A Knowledge-Grounded Conversational Agent for Object Grounding and Analysis in Remote Sensing

Ruiqian ZHANG, Wenjie LIU, Xiaogang NING and Hanchao ZHANG, China

Keywords: Large Vision-Language Models (VLMs); Remote Sensing (RS); Object Grounding; Knowledge-Grounded Reasoning;

1. SUMMARY

Large Vision-Language Models (VLMs) show promise for general-domain understanding but face significant limitations in specialized fields like remote sensing (RS). The unique characteristics of RS imagery, such as overhead perspectives, vast scale variations, complex backgrounds, and distinct object morphologies, create a significant domain gap. These limitations constitute a dual gap: 1) a perception gap, where models like Qwen-VL struggle with fine-grained object grounding, and 2) a knowledge gap, lacking the specialized, real-world context required for insightful analysis. In this paper, we propose a novel framework that addresses both challenges. First, to resolve the perception gap, we construct a specialized remote sensing dataset for VLM fine-tuning and employ Supervised Fine-Tuning to successfully develop a model specialized for high-accuracy object grounding in the RS domain. This fine-tuned model functions as the system's core perception module, providing high-accuracy visual grounding capabilities. Building upon this, our primary contribution is the development of an interactive conversational agent to bridge the knowledge gap. This agent facilitates specialized, knowledge-grounded reasoning and analysis for diverse scenarios. It enables the execution of complex analytical tasks—correlating visual findings with factual, domain-specific information—through a conversational dialogue and question-answering system. Experimental results demonstrate that our framework achieves high-precision object grounding and provides insightful, context-aware analysis, showcasing an effective approach for integrating large-scale VLM pre-training with domain-specific RS applications.

A Knowledge-Grounded Conversational Agent for Object Grounding and Analysis in Remote Sensing

Ruiqian ZHANG, Wenjie LIU, Xiaogang NING and Hanchao ZHANG, China

2. INTRODUCTION

Vision – Language Models (VLMs)[1] have rapidly advanced and demonstrated strong capabilities in multimodal understanding tasks, including image captioning, visual question answering, and instruction-following interaction[2]. Large-scale models trained on diverse image – text corpora, such as Qwen-VL**Fejl! Henvisningskilde ikke fundet.** and LLaVA[3], provide a unified interface for jointly reasoning over visual content and natural language, making them attractive foundations for building general-purpose visual intelligence systems.

Despite this progress, the direct application of general-purpose VLMs to remote sensing (RS) imagery remains challenging[5]. RS images differ fundamentally from natural images in acquisition geometry and semantic structure[6][7]. Typical characteristics include nadir or oblique overhead viewpoints, large intra-class scale variations, visually complex backgrounds, and subtle morphological differences between object categories. These factors lead to a pronounced domain gap that limits the effectiveness of models trained primarily on natural image distributions.

From a system perspective, these limitations can be analyzed along two complementary dimensions: a perception gap and a knowledge gap. The perception gap refers to the difficulty of achieving reliable fine-grained visual grounding in RS imagery. Many RS targets are small, densely distributed, or visually similar, which makes instance-level localization particularly challenging in language-guided settings. Queries that involve relative position, attribute comparison, or implicit constraints further amplify this difficulty and require precise spatial reasoning across multiple candidate instances.

The knowledge gap arises from the lack of domain-specific contextual understanding in general VLMs. In practical RS applications—such as land-use surveys, urban planning, and infrastructure monitoring—visual recognition alone is insufficient. Interpretation often relies on professional definitions, regulatory standards, or auxiliary knowledge sources[8][9]. Without explicit access to such information, model outputs are typically limited to generic descriptions and cannot support application-oriented reasoning.

To address these challenges, we propose TerraSense, a remote sensing‐oriented vision‐language framework designed to explicitly bridge both the perception and knowledge gaps. We first construct a dedicated RS dataset and apply supervised fine-tuning to adapt a general-purpose VLM into a domain-specific perception model capable of high-precision object grounding. Building upon this perception backbone, we further develop TerraSense Agent, a knowledge-grounded conversational agent that integrates visual grounding results with domain knowledge to support interactive, language-driven analysis. Together, these components form an end-to-end system for accurate localization and context-aware interpretation in realistic RS scenarios.

3. METHODOLOGY

3.1. System Overview

The proposed TerraSense framework operates via a modular architecture comprising three main components: (1) a system orchestration layer, (2) a domain-specific perception module, and (3) an interactive, knowledge-grounded conversational agent. We establish natural language as the unified interface, enabling instruction-driven analysis of RS image analysis. Upon receiving an input image and a user query, the system first interprets the task intent and determines whether visual grounding, knowledge-based reasoning, or both are required. The perception module is invoked when visual evidence is needed, and the resulting outputs are integrated with linguistic reasoning to generate the final response.

3.2. Domain-Specific Perception via Supervised Fine-Tuning

To bridge the perception gap, our proposed framework adapts a general VLM into a remote sensing perception expert using supervised fine-tuning (SFT)[3]. Rather than modifying the model architecture, we focus on task reformulation and data design to align visual representations with RS-specific semantics.

Unified Instruction Representation. We construct a specialized instruction-tuning dataset derived from public remote sensing benchmarks, specifically DOTA[10] and DIOR[11]. To enable language-driven analysis, we do not simply use the original classification labels; instead, we restructure the annotations into a unified natural language instruction format. Formally, we define each training sample as a triplet (I,T,y). Here, I denotes the input remote sensing image, T represents the user instruction (e.g., “Detect all vehicles”), and y corresponds to the target output sequence. The output y contains structured grounding results, including category labels and bounding box coordinates, tailored to the specific requirements of the instruction.

Fine-Tuning Procedure. We formulate the object grounding task as a conditional text generation problem[2]. During SFT, the model is trained to generate the target sequence y in an autoregressive manner, conditioned on both the visual features of image I and the linguistic semantics of T. The optimization objective is defined using the standard cross-entropy loss:

$$\mathcal{L}_{SFT} = -\sum_{t=1}^{N} \log P(y_t | I, T, y < t)$$

where y_t is the token to be predicted at step t, and $y < t$ denotes the sequence of all preceding tokens. This objective encourages the model to maximize the likelihood of the correct grounding coordinates given the instruction, ensuring precise visual-language alignment for domain-specific targets. After fine-tuning, this model serves as the core perception backbone of the TerraSense framework.

3.3. TerraSense Agent: Knowledge-Grounded Conversational Agent

Leveraging the domain-adapted perception backbone, the TerraSense Agent facilitates interactive analysis through natural language dialogue. A large language model (LLM) functions as the central reasoning unit, responsible for task decomposition and semantic coordination. Crucially, the perception model optimized via SFT (as detailed in Sec. 2.2) is encapsulated as a specialized tool within the agent's workflow[12]. For complex queries involving spatial constraints or visual comparison, the LLM invokes this tool to obtain grounded visual evidence. To further bridge the knowledge gap, domain-specific constraints are injected through structured prompts and external knowledge bases (e.g., official land-use codes), enabling advanced reasoning beyond visual recognition[5][9].

4. IMPLEMENTATION AND QUALITATIVE RESULTS

4.1. System Deployment

The TerraSense Agent is implemented using a decoupled, scalable architecture designed for high-availability inference. A high-throughput orchestration engine manages concurrent LLM requests, while the perception module operates as an independent service accessed via tool calls. This separation of concerns ensures that the heavy computational load of visual encoding does not impede the latency of conversational reasoning.

4.2. User Interface

To facilitate end-to-end analysis, the system integrates a unified graphical interface as shown in Figure 1. The layout consolidates model configuration controls, an interactive

conversational window, and a visualization panel. This design enables users, regardless of programming expertise, to visually verify grounding results and engage in multi-turn dialogue for refinement.

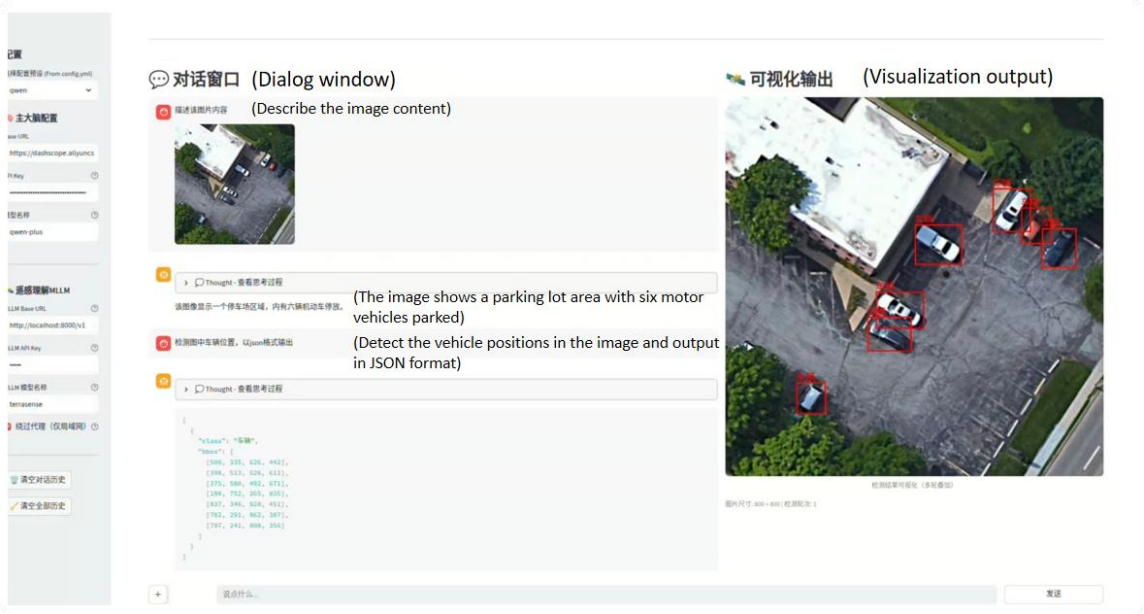

Figure 1. The interactive of user interface of the TerraSense Agent.

4.3. Visual Grounding and Reasoning Results

To validate the effectiveness of our domain-specific adaptation strategies, we conducted comparative qualitative experiments focusing on challenging remote sensing scenarios, such as dense object counting and small target localization.

As illustrated in Figure 2, we compare the performance of vanilla Qwen3-VL-8B[13] (Baseline) against our TerraSense Agent. Figure 2 (a) shows that the baseline model exhibits significant limitations when processing complex overhead imagery; it frequently fails to distinguish small, densely packed objects from cluttered backgrounds, resulting in substantial missed detections (indicated by red boxes) and spurious false positives (green boxes).

In contrast, as shown in Figure 2(b), TerraSense demonstrates superior robustness. Leveraging the perception module optimized via SFT (Sec. 2.2), our system achieves high-precision grounding even in dense scenes. It consistently localizes small-scale targets with accurate spatial alignment, effectively mitigating the perception gap. These results substantiate that supervised fine-tuning with RS-specific instructions significantly enhances fine-grained perception capabilities, providing a reliable foundation for downstream reasoning tasks.

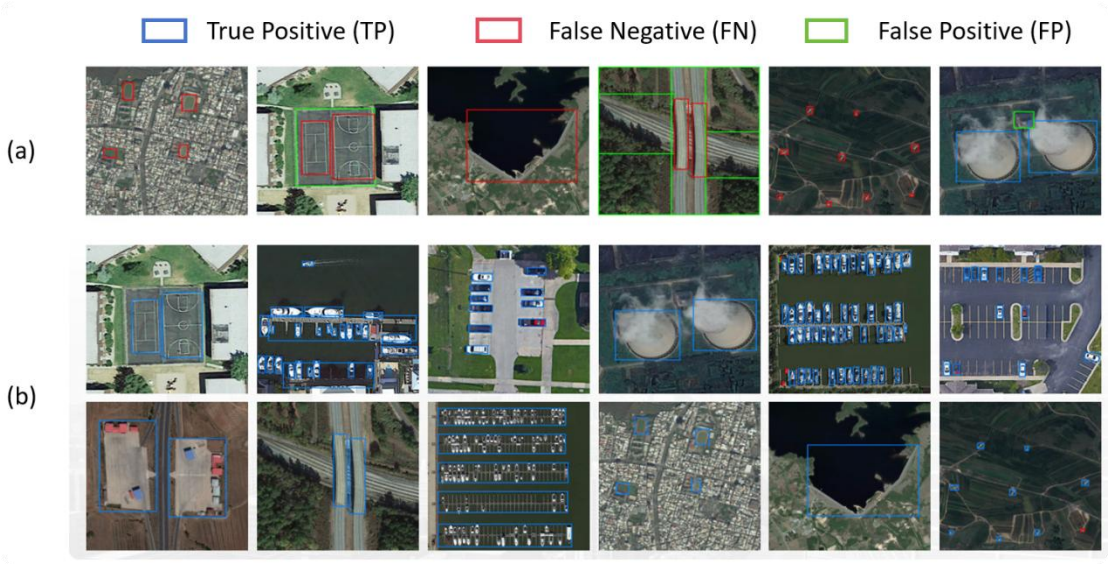

Figure 2. Qualitative comparison of object grounding performance in remote sensing imagery.

5. CONCLUSION

This paper presents TerraSense Agent, a knowledge-grounded vision‑language system for object grounding and analysis in remote sensing imagery. By explicitly addressing both perception reliability and domain knowledge integration, the proposed framework enables accurate, instruction-driven localization and context-aware interpretation through natural language interaction.

The modular design of TerraSense facilitates flexible deployment and extension, making it suitable for practical RS workflows that require interactive analysis and interpretability. Overall, this work provides a system-oriented reference for integrating domain-specific perception and conversational reasoning within vision‑language frameworks for remote sensing applications.

REFERENCES

[1] Zhang, J., Huang, J., Jin, S., et al. Vision-language models for vision tasks: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2024, 46(8): 5625-5644.

[2] Alayrac, J. B., Donahue, J., Luc, P., et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 2022, 35: 23716-23736.

[3] Liu, H., Li, C., Wu, Q., et al. Visual Instruction Tuning. Advances in Neural Information Processing Systems. 2023, 36: 34892-34916.

[4] Bai, J., Bai, S., Chu, Y., et al. Qwen technical report. 2023, arXiv preprint arXiv:2309.16609.

[5] Kuckreja, K., Danish, M. S., Naseer, M., et al. GeoChat: Grounded large vision-language model for remote sensing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, 27831-27840.

[6] Lu, X., Wang, B., Zheng, X., et al. Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing, 2017, 56(4): 2183-2195.

[7] Silva, J. D., Magalhães, J., Tuia, D., et al. Remote sensing visual question answering with a self-attention multi-modal encoder. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery. 2022, 40-49.

[8] Zhan, Y., Xiong, Z., & Yuan, Y. SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model. ISPRS Journal of Photogrammetry and Remote Sensing, 2025, 221: 64-77.

[9] Zhang, W., Cai, M., Zhang, T., et al. EarthGPT: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-20.

[10] Xia, G., Bai, X., Ding, J., et al. DOTA: A large-scale dataset for object detection in aerial images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3974-3983.

[11] Li, K., Wan, G., Cheng, G., et al. Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS journal of photogrammetry and remote sensing, 2020, 159: 296-307.

[12] Topsakal, O., Akinci, T C. Creating large language model applications utilizing langchain: A primer on developing llm apps fast[C]//International conference on applied engineering and natural sciences. 2023, 1(1): 1050-1056.

[13] Yang, A., Li, A., Yang, B., et al. Qwen3 technical report[J]. arXiv preprint arXiv:2505.09388, 2025.

BIOGRAPHICAL NOTES

Ruiqian Zhang received her Ph.D. degree in Photogrammetry and Remote Sensing from Wuhan University in 2021. She is currently an Associate Research Fellow at the Institute of

Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping (CASM). Her research interests include high-resolution image change detection, pattern recognition, and remote sensing applications.

Wenjie Liu received the B.Eng. degree in Surveying and Mapping Engineering from Central South University, Changsha, China, in 2024. He is currently pursuing the M.Sc. degree with the Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping. His research interests include remote sensing change detection.

Xiaogang Ning received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006. He is currently a Professor of the Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping (CASM). His research interests include high-resolution image processing, pattern recognition, and urban applications of remote sensing.

Hanchao Zhang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2019. He is currently an Associate Research Fellow with the Chinese Academy of Surveying and Mapping (CASM). His primary research interests cover remote sensing interpretation, change detection, and deep learning.

CONTACTS

Dr. Ruiqian Zhang
Chinese Academy of Surveying and Mapping,
No.28 Lianhuachi West Road, Haidian District,
Beijing, 100036 Beijing, China.
Email: zhangrq@casm.ac.cn