Visual Agentic System for Spatial Metric Query Answering in Remote Sensing Images

Yinghao Wang[1] and Cheng Wang[2]

[1]University of Cambridge, United Kingdom [2]University of East Anglia, United Kingdom

Abstract

Accurately measuring real-world object dimensions from Remote Sensing (RS) images is crucial for applications in geospatial analysis and urban planning. Traditional Vision-Language Models (VLMs) struggle with spatial reasoning, while end-to-end remote sensing VLMs are often limited to predefined tasks such as image captioning. In this paper, we propose a visual agentic system for spatial metric query answering, dynamically integrating code-generation agents with a grounded remote sensing VLM and a Vision Specialist. Our system autonomously identifies reference objects, infers scale factors, and performs spatial measurements through structured subroutines. Experiments demonstrate that our approach achieves higher accuracy in footprint area estimation compared to state-of-the-art large language models with vision capabilities.

CCS Concepts

• ***Computing methodologies*** → *Scene Understanding; Image Segmentation; Object Identification;*

1. Introduction

Recent developments in agentic AI, where AI agents autonomously interact with their environment, query multiple models, and synthesize multi-step reasoning, offer a promising new paradigm for spatial metric query answering (SMQA) in remote sensing images. In this work, we propose a novel agent-driven pipeline (Figure 1) that leverages GPT-4o to construct an agentic system capable of autonomously interacting with GeoChat [KDN*24], a grounded remote sensing VLM, for reference object identification in RS images. The system further integrates Segment Anything Model (SAM2) [KMR*23] as a Vision Specialist for precise segmentation and computes spatial metrics through API-based dynamic code generation. This multi-step approach enables the agentic system to answer complex geospatial queries, such as "What is the building's real-world footprint?" without requiring explicit manual input.

2. Related Work

Foundation Vision-Language Models (VLMs) have been used for Remote Sensing image analysis, particularly in tasks such as scene understanding, object localization, counting, and change detection. It has been discussed in [ZW24] that models like GPT-4V exhibit strong performance in open-ended tasks but their limited spatial reasoning capabilities reduce effectiveness in tasks requiring precise object localization and counting. To address these limitations, recent work such as GeoChat [KDN*24] has introduced grounded VLMs trained on multimodal remote sensing datasets. These specialized models enhance zero-shot image and region captioning with improved spatial reasoning and geospatial adaptability.

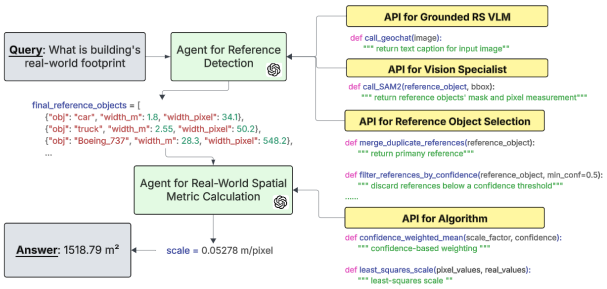

Figure 1: *Workflow of the visual agentic system for SMQA.*

Despite these advancements, a fundamental task, scale factor estimation, remains an underexplored challenge in remote sensing, crucial for answering real-world spatial queries. Its complexity stems from the need for both visual processing and reasoning, which specialized end-to-end RS models fail to separate, limiting interpretability and generalization. To address this, our agentic system draws inspiration from ViperGPT [SMV23], which leverages code-generation models to structure vision and language tasks into subroutines for complex visual queries. By dynamically integrating models and generating code from APIs, our visual agentic system autonomously infers scale factors and enhances spatial measurement accuracy, improving both interpretability and precision for real-world object measurement in remote sensing images.

DOI: 10.2312/egp.20251028

3. Method and Implementation

Our visual agentic system is designed to process spatial metric queries in RS images. The system comprises two GPT-4o agents, each capable of dynamically generating code based on available APIs and user query. These agents interact with a grounded remote sensing VLM for semantic scene understanding and a Vision Specialist for precise object segmentation and pixel-level analysis.

3.1. Agent for reference detection

When querying spatial metrics in remote sensing images, the key challenge is to determine the scale factor. To achieve this, we design a GPT-4o agent for the detection and selection of reference objects. The agent operates in a multi-step process:

- **Reference Object Identification** – The agent first calls an API to leverage GeoChat's strong RS image captioning capabilities, generating a list of potential reference objects along with their bounding box coordinates in the image.
- **Reference Object Selection** – The agent utilizes a reference selection API to dynamically generate code for filtering and prioritizing reference objects based on query-specific criteria, such as bounding box area and proximity to the object of interest. It also verifies the availability of external knowledge on the reference object's real-world dimensions.
- **Segmentation and Pixel-Level Analysis** – The agent then calls the Vision Specialist API to utilize SAM2 [KMR*23] for precise segmentation of the selected reference objects, extracting pixel-level metrics essential for scale estimation.

This modular API enables autonomous and interpretable reference selection, ensuring the agent identifies the most reliable references for spatial measurement.

3.2. Agent for real-world spatial metric calculation

The spatial metric calculation agent is responsible for inferring the scale factor based on the selected reference objects. To achieve this, the agent dynamically generates code based on APIs tailored to various object of interest and scene settings.

The approach used in Figure 2 utilizes a weighted confidence score based on the pixel-level area of reference objects. This weighting mechanism accounts for the fact that objects with small pixel areas are more susceptible to imaging artifacts and environmental noise, potentially leading to inaccuracy. For a given reference object i, the scale factor S_i is defined as: $S_i = R_i \,/\, P_i$ where R_i is the known real-world size of the reference object, and P_i is the pixel size of the segmented reference object in the image.

Given N reference objects, let A_i be the pixel area of the i-th reference object in the image and W_i be the weight assigned to the i-th reference object. The final weighted *scale factor* is: $S = \sum_{i=1}^{N} W_i S_i = \sum_{i=1}^{N} \left(A_i / \sum_{j=1}^{N} A_j \right) S_i$

4. Results and Conclusion

Unlike directly querying foundation large language models with vision capabilities, our visual agentic system integrates code-generation agents to dynamically orchestrate a grounded specialized VLM and a Vision Specialist into structured subroutines. As demonstrated in Figure 2, our system achieves higher accuracy in measuring the footprint area of buildings in remote sensing images, highlighting its potential for autonomous spatial measurement and metric query answering.

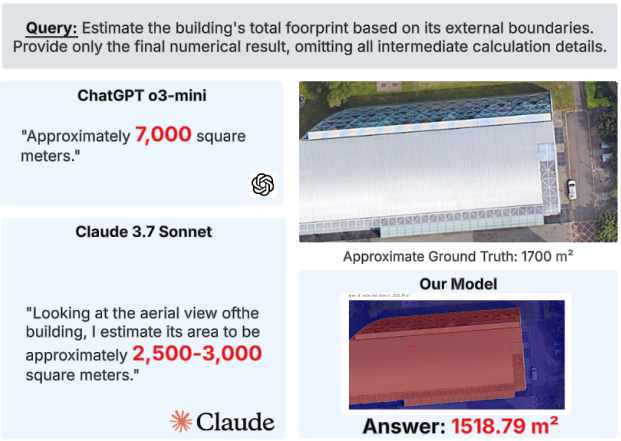

Figure 2: *Example demonstrating how our visual agentic system provides a more reliable answer compared to baseline responses from state-of-the-art multimodal models, including ChatGPT o3-mini and Claude 3.7 Sonnet. Test image source: Google Maps.*

Unlike general-purpose VLMs, which excel at question answering but lack spatial reasoning, and end-to-end remote sensing VLMs, which are limited to predefined tasks, our agentic system bridges this gap by enabling spatial metric query answering through dynamic code-generation agents. This approach enhances interpretability and adaptability, unlocking new possibilities for autonomous geospatial analysis and urban planning.

Future work includes integrating additional Vision Specialists into the APIs to expand capabilities, such as shadow detection for object height estimation, and extending spatial metric analysis to 3D measurements for enhanced real-world applicability.

References

[KDN*24] KUCKREJA K., DANISH M. S., NASEER M., DAS A., KHAN S., KHAN F. S.: Geochat: Grounded large vision-language model for remote sensing. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024). 1

[KMR*23] KIRILLOV A., MINTUN E., RAVI N., MAO H., ROLLAND C., GUSTAFSON L., XIAO T., WHITEHEAD S., BERG A. C., LO W.-Y., DOLLAR P., GIRSHICK R.: Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2023), pp. 4015–4026. 1, 2

[SMV23] SURÍS D., MENON S., VONDRICK C.: Vipergpt: Visual inference via python execution for reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)* (2023). 1

[ZW24] ZHANG C., WANG S.: Good at captioning, bad at counting: Benchmarking GPT-4V on Earth observation data . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Los Alamitos, CA, USA, June 2024), IEEE Computer Society, pp. 7839–7849. 1