GeoMMBench and GeoMMAgent: Toward Expert-Level Multimodal Intelligence in Geoscience and Remote Sensing

Supplementary Material

1. More Descriptions on GeoMMBench

1.1. Dimensions and Tasks in GeoMMBench

Below we provide explanations for the abbreviations of evaluation dimensions in GeoMMBench, as listed in Tables 1 and 3 of the paper, along with their corresponding tasks.

Disciplines: "RS" (Remote Sensing), "Ph." (Photogrammetry), "GIS" (Geographic Information System), and "GNSS" (Global Navigation Satellite System). Please note that in this paper, RS in "geoscience and remote sensing" is used in a broad sense, referring to the observation of objects without physical contact. At the disciplinary level, photogrammetry focuses on geometric measurements, whereas RS is primarily concerned with radiation-based measurements.

Sensor Modalities: "Opt.": Optical RGB imagery, "HSI": Multispectral/Hyperspectral imagery (MSI can be regarded as a special case of HSI with fewer spectral bands, so we simplify here for briefly), "SAR": Synthetic Aperture Radar, "LiD.": Light Detection and Ranging, "DEM": Digital Elevation Model, "The.": Thermal imagery.

Task Spectrum:

- "Pri." (Principles) refers to fundamental principles analysis and theoretical understanding.
- "Per." (Perception) covers perception tasks across different sensor modalities, including: *Basic perception tasks* (scene classification, object referring, object grounding, object counting, and sensor-level recognition such as modality discrimination); *Fine-level object recognition* and *complex interpretation of sensor imagery*, including attribute analysis and identification/interpretation of spectral curves.
- "Spa." (Spatial) refers to spatial relation analysis, including relative direction, distance estimation, spatial positioning, and other geospatial interpretations such as estimating ocean depth or mountain height.
- "Qua." (Quality) includes data quality tasks such as image quality inspection and corrections (atmospheric, radiometric, and geometric).
- "Tim." (Time Series) refers to temporal analysis across multiple images with various formats, including sensor imagery, maps, and RS mapping products.
- "App." (Applications) covers RS applications, focusing on the analysis of RS products across various domains, such as economics, environment, oceanography, and climate studies.

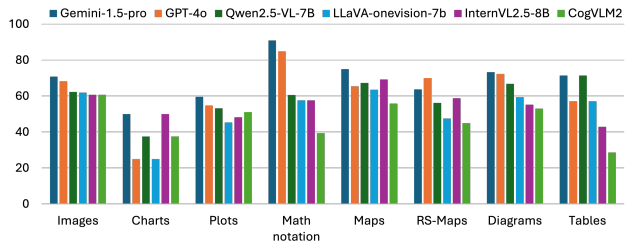


Figure A1. Performance of MLLMs on different types of images in GeoMMExpert.

1.2. Different Image Types.

We analyze model performance across various image types in Fig. A1, including sensor imagery, charts, plots, mathematical notations, maps, RS-map products, diagrams, and tables. Among all models, Gemini-1.5 Pro demonstrates the highest accuracy across all categories, suggesting its robust generalization to diverse visual formats commonly used in geoscience and RS. Open-source models, while improving in some areas, exhibit notable weaknesses in mathematical notation interpretation and show relatively lower performance in charts and RS-map products. This indicates that they face difficulties with these specialized representations, which require both multimodal understanding and domain-specific reasoning. These results highlight the importance of enhancing model adaptability to structured and geospatial-specific visual information for improved expert-level AGI performance.

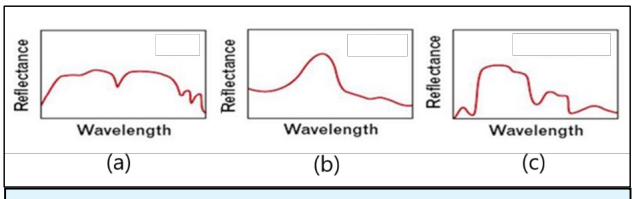
1.3. More Error Cases

In this subsection, we provide additional error cases from GPT-4o [5], as shown in Figs. A4 to A16. These cases span diverse geoscience and RS dimensions, illustrating common failure patterns in SOTA MLLMs when interpreting geospatial data and tasks.

2. More Descriptions on GeoMMAgent

2.1. Toolkit Library

We present the tools integrated into GeoMMAgent. As shown in Fig. 4 and Section 3 of the manuscript, the toolkit library is organized into four categories: *general toolkit*, *knowledge toolkit*, *perception toolkit*, and *reasoning toolkit*. GeoMMAgent is designed as a fully training free and extensible framework, where new tools can be seamlessly added



Question: In the provided diagram, which surface type has the highest reflectance variability over wavelengths?

Options:

A: (a) B: (b) C: (c) D: None of the above

GPT-4o Prediction C ✗

Figure A2. A sample error case of remote sensing principle understanding. Subfield: Spectral curve interpretation.

Question: Which material has the highest reflectance in the visible wavelength region?

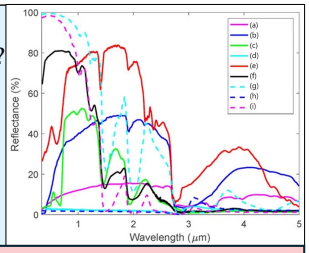
Options:

A: (b)

B: (c)

C: (i)

D: (e)



GPT-4o Prediction D ✗

Figure A3. A sample error case of remote sensing principle understanding. Subfield: Spectral curve interpretation.

Question: What does ② represent in the context of atmospheric windows in remote sensing?

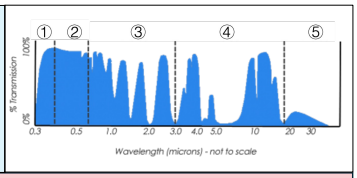
Options:

A: Radio Window

B: Ultra-Violet Window

C: Infrared Window

D: Visible Window



GPT-4o Prediction C ✗

Figure A4. A sample error case in remote sensing principles. Subfield: Atmospheric window recognition.

Question: Which process is most likely represented by 'B' in the image?

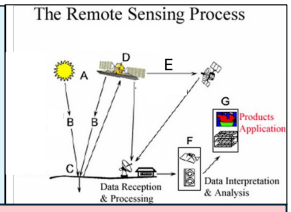
Options:

A: Atmospheric Interaction

B: Target Interaction

C: Data Processing

D: Energy Reflection



GPT-4o Prediction B ✗

Figure A5. A sample error case in remote sensing principles. Subfield: Remote sensing process.

without any model fine tuning or architectural modification. This plug and play design ensures superior flexibility, adaptability to emerging geospatial tasks, and long term upgradeability as external tools and APIs evolve. Below, we provide

Question: Which satellite in the image has a greater number of spectral bands?

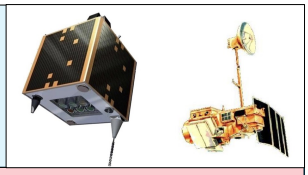
Options:

A: Left

B: Right

C: Both have the same number

D: Cannot determine



GPT-4o Prediction D ✗

Figure A6. A sample error case in remote sensing platforms. Subfield: Remote sensing satellite identification.

Question: Which component in the diagram corresponds to spectral resolution?

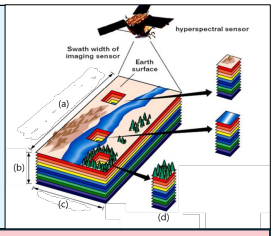
Options:

A: (a)

B: (b)

C: (c)

D: (d)



GPT-4o Prediction C ✗

Figure A7. A sample error case in remote sensing principles. Subfield: Resolution type recognition.

Question: The image shows ice thickness maps of a lake for the months of January to April. How does the ice thickness generally change across these months?

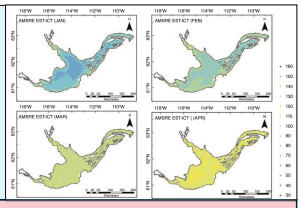
Options:

A: Increases steadily

B: Decreases steadily

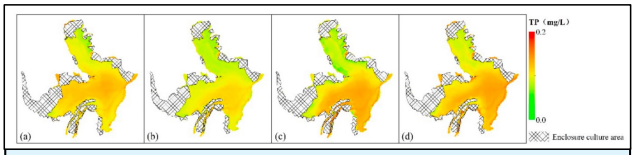
C: Remains constant

D: Peaks in March and decreases in April



GPT-4o Prediction D ✗

Figure A8. A sample error case in time series analysis. Subfield: Environmental monitoring.



Question: Which season shows the highest total phosphorus (TP) concentrations across most areas of the lake in the seasonal average images: (a) Spring, (b) Summer, (c) Autumn, and (d) Winter?

Options:

A: Spring B: Summer C: Autumn D: Winter

GPT-4o Prediction C ✗

Figure A9. A sample error case in time series analysis. Subfield: Pollution analysis.

detailed descriptions of each toolkit category.

General Toolkit provides essential preprocessing and post-processing utilities that ensure proper data formatting, quality control, and task specific output handling. These tools serve as the foundation for downstream specialized agents and enable robust execution across diverse geospatial tasks.

- *Format conversion*: Converts between different data for-

Question: Which LiDAR platform is most likely represented by ④ in the image?

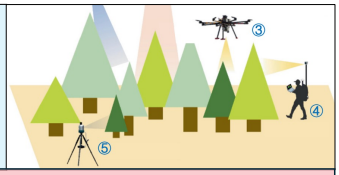
Options:

A: Spaceborne LiDAR

B: Mobile LiDAR

C: Terrestrial LiDAR

D: UAV LiDAR



GPT-4o Prediction C ✗

Figure A10. A sample error case in remote sensing principles. Subfield: LiDAR platforms.

Question: Which category does Region B represent?

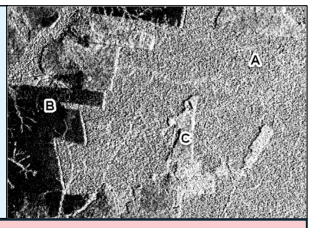
Options:

A: Forested areas

B: Bare land

C: Recently deforested areas

D: Water bodies



GPT-4o Prediction D ✗

Figure A11. A sample error case in remote sensing perception. Subfield: SAR imagery perception.

Question: Which of the following conditions will most likely result in high positive values for this formula?

$$\frac{(NDVI - NDWI)}{(NDVI + NDWI)}$$

Options:

A: Saturated soil and over-irrigated vegetation

B: Dry conditions with stressed vegetation and limited water availability

C: Dense forest cover with high biomass

D: Water-dominated surfaces like lakes and wetlands

GPT-4o Prediction C ✗

Figure A12. A sample error case in hyperspectral theoretical understanding. Subfield: Spectral index understanding.

Question: What is the OBJECTID of the selected region highlighted on the map?

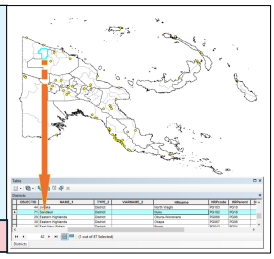
Options:

A: 71

B: 30

C: 44

D: 62



GPT-4o Prediction B ✗

Figure A13. A sample error case in GIS interpretation. Subfield: GIS map recognition.

mats to ensure compatibility across data sources and models.

- *Patch tiling and merging*: Divides large images into manageable tiles for efficient processing and later aggregates individual predictions into a unified output.
- *Filtering*: Applies smoothing, denoising, or sharpening operations to improve data quality and enhance downstream perception performance.
- *Cropping*: Extracts user specified or automatically determined regions of interest from large scale imagery. This reduces irrelevant spatial context, lowers computational cost, and ensures that subsequent modules focus on the most relevant areas.
- *Scaling*: Resizes imagery to the resolution or aspect ratio required by downstream modules. It supports both upsampling and downsampling and maintains consistent data formats when combining multi sensor or multi scale inputs.
- *Super resolution*: Improves the spatial resolution of remote sensing images using learning based or model based algorithms. This enhances visibility of fine grained structures.
- *Area counting*: Measures the total surface area of a specific region or object class after segmentation or thresholding.
- *Box counting*: Counts the number of detected bounding

Question: Which remote sensing quality issue is present in the image?

Options:

A: Twisted Object

B: Stretching Blur

C: Stitching Misalignment

D: Seam Line

GPT-4o Prediction C ✗

Figure A14. A sample error case in data quality examination. Subfield: Image quality inspection.

Question: What is the approximate geospatial location of point 1 (marked by the red dot in the image)?

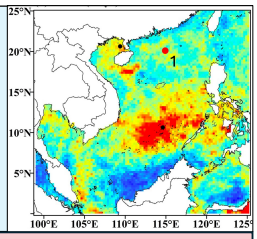
Options:

A: 115° E, 20° N

B: 120° E, 25° N

C: 110° E, 20° N

D: 115° E, 15° N



GPT-4o Prediction B ✗

Figure A15. A sample error case in spatial relation analysis. Subfield: Geospatial location identification.

Question: What is the ground sampling distance (GSD) of the remote sensing image?

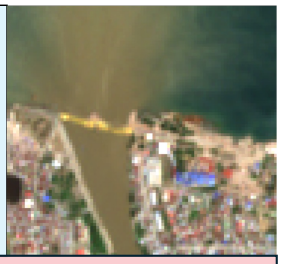
Options:

A: 100m

B: 50m

C: 10m

D: 1m



GPT-4o Prediction A ✗

Figure A16. A sample error case in remote sensing image recognition. Subfield: GSD recognition.

boxes for objects of interest.

Knowledge Toolkit enables GeoMMAgent to retrieve specialized information from external sources and augment its internal knowledge. This capability is crucial for tasks that involve dynamic environments, region specific context, or domain specific terminology or data that general purpose pre training does not fully capture. By integrating diverse knowledge bases, the agent can access factual, up to date, and geographically grounded information to support reliable geospatial reasoning.

- *Google API*: Provides access to open domain information and real time web content. It is used to verify geographical facts, gather updates on recent environmental events, and supplement internal knowledge with external data. This strengthens the agent's ability to handle dynamic scenarios.
- *Wikimedia API*: Offers structured encyclopedic knowledge about geospatial entities, geological terms, and landmarks. It helps the agent interpret technical vocabulary, understand the characteristics of landforms, and retrieve detailed descriptions of points of interest.
- *GME* [8]: Acts as a semantic alignment engine for queries that combine image inputs with text descriptions. It maps both the fused query and the retrieval candidates into a unified embedding space to evaluate semantic similarity. Unlike traditional text based or image based retrieval methods, this multimodal module identifies evidence that matches the joint visual and textual context, reduces hallucination, and improves the reliability of image grounded reasoning.

Perception toolkit includes expert remote sensing models at the scene level, object level, and pixel level. These models support core geospatial perception tasks and are robust to variations in image resolution. They provide accurate and interpretable outputs that anchor the agent's reasoning in reliable visual evidence.

- *Scene classification model*: We train a Yolo11 based classifier [3] with backbone CSPNet on the Million-AID [4] dataset. The model recognizes 51 scene categories and land cover types, covering the major classes commonly used in remote sensing scene understanding. The toolkit outputs top five predictions with confidence scores to support precise interpretation of scene semantics.
- *Detection model*: We deploy a pre trained Yolo11 detector [3] with backbone CSPNet trained on the DOTA-v2 [7] dataset. It employs oriented bounding boxes to detect and localize diverse geospatial objects such as aircraft, vehicles, and buildings. The toolkit outputs object counts, spatial distributions, and detection reports that include class labels and confidence values.
- *Segmentation model*: We train a DeepLabv3 plus model with Xception backbone [2] on the LoveDA dataset [6] to perform semantic segmentation and pixel level classification of remote sensing imagery. The model delineates land cover types, urban structures, and natural features with precise boundaries. The toolkit provides segmentation masks with per pixel class labels that support area measurement and spatial analysis of heterogeneous landscapes.

Reasoning toolkit includes advanced multimodal language models designed for complex logical inference, spatial and temporal understanding, and knowledge integration. These models operate on the outputs of previous agents and generate reliable final answers for geospatial decision making.

- *Reasoning Agent*: We employ Qwen VL Max [1], which combines high quality visual understanding with strong textual reasoning. It integrates perception outputs, retrieved knowledge, the original image, and the question context to perform multi step inference. The agent conducts semantic matching, consistency checking, option filtering, and final answer generation for tasks that require advanced logical or spatial reasoning.

The framework is model agnostic, and other multimodal reasoning models can be incorporated without additional training, maintaining full compatibility with the overall agent pipeline.

2.2. Agent and Tool Prompts

We summarize the system prompts governing each agent in Table A1. Furthermore, Table A2 details the functional descriptions of the available tools; these descriptions serve as a reference for the Coordinator Agent to assess tool capabilities and facilitate precise tool invocation. Collectively, these specifications define role-specific behaviors and establish a unified guideline for coordinated multi-agent execution.

2.3. Example Cases

We provide example cases to demonstrate how GeoMMAgent operates as a professional expert in geoscience and remote sensing. Each case is presented in a step by step manner to illustrate how the agent coordinates its toolkits, integrates multimodal evidence, and performs reliable geospatial reasoning.

2.3.1. Example Case #1

This example showcases the GeoMMAgent's multi-phase reasoning process for analyzing a spectral band identification question (Figure A17).

Phase 1: Task Specification and Decomposition. The agent analyzes the query and recognizes it as a multiple-choice question about spectral band properties in remote sensing. The image shows an electromagnetic spectrum diagram with Band ① labeled at a specific wavelength region.

- **Input:** User query and image pair.

Agent Role	System Prompt Content
Coordinate Agent	You are an Intelligent Orchestration Expert in the field of Remote Sensing that analyzes images and queries, creating execution plans through the coordination of multiple specialized agents. You provide a detailed description of the inputs and deliver professional task decomposition informed by a meticulous review of the agents' and toolkits' documentation.
Perception Agent	You are a Specialized Perception Expert in the field of Remote Sensing, responsible for extracting reliable visual evidence from multi-sensor imagery. You perform scene classification, object detection, and semantic segmentation, and provide calibrated predictions for downstream reasoning.
Knowledge Agent	You are a Geospatial Knowledge Retrieval Expert in the field of Remote Sensing, specialized in querying external knowledge bases and web resources. You retrieve, filter, and summarize factual information about remote sensing, geophysics, and geographic entities.
Reasoning Agent	You are an Expert Reasoning Agent in the field of Remote Sensing, specialized in multimodal geospatial reasoning. You integrate visual features, retrieved knowledge, and task context to perform step-by-step analysis and produce logically consistent answers.
Self-Evaluation Agent	You are a Professional Assessment Expert in the field of Remote Sensing, specialized in evaluating the correctness of image analysis results. You assess logic, consistency, completeness.

Table A1. Representative system prompts assigned to different agents within the GeoMMAgent framework.

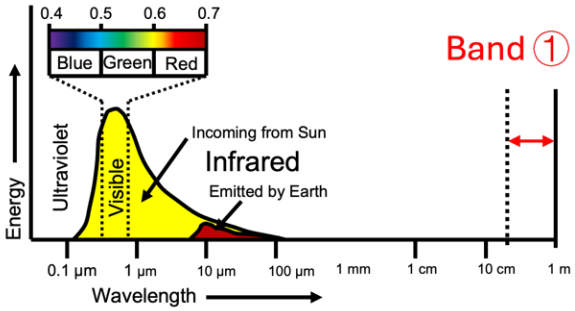


Figure A17. The multimodal query for our case study. The question is: "Which of the following statements about Band ① is accurate?" The options are: (A) It is suitable for thermal remote sensing, (B) It is used in passive remote sensing, (C) It operates only during the day and is affected by weather conditions, and (D) It can penetrate vegetation to detect subsurface targets. The answer is D.

- **Action (Task Decomposition):** The agent decomposes the task into: (1) Visually analyze the spectrum diagram to identify the wavelength range and position of Band ①. (2) Retrieve knowledge about the properties of bands in that spectral region. (3) Match the retrieved properties with the given options to determine which statement is accurate.
- **Output:** Task specification: identify Band ①'s spectral characteristics and match with options. Required agents: `Retrieval Agent` and `Reasoning Agent`.

Phase 2: Initial Task Execution. The agent executes the plan. The `Reasoning Agent` analyzes the image and identifies Band ① as located in the microwave region of the spectrum (approximately 1 mm to 1 m wavelength). The `Knowledge Toolkit` is invoked to search for microwave band properties, but the initial search query "Band ① properties" returns limited results. The agent generates an initial answer B based on partial reasoning that microwave sensing can be passive.

- **Input:** The task specification, query, and spectral diagram image.
- **Execution Log:** Visual analysis: Band ① identified in microwave region. Knowledge retrieval: limited results. Reasoning: partial match with option B.
- **Output:** Initial answer B: "It is used in passive remote sensing."

Phase 3: Self-Evaluation and Error Analysis. The self-evaluation agent reviews the answer and execution log. It identifies that while option B is partially correct (microwave can be passive), the answer lacks sufficient justification and does not fully leverage the visual information. The evalua-

Tool	System Prompt Content
General Toolkit	
Format conversion	Use this tool to convert inputs between different image and geospatial data formats so that downstream models and tools can directly consume the data.
Patch tiling and merging	Use this tool to split large images into tiles for efficient processing and then merge tile-level predictions back into a spatially consistent full-scene result.
Filtering	Use this tool to denoise, smooth, or sharpen imagery in order to improve data quality before perception or reasoning.
Cropping	Use this tool to crop user-specified or automatically selected regions of interest, removing irrelevant areas and reducing computational cost.
Scaling	Use this tool to resize imagery to the resolution or aspect ratio required by subsequent models, supporting both upsampling and downsampling.
Super resolution	Use this tool to enhance the spatial resolution of remote sensing imagery and reveal fine-grained structures that are important for detailed analysis.
Area counting	Use this tool to compute the surface area of a given region or semantic class based on segmentation or thresholding results.
Box counting	Use this tool to count detected objects from bounding-box outputs and summarize object statistics by category.
Knowledge Toolkit	
Google API	Use this tool to query open-domain web information related to a geospatial question, verify geographic facts, and obtain up-to-date context about environmental events.
Wikimedia API	Use this tool to retrieve structured encyclopedic knowledge about places, landforms, and technical terminology in remote sensing and geoscience.
GME	Use this tool to perform multimodal semantic retrieval: given an image–text query, rank candidate documents or patches by similarity in a unified embedding space and return the most relevant evidence.
Perception Toolkit	
Scene classification model	Use this tool when a scene-level category is needed: it classifies remote sensing images into 51 scene and land-cover types, including Dry land, Greenhouse, Paddy field, Terraced field, Meadow, Forest, Orchard, Commercial area, Storage tank, Wastewater tank, Works, Oil field, Mine, Quarry, Solar, Wind, Substation, Swimming pool, Church, Cemetery, Basketball court, Tennis court, Baseball field, Ground track field, Golf course, Stadium, Detached house, Apartment, Mobile home park, Apron, Helipad, Runway, Road, Viaduct, Bridge, Intersection, Parking lot, Roundabout, Pier, Railway, Train station, Rock land, Bare land, Ice land, Island, Desert, Sparse shrub land, Lake, River, Beach, and Dam, and returns top-k labels with confidence scores.
Detection model	Use this tool when object instances are required: it detects and localizes oriented objects, including Plane, Ship, Storage Tank, Baseball Diamond, Tennis Court, Basketball Court, Ground Track Field, Harbor, Bridge, Large Vehicle, Small Vehicle, Helicopter, Roundabout, Soccer Ball Field, and Swimming Pool, and outputs bounding boxes, categories, and confidence scores.
Segmentation model	Use this tool when pixel-wise masks are needed: it produces semantic segmentation maps that delineate land-cover types and structures, enabling area measurement and spatial pattern analysis.
Reasoning Toolkit	
Spatial-Temporal Analysis	Use this tool to analyze multi-temporal or multi-sensor remote sensing data, characterize spatial–temporal patterns, and reason about changes, trends, and dynamic processes across time.
Multiple Choice Matching	Use this tool to align free-form model answers with discrete options in multiple-choice questions, selecting the option that is most semantically consistent with the reasoning outcome.

Table A2. Representative system prompts assigned to different tools within the GeoMMAgent framework.

tion notes that the search query was too generic and did not retrieve specific properties that would distinguish between options.

- **Input:** The image-query pair, all four options (A, B, C, D), the initial answer (B), and the reasoning trace log from Phase 2.
- **Action (Error Analysis):** The evaluation log states: "Low confidence. Answer B is partially correct but lacks strong evidence. The knowledge retrieval was insufficiently specific. Need to refine search strategy to retrieve distinguishing properties of microwave bands, particularly their penetration capabilities."
- **Output:** Failure status with specific error analysis: insufficient knowledge retrieval depth.

Phase 4: Re-execution with Revised Plan. Based on the evaluation feedback, the agent revises its knowledge retrieval strategy. It searches for more specific queries: "microwave band penetration vegetation subsurface" and "microwave remote sensing properties." The refined search successfully retrieves the key property: "Microwave bands can penetrate vegetation and detect subsurface features." The `Reasoning Agent` then synthesizes this knowledge with the visual confirmation that Band ① is indeed in the microwave region, leading to a confident match with option D.

- **Input:** The revised plan, error analysis, original query, and image.
- **Action (Revised Execution):** Refined knowledge retrieval with specific queries about microwave penetration properties. Visual confirmation of Band ①'s position. Synthesis of knowledge and visual evidence.
- **Output:** Correct answer D: "It can penetrate vegetation to detect subsurface targets."

Phase 5: Final Self-Evaluation. The agent performs a final evaluation. The answer D is confirmed as correct: it is supported by (1) visual evidence showing Band ① in the microwave region, (2) retrieved knowledge about microwave penetration properties, and (3) logical matching with option D's description. The evaluation yields high confidence.

- **Input:** The final answer (D) and the complete execution log from Phase 4.
- **Output:** Success status with high confidence, confirming the answer is well-supported by both visual analysis and domain knowledge.

2.4. Example Case #2

This case illustrates how GeoMMAgent identify and count specific objects.

Phase 1: Task Specification and Decomposition. The agent analyzes the query and recognizes it as an object counting task in remote sensing imagery. The image contains multiple aircraft that need to be detected and counted.

- **Input:** User query and image pair.

Figure A18. The multimodal query for aircraft counting. The question is: "How many aircraft are there in the image?" The options are: (A) 13, (B) 10, (C) 12, (D) 9. The answer is C.

- **Action (Task Decomposition):** The agent decomposes the task into: (1) Use the `Detection Toolkit` to detect all aircraft in the image. (2) Count the number of detected aircraft. (3) Match the count with the given options to determine the correct answer.
- **Output:** Task specification: detect and count aircraft in the image. Required agents: `Detection Agent`.

Phase 2: Task Execution. The agent executes the plan. The `Detection Toolkit` processes the image and detects aircraft using oriented bounding boxes. The detection successfully identifies 12 aircraft in the image. The count is then matched with the given options, leading to answer C.

- **Input:** The task specification, query, and image.
- **Execution Log:** Detection: 12 aircraft detected with bounding boxes. Count: 12. Match with option C.
- **Output:** Answer C: 12 aircraft.

Phase 3: Self-Evaluation. The agent performs a final evaluation. The answer C is confirmed as correct: it is supported by (1) detection results showing 12 aircraft with bounding boxes, (2) accurate counting of detected objects, and (3) logical matching with option C. The evaluation yields high confidence.

- **Input:** The final answer (C) and the complete execution log from Phase 2.
- **Output:** Success status with high confidence, confirming the count is accurate and well-verified.

	val	test
w/o Knowledge	83.8	87.4
w/o Perception	83.8	80.3
w/o Reasoning	59.5	67.3
w/o Self-evaluation	81.1	80.1
full GeoMMAgent	86.5	88.4

Table A3. Component-wise ablation study of GeoMMAgent on the val and test sets.

2.5. Ablation Study

To systematically evaluate the contribution of each component in GeoMMAgent, we perform a component-wise ablation study. The results are summarized in Table A3. Overall, all components contribute positively to the final performance. In particular, removing the reasoning module leads to the most significant degradation, highlighting its critical role in the framework. The knowledge and perception modules also provide consistent improvements, while self-evaluation contributes to further refinement, albeit with a comparatively smaller impact.

3. Limitation

Like any benchmark, GeoMMBench has limitations despite its comprehensive design. The manual curation process may introduce selection biases, and the chosen knowledge points, while diverse, cannot fully represent the complete breadth and depth required for evaluating an Expert AGI in geoscience and remote sensing. Even so, we argue that strong performance on GeoMMBench remains a necessary criterion for an Expert AGI, since it reflects broad domain knowledge, deep subject understanding, and expert level reasoning capabilities.

In terms of system design, the tools integrated into GeoMMAgent are primarily targeted toward the tasks represented in GeoMMBench. Although the toolkit covers many important functionalities, it cannot fully encompass the entire range of tasks present in the wider geoscience and remote sensing community. Nonetheless, GeoMMAgent is built as a fully training free and extensible framework in which new tools can be added seamlessly without model fine tuning or architectural modification. This plug and play design enables high flexibility, adaptability to emerging geospatial applications, and long term upgradeability as external tools, APIs, and models evolve. We plan to incorporate more tools in the future to support an even broader spectrum of geospatial tasks.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 4

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4

[3] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 4

[4] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230, 2021. 4

[5] OpenAI. Hello gpt-4o. Technical report, OpenAI, 2024. Accessed: 2025-02-26. 1

[6] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 4

[7] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[8] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms, 2025. 4