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

Ruiqian Zhang, Wenjie Liu, Xiaogang Ning and Hanchao Zhang (China, PR)

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

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.