

VISUAL REASONING AGENT: ROBUST VISION SYSTEMS IN REMOTE SENSING VIA INFERENCE-TIME SCALING

Chung-En Johnny Yu
University of West Florida
Pensacola, Florida, USA
cy31@students.uwf.edu

Brian Jalaian
University of West Florida
Pensacola, Florida, USA
bjalaian@uwf.edu

Nathaniel D. Bastian
United States Military Academy
West Point, New York, USA
nathaniel.bastian@westpoint.edu

ABSTRACT

Building robust vision systems for high-stakes domains such as remote sensing requires stronger visual reasoning than what single-pass inference typically provides; yet, retraining large models is often computationally expensive and data intensive. We present **Visual Reasoning Agent (VRA)**, a training-free agentic visual reasoning framework that orchestrates off-the-shelf large vision-language models (LVLMs) with a large reasoning model (LRM) through an iterative *Think-Critique-Act* loop for cross-model verification, self-critique, and recursive refinement. On the remote sensing benchmark VRSBench VQA dataset, VRA consistently outperforms multiple standalone LVLM baselines and achieves up to 40.67% improvement on challenging question types spanning both perception and reasoning tasks. In addition, integrating three LVLMs with VRA improves the overall accuracy of the standalone LVLMs from 52.8% to 78.8%, demonstrating the effectiveness of agentic reasoning with increased inference-time compute.

Keywords Agentic AI, Vision-Language Model, AI Robustness, Visual Reasoning, Remote Sensing

1 Introduction

Artificial intelligence (AI)-enabled vision systems, including pure vision models like object detectors, and large vision-language models (LVLMs), often struggle to generalize reliably in high-stakes domains such as remote sensing, where tasks frequently require both fine-grained perception and multi-step visual reasoning [1]. While fine-tuning can improve performance, it is often computationally expensive, requires extensive labeled data, and offers no guarantee of improved robustness. Consequently, many practitioners lack the resources or domain-specific datasets needed to retrain models for operational deployment.

Recent research has explored improving LVLM factuality through post-hoc correction methods such as Visual Fact Checker [2], LogicCheckGPT [3]. However, these approaches typically rely on rigid, predefined pipelines that are difficult to adapt to diverse visual reasoning tasks. More broadly, current LVLM systems frequently generate single-pass outputs, which can lead to brittle reasoning and hallucinated interpretations when confronted with complex visual scenes. Agentic reasoning frameworks offer a promising alternative. Recent advances in large language model (LLM)-based agentic reasoning, e.g., ReAct [4] and CRITIC [5], demonstrate that iterative reasoning and self-critique can significantly improve robustness and reasoning quality. These approaches allow systems to refine model outputs through multi-step reasoning and tool interaction. However, while agentic methods have been widely studied in text-based tasks, their application to improving robustness across diverse LVLM visual reasoning tasks remains relatively underexplored [6].

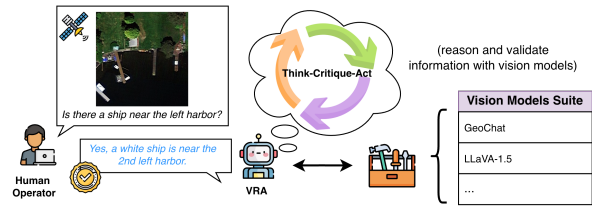


Figure 1: Overview of the Visual Reasoning Agent (VRA) framework. VRA performs iterative *Think-Critique-Act* reasoning with cross-model verification to improve reasoning on remote sensing tasks.

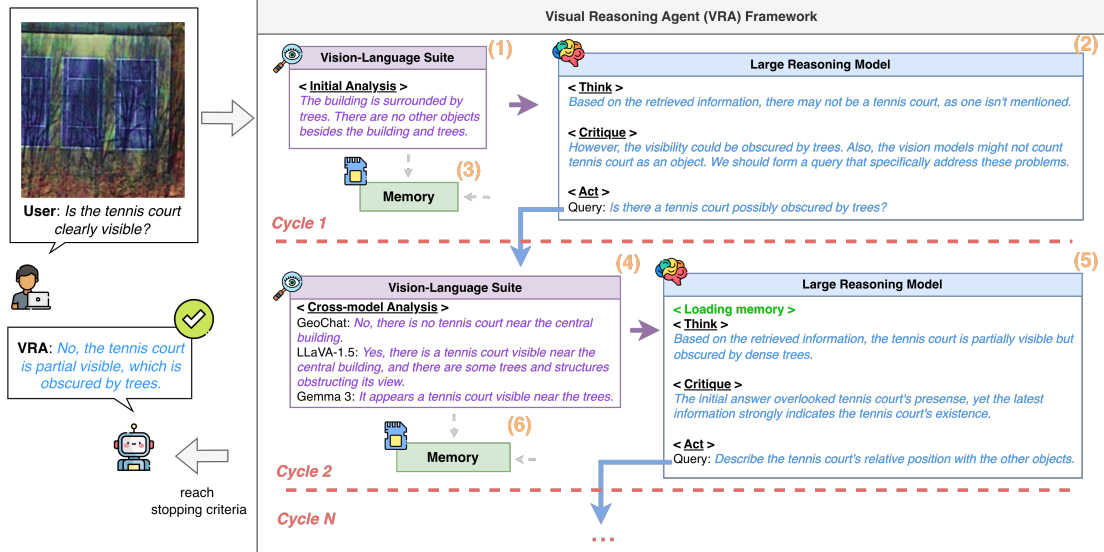


Figure 2: Workflow of Visual Reasoning Agent (VRA) framework. VRA performs iterative visual reasoning via a Think-Critique-Act loop, where heterogeneous LVLMs provide diverse visual evidence, a Large Reasoning Model (LRM), e.g., QwQ, critiques and refines reasoning and generates follow-up queries, and a memory module stores visual evidence and reasoning trajectory across iterations.

To address these challenges, we introduce the **Visual Reasoning Agent (VRA)**, an agentic framework that enhances LVLM reasoning through an iterative *Think-Critique-Act* loop. As illustrated in Figure. 1, VRA orchestrates multiple LVLMs to provide diverse visual evidence while a large reasoning model (LRM) performs structured reasoning, self-critique, and follow-up querying. This design enables cross-model verification and iterative refinement of visual understanding, improving robustness without requiring model retraining.

Remote sensing presents particularly demanding conditions for visual reasoning, including high-resolution imagery, dense small objects, long-range spatial dependencies, and frequent domain shifts across sensors and geographic regions. Recent LVLMs for remote sensing, such as RSGPT [7] and GeoChat [8], demonstrate promising capabilities but still rely primarily on single-pass inference. Benchmarks such as VRSBench [9], XLRS-Bench [10], and SkySense-O [11] further highlight persistent challenges in fine-grained perception and multi-step reasoning for satellite imagery understanding. Our proposed framework addresses these limitations by enabling iterative reasoning with cross-model validation, producing more reliable outputs than conventional single-pass LVLM inference. In addition, VRA records reasoning traces and visual evidence throughout the reasoning process, providing transparent and auditable decision flows for reliable AI deployment.

Our key contributions are summarized as follows:

1. We introduce **Visual Reasoning Agent (VRA)**, an iterative *Think-Critique-Act* framework that coordinates multiple LVLMs and a reasoning model to improve visual reasoning robustness.
2. VRA integrates heterogeneous LVLMs to provide complementary visual evidence, enabling cross-model validation that reduces hallucination and model-specific bias.
3. Experiments on the VRSBench VQA benchmark demonstrate up to 40% accuracy improvements across challenging remote sensing reasoning tasks.

2 Visual Reasoning Agent

We propose the Visual Reasoning Agent (VRA) framework to improve the robustness of large vision-language models (LVLMs) across diverse visual tasks, encompassing both perception and reasoning-intensive visual tasks. VRA adopts an iterative *Think-Critique-Act* loop, inspired by ReAct [4] and Reflexion [12], enabling structured self-refinement and cross-model validation at inference time. As illustrated in Figure 2, the framework consists of three core components: (i) a Vision-Language Suite, (ii) a Large Reasoning Model (LRM), and (iii) a memory module.

2.1 Framework Components

Vision-Language Suite. The Vision-Language Suite comprises a set of heterogeneous LVLMs that independently analyze the input image and provide diverse natural-language descriptions. By aggregating multiple model perspectives, the system mitigates reliance on any single LVLM and reduces model-specific bias or hallucination. **Large Reasoning Model (LRM).** The LRM functions as the central agent responsible for reasoning over the collective visual evidence. It is selected for strong agentic reasoning capability and structured multi-step reasoning outputs. Beyond answer generation, the LRM performs self-critique to detect inconsistencies across model responses and contradictions within its own reasoning trajectory. For example, we utilized QwQ [13] in our experiments. **Memory Module.** The memory module logs the visual observations and reasoning traces across iterations. This module serves three purposes: (1) maintaining logical consistency across cycles, (2) preventing redundant queries or reasoning steps, and (3) providing auditable reasoning traces for post-hoc verification by human operators.

2.2 VRA Workflow

As illustrated in Figure 2, the workflow proceeds iteratively: **(1) Initial Visual Context.** An LVLM generates an initial image analysis, providing foundational visual context and scene interpretation. **(2) Think-Critique-Act (Cycle 1).** The LRM performs three structured operations: *Think*: Reasons out an initial answer based on the current visual information. *Critique*: Identifies potential inconsistencies in visual information and contradictions in its reasoning. *Act*: Generates a targeted follow-up query to retrieve additional visual evidence from the Vision-Language Suite. **(3) Memory Update.** The visual information generated by the LVLM and the Think-Critique-Act outputs by the LRM are stored in the memory module. The memory is used for the LRM reasoning process in Think-Critique-Act loop in the subsequent cycles. **(4) Cross-Model Analysis (Cycle ≥ 2).** The heterogeneous LVLMs respond to the LRM-generated query. This step performs cross-model validation, capturing diverse interpretations of the same visual evidence. **(5) Iterative Think-Critique-Act with Memory Loading.** The LRM retrieves the entire reasoning trajectory from memory module and performs another Think-Critique-Act cycle based on the updated visual evidence. By conditioning on the complete reasoning trajectory, the agent enforces consistency and refines its answer across iterations. **(6) Memory Update and Continuation.** The latest visual information and reasoning trace are stored in the memory module. The iterative process continues until a stopping criterion is met, such as reaching a maximum number of iterations.

VRA is motivated by the hypothesis that increasing inference-time compute through structured agentic reasoning improves robustness for vision systems. Rather than relying on single-pass inference from one LVLM, VRA performs: cross-model verification, structured self-critique, iterative evidence refinement. This design is particularly beneficial for complex visual reasoning tasks, where superficial single-model inference is insufficient. Although computational cost increases, the trade-off is justified in high-stakes applications such as remote sensing, surveillance analysis, and safety-critical decision support, where reliability and robustness outweigh throughput.

The detailed procedure of VRA is summarized in Algorithm 1. The detailed prompt templates used in VRA are provided in Appendix *Prompt Templates*.

Algorithm 1: Visual Reasoning Agent (VRA)

Require: Image I , question Q , Vision-Language Suite $\mathcal{V} = \{V_1, \dots, V_M\}$, LRM R , memory $\mathcal{M}$, max iterations K

Ensure: Final answer $\hat{a}$ and trace $\mathcal{M}$

```

1:  $\mathcal{M} \leftarrow \emptyset$ 
2: Initial visual context:  $o^{(0)} \leftarrow V_1(I)$ 
3:  $\mathcal{M} \leftarrow \mathcal{M} \cup \{o^{(0)}\}$ 
4: for  $t = 1$  to  $K$  do
5:   Think-Critique-Act:  $(\hat{a}^{(t)}, c^{(t)}, q^{(t)}) \leftarrow R(Q, \mathcal{M})$ 
      $\triangleright$  Reason out answer, critique, and follow-up query
6:    $\mathcal{M} \leftarrow \mathcal{M} \cup \{\hat{a}^{(t)}, c^{(t)}, q^{(t)}\}$ 
7:   Cross-model visual evidence:  $O^{(t)} \leftarrow \{V_m(I, q^{(t)})\}_{m=1}^M$ 
8:    $\mathcal{M} \leftarrow \mathcal{M} \cup \{O^{(t)}\}$ 
9: end for
10:  $\hat{a} \leftarrow \hat{a}^{(t)}$ 
11: return  $\hat{a}, \mathcal{M}$ 

```

3 Preliminary Experiments and Results

3.1 Experiment Setup

To validate the effectiveness of VRA, we conducted preliminary experiments on visual reasoning tasks in the remote sensing domain, a high-stakes application area where robustness is critical for applications like disaster response and surveillance. **Dataset.** We evaluated VRA on the VRSBench VQA dataset [9], which offers comprehensive satellite imagery understanding tasks that evaluate the capabilities of both fine-grained perception and complex visual reasoning. The question types in VRSBench are meticulously categorized to cover both low-level visual attributes—such as object category, count, color, and shape—and high-level visual reasoning, including spatial relationships and scene-level understanding. This hierarchical structure allows for a robust assessment of a model’s ability to transition from basic feature identification to complex relational logic in a geo-spatial context. We sampled 50 questions for each of the ten question types, resulting in 500 questions in total. Performance is measured by accuracy per question type and overall accuracy is computed as the average across all evaluated question types. **Models.** To demonstrate VRA’s adaptability and robustness, we integrated it with multiple established LVLMs as plug-in components. These baseline LVLMs include: GeoChat (7B) [8], LLaVA-v1.5 (7B) [14], and Gemma 3 (12B) [15]. Note that GeoChat is a specialized RS-LVLM, fine-tuned LLaVA-v1.5 on remote sensing datasets. The LRM is QwQ (32B) [13], a text-based reasoning model that enables agentic operations. Also, Phi-4 [16] is used in VRSBench VQA dataset for evaluating free-form answers. The prompt template used for evaluation is provided in Appendix *Prompt Templates*. **VRA Variants.** VRA is designed to be modular, allowing for the integration of single or multiple off-the-shelf LVLMs as visual information providers. We define the following VRA variants for comparison: *VRA (LVLM_NAME)*, which denotes VRA with only the specified LVLM as the visual information source, e.g., GeoChat, LLaVA-1.5, Gemma 3. *VRA (+1 LVLM avg)* represents the average of VRA (GeoChat), VRA (LLaVA-1.5), and VRA (Gemma 3), which is used in Figure 5 and 4. *VRA (+2 LVLMs)* utilizes GeoChat and LLaVA-1.5 in VRA. *VRA (+3 LVLMs)* integrates all three baseline LVLMs (GeoChat, LLaVA-1.5, and Gemma 3) concurrently. Finally, *LVLM alone (avg)* in the figures refers to the average performance of the three baseline LVLMs (GeoChat, LLaVA-1.5, and Gemma 3) without VRA.

3.2 Results Analysis

Overall Accuracy Improvement. As illustrated in Figure 3, VRA consistently improves the overall accuracy across all evaluated LVLMs. GeoChat improves from 46.40% to 66.80% (+20.40%), LLaVA-1.5 increases from 57.60% to 71.80% (+14.20%), and Gemma 3 rises from 54.40% to 64.60% (+10.20%). These consistent gains demonstrate that the proposed agentic reasoning framework effectively enhances the reliability of LVLM predictions. By iteratively refining reasoning through the Think-Critique-Act loop and incorporating cross-model visual evidence, VRA mitigates errors produced by single-pass LVLM inference.

Enhanced Performance Across Question Types. Figure 5 and Table 1 further analyze performance across ten visual reasoning categories. VRA consistently outperforms the baseline LVLM average across most question types, particularly in tasks requiring precise spatial understanding and reasoning. The most significant improvement is observed in *object direction* question type, where accuracy increases from 39.33% to 80.00% (+40.67%) when using three LVLMs. Substantial gains are also observed in *object quantity*, where accuracy increases from 16.67% to 56.00% (+39.33%) and *scene type*, where increases from 65.33% to 98.00% (+32.67%). These improvements suggest that cross-model verification and iterative reasoning are particularly beneficial for tasks requiring spatial interpretation and quantity understanding. Performance improvements are also consistent in reasoning-heavy categories such as *object position*, *object size*, and *reasoning*, indicating that VRA enhances both perception-level and reasoning-level visual question answering capabilities. Overall accuracy increases from 52.80% for standalone LVLMs to 78.80% (+26%) when VRA integrates three LVLMs.

Impact of Increasing Multiple LVLMs. Increasing the number of LVLMs in VRA further improves performance. As shown in Table 1, overall accuracy increases from 67.73% (+1 LVLM) to 75.60% (+2 LVLMs) and 78.80% (+3 LVLMs) with three LVLMs. This trend highlights the benefit of cross-model analysis, where diverse visual interpretations provide complementary information that reduces individual model biases and hallucinations.

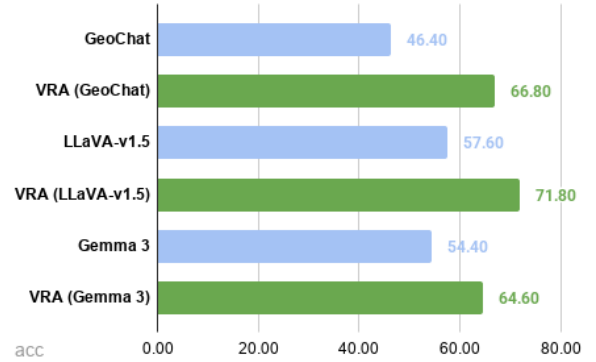


Figure 3: Overall accuracy comparison between standalone LVLMs and VRA-augmented variants.

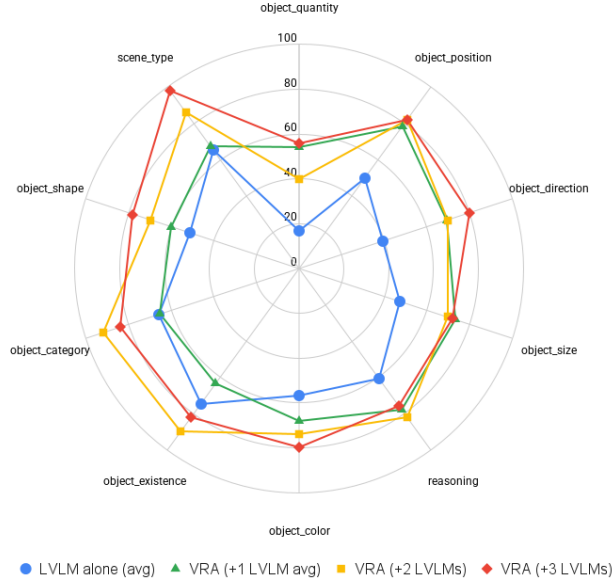


Figure 5: Performance of averaged standalone LVLms versus VRA-augmented variants across different question types. Refer to Section 3.1 for the definitions of VRA variants.

Table 1: Performance comparison between average standalone LVLms and VRA variants. The metric is accuracy (%)

Model	obj_quantity	obj_position	obj_direction	obj_size	reasoning	obj_color	obj_existence	obj_category	obj_shape	scene_type	overall
LVLm alone (avg)	16.67	50.00	39.33	47.33	60.67	56.67	74.67	66.00	51.33	65.33	52.80
VRA (+1 LVLm avg)	54.00	78.67	69.33	73.33	78.00	68.00	63.33	65.33	60.00	67.33	67.73
VRA (+2 LVLms)	40.00	82.00	70.00	70.00	82.00	74.00	90.00	92.00	70.00	86.00	75.60
VRA (+3 LVLms)	56.00	82.00	80.00	72.00	76.00	80.00	82.00	84.00	78.00	98.00	78.80

Runtime Analysis

Figure 4 presents the runtime cost of the proposed framework. While standalone LVLm inference requires only 0.03 minutes per sample on average, incorporating the agentic reasoning loop increases the runtime to approximately 3.12 minutes with one LVLm and 3.78 minutes with three LVLms. The increase arises from iterative reasoning cycles and multiple LVLm queries required for cross-model validation. Although VRA introduces higher inference-time cost, the trade-off is justified in applications where robustness and reliability are critical, such as remote sensing and safety-critical decision support.

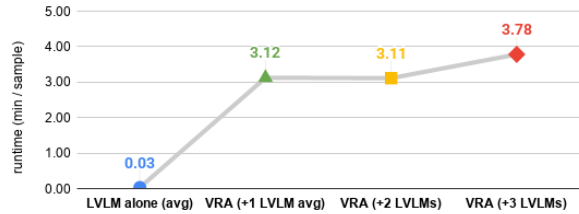


Figure 4: Overall runtime (minute/sample) comparison between averaged standalone LVLms and VRA-augmented variants. Refer to Section 3.1 for VRA variants.

4 Discussion

Our preliminary findings from VRA provide compelling evidence that agentic reasoning significantly enhances the general robustness of LVLms across diverse visual tasks. The substantial accuracy improvements in remote sensing for both perceptual and reasoning tasks, coupled with VRA’s architectural components, underscore the potential for more reliable LVLm deployment in safety-critical domains.

4.1 Iterative Agentic Reasoning and Diverse Models

Unlike static inference, the *Think-Critique-Act* loop allows VRA to recursively refine its understanding and outputs, which enables it to navigate complex visual information and resolve inconsistencies. This dynamic adaptation is crucial for high accuracy across visual reasoning challenges, as evidenced by VRA’s consistent outperformance of standalone LVLms. The robustness gained is from iterative self-correction.

By leveraging multiple heterogeneous LVLMs, VRA inherently mitigates risks associated with reliance on a single model. Each LVLM in VRA offers a distinct visual encoder and training bias, enabling cross-model validation by the LRM. This enhances factual consistency and reducing susceptibility to individual LVLM failures. Moreover, the LRM plays a pivotal role by contributing its extensive reasoning capabilities. VRA leverages its powerful “thinking” to synthesize information from disparate sources, critiques, and formulate necessary follow-up questions. This enables effective visual reasoning even with basic initial prompts, highlighting that robustness stems from the agentic workflow, the underlying reasoning model, and effective orchestration of diverse vision models, also contributing to make the agent’s decision-making process more transparent and auditable.

4.2 Limitations and Future Directions

Our VRA framework is not without limitations, primarily related to computational overhead. As shown in the runtime analysis in Figure 4, leveraging a long-thinking process from a LRM and increasing the number of integrated LVLMs leads to a substantial increase in inference time, which could be a barrier for real-time applications.

Future work will focus on: (1) Investigating more efficient agentic frameworks that retain robustness while significantly reducing computational overhead, e.g., dynamic agent routing techniques or early stopping criterion when VRA reaches high confidence. (2) Evaluating VRA on hallucination benchmarks, e.g., Throne [17] for intrinsic errors assurance. (3) Rigorously validating VRA’s adversarial robustness against a wider array of state-of-the-art adversarial attacks, exploring agentic defense mechanisms. (4) Further evaluating VRA’s visual capabilities on diverse safety-critical domains to assess VRA’s generalizability for real-world reliability.

5 Conclusion

Robust vision systems are essential for high-stakes domains such as remote sensing. We have introduced **Visual Reasoning Agent (VRA)**, a training-free agentic reasoning framework that wraps off-the-shelf LVLMs in a *Think-Critique-Act* loop. VRA’s training-free design democratizes access to robust vision deployments, even for teams lacking fine-tuning resources, such as labeled data and exhaustive compute. Moreover, its cross-model validation and iterative self-correction principles point toward promising robust visual reasoning capabilities. Empirical results on the remote sensing benchmark VRSBench VQA dataset demonstrate up to 40% accuracy improvements, confirming that increased inference-time compute via agentic reasoning is a justifiable trade-off for reliability. Future work will focus on dynamic agent routing and confidence-based early stopping to reduce inference overhead.

Acknowledgment

This work was supported in part by the U.S. Military Academy (USMA) under Cooperative Agreement No. W911NF-23-2-0108. The views and conclusions expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Military Academy, U.S. Army, U.S. Department of War, or U.S. Government.

References

- [1] Zijiang Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodal large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024.
- [2] Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. Visual fact checker: enabling high-fidelity detailed caption generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14033–14042, 2024.
- [3] Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*, 2024.
- [4] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.
- [5] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.
- [6] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

- [7] Zehan Wu, Zexin Hu, Xinyu Hou, Yuxuan Xu, Zhenghao Wang, Yuxuan Yang, Fan Zhang, Lingbo Li, Yilei Tong, Jiaming Wu, Li Zhao, Qingni Li, Xiao Huang, Liang Peng, Guofan Chen, Kai Jiang, Yumin Deng, Junwei Liu, Xiaoyu Wang, Jijun Li, Xin Liu, Liyue Song, Lei Zhang, Weida Wang, Yiyuan Wang, Kewei Zhang, Yahui Liu, Jing Xu, Jiangeng Cao, Qing Liu, Biao Zhao, Jiaqi Zhang, Zhiheng Liu, Yitian Li, You Zhang, Jintao Xu, Wei Zhou, Jun Zhu, Ji Zhang, Xue Fan, Sicong Wang, Xi Jin, Xiao Li, Jiankun Cui, Dayan Wang, Congwei Liu, Chunliang Fang, Chao Ma, Deng Zhang, Jiale Wang, Chenshuo Wu, Jiahao Chen, Yong Zhao, Hang Zhang, Pan He, Ya Yao, Qiang Liu, Chao Dong, and Kun Fu. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023.
- [8] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024.
- [9] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12384*, 2024.
- [10] Siyuan Liu, Weikang Wang, Jingting Han, Xiangyun Jia, Hang Yang, Ge Lin, Hao Ma, Yidong Wang, Yiqun Wang, Jianwei Yin, Qing Gao, Zhen Wang, Yanfeng Huang, Weidong Yang, Xudong Zhu, Yansong Wang, Yuxuan Kang, Lei Song, Fan Zhang, Cunjian Yuan, Qiang Wang, Jie Feng, Yanning Zhao, Dacheng Tao, Tong Lu, and Yuan Gao. Xlrs-bench: Evaluating large vision-language models on heterogeneous extreme-resolution remote sensing imagery. *arXiv preprint arXiv:2501.13441*, 2025.
- [11] Siyuan Liu, Yuan Gao, Weikang Wang, Xudong Zhu, Yanfeng Huang, Yansong Wang, Lei Song, Xin Deng, Xiangyun Jia, Fan Zhang, Cunjian Yuan, Qiang Wang, Jie Feng, Tong Lu, Dacheng Tao, and Yanning Zhao. Skysense-o: Towards open-world remote sensing understanding with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [12] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- [13] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025.
- [14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [15] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- [16] Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025.
- [17] Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27228–27238, 2024.

6 Appendix

6.1 Comprehensive Results Details

The following tables present the comprehensive evaluation results on the VRSBench VQA subset, consisting of 50 questions per type and overall performance of all question types. Table 2 reports the accuracy results (%) across different question types, comparing various baseline LVLMs and their corresponding VRA-augmented variants. Table 3 summarizes the runtime statistics, illustrating the computational costs associated with each model configuration. These results provide a holistic view of the performance trade-offs between accuracy and efficiency across diverse vision tasks.

Table 2: VRSBench VQA Subset (50 questions per type) Accuracy in %

Model	obj_quantity	obj_position	obj_direction	obj_size	reasoning	obj_color	obj_existence	obj_category	obj_shape	scene_type	overall
GeoChat	10.00	68.00	34.00	22.00	60.00	24.00	84.00	76.00	34.00	52.00	46.40
VRA (GeoChat)	52.00	76.00	70.00	74.00	86.00	50.00	78.00	70.00	62.00	50.00	66.80
LLaVA-1.5	30.00	42.00	36.00	60.00	56.00	68.00	78.00	74.00	60.00	72.00	57.60
VRA (LLaVA-1.5)	48.00	88.00	72.00	72.00	72.00	80.00	70.00	74.00	62.00	80.00	71.80
Gemma 3	10.00	40.00	48.00	60.00	66.00	78.00	62.00	48.00	60.00	72.00	54.40
VRA (Gemma 3)	62.00	72.00	66.00	74.00	76.00	74.00	42.00	52.00	56.00	72.00	64.60
VRA (+2 LVLMS)	40.00	82.00	70.00	70.00	82.00	74.00	90.00	92.00	70.00	86.00	75.60
VRA (+3 LVLMS)	56.00	82.00	80.00	72.00	76.00	80.00	82.00	84.00	78.00	98.00	78.80

Table 3: VRSBench VQA Subset (50 questions per type) Runtime in minutes

Model	obj_quantity	obj_position	obj_direction	obj_size	reasoning	obj_color	obj_existence	obj_category	obj_shape	scene_type	overall
GeoChat	0.79	0.83	0.79	0.81	1.05	0.84	0.81	0.91	0.78	1.24	0.89
VRA (GeoChat)	135.26	145.62	144.87	133.42	142.09	144.86	143.71	141.49	142.27	134.08	140.77
LLaVA-1.5	0.56	0.71	0.84	0.80	1.08	0.68	0.74	0.79	0.73	0.93	0.79
VRA (LLaVA-1.5)	149.32	155.86	155.37	157.06	142.53	139.65	143.76	145.92	138.23	143.66	147.14
Gemma 3	2.29	2.63	2.64	4.46	3.28	1.72	2.40	3.08	2.46	3.92	2.89
VRA (Gemma 3)	170.06	163.17	183.04	189.95	185.91	194.74	174.01	179.46	179.34	186.15	180.58
VRA (+2 LVLMS)	156.53	166.21	155.29	160.79	152.47	159.94	154.45	160.59	140.46	149.14	155.59
VRA (+3 LVLMS)	190.03	196.92	195.75	212.28	197.24	184.62	174.63	197.11	164.74	177.57	189.09

6.2 Prompt Templates

Table 4: Prompt template for VRA – LRM performs Think-Critique-Act process in Cycle 1.

System Prompt

You are expert in remote sensing and geospatial image analysis.

Task:

1. Provide a ~50 word answer to the user’s question based on the conversation.
2. Reflect and critique your answer.
3. Provide one question to ask vision model for retrieving more visual information. Your question should be straightforward and relevant to the answer and user question.

User Prompt

Reflect on the user’s original question and the actions taken thus far.

Table 5: Prompt template for VRA – LVLMs in the Vision-Language Suite.

System Prompt

(No system prompt)

User Prompt

<image> {LRM_generated_question}

Table 6: Prompt template for for VRA – LRM performs Think-Critique-Act process in Cycle ≥ 2 .

System Prompt

You are expert in remote sensing and geospatial image analysis. In the preceding messages, you will find multiple tools' outputs providing visual information.

{history}

Task:

Revise your previous answer using the new visual information provided by multiple tools' outputs.

- You should use the previous critique to add important information to your answer.
 - You MUST include numerical citations in your revised answer to ensure it can be verified.
 - Add a "References" section to the bottom of your answer (which does not count towards the word limit). In form of:
 - [1] visual information here
 - [2] visual information here
 - More visual information here if there is any...
- You should use the previous critique to remove superfluous information from your answer and make SURE it is not more than 50 words.
- You should provide one question to ask vision model for retrieving more visual information. Your question should be straightforward without repeating previous questions.

User Prompt

Reflect on the user's original question and the actions taken thus far.

Table 7: Prompt template for the evaluation model used in VRSBench VQA Dataset.

System Prompt

(No system prompt)

User Prompt

Question: {question}

Ground Truth Answer: {ground_truth}

Predicted Answer: {prediction}

Does the predicted answer match the ground truth? Answer 1 for match and 0 for not match.

Use semantic meaning not exact match. Synonyms are also treated as a match, e.g., football and soccer, playground and ground track field, building and rooftop, pond and swimming pool.

Do not explain the reason.