Leveraging Vision-Language Models for Remote Sensing VQA: Analysis, Gaps, and Directions

□ **Maryam Qamar, Sung-Ho Bae** / Kyung Hee University

Abstract

Vision-language models like LLaVA have propelled visual reasoning to new heights on natural image datasets, yet the domain of remote sensing (RS) poses unique challenges that often undermine their performance. In this work, we critically evaluate LLaVA and its remote-sensing-tuned variant, RS-LLaVA, on a comprehensive remote sensing visual question answer benchmark comprising presence, counting, comparison, and rural/urban classification queries. Our detailed analysis uncovers both unexpected strengths of vanilla LLaVA in certain RS scenarios and persistent failure modes that persist even after fine-tuning. We highlight the domain-specific characteristics of RS data that break assumptions intrinsic to mainstream VLM design, quantify category and error-type gaps, and propose future directions for robust multidisciplinary VLMs.

I . Introduction

Vision-language models (VLMs) have achieved remarkable milestones by connecting deep vision backbones with large language models, yielding state-of-the-art results on a variety of visual reasoning tasks[1]. The pretraining datasets that power these models, however, consist predominantly of natural images, rich in common objects, everyday scenarios, and culturally familiar scenes.

Remote sensing (RS) imagery presents a fundamentally different regime. Captured from satellites or aerial platforms, these data are characterized by wide-area coverage, extreme variation in object scale, high object density in urban environments, unique spectral signatures, and

forms of contextual ambiguity absent from natural photography. Such differences mean that VLMs tuned on natural image distributions face a significant domain shift when applied to RS tasks. A pressing research question therefore arises: can generic VLMs suffice for RS applications, or do the unique demands of Earth observation necessitate domain-specific fine-tuning, dataset design, and architectural adaptation?

The answer has important implications. Remote sensing plays a central role in climate monitoring, urban growth analysis, disaster response, and security applications-domains where reliable automated interpretation is crucial. Unlike natural image tasks, RS queries often demand fine-grained spatial reasoning (e.g., counting thousands of small objects across wide scenes), sensitivity to subtle semantic cues (e.g., distinguishing cropland from grassland), and robustness against class imbalance where rare but critical categories, such as runways or military installations, may appear only sporadically. These requirements call for VLMs that are not only linguistically capable but also tailored to the geospatial context.

In this study, we benchmark the base LLaVA[2] model and its remote sensing fine-tuned variant, RS-LLaVA[3], dissecting their strengths and limitations across diverse question types. Our analysis offers critical insights into where generic VLMs transfer successfully, where fine-tuning yields the most substantial gains, and where persistent shortcomings remain. Beyond performance reporting, we map the reasons behind successes and failures, examining issues of scale, scene composition, semantic ambiguity, and confidence calibration. In doing so, we aim to inform both practitioners deploying current models for Earth observation and researchers designing the next generation of robust, domain adapted VLMs for geospatial intelligence.

Ⅱ. Related Work

The popularity of VLMs such as CLIP[4], Flamingo[5], BLIP[6], and Large Language and Vision Assistant (LLaVA)[2] stems from their ability to couple data rich vision backbones with generative LLMs, enabling open ended visual question answering (VQA) across a spectrum of natural image tasks. These models exploit coarse and fine image-language alignments, developing powerful representations over millions of internet images where objects are salient, well-captioned, and relate naturally to language. Tasks emphasize common-sense, spatial reasoning, and object recognition.

In detail, LLaVA[2] explored visual instruction tuning to make large multimodal models learn enhanced reasoning and comprehension capabilities by integrating a pre-trained vision encoder with a large language model (LLM) and fine-tuning end-to-end on multimodal instruction-following data. Authors synthesized GPT-4 based diverse instruction-response pairs that pertain to visual content, spatial relations, and complex reasoning from existing image-caption and annotation datasets, enabling the model to learn to follow human instructions on image understanding and reasoning tasks. The instruction tuning process occurs in two stages. The first stage aligns visual features from the vision encoder to the language embedding space via a lightweight linear projection, allowing efficient adaptation without modifying the original encoders. The second

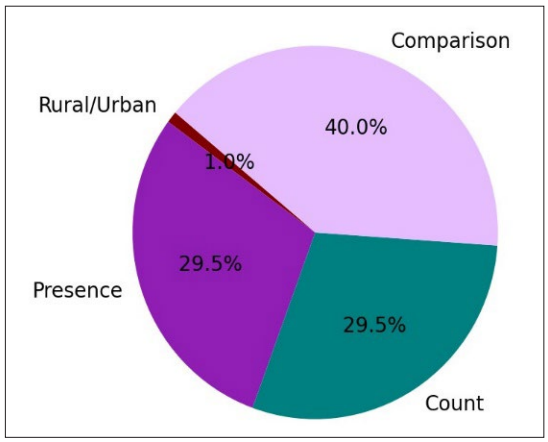
stage fine-tunes the full multimodal model on the generated instruction-following dataset containing conversational Q&A, detailed descriptions, and complex reasoning examples. This comprehensive tuning significantly improves the model's zero-shot and few-shot ability to interpret and verbally respond to image-related queries.

However, the training dataset is primarily composed of natural images, and remote sensing data diverges from natural images in crucial ways, for instance, they show vastly different spatial scale, images may cover kilometers, with objects ranging from sub-meter (cars) to hundreds of meters (fields). They are also characterized by high density images, Urban scenes have thousands of objects, often overlapping and distributed non-uniformly. The semantic cues, for example, texture, layout, spectral reflectance etc. defining overall context differ drastically from those in natural photography. Thus, the domain gap may limit LLaVA's effectiveness on remote sensing-specific tasks.

To address this domain gap between natural images and remote sensing (RS) data, recent work introduced RS-LLaVA, an adaptation of the LLaVA model specifically designed for RS imagery. RS-LLaVA leverages low-rank adaptation (LoRA)[7] fine-tuning on a multi-task RS vision-language instruction dataset constructed from heterogeneous captioning and visual question answering (VQA) datasets, tailored to the nuanced challenges of RS image analysis. The model architecture keeps the image encoder and language decoder frozen while fine-tuning the projection network and LLM weights through LoRA, enabling efficient adaptation despite the large parameter sizes.

The RS-instructions dataset integrates four domain-specific datasets, UCM-captions[8], UAV[9], RSVQA-LR[10], and RSIVQA-DOTA[11], converted into instructional formats suited for joint captioning and VQA tasks. This multi-task instruction tuning significantly improves generalization to RS-specific queries, outperforming previous state-of-the-art RS captioning and VQA models.

In this study, we systematically compare off-the-shelf LLaVA against its RS-finetuned counterpart RS-LLaVA, with breakdowns by question category and error type analysis, highlighting the effectiveness of base LLaVA in certain scenarios and also showing where RS-LLaVA still lacks. Category distributions pie chart in Figure 1 reflects the percentages of questions in each category which can be meaningful to understand the overall results given the deeper category-wise success and failure modes.



<Figure 1> Distribution of the dataset's questions across the four main categories.

Ⅲ. Experiments, Results and Analysis

1. Datasets and Question Types

<Table 1> Questions' Categories and corresponding examples in the dataset.

Category	Example Question	Example Answer
Presence	Is there a grass area?	Yes
Count	What is the number of roads?	13
Comparison	Are there more residential buildings than water areas?	No
Rural/Urban	Is it a rural or an urban area?	Rural

<Table 2> Overall and Category-wise Accuracy (%) of LLaVA vs. RS-LLaVA.

Model	Overall	Presence	Comparison	Counting	Rural/Urban
LLaVA	46.5	73.1	62.3	0	0
RS-LLaVA	67.8	90.1	79.9	24.7	86

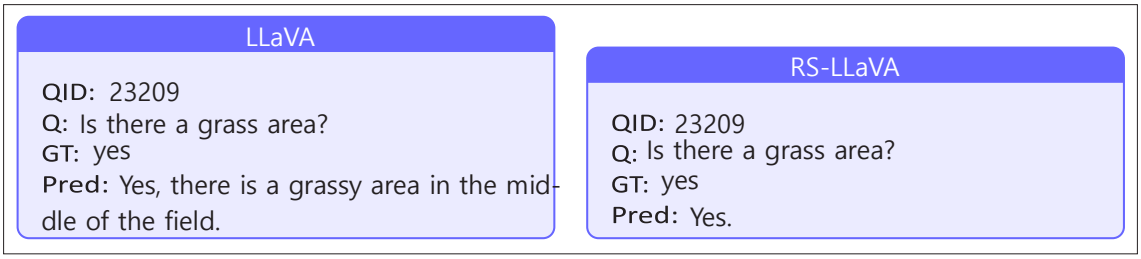
We evaluate both LLaVA[2] and RS-LLaVA[3] on a dataset of 10,004 remote sensing questions[10]. Question categories and some examples are shown in Table 1.

2. Results and Analysis

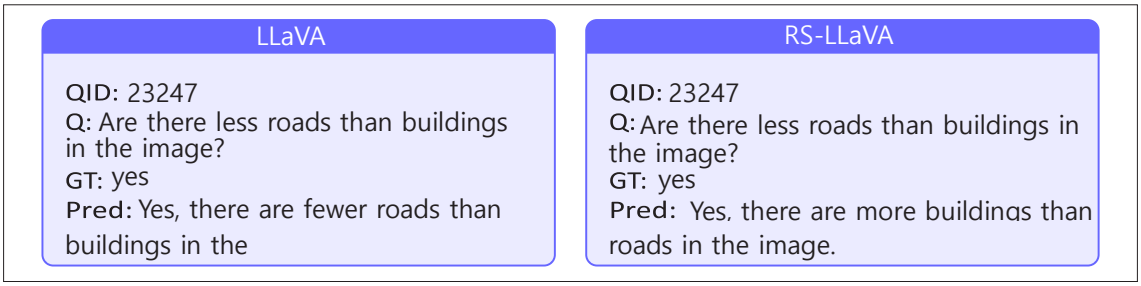
Overall RS-LLaVA delivers dramatic gains in counting, comparisons, and rural/urban tasks as observed in Table 2. Next, we delve into the details of success and failure modes of both models.

3. Success and Failure Modes

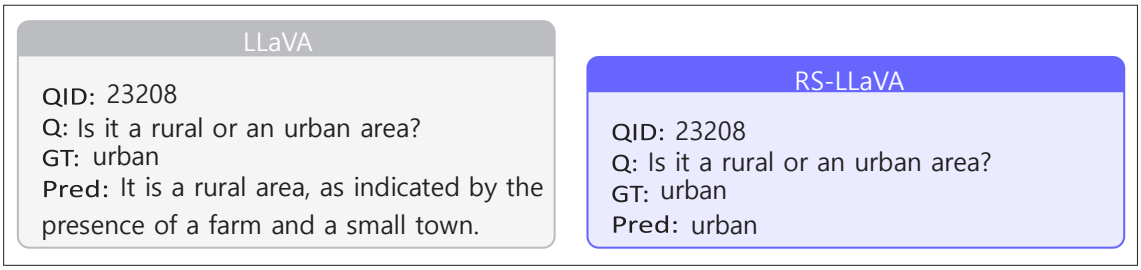
While LLaVA show weaknesses in the raw metrics comparison, favorable patterns can still be observed



Presence in Familiar Contexts



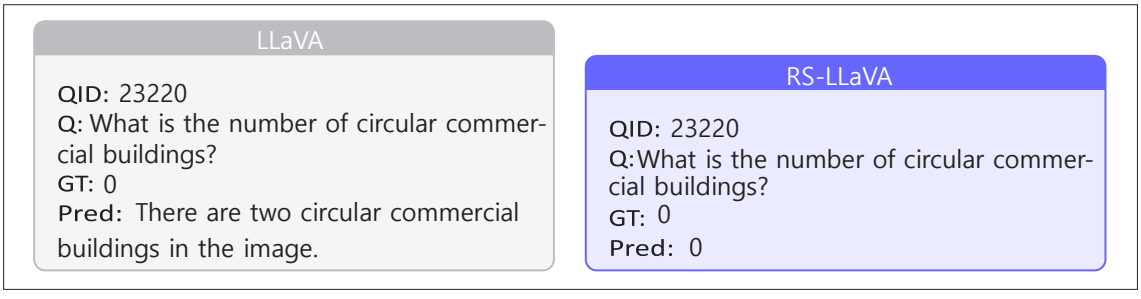
Comparison: Relative Quantity without Precise Counting



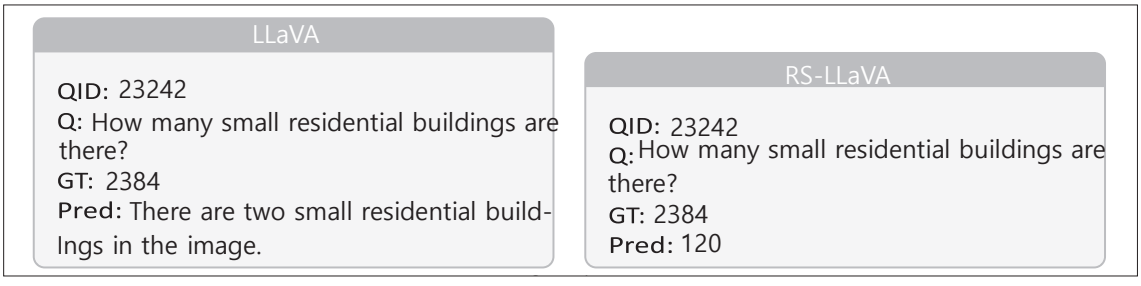
Rural/Urban Contrast

when looking at the categories in detail. LLaVA reliably detects highly common objects and textures, for instance, grass, roads, water areas, farmland in remote sensing images, even without fine-tuning. This could be credited to either learning of common textures from natural images or inclusion of general aerial imagery through datasets like LAION, during pre-training.

Moreover, LLaVA captures directionality ("more", "less") decently when visual contrast is strong, despite failing at exact counts. However, LLaVA completely fails all Rural/Urban and counting questions, by either refusing to answer ("cannot determine"), confidently guessing wrong or outputting irrelevant inflated counts to tricky queries where the object in question actually does not exist. Very likely reasons for such response are crowded repetitive structures confusing the rural versus urban understanding and remote sensing object scales being drastically different from natural image objects making counting, an already



Counting



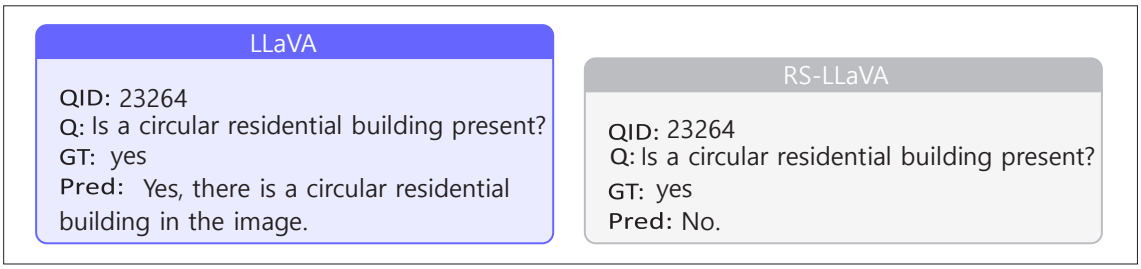
Counting

hard task for VLMs, even harder. In contrast RS-LLaVA accurately counts simple scenes or small objects, and gained 86% for. Fine-tuning provided RS-LLaVA with more accurate cues from road density, building texture, and layout typical to RS imagery. But it still underestimates for large counts due to inherent limitations of instance recognition in dense RS imagery.

Fine-tuning yields substantial improvements, particularly in counting, comparison, and distinguishing rural versus urban scenes, as the model adapts to RS-specific scene structure, semantics, and spatial priors. Counting large numbers, however, appears to remain a systemic challenge rooted in the architecture itself rather than solely in data availability. Furthermore, RS-LLaVA tends to be more cautious and less confident, often producing false negatives for uncommon classes, likely due to the limited diversity of such categories in the fine-tuning data. Presence-type questions in RS frequently require nuanced judgment about what qualifies as present, for small, ambiguous, or camouflaged objects, vanilla models tend to hallucinate (and might answer correctly simply due to overestimating), whereas RS-LLaVA, after tuning, may underestimate. Refer to the illustrative examples for different cases discussed, blue indicates correct prediction while gray indicates a failure case.

Ⅳ. Future Directions

Our analysis highlights the need for more robust remote sensing VLMs. Promising directions include hybrid counting heads, where architectures explicitly parse objects at the patch or region level, potentially combining segmentation with object-counting priors. Ambiguity-aware training is also critical, requiring explicit annotation of marginal presence cases, ideally with multiple annotators and hard negative mining to better capture borderline examples. To move beyond pixel-level interpretation, scene context modules could integrate cues from urban form and landmarks, and further leverage external modalities such as additional spectral bands. The creation of synthetic datasets offers another path, enabling large-scale, controlled scene generation to support curve-fitting for large object counting and rare presence detection. Finally, there is a need to balance, rather than over-conservatism emerging after fine-tuning, models should be trained with diverse presence annotations and objective functions that explicitly manage the trade-off between false positives and false negatives.



Counting

V. Conclusion

Our systematic study demonstrates that, while vision-language models pretrained on natural data do transfer some capability to remote sensing scenes (especially obvious presence/comparison queries), substantial domain-specific challenges cripple them for complex and high-density RS analytics. Fine-tuning offers dramatic improvements in core tasks but does not fully close the gap, counting at scale and marginal object detection remain key issues. Progress in this space depends not only on more and better RS annotations, but also on creative new architectural directions that bridge the gap between generic visual-language intelligence and the nuanced, ambiguous, and richly structured world of geospatial imagery.

References

[1] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5625-5644, 2024.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34892-34916, 2023.

[3] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, and F. Melgani, "Rs-llava: A large vision language model for joint captioning and question answering in remote sensing imagery," *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748-8763, PmLR, 2021.

[5] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23716-23736, 2022.

[6] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*, pp. 12888-12900, PMLR, 2022.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.

[8] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International conference on computer, information and telecommunication systems (Cits)*, pp. 1-5, IEEE, 2016.

[9] G. Hoxha and F. Melgani, "A novel svm-based decoder for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-14, 2021.

[10] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "Rsvqa: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555-8566, 2020.

[11] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-14, 2021.

Authors

Maryam Qamar

Maryam Qamar received her BS degree with gold medal in computer science from University of AJ&K, Muzaffarabad, AJK, Pakistan in 2013. She earned her MS degree in computer science from National University of Sciences and Technology, Islamabad, Pakistan in 2017. She is currently doing her PhD in artificial intelligence at Kyung Hee University, South Korea. Since 2019, she has been a Lecturer with the Department of Computer Science and Information Technology, University of AJ&K, Muzaffarabad, AJK, Pakistan. Her research interests include artificial intelligence, machine learning, image and video processing, and computer vision.

Sung-Ho Bae

Sung-Ho Bae (Member, IEEE) received the B.S. degree from Kyung Hee University, South Korea, in 2011, and the M.S. and Ph.D. degrees from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012 and 2016, respectively. From 2016 to 2017, he was a Postdoctoral Associate with the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), MA, USA. Since 2017, he has been an Assistant Professor with the Department of Computer Science and Engineering, Kyung Hee University. He has been involved in model compression/interpretation for deep neural networks and inverse problems in image processing and computer vision.