj. The LoRA rank is set to 8 and the LoRA alpha is set to 16. We train the model with bfloat16 precision, a learning rate of  $5 \times 10^{-5}$ , a warmup ratio of 0.03, a weight decay of 0.01, a global batch size of 64, and a maximum sequence length of 4096 for 2 epochs. AdamW is used as the optimizer. The detailed hyperparameters for cold-start SFT are reported in Table A1. After cold-start SFT, we merge the LoRA weights into the original model weights, obtaining the supervised fine-tuned model used to initialize the subsequent reinforcement learning stage. The cold-start SFT stage is conducted on 4 NVIDIA A800-80GB NVLink GPUs using the ms-swift training framework. The training environment consists of Ubuntu 22.04, an Intel Xeon Platinum 8374C CPU, CUDA 12.8, Python 3.12, and PyTorch 2.8.0.

Table A1. Hyperparameters for cold-start SFT training.

| Training Hyperparameter | Value                          |
|-------------------------|--------------------------------|
| Freeze ViT              | true                           |
| LoRA rank               | 8                              |
| LoRA alpha              | 16                             |
| LoRA target modules     | q_proj, k_proj, v_proj, o_proj |
| Precision               | bfloat16                       |
| Learning rate           | $5 \times 10^{-5}$             |
| LR scheduler type       | cosine                         |
| Warmup ratio            | 0.03                           |
| Weight decay            | 0.01                           |
| Global batch size       | 64                             |
| Optimizer               | AdamW                          |
| Training epochs         | 2                              |
| Max sequence length     | 4096                           |

After supervised fine-tuning, we further optimize the model with GRPO-based reinforcement learning. To improve the effectiveness of the reinforcement learning data, we first apply rejection sampling to filter the training samples. Specifically, the supervised fine-tuned model from the previous stage is used to generate multiple responses for each training sample, and the samples are then categorized according to the correctness of the generated answers. Samples for which all responses are correct are considered too easy to provide informative optimization signals, whereas samples for which all responses

are incorrect are unlikely to provide stable relative preference information for the current model. Therefore, we discard both types of samples and retain only those for which the model produces both correct and incorrect responses. These retained samples are then used for GRPO training.

In the GRPO stage, we adopt full-parameter fine-tuning while keeping the vision encoder frozen, such that only the language model parameters are updated. The supervised fine-tuned model is used as the initial policy model, and training is performed with bfloat16 precision. For each question, we sample 8 candidate responses with a sampling temperature of 1.0 and a top-p value of 0.9. The KL coefficient is set to 0.001 to constrain the policy update with respect to the reference model. We use AdamW as the optimizer, with a learning rate of  $1 \times 10^{-6}$ , a cosine learning-rate scheduler, a warmup ratio of 0.03, a weight decay of 0.01, a gradient clipping threshold of 1.0, a global batch size of 128, a total of 200 RL training steps, and a maximum sequence length of 4096. The detailed hyperparameters for GRPO training are summarized in Table A2. The GRPO stage is conducted on 8 NVIDIA A800-80GB NVLink GPUs using the same ms-swift framework and training environment as in the cold-start SFT stage.

**Table A2.** Hyperparameters for RL training.

| Training Hyperparameter     | Value              |
|-----------------------------|--------------------|
| Freeze ViT                  | true               |
| Precision                   | bfloat16           |
| Learning rate               | $1 \times 10^{-6}$ |
| LR scheduler type           | cosine             |
| Number of generations       | 8                  |
| Sampling temperature        | 1.0                |
| Top-p                       | 0.9                |
| KL coefficient              | 0.001              |
| Warmup ratio                | 0.03               |
| Weight decay                | 0.01               |
| Gradient clipping threshold | 1.0                |
| Global batch size           | 128                |
| Optimizer                   | AdamW              |
| Total RL steps              | 200                |
| Max sequence length         | 4096               |

#### Appendix A.2. Prompt Templates for Active Perception and Evaluation

This appendix presents the main prompt templates used in this paper for remote sensing visual question answering. Table A3 shows the system prompt adopted during the model inference stage. This prompt constrains the model to follow a reasoning procedure consisting of image description, cropping decision-making, local detail analysis, or direct answering, and specifies special output formats such as <think>, <crop>, and <answer>. With this prompt, the model can actively output a cropping region when fine-grained visual information is required, while directly answering based on the global image when cropping is unnecessary.

Table A4 presents the evaluation prompt used for LLM-as-a-judge and answer correctness assessment. This prompt requires the evaluator model to strictly compare the predicted answer with the ground-truth answer in terms of semantic and factual consistency, and to output only a binary judgment. Specifically, a prediction is marked as correct when its core meaning is consistent with the ground-truth answer, the key information is complete, and there are no factual errors. In contrast, a prediction is marked as incorrect if

it contains factual errors, misses key information, is irrelevant to the question, or is vague and indeterminable.

**Table A3.** System prompt used for remote sensing visual question answering.

| Item                        | Content                                                                                                                                                                                      |
|-----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Role                        | You are an expert in remote sensing visual question answering, specializing in answering questions based on remote sensing images provided by the user.                                      |
| Reasoning format            | Please reason step by step and output your reasoning process in the format of <code>&lt;think&gt;...&lt;/think&gt;</code> .                                                                  |
| Cropping tool               | When solving the problem, you may choose to call the image cropping tool to obtain fine-grained details from the original image.                                                             |
| Decision process            | First, briefly describe the image, and then determine whether the image cropping tool is needed.                                                                                             |
| If cropping is required     | If cropping is required, first output the reasoning process, and finally output the target bounding box in the format of <code>&lt;crop&gt;...&lt;/crop&gt;</code> and stop answering.       |
| If cropping is not required | If cropping is not required, continue reasoning until you can fully answer the user's question, and output the final result in the format of <code>&lt;answer&gt;...&lt;/answer&gt;</code> . |

**Table A4.** Evaluation prompt for answer correctness judgment.

| Prompt                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>You are an objective, rigorous, and professional evaluator responsible for assessing the correctness of the model's answer. Your task is to strictly compare the predicted answer with the ground-truth answer and determine whether they are semantically and factually consistent. Please strictly follow the rules below:</p> <ol style="list-style-type: none"> <li>The predicted answer should be judged as correct if all of the following conditions are satisfied: <ul style="list-style-type: none"> <li>The core meaning is completely consistent with the ground-truth answer.</li> <li>The key information is complete and not omitted.</li> <li>There are no factual errors.</li> </ul> </li> <li>The predicted answer should be judged as incorrect if any of the following conditions are satisfied: <ul style="list-style-type: none"> <li>There is a factual error.</li> <li>Key answer information is missing.</li> <li>The answer is irrelevant to the original question.</li> <li>The content is vague or cannot be determined.</li> </ul> </li> </ol> <p>Please output your judgment strictly in the following format, without any additional explanation:</p> <pre>{'label': 'YES/NO'}</pre> <p>Original question: {Question}<br/> Ground-truth answer: {Ground-Truth}<br/> Predicted answer: {Prediction}</p> |

Table A5 shows the teacher-model prompt used to generate high-quality chain-of-thought trajectory data. This prompt takes the image, question, crop-needed label, crop bounding box, and annotated answer as inputs, and guides the teacher model to generate a reasoning process that is consistent with the annotated answer and grounded in visual

evidence. This template is mainly used for constructing cold-start supervised fine-tuning data, enabling the model to learn the remote sensing visual question answering process from global scene understanding to local detail analysis and, finally, to answer generation.

**Table A5.** Teacher prompt for generating chain-of-thought trajectories.

---

### Prompt

---

You are an expert in remote sensing visual question answering, specializing in answering questions based on remote sensing images. You now need to serve as a teacher model and generate high-quality chain-of-thought training data based on the given image, question, crop-needed label, crop bbox, and annotated answer.

Please note that your output is not intended to answer the question from scratch. Instead, you should generate a reasonable, clear, and visually grounded reasoning trajectory based on the known annotated answer. The reasoning process must explain why the final answer is the annotated answer.

Please strictly follow the procedure below:

1. First, briefly describe the overall scene of the remote sensing image, including the main land-cover or object types and their spatial layout.
2. Next, determine the visual information required to answer the question:
  - If Need\_Crop is true, it indicates that the question requires local fine-grained details. You should state that the given bbox region needs to be examined, and analyze the visible objects, textures, shapes, colors, and spatial relationships within that region.
  - If Need\_Crop is false, it indicates that the entire image already provides sufficient information. You should reason directly based on the global image content without mentioning additional cropping.
3. If Need\_Crop is true and a BBox is provided, you must use the information from the corresponding bbox region in your reasoning. Do not modify the bbox, do not output a new bbox, and do not claim that the bbox is inaccurate.
4. The reasoning should reflect the characteristics of remote sensing visual question answering. It may involve the following aspects:
  - land-cover or object category recognition;
  - the number, shape, size, color, or texture of targets;
  - the relationship between targets and surrounding roads, buildings, water bodies, vegetation, farmland, or other geographic objects;
  - the relative position of targets in the image;
  - how local details support the final judgment.
5. The reasoning process must be fully consistent with Ground\_Truth. It must not contain any information that contradicts the annotated answer. The final answer must directly use Ground\_Truth.
6. Do not include irrelevant explanations, additional comments, or any content outside the required format.

Please strictly output in the following format:

<think>

[Generate the complete reasoning process here. The reasoning process should include a brief image description, the judgment of whether cropping is needed, the analysis of relevant visual evidence, and the reason for deriving the final answer.]

</think>

<answer>{Ground\_Truth}</answer>

The input information is as follows:

Image: {Image}

Question: {Question}

Need crop: {Need\_Crop}

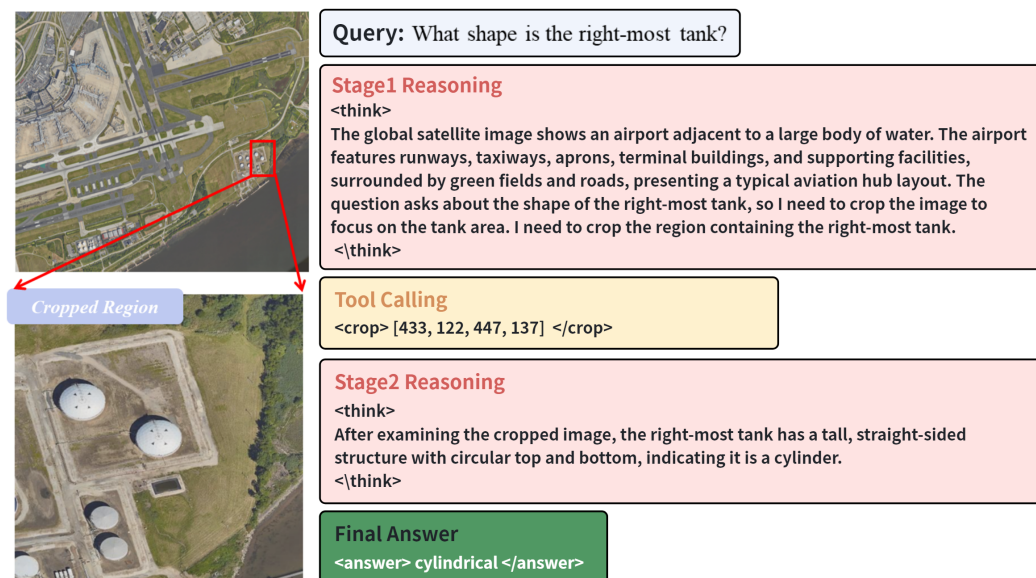
Crop bbox: {BBBox}

Annotated answer: {Ground\_Truth}

---

### Appendix A.3. Case Study

Figure A1 shows a fine-grained object recognition case that requires crop-tool invocation. In this example, the question asks about the shape of the rightmost tank. Since the target tank occupies only a small region in the original high-resolution remote sensing image, directly answering from the global view may be unreliable. The model first analyzes the global scene and determines that local visual evidence is needed. It then generates a structured crop instruction to focus on the region containing the target tank. After inspecting the cropped region, the model observes that it has a circular top and bottom with straight sides, and finally predicts the answer as “cylindrical”. This case demonstrates that the crop tool can help the model acquire fine-grained local evidence for accurate object attribute recognition.



**Figure A1.** A fine-grained object recognition case with crop-tool invocation.

Figure A2 presents a global scene understanding case without crop-tool invocation. The question asks about the location of the fountain in the image. Since the fountain and the surrounding green lake are already visible in the global image, the model can answer the question by directly analyzing the overall scene layout without performing additional cropping. The reasoning process shows that the model identifies the urban residential area, recognizes the green lake in the lower-left part of the image, and locates the fountain near the center of the lake. It therefore selects option C as the final answer. This case indicates that the proposed active perception framework can not only invoke the crop tool when fine-grained evidence is required, but also avoid unnecessary tool use when global visual information is sufficient.



**Query**

: Where is the fountain in the picture?

(A) At the top left side of the picture.

(B) In the middle of the picture.

(C) On the lower left side of the picture, in the center of the green lake.

(D) On the lower right side of the picture.

(E) This image doesn't feature the position.

**Stage1 Reasoning**

<think>

The user provides an image. First, give a brief description: This is a satellite top-down view of an urban residential area. Dense multi-story residential buildings are neatly arranged along streets, with scattered trees and small green spaces inside the area. The road network is clear, showing a compact urban residential layout. The question asks for the location of the fountain in the image. The fountain can be directly recognized from the image, so cropping is not needed. Then locate the fountain: it is in the lower left of the image, there is a green lake in the center, and the fountain is located in the central area of the lake.

</think>

**Final Answer**

<answer> C </answer>

**Figure A2.** A global scene understanding case without crop-tool invocation.

## References

1. Zhang, L.; Zhang, L. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294.
2. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36.
3. Zhou, G.; Qian, L.; Gamba, P. Advances on multimodal remote sensing foundation models for Earth observation downstream tasks: A survey. *Remote Sens.* **2025**, *17*, 3532.
4. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
5. Sarkar, A.; Chowdhury, T.; Murphy, R.R.; Gangopadhyay, A.; Rahnemounfar, M. SAM-VQA: Supervised attention-based visual question answering model for post-disaster damage assessment on remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16.
6. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual attention inception network for remote sensing visual question answering. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14.
7. Lobry, S.; Marcos, D.; Murray, J.; Tuia, D. RSVQA: Visual question answering for remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8555–8566.
8. Liu, C.; Zhang, J.; Chen, K.; Wang, M.; Zou, Z.; Shi, Z. Remote sensing spatiotemporal vision-language models: A comprehensive survey. *IEEE Geosci. Remote Sens. Mag.* **2026**, *14*, 383–423.
9. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023; pp. 34892–34916.
10. Zhang, Y.; Zhang, H.; Tian, H.; Fu, C.; Zhang, S.; Wu, J.; Li, F.; Wang, K.; Wen, Q.; Zhang, Z.; et al. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In Proceedings of the 13th International Conference on Learning Representations (ICLR 2025), Singapore, 24–28 April 2025.
11. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; Khan, F.S. GeoChat: Grounded large vision-language model for remote sensing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), Seattle, WA, USA, 17–21 June 2024; pp. 27831–27840.
12. Zhan, Y.; Xiong, Z.; Yuan, Y. SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS J. Photogramm. Remote Sens.* **2025**, *221*, 64–77.
13. Luo, J.; Pang, Z.; Zhang, Y.; Wang, T.; Wang, L.; Dang, B.; Lao, J.; Wang, J.; Chen, J.; Tan, Y.; et al. SkySenseGPT: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv* **2024**, arXiv:2406.10100.
14. Danish, M.; Munir, M.A.; Shah, S.R.A.; Kuckreja, K.; Khan, F.S.; Fraccaro, P.; Lacoste, A.; Khan, S. GeoBench-VLM: Benchmarking vision-language models for geospatial tasks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2025), Honolulu, HI, USA, 19–23 October 2025; pp. 7132–7142.

15. Yang, C.; Li, Z.; Zhang, L. MGIMM: Multi-granularity instruction multimodal model for attribute-guided remote sensing image detailed description. *arXiv* **2024**, arXiv:2406.04716.
16. Wang, F.; Chen, M.; Li, Y.; Wang, D.; Wang, H.; Guo, Z.; Wang, Z.; Boqi, S.; Lan, L.; Wang, Y.; et al. GeoLLaVA-8K: Scaling remote-sensing multimodal large language models to 8K resolution. *arXiv* **2025**, arXiv:2505.21375.
17. Lin, Z.; Leng, B. SSN: Scale selection network for multi-scale object detection in remote sensing images. *Remote Sens.* **2024**, *16*, 3697.
18. Wang, Z.; Liu, L.; Wan, G.; Zhang, W.; Zhong, B.; Chang, H.; Li, X.; Liu, X.; Sun, G. A review of visual grounding on remote sensing images. *Electronics* **2025**, *14*, 2815.
19. Wang, F.; Wang, H.; Guo, Z.; Wang, D.; Wang, Y.; Chen, M.; Ma, Q.; Lan, L.; Yang, W.; Zhang, J.; et al. XLRs-Bench: Could your multimodal LLMs understand extremely large ultra-high-resolution remote sensing imagery? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025), Nashville, TN, USA, 11–15 June 2025; pp. 14325–14336.
20. Zheng, G.; Yang, B.; Tang, J.; Zhou, H.Y.; Yang, S. DDCOT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023; pp. 5168–5191.
21. Shao, R.; Li, Z.; Zhang, Z.; Xu, L.; He, X.; Yuan, H.; He, B.; Dai, Y.; Yan, Y.; Chen, Y.; et al. Asking like Socrates: Socrates helps VLMs understand remote sensing images. *arXiv* **2025**, arXiv:2511.22396.
22. Faria, F.T.J.; Baniata, L.H.; Choi, A.; Kang, S. Towards robust chain-of-thought prompting with self-consistency for remote sensing VQA: An empirical study across large multimodal models. *Mathematics* **2025**, *13*, 3046.
23. Wang, J.; Ma, A.; Chen, Z.; Zheng, Z.; Wan, Y.; Zhang, L.; Zhong, Y. EarthVQANet: Multi-task visual question answering for remote sensing image understanding. *ISPRS J. Photogramm. Remote Sens.* **2024**, *212*, 422–439.
24. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023.
25. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing reasoning and acting in language models. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023.
26. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; pp. 24824–24837.
27. Yang, S.; Li, J.; Lai, X.; Wu, J.; Li, W.; M.A.; Z.; Yu, B.; Zhao, H.; Jia, J. VisionThink: Smart and efficient vision-language model via reinforcement learning. *arXiv* **2025**, arXiv:2507.13348.
28. Zheng, Z.; Yang, M.; Hong, J.; Zhao, C.; Xu, G.; Yang, L.; Shen, C.; Yu, X. DeepEyes: Incentivizing “thinking with images” via reinforcement learning. *arXiv* **2025**, arXiv:2505.14362.
29. Liu, R.; Fu, B.; Song, J.; Li, K.; Li, W.; Xue, L.; Qiao, H.; Zhang, W.; Meng, D.; Cao, X. ZoomEarth: Active perception for ultra-high-resolution geospatial vision-language tasks. *arXiv* **2025**, arXiv:2511.12267.
30. Liu, J.; Sun, L.; Fu, R.; Yang, B. Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. *arXiv* **2025**, arXiv:2509.22221.
31. Wang, F.; Chen, M.; Li, Y.; Yang, Y.; Zhang, Y.; Lan, L.; Yang, X.; Sun, H.; Wang, Y.; Wang, D.; et al. GeoEyes: On-demand visual focusing for evidence-grounded understanding of ultra-high-resolution remote sensing imagery. *arXiv* **2026**, arXiv:2602.14201.
32. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130.
33. Li, Y.; Luo, J.; Zhang, Y.; Tan, Y.; Yu, J.G.; Bai, S. Learning to holistically detect bridges from large-size VHR remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *46*, 11507–11523.
34. Li, Y.; Wang, L.; Wang, T.; Yang, X.; Luo, J.; Wang, Q.; Deng, Y.; Wang, W.; Sun, X.; Li, H.; et al. STAR: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *arXiv* **2024**, arXiv:2406.09410.
35. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **2025**, *645*, 633–638.
36. Snell, C.; Lee, J.; Xu, K.; Kumar, A. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In Proceedings of the 13th International Conference on Learning Representations (ICLR 2025), Singapore, 24–28 April 2025.
37. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2.5-VL technical report. *arXiv* **2025**, arXiv:2502.13923.
38. ByteDance Seed. Introduction to Techniques Used in Seed 1.6. ByteDance Seed Blog. 2025. Available online: <https://seed.bytedance.com/blog/introduction-to-techniques-used-in-seed1-6> (accessed on 28 May 2026).
39. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtual, 25–29 April 2022.

40. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
41. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338.
42. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023; pp. 46595–46623.
43. Luo, J.; Zhang, Y.; Yang, X.; Wu, K.; Zhu, Q.; Liang, L.; Chen, J.; Li, Y. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2025), Honolulu, HI, USA, 19–23 October 2025; pp. 9206–9217.
44. Li, X.; Ding, J.; Elhoseiny, M. VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024; pp. 3229–3242.
45. OpenAI; Hurst, A.; Lerer, A.; Goucher, A.P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.J.; Welihinda, A.; Hayes, A.; et al. GPT-4o system card. *arXiv* **2024**, arXiv:2410.21276.
46. OpenAI. GPT-4V(ision) System Card. OpenAI Technical Report. 2023. Available online: <https://openai.com/research/gpt-4v-system-card> (accessed on 28 May 2026).
47. Comanici, G.; Bieber, E.; Schaeckermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv* **2025**, arXiv:2507.06261.
48. Anthropic. Claude 3.5 Sonnet Model Card Addendum. Anthropic Technical Report. 2024. Available online: <https://www.anthropic.com/news/claude-3-5-sonnet> (accessed on 28 May 2026).
49. Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv* **2024**, arXiv:2412.05271.
50. Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv* **2025**, arXiv:2504.10479.
51. Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. LLaVA-OneVision: Easy visual task transfer. *arXiv* **2024**, arXiv:2408.03326.
52. Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. Qwen3-VL technical report. *arXiv* **2025**, arXiv:2511.21631.
53. Fiaz, M.; Debary, H.; Fraccaro, P.; Paudel, D.; Van Gool, L.; Khan, F.; Khan, S. GeoVLM-R1: Reinforcement fine-tuning for improved remote sensing reasoning. *arXiv* **2025**, arXiv:2509.25026.
54. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv* **2024**, arXiv:2409.12191.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.