

VLM-ReG : Vision-Language Models Enhanced via Reward-Refined GRPO in Remote Sensing Reasoning

YanZhong Zhou

College of Intelligence Science and Technology, National University of Defense Technology
Changsha, China
zhouyanzhong18@nudt.edu.cn

BiaoXin Li

College of Intelligence Science and Technology, National University of Defense Technology
Changsha, China
libiaoxin@nudt.edu.cn

ZhangLing Wang

College of Intelligence Science and Technology, National University of Defense Technology
Changsha, China
wangzhangling19@nudt.edu.cn

Chao Wang

College of Intelligence Science and Technology, National University of Defense Technology
Changsha, China
wangchao@nudt.edu.cn

WanPeng Zhang

College of Intelligence Science and Technology, National University of Defense Technology
Changsha, China
wpzhang@nudt.edu.cn

ABSTRACT

Multimodal large language models (MLLMs) have achieved remarkable success in natural-image vision-language tasks; however, their adaptation to remote sensing (RS) imagery remains limited due to a substantial domain gap. To advance MLLM capabilities in RS, we present two core contributions. First, we introduce the Remote Sensing Image Reasoning Dataset (RS-IRD), a high-quality dataset comprising 20k RS images paired with detailed reasoning chains. Leveraging a novel "Answer-Guided Modality Bridging" pipeline, each image is automatically converted into "Image-CoT" (Image-based Chain of Thought) that encapsulates the full reasoning process. Second, We enhance Group Relative Policy Optimization (GRPO) by integrating Adaptive Advantage Prioritization (AAP) to eliminate gradient stagnation and Refined Quality Rewards (RQPO) to sharpen reward signals, establishing an end-to-end reasoning-boosting pipeline for RS-MLLM. The resulting Vision-Language Model trained via refined GRPO rewards (VLM-ReG) achieves significant improvements in both the quality and efficiency of RS image reasoning.

KEYWORDS

Multimodal Large Language Model (MLLM); Remote Sensing (RS); Chain-of-Thought (CoT) Reasoning; Group Relative Policy Optimization (GRPO)

ACM Reference Format:

YanZhong Zhou, BiaoXin Li, ZhangLing Wang, Chao Wang, and WanPeng Zhang. 2026. VLM-ReG : Vision-Language Models Enhanced via Reward-Refined GRPO in Remote Sensing Reasoning. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 10 pages. <https://doi.org/10.65109/GMIK4254>



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). <https://doi.org/10.65109/GMIK4254>

1 INTRODUCTION

Remote sensing (RS) image interpretation serves as a cornerstone for Earth observation, underpinning vital applications such as environmental monitoring, urban sprawl analysis, disaster response, and agricultural yield estimation[27, 49]. A large number of these practical tasks call for complex reasoning capabilities—for example, deducing the extent of flood-affected areas by fusing synthetic aperture radar (SAR) and optical image data[54], or calculating farm crops coverage changes based on multi-temporal infrared imagery[46]. Such processes demand seamless integration of visual information from diverse RS sensors and logical language reasoning, making the enhancement of multimodal reasoning ability in RS-oriented models a key prerequisite for advancing intelligent RS systems. However, achieving this goal remains a prominent challenge in the field[20].

Conventional RS reasoning methods mostly rely on task-specific specialist models, such as those designed solely for RS image captioning or RS visual question answering (VQA). These models typically adopt a "direct prediction" approach, generating concise final results without explicit intermediate reasoning steps[5, 18]. This paradigm struggles to handle the intricate logical relationships inherent in complex RS tasks. In recent years, researchers have begun exploring Multimodal Large Language Models (MLLMs) for RS scenarios, attempting to introduce CoT reasoning through supervised fine-tuning (SFT) with manually constructed datasets[23, 50]. Unfortunately, these manually designed datasets are mostly confined to basic vision-language alignment tasks (lacking data supporting RS-specific visual reasoning) and rely heavily on extensive, time-consuming, and labor-intensive manual annotation. This dual limitation directly leads to the "Pseudo-CoT" phenomenon in RS-oriented Multimodal Large Language Models (MLLMs) trained via supervised fine-tuning—their reasoning only involves rigid step-by-step descriptions, lacking human-like cognitive processes essential for complex RS tasks[13, 15].

Recently, DeepSeek-R1[10] made a breakthrough in the field of text-only Large Language Models (LLMs): it proved that Reinforcement Learning (RL) can induce the spontaneous emergence

of complex cognitive reasoning capabilities. This achievement offers a new paradigm for enhancing reasoning ability without relying on extensive human annotations, prompting a critical question: Can RL be applied to stimulate high-quality reasoning in RS-oriented MLLMs? Early attempts to directly apply the DeepSeek-R1-Zero paradigm[10] (pure RL training) to base RS-MLLMs have achieved considerable progress and results[30, 38, 43]. Without access to large-scale, high-quality multimodal RS reasoning data, these models fail to generate coherent and complex CoT tailored to RS characteristics[4, 35]. Moreover, prolonged RL training often leads to a sharp decline in model performance, making it difficult to achieve effective reasoning enhancement[38, 45].

To address the aforementioned dataset limitations and model training challenges in multimodal reasoning within the RS domain, this study focuses on enhancing the reasoning capability of Vision-Language Models (VLMs) in RS scenarios. It explores the use of Reinforcement Learning (RL) to stimulate high-quality reasoning behavior in models and enable interpretable reasoning process outputs.

Our main contributions are summarized as follows:

- We introduce remote sensing image reasoning dataset (RS-IRD), a high-quality remote sensing reasoning dataset comprising 20k samples, constructed through a novel process that converts multiple remote sensing images into text-based “Image-CoT” that incorporates detailed reasoning processes. This dataset enables effective cold-start initialization for RS-MLLMs without manual annotation, providing a foundation for learning preliminary “human-like” reasoning patterns.
- we integrate Adaptive Advantage Prioritization (AAP) into Group Relative Policy Optimization (GRPO) to mitigate gradient stagnation and employ Refined Quality Rewards (RQPO) to refine reward signals, ultimately improving reasoning quality and training efficiency.
- Extensive experiments on multiple RS reasoning tasks demonstrate the effectiveness of our approach, showing significant improvements in reasoning quality and efficiency, with clear and logical reasoning outputs that enhance the model’s ability to handle complex remote sensing reasoning challenges.

2 RELATED WORK

2.1 RS Explanation and Reasoning Task

Remote sensing (RS) explanation and reasoning task requires models to transcend outputting task results (e.g., VQA answers, image captions, object detection boxes) and generate interpretable inference logic (e.g., associating RS image features with semantic conclusions), which is essential for high-reliability scenarios like disaster assessment[27, 49]. However, existing specialist models for these sub-tasks prioritize accuracy but lack reasoning transparency: RSVQA[25], MQVQA[48] and OKVQA[28] output final answers without explaining how image content maps to textual queries; RS image captioning methods MLCA-Net[7] and FC-Att[51] generate coherent scene descriptions without illustrating links between visual elements and text; RS visual grounding models TransVG[8] and VLTVG[42] realize target localization via language guidance but fail to articulate spatial reasoning logic; RS object detection models S2A-Net[11] and Oriented R-CNN[40] excel at oriented

target recognition yet only output bounding box coordinates and categories without justifying detection decisions. Moreover, current RS datasets also fail to support reasoning capability training: UCM-Captions[34] (for RS image captioning), CRSVQA[48] (for RS VQA), and DIOR[12] (for RS object detection) only provide final labels without annotated reasoning steps. This lack of reasoning-centric supervision prevents models from learning to generate interpretable logical chains, resulting in insufficient reasoning capabilities.

2.2 Reinforcement Learning for Multimodal LLM Reasoning

To enhance reasoning capabilities of multimodal large language models (MLLMs), recent studies have explored reinforcement learning (RL) approaches, drawing inspiration from effective RL-based reasoning techniques in text-only LLMs. For instance, LMM-R1[33] proposes a two-stage rule-based RL framework to boost reasoning in 3B-scale MLLMs, though its performance gains are primarily limited to textual scenarios. Reason-RFT[37] leverages supervised fine-tuning (SFT) with Chain-of-Thought (CoT) data to initialize the RL process, laying a foundation for multimodal reasoning. Vision-R1[14] enhances multimodal CoT datasets using RL strategies from text LLMs and introduces progressive thinking suppression during Group Relative Policy Optimization (GRPO) training. MM-Eureka[29] constructs the MMK12 dataset and adopts a two-stage RL method to explore multimodal reasoning frontiers, while VL-Rethinker[38] employs selective sample replay and explicit textual rethinking triggers to refine reasoning outcomes. R1-V[47] focuses on RL for visual-centric reasoning tasks but lacks generalization to broader multimodal domains. Despite these efforts, most existing RL methods for multimodal LLM reasoning often result in redundant or misleading reasoning steps, while poorly designed reward functions further exacerbate issues like output verbosity, limiting overall reasoning performance.

3 METHOD

The overall pipeline of VLM-ReG is illustrated in Figure 1. First, we construct the Remote Sensing Image Reasoning Dataset (RS-IRD) (20 k samples) via Answer-Guided Modality Bridging to generate Image-CoT examples, providing a cold-start initialization. Second, a task-specific loss function pre-trains the policy model to elicit preliminary reasoning ability. Finally, we optimize Group Relative Policy Optimization (GRPO) with two components: Adaptive Advantage Prioritization (AAP), which alleviates gradient stagnation, and Enhanced GRPO with Refined Quality Rewards (RQPO), yielding an end-to-end reasoning-enhancement path for RS-VLM.

3.1 Remote Sensing Image Reasoning Dataset(RS-IRD) Construction

The ability to generate high-quality, complex Chain of Thought (CoT) data that is closely aligned with human thinking patterns has become a key factor in training Multimodal Large Language Models to competently tackle complex multimodal reasoning tasks[41, 44]. To address this, a “Answer-Guided Modality Bridging” method is proposed, as illustrated in Figure 2. By integrating the capabilities of existing multimodal large models for reasoning, this method enables the indirect conversion of multimodal information, thereby

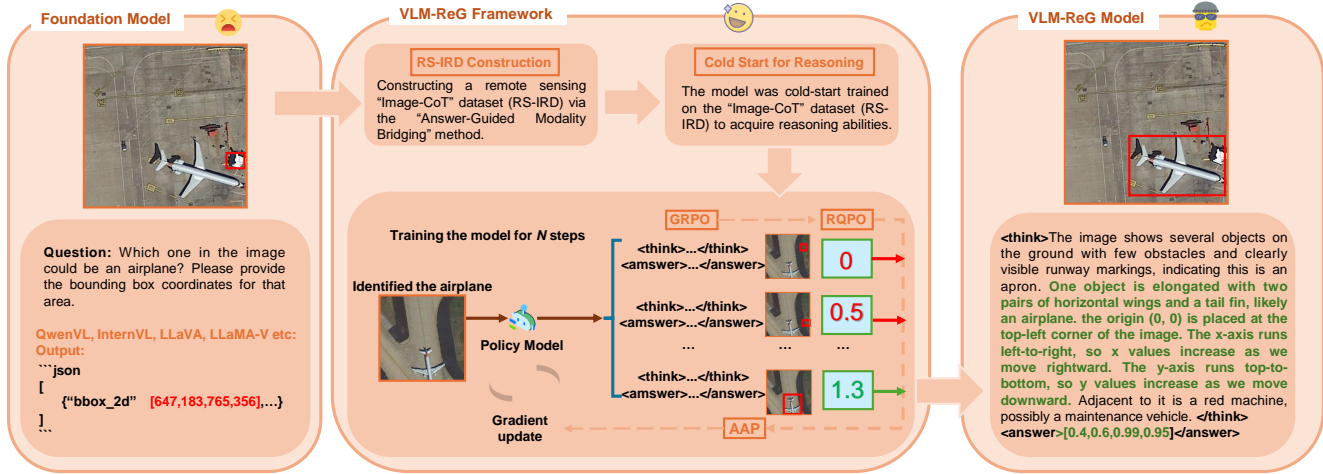


Figure 1: Overall pipeline of VLM-ReG framework.

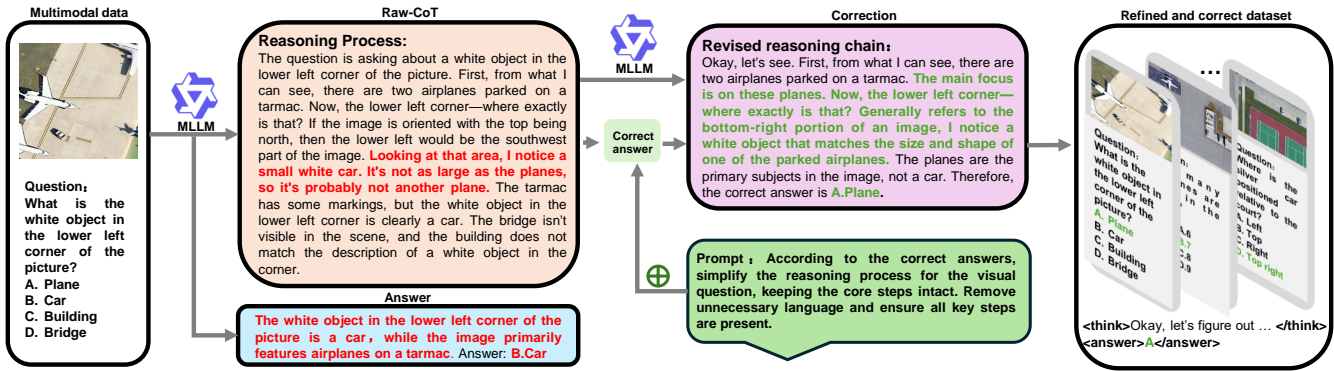


Figure 2: The overall data generation pipeline incorporating our Answer-Guided Modality Bridging method. The multimodal data (e.g., airport images, questions, options), input to MLLM for initial image description and reasoning. Concatenate this reasoning with ground-truth answers, input to another MLLM for CoT correction. Filter consistent samples to build dataset.

capturing complex reasoning processes. The specific workflow is as follows:

First, multimodal inputs—including airport images, textual queries (e.g., "What is the white object in the lower left corner of the picture?"), and candidate answers (e.g., Plane, Car)—are fed into an MLLM (in this paper, we selected the Qwen2-VL-72B), a state-of-the-art multimodal model designed for vision-language tasks. This process generates a "Raw-Chain of Thought (Raw-CoT)" that encompasses image descriptions and reasoning processes. During this stage, the model analyzes various options: for cars, it considers color attributes and visual features presented from the photo's perspective; for building, it determines that their scale characteristics make them incompatible with the description of a "small white object"; for bridges, it isn't visible in the scene. Ultimately, the model outputs a preliminary reasoning conclusion that tends to select "Car" along with the corresponding reasoning process.

Second, the "Raw-CoT" and the correct answer are concatenated, and this combined input is re-fed into a suitable multimodal reasoning large model. Incorporating the correct answer, the model

performs Chain of Thought correction: it refocuses on the matching between the white object in the image and the characteristics of a plane, and optimizes the reasoning logic by considering factors such as the generally white appearance of planes in both commercial and military models. Through this optimization, the model ultimately confirms the answer as "Plane."

Finally, reasoning processes where the answer aligns with the actual scenario are manually selected and verified. Via the filtering mechanism, these valid processes are integrated to form a concise and accurate dataset. This dataset includes images, questions, revised reasoning processes, and correct answers. It can be applied in the model training phase to support cold start: by leveraging "Image-CoT" data that aligns with human cognitive patterns, the model's reasoning process exhibits a natural and logically coherent thinking paradigm, thereby enhancing the model's reasoning capability in multimodal visual question answering tasks.

Table 1: Comparison of the occurrence frequency of self-reflective indicators between llava-cot, mulberry and our RS-IRD. The higher frequency of these reflective markers in our dataset demonstrates its distinctive self-reflection and self-correction characteristic.

Word	llava-cot (100k)	Mulberry (260k)	RS-IRD (20k)
Hmm	1	0	6,735
Alternatively	251	68	8,741
Mistake	183	8,784	12,366
Wait	2,300	1,122	32,756
Check	8,332	26,421	20,511

We present an assessment of our newly developed Remote Sensing Image Reasoning Dataset. The development of RS-IRD addresses critical limitations in current multimodal Chain-of-Thought resources, which exhibit insufficient cognitive complexity. This dataset strategically integrates MLLM refined Chain-of-Thought framework as foundational training material. As shown in Table 1, our quantitative evaluation compares key cognitive elements questioning, reflection, and inspection—across Mulberry[44], LLaVA-CoT[41], and RS-IRD. Findings reveal that RS-IRD exhibits a substantially greater prevalence of cognitive patterns reflecting human reasoning compared to prior approaches. This enhanced structural complexity enables base multimodal large language models to effectively acquire reasoning capabilities, establishing a high-quality foundation for subsequent reinforcement learning optimization.

3.2 Cold Start of Policy Model for Reasoning

The initialization phase of the policy model is a critical prerequisite for enabling subsequent reinforcement learning processes, as it equips the initial multi-modal large language model with fundamental self-reflective reasoning capabilities. This phase ensures that the policy model, denoted as π_θ , can generate reasoning paths that incorporate reflection awareness before entering the RL loop, thereby laying a solid foundation for high-quality reasoning and self-correction.

The core objective of this initialization is twofold. First, it trains the model to revise its initial response toward the ground truth answer with the guidance of external reflection generated by a more advanced LLM. Second, it enables the model to leverage the reasoning logic and knowledge embedded in the ground truth to optimize its future prediction trajectories. To achieve these goals, a dedicated loss function for cold-start initialization is formulated, which is defined as follows:

$$\mathcal{L}_{\text{cold-start}} = -\mathbb{E}_{\tau \sim \mathcal{D}} \left[\sum_{t=1}^T \log \left(\pi_{\text{initial}}(a_1, \langle \text{think} \rangle \dots \langle \text{/think} \rangle, a_2 \mid q) \right) \right] \quad (1)$$

In this equation, a_1 represents the initial response generated by the policy model when given the input prompt q ; the segment " $\langle \text{think} \rangle \dots \langle \text{/think} \rangle$ " stands for the think content produced by a

larger-scale LLM to identify flaws in a_1 or streamline redundant expressions; and a_2 denotes the ground truth answer corresponding to the prompt q .

During the initialization training process, the policy model learns to establish a connection between its initial response, external thinking, and the ground truth. This learning process not only enables the model to acquire basic self-correction skills but also aligns its reasoning direction with the correct logic. By the end of this phase, the policy model is no longer limited to generating simple responses—it can integrate thinking information to optimize reasoning paths, thus avoiding the problem of low-quality outputs caused by local dependency in reasoning and providing reliable initial capabilities for the subsequent RL phase.

3.3 Group Relative Policy Optimization (GRPO) Algorithm

In the reinforcement learning framework designed for complex reasoning tasks, the Group Relative Policy Optimization algorithm serves as the core policy optimization method[10], addressing limitations of traditional approaches such as token level loss-based supervised fine-tuning (SFT). Unlike SFT, which focuses on matching individual tokens to reference outputs, GRPO optimizes the policy by leveraging policy gradients calculated from reward signals, enabling more effective exploration of high-quality reasoning paths.

The key design idea of GRPO lies in comparing multiple generated responses within a sampled group to estimate relative advantages, thereby promoting the exploration of richer and more diverse reasoning solutions. Formally, let Q represent the set of input questions, $\pi_{\theta_{\text{old}}}$ be the policy model before update (used as the baseline policy), and $\{o_1, o_2, \dots, o_G\}$ denote a group of responses generated by $\pi_{\theta_{\text{old}}}$ for a specific question $q \in Q$. Additionally, $\pi_{\theta_{\text{ref}}}$ is a frozen reference model introduced to prevent excessive deviation of the updated policy from the initial reasonable reasoning distribution. The optimization objective of GRPO is defined as:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{q \sim Q, \\ \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}}} \left[\frac{1}{G} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left(\frac{\pi_\theta(o_{i,t} \mid q)}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q)} A_i, \text{clip} \left(\frac{\pi_\theta(o_{i,t} \mid q)}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right] \quad (2)$$

Here, ϵ is a clipping hyperparameter (typically set to a small value like 0.2) that limits the excessive update of the policy ratio to ensure training stability; β is the penalty coefficient for the Kullback-Leibler (KL) divergence between the updated policy π_θ and the reference policy $\pi_{\theta_{\text{ref}}}$, which prevents the policy from drifting too far from the initial valid reasoning space. The advantage A_i for each response o_i in the group is calculated based on the relative difference between the reward of o_i and the statistical characteristics of the group's rewards:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})} \quad (3)$$

where $\{r_1, r_2, \dots, r_G\}$ are the rewards obtained by each response in the group. This intra-group advantage estimation method eliminates the need for a separate critic model (as required by traditional Proximal Policy Optimization, PPO), reducing computational complexity while effectively capturing the relative quality of different reasoning paths. By optimizing the above objective, GRPO guides the policy model to gradually generate reasoning paths that are more in line with task requirements, laying the groundwork for further enhancing reasoning quality through subsequent improvements.

3.4 Enhanced GRPO with Refined Quality Rewards(RQPO)

Within the reinforcement learning paradigm, conventional GRPO training, which depends predominantly on final-answer correctness signals, often fails to address the internal quality of the generated reasoning sequences. This can lead to verbose, logically inconsistent, or inefficient step-by-step explanations, ultimately impairing the model’s generalization and final task performance. To mitigate these issues, we introduce **RQPO**, an augmented GRPO framework. In RQPO, our primary objectives are twofold: (1) to strengthen the policy model’s core reasoning proficiency via reinforcement learning, and (2) to explicitly encourage the production of high-caliber reasoning steps that are both succinct and logically sound. To achieve these aims, we devise a composite reward function:

$$R_{\text{total}} = R_{\text{task}} + R_{\text{cal}} \quad (4)$$

This formulation encourages the model to not only produce correct answers but also to engage in insightful and concise reflection.

3.4.1 Task-Oriented Reward. The task reward R_{task} comprises two components: a formatting reward and an accuracy reward. The formatting reward R_{format} promotes adherence to structural guidelines by rewarding the model when the reasoning process is correctly enclosed within dedicated tags, such as "`<think>...</think>`". The accuracy reward R_{accuracy} evaluates whether the initial solution matches the ground-truth answer. This dual-component design reinforces both structural discipline and logical correctness in the reasoning process:

$$R_{\text{task}} = R_{\text{format}} + R_{\text{accuracy}} \quad (5)$$

where:

$$R_{\text{format}} = \begin{cases} 0.5, & \text{if formatting rules are satisfied} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

$$R_{\text{accuracy}} = \begin{cases} 0.5, & \text{if the initial answer is correct} \\ 0, & \text{otherwise} \end{cases}$$

3.4.2 Adaptive Length-Calibrated Reward. In reasoning, maintaining an equilibrium between response comprehensiveness and conciseness is non-trivial. To address this, we assign each input a dynamic target length L_{adaptive} retrieved from a Reference Length Map constructed before reinforcement learning: the token-wise lengths of ground-truth reasoning chains extracted from the human-annotated subset are grouped by K-means clusters of textual and visual features, and the median length of each cluster is stored as its representative L_{cluster} .

During training, for a given input x , we first determine its cluster assignment and retrieve the corresponding $L_{\text{adaptive}} = L_{\text{cluster}}$. The reward is then computed based on the deviation of the generated length L_y from this adaptive target. The deviation ratio is defined as:

$$\Delta = \frac{|L_y - L_{\text{adaptive}}|}{L_{\text{adaptive}}} \quad (7)$$

A raw reward score is calculated using a Gaussian kernel, which provides a smooth penalty:

$$R_{\text{raw}} = \delta \cdot \exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \quad (8)$$

Here, δ is a baseline reward (set to 0.5) granted when the output length perfectly matches the target, and σ is a bandwidth parameter controlling the strictness of the length tolerance. The final calibrated reward:

$$R_{\text{cal}} = \max(0.0, \min(1.0, R_{\text{raw}})) \quad (9)$$

This formulation delivers a continuous, context-sensitive signal that encourages the model to tailor its response length to the intrinsic demands of the problem. By penalizing significant deviations from a dynamically assigned, semantically-grounded target, the reward function effectively promotes responses that are both sufficiently detailed and efficiently expressed, thereby enhancing the overall quality and stability of the reasoning output.

3.5 Adaptive Advantage Prioritization (AAP)

While the Group Relative Policy Optimization (GRPO) algorithm provides effective support for policy optimization in complex reasoning tasks, it encounters a critical challenge termed "gradient stagnation" when deployed in multimodal reasoning scenarios. This issue primarily stems from the sparse binary rewards that are common in multimodal tasks: when all responses within a sampled group yield identical rewards (e.g., all responses are correct or all are incorrect), the advantage estimates for these responses become near-zero. Consequently, the policy gradients derived from these advantages are close to zero, which halts the model’s learning progress for such samples and severely undermines the overall efficiency of the training process.

To mitigate the gradient stagnation problem of GRPO and enhance the efficiency of sample utilization, we introduce Adaptive Advantage Prioritization (AAP), a method that dynamically identifies and emphasizes training samples with the most reliable and substantial learning signals. AAP operates through the following detailed procedure:

3.5.1 Dynamic Advantage Thresholding: For each training batch $\mathcal{B} = \{(x_i, y_i, \hat{A}_i)\}_{i=1}^N$, AAP first analyzes the distribution of advantage values \hat{A}_i . Instead of using fixed, pre-defined thresholds, it calculates a dynamic cutoff point γ based on the batch’s own statistics:

$$\gamma = \mu_{|\hat{A}|} + \beta \cdot \sigma_{|\hat{A}|} \quad (10)$$

Here, $\mu_{|\hat{A}|}$ and $\sigma_{|\hat{A}|}$ are the mean and standard deviation of the absolute advantages $|\hat{A}_i|$ within the batch. The hyperparameter β controls the selectivity of the filtering process. A candidate set

Table 2: Comprehensive comparison with SoTA MLLMs (closed-source, open-source general/reasoning MLLMs) across diverse multimodal reasoning benchmarks. “ALL” represents the average score of the four LRS sub-tasks (FAIR, Bridge, STAR). “Avg.” denotes the average performance over all benchmarks. The best results are bolded and the second best results are underlined in all following tables. The best results are bolded. Our VLM-ReG outperforms baselines by a large margin.

Model	Param.	LRS				MME-RW-RS	VRSBench-val	DDFAV	RSVLM-QA	Avg.
		FAIR	Bridge	STAR	ALL					
<i>Closed-source</i>										
GPT-4o	–	22.15	31.84	27.40	27.13	28.92	52.32	51.88	51.12	42.27
GPT-4o-mini	–	18.67	31.99	25.85	25.50	6.69	40.33	46.23	48.72	33.49
Claude-3.5-Sonnet	–	12.95	26.69	13.29	17.64	25.74	48.24	48.66	53.85	38.83
<i>Open-source</i>										
Qwen2.5-VL-Instruct	7B	25.62	<u>40.28</u>	<u>29.89</u>	<u>31.93</u>	52.93	59.81	56.32	55.32	51.26
InternVL2.5-Instruct	8B	<u>27.25</u>	36.52	26.11	29.96	45.65	45.41	51.22	<u>56.91</u>	45.83
LLaVA-1.5-Instruct	7B	18.76	30.70	22.63	24.03	26.38	59.32	62.22	52.18	44.83
LLaVA-Next	7B	21.14	37.25	26.10	28.16	40.35	62.36	64.94	54.25	50.01
LLaVA-OV	7B	20.61	35.11	26.08	27.27	<u>53.53</u>	64.26	<u>66.77</u>	54.84	<u>53.33</u>
Mini-Gemini	7B	17.90	35.92	20.13	24.65	31.51	52.33	49.52	50.09	41.62
Geochat	7B	20.18	24.54	13.75	19.49	28.62	<u>67.39</u>	55.12	52.36	44.60
<i>Our model</i>										
VLM-ReG (Ours)	7B	32.14	45.11	35.65	37.63	60.12	78.11	74.12	63.12	62.62
Δ (Ours – SoTA)	–	+4.89	+4.83	+5.76	+5.70	+6.59	+10.72	+7.35	+6.21	+9.29

\mathcal{S} is then formed, containing indices of samples whose absolute advantage meets or exceeds this cutoff:

$$\mathcal{S} = \{i \mid |\hat{A}_i| \geq \gamma\} \quad (11)$$

This adaptive thresholding ensures that the selection criteria remain relevant to the current training distribution, filtering out samples with negligible learning signals while retaining those with meaningful policy gradients.

3.5.2 Entropy-Aware Sample Weighting. Samples within the candidate set \mathcal{S} are not treated equally. AAP assigns a relative importance weight w_i to each sample $i \in \mathcal{S}$ based on both the magnitude of its advantage and the associated policy entropy. The weight is computed using a modified softmax distribution:

$$w_i = \frac{\exp(|\hat{A}_i|/\tau_i)}{\sum_{j \in \mathcal{S}} \exp(|\hat{A}_j|/\tau_j)} \quad (12)$$

The key innovation is the sample-specific temperature τ_i , which is adaptively set as:

$$\tau_i = \tau_0 \cdot (1 + \eta \cdot H(\pi_\theta(\cdot|x_i))) \quad (13)$$

In this equation, $H(\pi_\theta(\cdot|x_i))$ represents the entropy of the policy for the given input x_i , τ_0 is a base temperature, and η is a scaling factor. This mechanism increases the temperature for high-entropy samples (where the model is uncertain), effectively flattening the weight distribution and encouraging exploration. In our experiments, we adopt $\beta = 1.0$ and $\tau_0 = 1.0$ as default values, performing coarse grid searches around these defaults when computational resources allow. Monitoring the retained sample proportion in \mathcal{S} serves as a reliable diagnostic for β tuning.

3.5.3 Weighted Policy Optimization Loss. The final distilled batch for policy update is constructed from the candidate set \mathcal{S} . The standard GRPO objective is then modified to incorporate the computed sample weights, focusing the gradient update on the most valuable samples:

$$\mathcal{L}_{\text{AAP}} = \mathbb{E}_{i \in \mathcal{S}} \left[w_i \cdot \min \left(r(\theta) \hat{A}_i, \text{clip} \left(r(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_i \right) \right] \quad (14)$$

where $r(\theta) = \frac{\pi_\theta(y_i|x_i)}{\pi_{\theta_{\text{ref}}}(y_i|x_i)}$ is the probability ratio. This weighted loss ensures that policy updates are dominated by samples that provide strong, reliable advantage signals and are appropriately balanced between exploration and exploitation based on model confidence.

To ensure numerical stability, especially in early training stages where the candidate set \mathcal{S} might be small, AAP employs a fallback strategy. If $|\mathcal{S}|$ is below a minimal size k_{min} , the batch is supplemented with the top- k_{min} samples from \mathcal{B} ranked by $|\hat{A}_i|$. This prevents overly aggressive filtering when advantage signals are initially weak or homogeneous. By implementing this multi-stage process, AAP provides a robust and adaptive framework for sample prioritization. It enhances learning efficiency and training stability by dynamically focusing on high-value experiences and intelligently balancing the exploration-exploitation trade-off throughout the training process.

4 EXPERIMENTAL RESULTS

4.1 Experiment Settings

4.1.1 Training Dataset. (1) SFT: To construct the reasoning SFT dataset (RS-IRD) for the cold-start initialization phase, we first curate samples from several established multimodal RS sources, including the FARIM[36], RSIVQA[55], VRSBench-train[19], and DOTA[39]

Table 3: Ablation study on the contribution of individual training components to the VLM-ReG model. “AAP” denotes the improvement made to the original GRPO algorithm via “Adaptive Advantage Prioritization”, while “RQPO” denotes the improvement via “Refined Quality Rewards”. The best result in each column is highlighted in bold.

Model Components	LRS			ALL	MME-RW-RS	VRSBench-val	Avg.
	FAIR	Bridge	STAR				
Cold Start (CS) only	23.12±0.19	38.26±0.21	27.12±0.14	29.50±0.18	50.31±0.15	57.12±0.24	45.64±0.19
Original GRPO only	27.16±0.12	41.22±0.13	31.31±0.11	33.23±0.12	54.22±0.16	65.12±0.17	50.86±0.15
AAP+RQPO	28.42±0.19	41.92±0.21	32.27±0.11	34.20±0.17	55.53±0.13	67.26±0.18	52.33±0.16
CS + AAP	30.11±0.19	42.32±0.15	34.11±0.16	35.51±0.17	57.81±0.18	70.12±0.12	54.48±0.16
CS + RQPO	31.01±0.14	43.88±0.16	34.91±0.18	36.60±0.16	58.42±0.13	74.25±0.14	55.98±0.14
VLM-ReG (Ours)	32.14±0.22	45.11±0.14	35.65±0.15	37.63±0.17	60.12±0.12	78.11±0.13	58.62±0.14

dataset. We subsequently implement the data construction methodology outlined in Section 3.1, which yields a refined supervised fine-tuning (SFT) dataset containing roughly 20k samples. (2) RL: The reinforcement learning phase utilizes a heterogeneous ensemble of multimodal remote sensing inference examples, drawn from various RS-specialized corpora such as RSVQA-LR[53], CRSVQA[48], WHU-RS19[3] and DIOR[12]. These resources comprehensively cover RS-oriented reasoning paradigms, including multi-layered scene interpretation, color-based query resolution, numerical quantity estimation, and object localization query processing. The resulting RL training dataset comprises heterogeneous, inter-scenario remote sensing inference examples.

4.1.2 Baselines and Benchmarks. To comprehensively evaluate VLM-ReG, we compare against three groups of baselines: (1) Closed-source MLLMs: General-purpose models GPT-4o[31], GPT-4o-mini[32], Claude-3.5-Sonnet[1]; (2) Open-source MLLMs: InternVL2.5[6], Qwen-2.5-VL[2], LLaVA-1.5[24], LLaVA-Next[22], LLaVA-OV[21], Mini-Gemini[9] and Geo-chat[16]. The parameters of the above open-source models are all around 7B. We evaluate VLM-ReG across three categories of multimodal RS reasoning benchmarks: LRS[26], MME-RW-RS[52], VRSBench-validation[19], DDFAV[17] and RSVLM-QA[56]. These benchmarks comprehensively encompass a diverse spectrum of complex reasoning challenges, integrating multi-scale visual data derived from both orbital satellite observations and low-altitude drone remote sensing platforms.

4.1.3 Implementation Setup. The VLM-ReG model leverages Qwen-2.5-VL-7B-Instruct as its foundational framework, employing a two-phase training methodology. Initial model development utilizes LLaMA Factory in conjunction with RS-IRD to establish foundational language understanding and reasoning capabilities, with supervised fine-tuning executed over a single epoch to mitigate overfitting risks. Subsequent reinforcement learning employs the Verl framework, processing 3,000 samples across three epochs. This phase implements batch dimensions of 128 for both rollout and training procedures, incorporating eight rollouts per input instance, a temperature parameter set to 1.0, and Adam-based optimization with a learning rate of 1×10^{-6} . Notably, the GRPO algorithm excludes KL divergence regularization to promote expansive policy exploration. All computational evaluations were conducted using eight NVIDIA A100-80G GPUs.

4.2 Main Results of Benchmarks

Table 2 comprehensively compares our proposed model, VLM-ReG, against state-of-the-art multimodal large language models, including both closed-source and open-source general and reasoning-oriented MLLMs—across diverse and challenging multimodal reasoning benchmarks. The “LRS” metric systematically aggregates performance across three key subtasks (FAIR, Bridge, and STAR), with “ALL” denoting their average and “Avg.” representing the mean performance across all benchmarks. Best results are presented in bold, while second-best results are underlined with asterisks for enhanced clarity and reproducibility.

Among closed-source MLLMs, GPT-4o achieves an “Avg.” score of 42.27, with notable performance on MME-RW-RS (28.92) and VRSBench-val (52.32). In contrast, GPT-4o-mini and Claude-3.5-Sonnet obtain lower “Avg.” scores of 33.49 and 38.83, respectively.

Among open-source 7B/8B MLLMs, Qwen2.5-VL (Avg. 51.26) and InternVL2.5 (Avg. 45.83) show competitive performance, while LLaVA-OV attains the second-best Avg. of 53.33. Notably, our VLM-ReG (7B) significantly outperforms all SoTA baselines. In LRS tasks, it achieves the highest “ALL” score of 37.63 (+5.70 improvement) by sweeping FAIR, Bridge, and STAR. VLM-ReG further dominates individual benchmarks—including MME-RW-RS (60.12), VRSBench-val (78.11), DDFAV (74.12), and RSVLM-QA (63.12)—surpassing previous bests by margins ranging from 6.21 to 10.72.

Collectively, VLM-ReG attains an “Avg.” score of 62.62, representing a +9.29 improvement over the next-best open-source model. This demonstrates that the intrinsic reasoning mechanism endows lightweight visual-language models with robust multimodal capabilities, enabling them to surpass all open-source counterparts and approach the performance of proprietary systems. The consistent superiority across benchmarks validates the efficacy and broad applicability of the proposed methodology.

4.3 Ablation Study

To systematically evaluate the contributions of different training components, we conduct an ablation study as shown in Table 3. The results demonstrate the effectiveness of each component in improving the performance of the VLM-ReG model across multiple metrics.

When only the Cold Start (CS) stage is used, the model achieves an average performance of 45.64, with relatively lower scores in metrics such as FAIR (23.12) and STAR (27.12). The Original GRPO

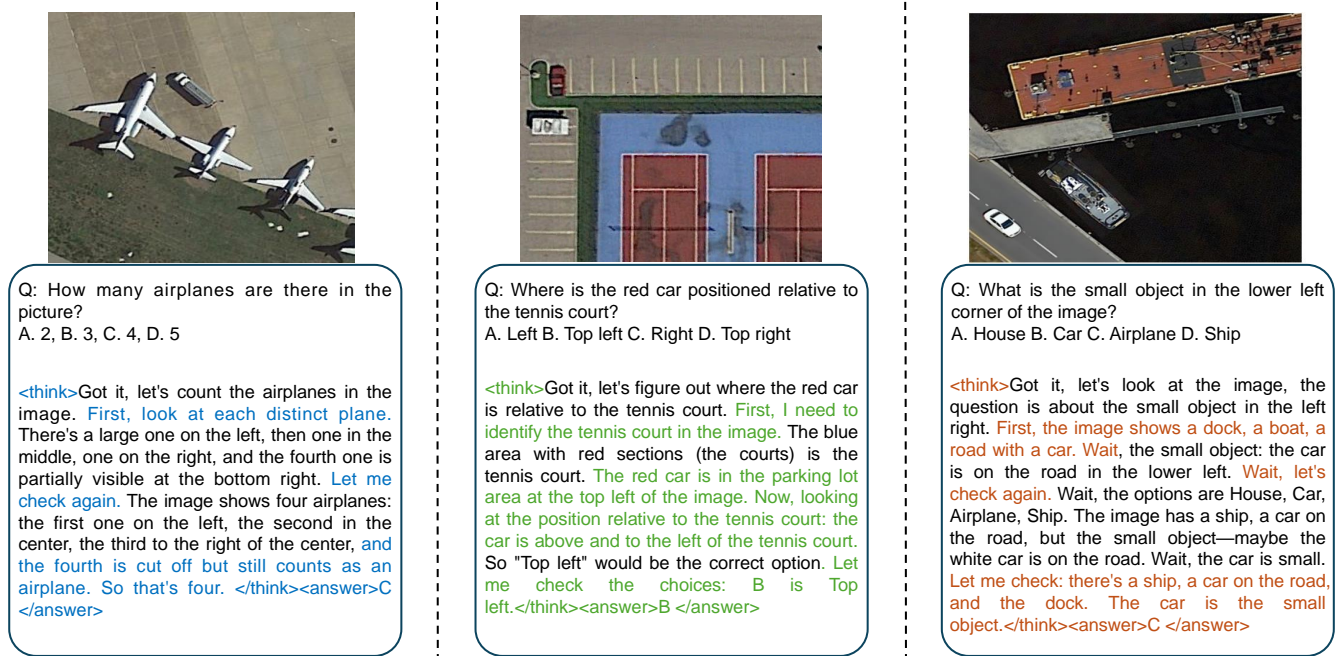


Figure 3: The output examples of VLM-ReG on VRSBench benchmark. VLM-ReG shows “human-like” questioning and self-reflective thought process when solving RS reasoning problems.

algorithm alone yields better results, with an average of 50.86, indicating that GRPO improves overall performance. However, the integration of the proposed components(AAP or RQPO) further enhances the model’s capabilities. The reason for the improvement is that the AAP or RQPO method effectively addresses the shortcomings of GRPO during training.

The addition of Adaptive Advantage Prioritization to the Cold Start stage (CS + AAP) leads to significant performance gains. The average metric score rises to 54.48, surpassing both the CS-only and Original GRPO-only models. Incorporating Refined Quality Rewards (RQPO) into the Cold Start stage (CS + RQPO) also results in performance improvements, with an average score of 55.98. RQPO enhances the model’s ability to distinguish high-quality rewards, as evidenced by better scores in metrics like Bridge (43.88) and VRSBench-val (74.25), which are critical for reward-based evaluation.

The complete VLM-ReG model, which combines CS, AAP, and RQPO, achieves the best performance across all metrics, with an average score of 58.62. It outperforms all ablation variants in key metrics: FAIR (32.14), Bridge (45.11), STAR (35.65), ALL (37.63), MME-RW-RS (60.12), and VRSBench-val (78.11). This demonstrates that the synergistic integration of AAP and RQPO into the Cold Start stage enables VLM-ReG to effectively optimize visual-linguistic reasoning and reward generation, confirming the necessity and efficacy of each component in the proposed framework.

4.4 Visualization

As shown in Figure 3, VLM-ReG demonstrates advanced reasoning capabilities through generating complex reasoning processes. It effectively manifests the "Aha moment" phenomenon[10], reflecting

human-like thought processes characterized by iterative questioning and reflective analysis. This cognitive mechanism directly elevates the system’s reasoning efficacy, resulting in marked progress when addressing advanced reasoning challenges. The emergent problem-solving capacity substantially boosts inferential accuracy across complex analytical scenarios.

5 CONCLUSION

This study aims to address the domain gap that limits multimodal large language models in remote sensing image reasoning and enhance the performance of RS-specific MLLMs (RS-MLLMs). Two core contributions are proposed: first, the Remote Sensing Image Reasoning Dataset (RS-IRD) with 20,000 samples, which generates text-based "Image-CoT" via a modality-bridging pipeline to enable cold-start training of RS-MLLMs without manual annotation, laying a foundation for learning human-like reasoning patterns; second, an enhanced reward design for Group Relative Policy Optimization, which improves reasoning quality and efficiency. Extensive experiments on multiple RS reasoning benchmarks show that the proposed VLM-ReG outperforms state-of-the-art models, verifying the efficacy of the framework. Future work may expand RS-IRD to more RS scenarios or further optimize the reward function for complex reasoning tasks.

ACKNOWLEDGMENTS

The research reported in this paper was supported by a grant from the National Natural Science Foundation of China, and the project ID is 61806212.

REFERENCES

- [1] Anthropic. 2024. Claude 3.5 Sonnet Model Card Addendum. <https://www.anthropic.com> Accessed: 2025-05-01.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] <https://arxiv.org/abs/2502.13923>
- [3] Mattia Balestra, Marina Paolanti, and Roberto Pierdicca. 2025. WHU-RS19 ABZSL: An Attribute-Based Dataset for Remote Sensing Image Understanding. *Remote Sensing* 17, 14 (2025), 2384. <https://doi.org/10.3390/rs17142384>
- [4] Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Mohamed Lamine Mekhalfi, Mansour Abdulaziz Al Zuair, and Farid Melgani. 2022. Bi-Modal Transformer-Based Approach for Visual Question Answering in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–11. <https://doi.org/10.1109/TGRS.2022.3192460>
- [5] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 18009–18019. <https://doi.org/10.1109/CVPR52688.2022.01750>
- [6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. InternVL: Scaling Up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 24185–24198.
- [7] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huaning Li, and Zhongyuan Wang. 2022. NWPU-captions dataset and MLCA-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–19.
- [8] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. TransVG: End-to-End Visual Grounding with Transformers. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Montreal, QC, Canada, 1769–1779.
- [9] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.LG] <https://arxiv.org/abs/2312.11805>
- [10] D. Guo, D. Yang, H. Zhang, et al. 2025. DeepSeek-R1: incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645 (2025), 633–638. <https://doi.org/10.1038/s41586-025-09422-z>
- [11] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. 2022. Align Deep Features for Oriented Object Detection. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–11. <https://doi.org/10.1109/TGRS.2021.3062048>
- [12] Junwei Han, Dingwen Zhang, Gong Cheng, Lei Guo, and Jinchang Ren. 2015. Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning. *IEEE Transactions on Geoscience and Remote Sensing* 53, 6 (2015), 3325–3337. <https://doi.org/10.1109/TGRS.2014.2374218>
- [13] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. 2025. Rsgpt: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* 224 (2025), 272–286.
- [14] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. arXiv:2503.06749 [cs.CV] <https://arxiv.org/abs/2503.06749>
- [15] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. 2024. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, New Orleans, LA, USA, 27831–27840.
- [16] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 27831–27840.
- [17] Haodong Li, Xiaofeng Zhang, and Haicheng Qu. 2025. DDFAV: Remote Sensing Large Vision Language Models Dataset and Evaluation Benchmark. *Remote Sensing* 17, 4 (2025), 719. <https://doi.org/10.3390/rs17040719>
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*. PMLR, PMLR, Honolulu, HI, USA, 19730–19742.
- [19] Xiang Li, Jian Ding, and Mohamed Elhoseiny. 2024. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. arXiv:2406.12384 [cs.CV] <https://arxiv.org/abs/2406.12384>
- [20] Xiang Li, Congcong Wen, Yuan Hu, Zhenghang Yuan, and Xiao Xiang Zhu. 2024. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine* 12, 2 (2024), 32–66.
- [21] Yuheng Li, Chunyuan Li, Haotian Liu, and Yong Jae Lee. 2024. LLaVA-OneVision: Easy Visual Task Transfer. arXiv preprint. arXiv:2408.03326 [cs.CV] <https://arxiv.org/abs/2408.03326>
- [22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. arXiv preprint. arXiv:2401.07120 [cs.CV] <https://arxiv.org/abs/2401.07120>
- [23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning: LLaVA-1.5. arXiv preprint. arXiv:2310.03744 [cs.CV] <https://arxiv.org/abs/2310.03744>
- [25] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. 2020. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing* 58, 12 (2020), 8555–8566.
- [26] Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. 2025. When Large Vision-Language Model Meets Long Remote Sensing Imagery: Coarse-to-Fine Text-Guided Token Pruning. arXiv:2503.07588 [cs.CV] <https://arxiv.org/abs/2503.07588>
- [27] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing* 152 (2019), 166–177.
- [28] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 3195–3204.
- [29] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. MM-Eureka: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning. arXiv:2503.07365 [cs.CV] <https://arxiv.org/abs/2503.07365>
- [30] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. 2025. MM-Eureka: Exploring Visual Aha Moment with Rule-based Large-scale Reinforcement Learning. arXiv:2503.07365 [cs.CV] <https://doi.org/10.48550/arXiv.2503.07365>
- [31] OpenAI, ; Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, and Aidan Clark et al. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] <https://arxiv.org/abs/2410.21276>
- [32] OpenAI. 2024. GPT-4o-mini. OpenAI product documentation. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/> Accessed: 2025-06-25.
- [33] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. LMM-R1: Empowering 3B LMMs with Strong Reasoning Abilities Through Two-Stage Rule-Based RL. arXiv:2503.07536 [cs.CL] <https://arxiv.org/abs/2503.07536>
- [34] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. 2016. Deep Semantic Understanding of High Resolution Remote Sensing Image. In *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, Kunming, China, 1–5. <https://doi.org/10.1109/CITS.2016.7546397>
- [35] Mohamad M. Al Rahhal, Yakoub Bazi, Sara O. Alsaleh, Muna Al-Razgan, Mohamed Lamine Mekhalfi, Mansour Al Zuair, and Naif Alajlan. 2022. Open-ended remote sensing visual question answering with transformers. *International Journal of Remote Sensing* 43, 18 (2022), 6809–6823. <https://doi.org/10.1080/01431161.2022.2145583>
- [36] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, and Jihao Li et al. 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 184 (2022), 116–130. <https://doi.org/10.1016/j.isprsjprs.2021.12.004>
- [37] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning. arXiv:2503.20752 [cs.CV] <https://arxiv.org/abs/2503.20752>
- [38] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. 2025. VL-Rethinker: Incentivizing Self-reflection of Vision-Language Models with Reinforcement Learning. arXiv:2504.08837 <https://arxiv.org/abs/2504.08837>
- [39] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. 2018. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Salt Lake City, UT, USA, 3974–3983.
- [40] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. 2021. Oriented R-CNN for Object Detection. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Montreal, QC, Canada, 3520–3529.
- [41] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. arXiv:2411.10440 [cs.CV] <https://arxiv.org/abs/2411.10440>
- [42] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. 2022. Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

- IEEE, New Orleans, LA, USA, 9499–9508.
- [43] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. 2025. R1-OneVision: Advancing Generalized Multimodal Reasoning through Cross-Modal Formalization. arXiv:2503.10615 <https://arxiv.org/abs/2503.10615>
- [44] Huanjin Yao, Jiaying Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. 2024. Mulberry: Empowering MLLM with o1-like Reasoning and Reflection via Collective Monte Carlo Tree Search. arXiv:2412.18319 [cs.CV] <https://arxiv.org/abs/2412.18319>
- [45] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476 [cs.LG] <https://arxiv.org/abs/2503.14476>
- [46] Hongguo Zhang, Binbin He, and Jin Xing. 2022. Mapping paddy rice in complex landscapes with Landsat time series data and superpixel-based deep learning method. *Remote sensing* 14, 15 (2022), 3721.
- [47] Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization. arXiv:2503.12937 [cs.AI] <https://arxiv.org/abs/2503.12937>
- [48] Meimei Zhang, Fang Chen, and Bin Li. 2023. Multistep question-driven visual question answering for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–12.
- [49] Tong Zhang, Peng Gao, Hao Dong, Yin Zhuang, Guanqun Wang, Wei Zhang, and He Chen. 2022. Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain. *Remote Sensing* 14, 22 (2022), 5675.
- [50] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. 2024. Earth-GPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–20.
- [51] Xiangrong Zhang, Xin Wang, Xu Tang, Huiyu Zhou, and Chen Li. 2019. Description generation for remote sensing images using attribute attention mechanism. *Remote Sensing* 11, 6 (2019), 612.
- [52] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2025. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? arXiv:2408.13257 [cs.CV] <https://arxiv.org/abs/2408.13257>
- [53] Ze Zhang, Enyuan Zhao, Ziyi Wan, Xinyue Liang, Min Ye, Jie Nie, and Lei Huang. 2025. Frequency domain transfer learning for remote sensing visual question answering. *Expert Systems with Applications* 291 (2025), 128395. <https://doi.org/10.1016/j.eswa.2025.128395>
- [54] Bofei Zhao, Haigang Sui, Junyi Liu, Weiyue Shi, Wentao Wang, Chuan Xu, and Jindi Wang. 2024. Flood inundation monitoring using multi-source satellite imagery: A knowledge transfer strategy for heterogeneous image change detection. *Remote Sensing of Environment* 314 (2024), 114373.
- [55] Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. 2022. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14. <https://doi.org/10.1109/TGRS.2021.3079918>
- [56] Xing Zi, Jinghao Xiao, Yunxiao Shi, Xian Tao, Jun Li, Ali Braytee, and Mukesh Prasad. 2025. RSVLM-QA: A Benchmark Dataset for Remote Sensing Vision Language Model-based Question Answering. arXiv:2508.07918 [cs.CV] <https://arxiv.org/abs/2508.07918>