

GEOBench-VLM: Benchmarking Vision-Language Models for Geospatial Tasks

Muhammad Sohail Danish^{*1} Muhammad Akhtar Munir^{*1} Syed Roshaan Ali Shah² Kartik Kuckreja¹

Fahad Shahbaz Khan^{1,3} Paolo Fraccaro⁴ Alexandre Lacoste⁵ Salman Khan^{1,6}

¹Mohamed bin Zayed University of Artificial Intelligence, ²University College London, ³Linköping University, Sweden

⁴IBM Research Europe, UK, ⁵ServiceNow Research, ⁶Australian National University



Figure 1. Examples of tasks from the GEOBench-VLM benchmark. Our benchmark is designed to evaluate VLMs on a diverse range of remote sensing applications. The benchmark includes over 10,000 questions spanning a range of tasks essential for Earth Observation, such as Temporal Understanding, Referring Segmentation, Visual Grounding, Scene Understanding, Counting, Detailed Image Captioning, and Relational Reasoning. Each task is tailored to capture unique domain-specific challenges, featuring varied visual conditions and object scales, and requiring nuanced understanding for applications like disaster assessment, urban planning, and environmental monitoring.

Abstract

While numerous recent benchmarks focus on evaluating generic Vision-Language Models (VLMs), they fall short in addressing the unique demands of geospatial applications. Generic VLM benchmarks are not designed to handle the complexities of geospatial data, which is critical for applications such as environmental monitoring, urban planning, and disaster management. Some of the unique challenges in geospatial domain include temporal analysis for changes, counting objects in large quantities, detecting tiny objects, and understanding relationships between entities occurring in Remote Sensing imagery. To address this gap in the geospatial domain, we present GEOBench-VLM, a comprehensive benchmark specifically designed to evaluate VLMs on geospatial tasks, including scene understanding, object counting, localization, fine-grained categorization, and temporal analysis. Our benchmark features over 10,000 manually verified instructions and covers a di-

verse set of variations in visual conditions, object type, and scale. We evaluate several state-of-the-art VLMs to assess their accuracy within the geospatial context. The results indicate that although existing VLMs demonstrate potential, they face challenges when dealing with geospatial-specific examples, highlighting the room for further improvements. Specifically, the best-performing GPT4o achieves only 40% accuracy on MCQs, which is only double the random guess performance. Our benchmark is publicly available at <https://github.com/The-AI-Alliance/GEO-Bench-VLM>.

1. Introduction

Deep learning has revolutionized computer vision [17, 42, 57, 67, 73] and natural language processing [1, 6, 48, 49], enabling significant advancements in integrating visual and textual data. VLMs integrate image and language understanding to handle tasks combining both visual and textual

^{*}Equally contributing first authors.

information [3, 9, 23, 40]. These models have achieved impressive results in applications such as image captioning [58], visual question answering [7], and object recognition [56] learning from large visual-text datasets.

VLMs have the potential to enhance analysis in fields such as remote sensing and environmental monitoring, which rely on satellite and aerial imagery for applications like urban planning, environmental assessment, and disaster management [19, 25, 32, 52, 63]. VLMs can improve the efficiency and accuracy of geospatial analysis by automating complex tasks that require substantial manual effort, such as land cover mapping, damage assessment, and detecting objects in high-resolution images. However, despite this potential, the application of VLMs to geospatial tasks has been limited. This is due to the unique challenges posed by geospatial data, e.g., objects in satellite imagery can vary significantly in scale, making it difficult for models trained on natural imagery datasets. Moreover, satellite images are captured under diverse resolutions and lighting conditions, affecting visibility and image quality. Many geospatial applications also require temporal analysis to detect changes over time, such as monitoring urban development or environmental degradation, introducing an additional layer of complexity. Furthermore, the absence of benchmarks explicitly designed for evaluating VLMs on geospatial data restricts performance assessment and makes it more difficult to pinpoint areas that require improvement [22, 33, 64].

Existing benchmarks, such as SEED-Bench [22] and MMMU [64], have significantly improved understanding of VLMs in tasks such as visual question answering and scene understanding. These benchmarks do not sufficiently address the needs of geospatial data, which involves spatial reasoning and temporal understanding. On the other hand, VLEO benchmarks geospatial tasks, evaluating VLM performance in earth observation applications [66]. This lacks benchmarks for extended temporal analysis, synthetic aperture radar (SAR), and segmentation, limiting its geospatial scope. Another limitation is its use of generic VLMs instead of geospatial-specific models. Though VLMs simplify data analysis, the lack of comprehensive evaluation highlights the need for geospatial-specific models and targeted benchmarks to fully address geospatial tasks.

To bridge this gap, we introduce GEOBench-VLM, a comprehensive benchmark suite designed to evaluate VLMs (*generic and geospatial-specific*) on geospatial tasks (Fig. 1). GEOBench-VLM offers a diverse dataset with varying visual conditions and object scales supplemented with automated and manual annotations. It encompasses important categories such as *scene understanding*, *object counting*, *visual grounding*, *image captioning*, *temporal understanding*, *non-optical*, *referring segmentation*, and *relational reasoning* (see Fig. 2). These tasks are essential for a wide range of applications, including but not limited to ur-

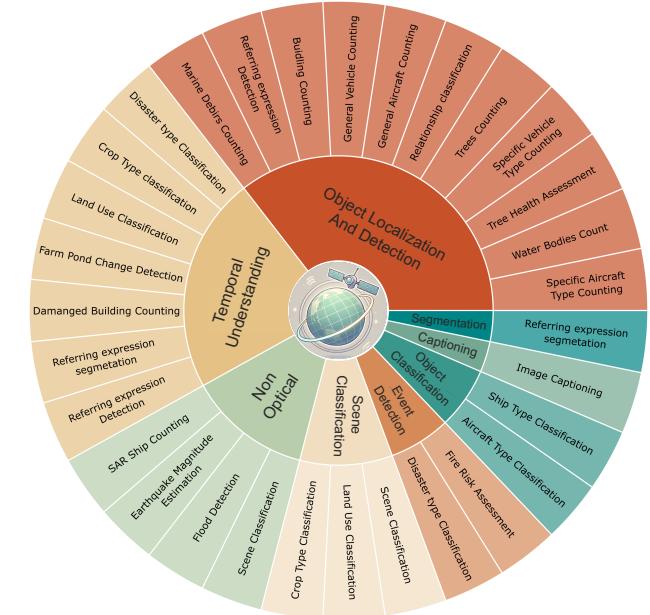


Figure 2. GEOBench-VLM comprehensively covers 31 fine-grained tasks categorized into 8 broad categories: scene and object classification, object detection, segmentation, captioning, event detection, non-optical and temporal understanding tasks.

ban planning, monitoring deforestation, assessing environmental changes, and managing natural disaster response.

We have evaluated several state-of-the-art VLMs using GEOBench-VLM to assess their performance in a geospatial context. Our findings suggest that while generic VLMs have potential, they struggle with geospatial tasks. We also include geospatial-specific VLMs in our evaluation to provide a comparative analysis of different models, including commercial ones like GPT-4o. Models perform differently across geospatial tasks; none excels in all areas. LLaVA-OneVision [23] leads in object localization and counting; GPT-4o excels in object classification and temporal understanding; Qwen2-VL [53] shows strength in event detection and interpreting non-optical SAR imagery. This comparison highlights the benefits of analyzing models for geospatial tasks and identifies areas needing improvement. We can summarize our contributions as follows:

1. We introduce GEOBench-VLM, a benchmark suite designed specifically for evaluating VLMs on geospatial tasks, addressing geospatial data challenges. It covers 8 broad categories and 31 sub-tasks with over 10,000 manually verified questions.
2. We provide a detailed evaluation of ten state-of-the-art VLMs, including generic (open and closed-source) and task-specific geospatial VLMs, highlighting their capabilities and limitations in handling geospatial tasks.
3. We analyze performance across a range of tasks, including scene classification, counting, change detection, relationship prediction, visual grounding, image caption-

Benchmark	Domain	Modalities	Data Sources	Answer Type	Annotation Type	Human Verify	Year	RS	Category
CulturalVQA[38]	General	O	Curated	FF	M	X	2024	N/A	
EXAMS-V[59]	General	O	Academic Exams	MCQ	M	-	2024	N/A	
M4U[51]	General	O	Academic Exams	MCQ	M	✓	2024	N/A	
MMMU[64]	General	O	Academic Exams	FF, MCQ	M	✓	2024	N/A	
MME[70]	General	O	Various Open-Source	Yes/No	M	✓	2024	N/A	
MMBench[33]	General	O	Various Open-Source	MCQ	A+M	X	2024	N/A	
MMStar[8]	General	O	Existing Benchmarks	MCQ	M	✓	2024	N/A	
LAMM[60]	General	O, PC	Various Open-Source	FF	A+M	X	2023	N/A	
SEED-Bench[20]	General	O, V	Various Open-Source	MCQ	A+M	✓	2023	N/A	
SEED-Bench2[22]	General	O, MI, V	Various Open-Source	MCQ	A+M	✓	2024	N/A	
SEED-Bench-H[22]	General	O, MI, V	Public Data / Curated	MCQ	M	✓	2024	N/A	
RSIEval[16]	RS	O, PAN	DOTA[55]	FF	M	✓	2023	6	
LHRS-Bench[36]	RS	O	GE + OSM	SC	M	✓	2024	4	
EarthGPT[68]	RS	O, IR, SAR	DRSD	FF, BBox	A+M	-	2024	5	
SkyEyeGPT[65]	RS	O, V	DRSD	FF, MCQ	A+M	✓	2024	6	
Fit-RSRC[34]	RS	O	DRSD	FF, MCQ	A+M	✓	2024	1	
Fit-RSFG[34]	RS	O	DRSD	FF, MCQ	A+M	X	2024	5	
VRSBench[26]	RS	O	DOTA[55], DIOR[24]	FF, BBox	A+M	✓	2024	3	
VLEO-Bench[66]	RS	O, BT	DRSD	FF, BBox, MCQ	A+M	✓	2024	6	
EarthVQA[52]	RS	O	DRSD	FF	A+M	X	2023	5	
RemoteCount[28]	RS	O	DOTA[55]	SC	A+M	✓	2024	1	
FineGrip[71]	RS	O	MAR20[54]	FF, Seg	A+M	✓	2024	5	
GeoChat-Bench[19]	RS	O	DRSD	FF, BBox	A+M	✓	2023	6	
GEOBench-VLM	RS	O, MS, SAR, BT, MT	DRSD	MCQ, BBox, Seg	A+M	✓	Ours	8	

Table 1. Overview of Generic and Geospatial-specific Datasets & Benchmarks, detailing modalities (O=Optical, PAN=Panchromatic, MS=Multi-spectral, IR=Infrared, SAR=Synthetic Aperture Radar, V=Video, MI=Multi-image, BT=Bi-Temporal, MT=Multi-temporal), data sources (DRSD=Diverse RS Datasets, OSM=OpenStreetMap, GE=Google Earth, answer types (MCQ=Multiple Choice, SC=Single Choice, FF=Free-Form, BBox=Bounding Box, Seg=Segmentation Mask), and annotation types (A=Automatic, M=Manual).

ing, segmentation, disaster detection, and temporal analysis, among others, providing key insights into improving VLMs for geospatial applications.

2. Benchmark Overview

Benchmarks for evaluating VLMs cover a wide range of tasks, including both generic and specialized geospatial applications. A detailed comparison of benchmarks, including their specific focus areas, is provided in Table 1.

Generic VLMs Benchmarks: Several benchmarks have been developed to evaluate multimodal models across diverse visual understanding domains, each with specific strengths and limitations. MMMU [64] assesses models across various disciplines with advanced visual formats and tests perceptual capabilities, but it lacks tasks unique to geospatial contexts. SEED-Bench [22] focuses on spatial and temporal understanding with diverse datasets for complex multimodal contexts. SEED-Bench-2 [21] specializes in text-rich visual scenarios like charts and maps but interprets structured, generalized maps rather than complex geospatial data. Also, MMBench [33], MM-Vet [62], and MMSTAR [8] evaluate VLMs on spatial reasoning and multimodal understanding; they lack specialized geospatial tasks relevant to earth observation applications.

Geospatial-specific: Few methods, alongwith benchmarks, including EarthVQA [52], LHRS-Bot [36], RS-LLaVA [5],

and GeoChat [19], evaluate VLMs in remote sensing. RS-LLaVA [5] evaluates VLMs on image captioning and VQA with high-resolution imagery but does not extend to tasks like change detection and temporal analysis. LHRS-Bot [36] focuses on high-resolution remote sensing tasks with an RS-specific image-text dataset but lacks multi-temporal diversity. EarthVQA [52] supports complex relational reasoning with a multi-modal dataset, though its emphasis on non-temporal imagery restricts broader applicability. GeoChat [19] also evaluates geospatial tasks like region captioning and spatially grounded responses; however, it lacks diverse temporal datasets and segmentation. VLEO [66] is an earth observation benchmark; however, it is limited to non-remote sensing specialized methods and lacks the diversity of tasks. Specifically, it lacks multi-temporal, segmentation, and non-optical data, limiting its analysis. A comprehensive geospatial benchmark is needed to evaluate VLM capabilities more effectively.

3. GEOBench-VLM

Our proposed benchmark is curated using a human-verified data pipeline. It covers 8 broad and 31 fine-grained geospatial task categories that are outlined below.

Categories: The diverse set of geospatial tasks in GEOBench-VLM are carefully designed to evaluate the capabilities of VLMs. We refer the reader to Fig. 3 for il-



Figure 3. Comprehensive benchmark for VLMs in numerous geospatial tasks. This benchmark evaluates VLMs across eight core task categories, assessing their ability to interpret complex spatial data, classify scenes, identify and localize objects, detect events, generate captions, segment regions, analyze temporal changes, and process non-optical data. Tasks span from classifying landscapes and objects (e.g., land use, crop types, ships, aircraft) to counting, detecting hazards, and assessing disaster impact, testing VLMs on spatial reasoning.

lustrations for these tasks explained next.

1) Scene Understanding: This category includes tasks like Scene and Land Use Classification, where models distinguish environments (e.g., airports, forests, bridges) and land types (e.g., agricultural, industrial, residential). Crop Type Classification further tasks models to identify crops using visual and environmental cues.

2) Object Classification: This category focuses on identifying specific objects, including Ship Type Classification (e.g., aircraft carrier, cargo ship) and Aircraft Type Classification (from civil transport to military bombers), assessing the model’s capability in fine-grained categories within high-resolution aerial views.

3) Object Localization and Detection: This includes Referring Expression Detection, where models predict bounding boxes based on text queries and Spatial Relationship Prediction. Counting tasks cover general and specific objects, examples include vehicle and aircraft counting, damaged and healthy tree counting, building and water body counting, and marine debris quantification.

4) Event Detection: This focuses on Fire Risk Detection and assesses fire hazards across forests, while Disaster-Type Classification involves inferring probable disaster causes from post-event imagery.

5) Caption Generation: This assesses image captioning by requiring models to capture both overall

scene context and specific object details.

6) Semantic Segmentation: Task includes Referring Expression Segmentation, where the model generates binary masks for specified objects or regions, such as urban and non-urban areas.

7) Temporal Understanding: This category includes tasks like Change Detection and Damage Assessment for identifying differences over time, such as post-disaster building conditions. Long Temporal Analysis includes tasks like Crop Type Classification.

8) Non-Optical Imagery: This category leverages non-optical imagery for tasks like SAR-based Ship Detection, Flood Detection, and Earthquake Magnitude Estimation, crucial for scenarios where optical imagery is insufficient.

Dataset pipeline: GEOBench-VLM is developed by integrating available open datasets, and manual annotation aided with automated tools. For diversity, each task samples images from multiple datasets. Fig. 4 shows the overall benchmark development pipeline.

For scene understanding tasks, including scene classification, land use classification, and crop type classification, we use classification datasets [10, 31, 35, 43, 45, 72]. We used GPT-4o to generate unique questions for each task, with five answer options: one correct answer, one semantically similar “closest” option (verified manually), and three

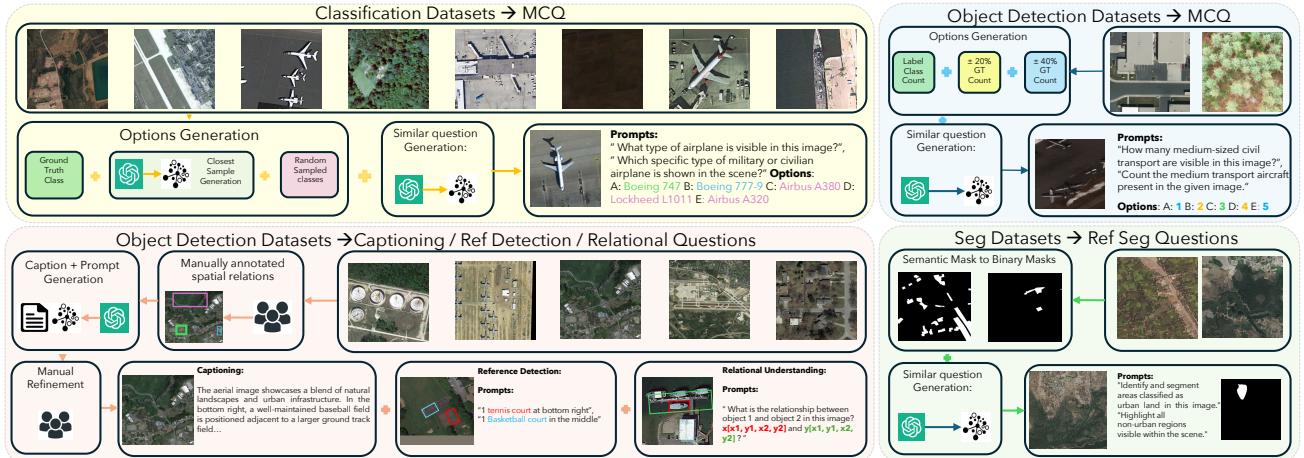


Figure 4. Data pipeline for the GEOBench-VLM: Our pipeline integrates diverse datasets, automated tools, and manual annotation. Tasks such as scene understanding, object classification, and non-optical analysis are based on classification datasets, while GPT-4o generates unique MCQs with five options: one correct answer, one semantically similar “closest” option, and three plausible alternatives. Spatial relationship tasks rely on manually annotated object pair relationships, ensuring consistency through cross-verification. Caption generation leverages GPT-4o, combining image, object details, and spatial interactions with manual refinement for high precision.

plausible alternatives. The answers for Object Classification, Non-Optical, and Fire Risk Assessment tasks were generated in a similar manner.

For counting tasks, we converted object detection data [12, 37, 46] into questions about the number of specific objects in an image, providing the correct count and alternatives with controlled deviations ($\pm 20\%$ and $\pm 40\%$) to maintain plausibility. For referring expression segmentation, we use segmentation datasets to create binary masks and prompts for identifying specific objects or regions. For spatial relationship tasks, we manually annotated relationships between object pairs using object locations from detection datasets [11, 47, 55], with cross-verification for consistency. Following our standard approach, five MCs were designed to assess these relationships.

For referring expression tasks, where existing datasets lack complex queries, our benchmark provides diverse questions. We collect basic object attributes (names, locations, colors) and annotated spatial relationships to generate queries that incorporate object details and interactions. Each query is reviewed for clarity and complexity. For caption generation, we use GPT-4o with image data, object attributes, and spatial relationships to create descriptions combining scene context and object details. Captions are manually refined for clarity, removing irrelevant or repetitive details.

4. VLMs Benchmarking

4.1. Selection of VLMs

Model Variability and Capabilities: We selected both generic and geospatial-specific VLMs, focusing on new models with advanced capabilities across various tasks.

Generic VLMs include LLaVA-1.5 [30], LLaVA-NeXT [29], LLaVA-OneVision [23], Sphinx [27], Ferret [61], InternVL2 [9], and Qwen2-VL [53]. These recent VLMs perform strongly in tasks like scene understanding, object recognition, and fine-grained visual classification. For geospatial-specific VLMs, we selected GeoChat [19] and RS-LLaVA [5], two models tailored for satellite and aerial image interpretation. Also, we added GPT-4o, a closed-source, commercially available model, to compare its adaptability to remote sensing tasks with open-source models.

Domain Specific Model Suitability: We align our model selection with intended application areas, assessing each VLM’s suitability for geospatial tasks. GeoChat [19] and RS-LLaVA [5], trained on satellite and aerial data, may face challenges with complex scene understanding and spatial relationships. However, can generic VLMs handle geospatial tasks like scene understanding, object detection, counting, and visual reasoning in remote sensing datasets? This targeted selection allows us to evaluate each VLM on complex tasks across various geospatial applications.

Open vs Closed Models: Open-source models like LLaVA [23, 29, 30], Qwen2-VL [53], and GeoChat [19] provide transparency, aiding in understanding their strengths and limitations. Closed models like GPT-4o, while lacking transparency, excel in generalization and complex tasks due to extensive training on proprietary datasets and advanced architectures. Including both open and closed models provides a comprehensive evaluation of VLM capabilities.

4.2. Benchmarking Approaches

Task Setup & Complexity Levels: To rigorously evaluate VLMs on diverse geospatial tasks, we curated a comprehensive dataset. Our benchmark includes tasks of varying

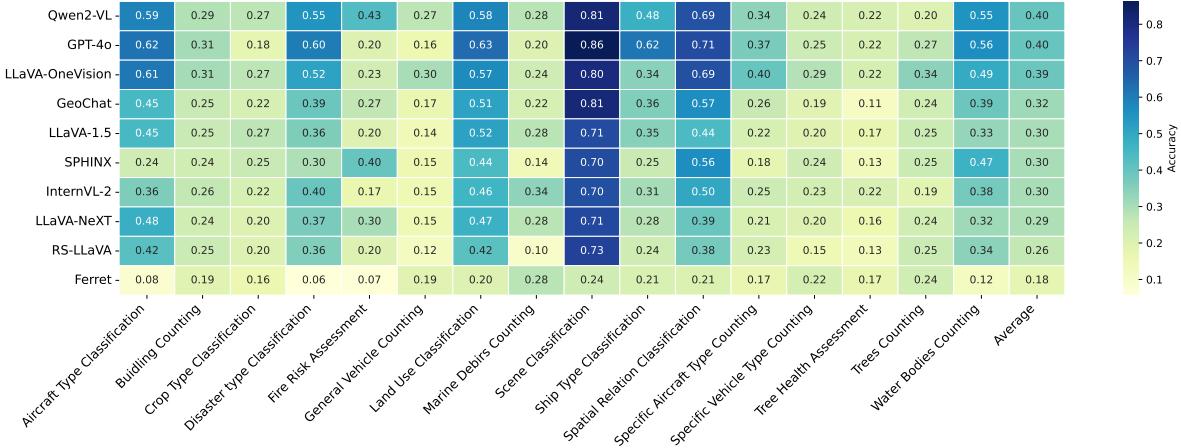


Figure 5. Performance Summary of VLMs Across Geospatial Tasks. GPT-4o achieves better accuracy in relatively easy tasks like Aircraft Type Classification, Disaster Type Classification, Scene Classification, and Land Use Classification. But, on average the best-performing GPT-4o achieves only **40%** accuracy on MCQs based on diverse geospatial tasks, which is only double the random guess performance. These results showcase the varying strengths of VLMs in addressing diverse geospatial tasks.

complexity (Fig. 3), from basic recognition assessing foundational capabilities to intermediate tasks like image captioning and object counting, which require descriptive and quantitative reasoning. Advanced tasks, like risk assessment, spatial relations, change detection, temporal classification, and temporal counting, demand interpretation of spatial dependencies, complex contextual cues, and temporal analysis. The dataset features rich annotations, temporal sequences, and segmentation data, enabling assessment of models across multiple dimensions and ensuring VLMs can handle the full spectrum of geospatial challenges.

Metrics: Our evaluation framework uses task-specific metrics to assess model performance comprehensively. For event detection, object classification, counting, and scene classification, we report accuracy based on multiple-choice questions (MCQs). Precision is used for referring expression detection, mean Intersection over Union (mIOU) for segmentation tasks, and ROUGE-L for evaluating the quality of generated image captions.

5. Evaluations and Results

In this section, we thoroughly evaluate and discuss the results for each task as follows.

Scene Understanding: Scene understanding in remote sensing encompasses agriculture, urban planning, and environmental monitoring. Models struggle with crop classification due to low-resolution imagery. GPT-4o excels in land use, and scene classification (see Fig. 5). Ferret [61] underperforms, reflecting limited geospatial focus.

Object Classification: Object type classification in remote sensing, such as identifying ship and aircraft types, is vital for maritime and airspace monitoring. GPT-4o likely achieves superior performance in object classification due

Model	Event Det	Object Cls	Counting	Scene Cls	Image Cap
Ferret [61]	0.0620	0.1463	0.2043	0.2032	0.1661
Sphinx [27]	0.3476	0.2456	0.2534	0.4665	0.1993
LLaVA-1.5 [30]	0.2806	0.3998	0.2534	0.5002	0.1880
LLaVA-NeXT [29]	0.3328	0.3808	0.2425	0.4613	0.1868
LLaVA-OneV [23]	0.3744	0.4773	0.3648	0.5478	0.1745
InternVL-2 [9]	0.2816	0.3329	0.2782	0.4585	0.1427
Qwen2-VL [53]	0.4920	0.5329	0.3413	0.5567	0.1672
RS-LLaVA [5]	0.2784	0.3289	0.2155	0.4483	0.0754
GeoChat [19]	0.3294	0.4046	0.2661	0.5142	0.1040
GPT-4o [39]	0.3996	0.6183	0.3373	0.5595	0.1784

Table 2. VLMs Accuracies Across Geospatial Tasks. Evaluation includes Event Detection, Object Classification, Counting, Scene Classification, and Image Captioning. Qwen2-VL excels in Event Detection, GPT-4o leads in Object and Scene Classification, LLaVA-OneV(ision) performs well in Counting, and Sphinx performs well in Image Captioning evaluated with ROUGE-L scores. Our fine-grained captions and the inherent complexity of the benchmark present significant challenges for VLMs.

to its optimized architecture for high-resolution images, while Ferret [61] performs the worst, focusing more on spatial localization and region-based tasks over object classification. For detailed results, we refer to Fig. 5.

Object Localization & Detection: In remote sensing, this category focuses on locating and counting objects, which is vital for identifying objects such as water bodies, buildings, trees, etc. This task is particularly challenging due to diverse scene content, varying object scales, and the high spatial resolution required for precise detection and counting. Here, we frame counting tasks as classification across multiple answer choices.

LLaVA-OneVision [23] consistently performs best in tasks such as *building/vehicle/tree counting, specific vehicle/aircraft type counting and tree health assessment*. It also performs well in *marine debris counting and spatial relation classification*, demonstrating its capability for high-

Model	Crop Type Cls	Disaster Type Cls	Farm Pond CD	Land Use Cls	Damaged Bldg Cnt
LLaVA-OneV [23]	0.1273	0.4493	0.1579	0.5672	0.2139
Qwen2-VL [53]	0.1273	0.5903	0.0921	0.5869	0.2270
GPT-4o	0.1818	0.6344	0.1447	0.6230	0.2420

Table 3. Results highlight the strengths of VLMs in handling temporal geospatial challenges. Evaluation across five tasks: Crop Type Classification, Disaster Type Classification, Farm Pond Change Detection (CD), Land Use Classification, and Damaged Building Counting. GPT-4o achieves the highest accuracy overall in classification and counting tasks.

resolution adaptation, and fine detail capture. These capabilities enable it to interpret complex scenes and spatial relationships. Despite its remote sensing focus, RS-LLaVA [5] underperforms in *general and specific vehicle counting tasks*, while GeoChat [19] shows lower performance in *tree health assessment*. This is likely because RS-LLaVA and GeoChat are optimized for region-based queries and spatially grounded dialogue rather than precise object counting and localization in more complex scenes. In addition to Fig. 5 results, we present the precision for the referring expression detection task in Table 4 within this category. These results demonstrate that Sphinx achieves the highest precision scores at *IoU: 0.25 & 0.50* compared to GeoChat, Qwen2-VL, Ferret, and GPT-4o. A lower IoU value captures the model’s approximate localization ability and its effectiveness in interpreting task-specific instructions, while an IoU of 0.5 reflects the standard setting for this task.

Event Detection: Event detection focuses on identifying and assessing disasters or hazards critical for environmental monitoring and risk management. Tasks like fire risk assessment and disaster type classification are essential for environmental management. Qwen2-VL [53] performs best in event detection, benefiting from advanced architectural and training optimizations. Its enhanced attention mechanisms, enable it to efficiently process complex, large-scale imagery and maintain contextual relevance over long inputs.

Caption Generation: In this category, we include an image captioning task in our benchmark, and the goal of this task is to generate descriptive captions that reflect visual content. Sphinx [27] achieves the highest ROUGE score in image captioning, followed closely by LLaVA-1.5 [30] and LLaVA-NeXT [29]. In contrast, RS-LLaVA [5] and GeoChat [19] have the lowest scores (Table. 2). Our captions include both fine-grained and coarser details, since Sphinx is trained with fine-grained details, it performs relatively better compared to other models.

Semantic Segmentation: Including the referring segmentation task in our benchmark enhances the evaluation of a model’s ability to perform fine-grained, query-driven segmentation in remote sensing. This task tests the models in interpreting specific object references in complex scenes, which is essential for land-use mapping and urban analy-

Model	Prec@0.5	Prec@0.25
Sphinx [27]	0.3408	0.5289
GeoChat [19]	0.1151	0.2100
Ferret [61]	0.0943	0.2003
Qwen2-VL [53]	0.1518	0.2524
GPT-4o [39]	0.0087	0.0386

Table 4. Referring expression detection. We report Precision on 0.5 IoU and 0.25 IoU.

Model	EQME	LU Cls
RS-LLaVA [5]	0.0863	0.3123
GeoChat [19]	0.1403	0.3156
Qwen2-VL [53]	0.2734	0.2525
GPT-4o [39]	0.0827	0.3256

Table 5. Performance comparison of different models on Non-Optical tasks. EQME: Earthquake Magnitude Estimation

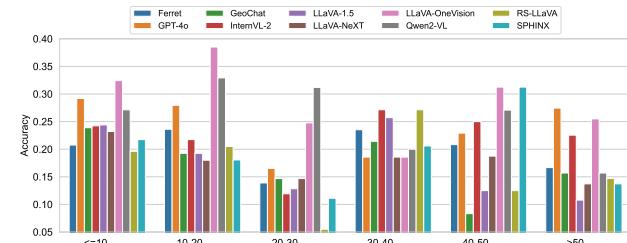


Figure 6. Object Density and Model Counting Accuracy. VLMs are assessed on their ability to maintain accuracy as object density increases. LLaVA-OneVision leads with balanced performance across all ranges. Qwen2-VL and InternVL2 show strengths in mid and high-density ranges, respectively

sis. While no remote sensing-specific models support this task yet, non-specialized models like GlaMM [40], with a baseline mIoU of 0.1411. By integrating referring segmentation, we improve our framework to identify models that can adapt to spatial queries in remote sensing segmentation.

Temporal Understanding: GEOBench-VLM also incorporates temporal tasks to analyze changes over time, an important aspect of real-world remote sensing applications. Here, we evaluate models with temporal capabilities as shown in Table 3, specifically GPT-4o, Qwen2-VL [53], and LLaVA-OneVision [23], selected for their versatility across tasks. GPT-4o emerges as the top performer across most temporal tasks in disaster type, crop type, and land use classification. LLaVA-OneVision demonstrates the highest performance in farm pond change detection, indicating its ability to perform tasks requiring more minute-level change detection over time. While Qwen2-VL also performs well, especially in disaster-type classification, it follows GPT-4o in overall temporal task performance.

Non-Optical: In the non-optical synthetic aperture radar (SAR) imagery, we include tasks for earthquake magnitude estimation and land use classification. In Table 5, results indicate that Qwen2-VL [53] achieves the highest score for earthquake magnitude estimation, showcasing its strength in interpreting SAR imagery, while GPT-4o ranks the lowest. For land use classification, GPT-4o performs best. GeoChat [19] and RS-LLaVA [5] perform relatively better in land use classification but show limitations in accurately estimating earthquake magnitudes.

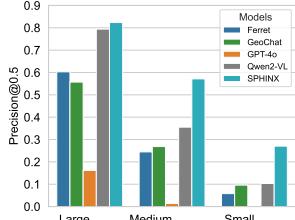


Figure 7. Performance comparison of different models on Referring Expression Detection across various object sizes.

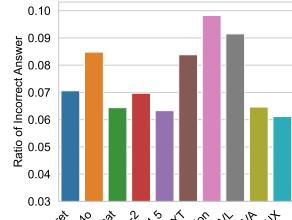


Figure 8. Percentage of incorrect answers falling within the 20% error range.

6. Analysis

Counting Accuracy by Object Count Range: The object count ranges, from ≤ 10 to > 50 , classify model performance based on object density within scenes, from sparse to highly populated areas (see Fig. 6). These ranges highlight each model’s capability to maintain counting accuracy as object density and scene complexity increase. LLaVA-OneVision [23] demonstrates robust adaptability across low and high object counts, while Qwen2-VL [53] shows competitive performance, particularly in mid-range counts. In the highest density ranges, GeoChat [19] shows limitations due to its focus on region-specific tasks rather than handling large object clusters. GPT-4o has variable results, performing well in mid-ranges but struggling in dense scenes. Results show the importance of robust spatial alignment and fine-grained detail, as seen in LLaVA-OneVision [23], for consistent counting accuracy across object densities.

Incorrect Answer Percentage by Option Range Distribution: To analyze incorrect answers by option range distribution, we observe that as alternative options deviate from the ground truth within 20% and beyond, the rate of incorrect responses increases significantly (Fig. 8). Models like GPT-4o, LLaVA-OneVision [23], and Qwen2-VL [53] show high sensitivity. It suggests that models that are trained on diverse datasets can sometimes focus too much on small details. This may result in an increased number of mistakes when the distribution of options varies.

Prompt-Based Performance Variance: We find that models exhibit different sensitivity levels to prompt variations, affecting task accuracy (Fig. 9). InternVL2 [9] and GPT-4o show slight instability across five prompts/task. This indicates degree of sensitivity to specific phrasing or structure of prompts. In contrast, GeoChat [19], Qwen2-VL [53], and LLaVA-OneVision [23] consistently perform better across prompts, indicating their robustness to prompt variation.

Single vs. Multi-Temporal Data: For single vs. multiple temporal data analysis (Fig. 10), we evaluate three models: GPT-4o, LLaVA-OneVision [19], and Qwen2-VL [53], across three tasks: (i) crop type classification, (ii) disaster type classification, and (iii) land use classification. For

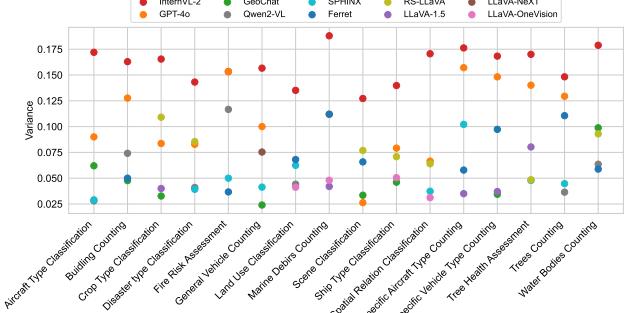


Figure 9. Accuracy Variance across different prompts.

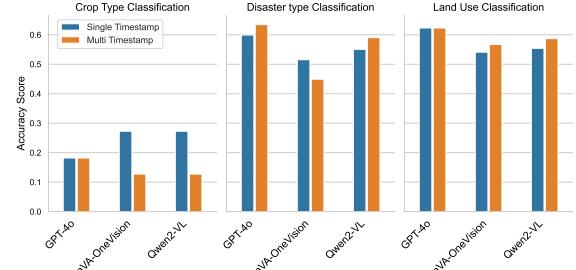


Figure 10. Single image vs Multi-temporal data comparison.

crop type classification, multi-timestamp data generally performs worse than single-timestamp data, likely due to variability in seasonal crop appearances. In disaster-type classification, multiple timestamps improve model performance, though LLaVA-OneVision shows lower performance. For land use classification, all models perform well with multi-timestamp data, suggesting that land use features are relatively stable and less affected by temporal changes. Unlike crop type classification, which is seasonally variable, land use categories (e.g., urban, forest) remain constant and have recognizable patterns across different timestamps.

Impact of Object Size on Detection Performance: In referring object detection across sizes, five models were evaluated (Fig. 7). GPT-4o performed worst, struggling with accurate detection and localization, while Sphinx [27] excelled due to its fine-grained processing. Qwen2-VL [53] followed, benefiting from robust attention mechanisms.

7. Conclusion

While generic VLMs have seen much progress in developing comprehensive benchmarks covering various reasoning aspects, the earth-observation domain lacks large-scale manually verified benchmarks covering diverse capabilities. Unlike existing benchmarks that lack the unique requirements of geospatial data, GEOBench-VLM is designed specifically for 31 remote-sensing tasks like flood detection, disaster monitoring, crop classification, marine debris counting, and temporal change segmentation. GEOBench-VLM covers a range of computer vision tasks, ranging

from referring expression segmentation to counting, detection, recognition, and temporal analysis. Our benchmark, featuring over 10,000 manually verified instructions across diverse conditions, highlights the limitations of current VLMs in geospatial contexts, even for state-of-the-art models. Through detailed experiments with the 10 best-performing VLMs, we show the current gaps that need to be addressed to unlock several remote-sensing applications.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Swedish Forest Agency. Forest damages – larch casebearer 1.0, 2021. National Forest Data Lab. Dataset. 14
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2
- [4] Gerald Baier, Antonin Deschamps, Michael Schmitt, and Naoto Yokoya. Synthesizing optical and sar imagery from land cover maps and auxiliary raster data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. 14
- [5] Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9):1477, 2024. 3, 5, 6, 7, 13
- [6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arxiv. arXiv preprint arXiv:2303.12712*, 2023. 1
- [7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 2
- [8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 3
- [9] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2, 5, 6, 8
- [10] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 4, 13, 14
- [11] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *CoRR*, abs/2110.01931, 2021. 5, 13, 14
- [12] CSE499DeforestationSatellite. Deforestation-satellite-imagery dataset. <https://universe.roboflow.com/cse499deforestationsatellite/deforestation-satellite-imagery-335n4>. 5,
- [13] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *CoRR*, abs/1805.06561, 2018. 14
- [14] Yanghua Di, Zhiguo Jiang, and Haopeng Zhang. A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sensing*, 13(4), 2021. 14
- [15] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric T. Heim, Howie Choset, and Matthew E. Gaston. xbd: A dataset for assessing building damage from satellite imagery. *CoRR*, abs/1911.09296, 2019. 13, 14
- [16] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023. 3
- [17] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. 1
- [18] Hannah Kerner, Snehal Chaudhari, Aninda Ghosh, Caleb Robinson, Adeel Ahmad, Eddie Choi, Nathan Jacobs, Chris Holmes, Matthias Mohr, Rahul Dodhia, Juan M. Lavista Ferres, and Jennifer Marcus. Fields of the world: A machine learning benchmark dataset for global agricultural field boundary segmentation, 2024. 13, 14
- [19] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 2, 3, 5, 6, 7, 8
- [20] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 3
- [21] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 3
- [22] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 2, 3
- [23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 5, 6, 7, 8, 13

- [24] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 3
- [25] Kaiyu Li, Xiangyong Cao, and Deyu Meng. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. 2
- [26] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12384*, 2024. 3
- [27] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 5, 6, 7, 8, 13
- [28] Fan Liu, Delong Chen, Zhangqiyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3
- [29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5, 6, 7
- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5, 6, 7, 13
- [31] Jinglei Liu, Weixun Zhou, Haiyan Guan, and Wenzhi Zhao. Similarity learning for land use scene-level change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:6501–6513, 2024. 4, 14
- [32] Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26658–26668, 2024. 2
- [33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 2, 3
- [34] Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024. 3
- [35] Gabriel L. S. Machado, Edemir Ferreira, Keiller Nogueira, Hugo N. Oliveira, Pedro H. T. Gama, and Jefersson A. dos Santos. Airound and cv-brct: Novel multi-view datasets for scene classification. *CoRR*, abs/2008.01133, 2020. 4, 13, 14
- [36] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. *arXiv preprint arXiv:2402.02544*, 2024. 3
- [37] T. Nathan Mundhenk, Goran Konjevod, Wesam A. Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *Computer Vision – ECCV 2016*, pages 785–800, Cham, 2016. Springer International Publishing. 5, 14
- [38] Shravan Nayak, Kanishk Jain, Rabiu Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024. 3
- [39] OpenAI. Hello gpt-4o, 2024. 6, 7
- [40] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdellah Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 2, 7
- [41] Daniele Rege Cambrin and Paolo Garza. Quakeset: A dataset and low-resource models to monitor earthquakes through sentinel-1. *Proceedings of the International IS-CRAM Conference*, 2024. 13, 14
- [42] Marc Rußwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pages 200–201, 2020. 1
- [43] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *ICCV*, 2021. 4, 13, 14
- [44] A. Shah, L. Thomas, and M. Maskey. Marine debris dataset for object detection in planetscope imagery, 2021. 14
- [45] Shuchang Shen, Sachith Seneviratne, Xinye Wanyan, and Michael Kirley. Firerisk: A remote sensing dataset for fire risk assessment with benchmarks using supervised and self-supervised learning, 2023. 4, 14
- [46] Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. Rareplanes dataset, 2020. 5, 14
- [47] Xian Sun, Peijin Wang, Zhiyuan Yan, F. Xu, Ruiping Wang, W. Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, M. Weinmann, S. Hinz, Cheng Wang, and K. Fu. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *Isprs Journal of Photogrammetry and Remote Sensing*, 2021. 5, 13, 14
- [48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1
- [50] Chintan Tundia, Rajiv Kumar, Om Damani, and G. Sivakumar. Fpcd: An open aerial vhr dataset for farm pond change detection. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, page 862–869.

- SCITEPRESS - Science and Technology Publications, 2023. 13, 14
- [51] Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*, 2024. 3
- [52] Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5481–5489, 2024. 2, 3
- [53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 5, 6, 7, 8, 13, 14
- [54] YU Wenqi, CHENG Gong, WANG Meijun, YAO Yanqing, XIE Xingxing, YAO Xiwen, and HAN Junwei. Mar20: A benchmark for military aircraft recognition in remote sensing images. *National Remote Sensing Bulletin*, 27(12):2688–2696, 2024. 3
- [55] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Beßlongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 3, 5, 13, 14
- [56] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 2
- [57] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 1
- [58] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [59] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the vision: Geo-diverse visual commonsense reasoning. *arXiv preprint arXiv:2109.06860*, 2021. 3
- [60] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [61] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 5, 6, 7, 13
- [62] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 3
- [63] Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. Rrsis: Referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2
- [64] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruopi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 2, 3
- [65] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *arXiv preprint arXiv:2401.09712*, 2024. 3
- [66] Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. *arXiv preprint arXiv:2401.17600*, 2024. 2, 3
- [67] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 1
- [68] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. 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. 3
- [69] Xiaokang Zhang, Weikang Yu, Man-On Pun, and Wenzhong Shi. Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:1–17, 2023. 13, 14
- [70] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qing-song Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. 3
- [71] Danpei Zhao, Bo Yuan, Ziqiang Chen, Tian Li, Zhuoran Liu, Wentao Li, and Yue Gao. Panoptic perception: A novel task and fine-grained dataset for universal remote sensing image interpretation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3
- [72] Weixun Zhou, Shawn D. Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *CoRR*, abs/1706.03424, 2017. 4, 14
- [73] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1
- [74] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle,

Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020. [13](#), [14](#)

GEOBench-VLM: Benchmarking Vision-Language Models for Geospatial Tasks

Supplementary Material

This supplementary material includes the dataset table (S1) and quantitative results to illustrate multiple cases for assessing model responses (S2). It also provides results on multispectral images (S3) and examines the influence of temperature variations on the model (S4), along with a comparison between bi-temporal and multi-temporal approaches (S5). Additionally, a geographical analysis is included (S6) to observe the span of data coverage across locations, followed by a detailed description of the word cloud (S7).

S1. Datasets

The datasets we use in our evaluation cover a wide range of geospatial tasks, showing the variety and depth of challenges in geospatial analysis. As shown in Table 6, these datasets include tasks like scene understanding, spatial relation, instance counting, temporal understanding, referring expression segmentation, and working with non-optical data. This diversity enables us to create versatile question-answer pairs tailored to each specific task. The inclusion of datasets from recent years ensures that our evaluation tackles recent challenges and uses up-to-date information.

These datasets also offer a rich variety of annotation types, sensor data, and spatial resolutions, reflecting the diverse nature of geospatial data. The annotation types range from class labels and bounding boxes to semantic and instance masks, giving different levels of detail for model evaluation. The sensor data includes RGB images, Multispectral Imaging (MSI), and Synthetic Aperture Radar (SAR), with resolutions from fine to coarse scales. This heterogeneity allows us to test models under different imaging conditions and resolutions, fostering robustness and generalizability. For example, datasets like FAIR1M [47], DIOR [11], and DOTA [55] provide high-resolution RGB images with bounding box annotations, which are critical for tasks like object detection and understanding spatial relationships in complex scenes. Temporal understanding datasets such as fMoW [18], xBD[15], PASTIS[43], FPCD[50], and GVLM[69] are crucial for tracking changes over time, helping in tasks such as disaster assessment and monitoring urban development. Non-optical datasets like So2Sat [74] and QuakeSet [41] introduce SAR data, expanding our analysis to situations where optical imagery isn't available due to weather or lighting conditions. Scene Understanding datasets like AiRound [35] and RESICS45 [10] offer class annotations that help categorize large-scale scenes, essential for land use and land cover classification.

S2. Qualitative Results

The images in Fig. 11 show patterns in how the models performed on geospatial tasks relevant to scene understanding. Models perform well in identifying scenes with distinct features, such as “interchange” and “solar farm”, where most models succeeded except Ferret [61], RS-LLaVA [5], and Sphinx [27], respectively. In the third image, all models except Ferret correctly identified the “stadium”, demonstrating notable contextual understanding. For the fourth image, only a few models correctly identified “mixed cereal” crops, with failures attributed to the ambiguous nature of crop patterns. The first image in the second row shows dense greenery, indicating a moist environment, with the fire risk correctly classified as “low”, LLaVA-1.5 [30] and Ferret misclassified the risk, likely overestimating fire hazard due to vegetation density. This highlights the need for better contextual reasoning in models. The second image in the second row benefits from clear context, aiding classification as a water treatment facility. In contrast, the third image in the same row lacks context, making it prone to misclassification. This comparison highlights the importance of contextual information for accurate scene classification. In the last image of Fig. 11, ambiguous scenes such as the “ferry terminal” where all models failed, the misclassification is likely due to overlapping visual cues. The visual similarity between a harbor and a ferry terminal makes it challenging for models to differentiate between these categories. For the counting tasks in Fig.12, Qwen 2-VL [53] and LLaVA-One [23] (overall, Fig. 5 main paper) performed relatively better by correctly recognizing objects such as a single pickup truck and a single water body. In contrast, most models struggled, with Ferret overestimating due to difficulty in differentiating objects.

Fig. 13 shows that the models performed well in the first two images, probably due to familiar contextual clues. The “atago-class destroyer” and “small civil transport/utility” aircraft are common object types with distinct characteristics, making them easier for models to recognize. However, in the last two images, none of the models successfully identified the “murasame-class destroyer” or “garibaldi aircraft carrier” which are rarer categories. The failure is likely due to insufficient exposure to these specific classes in training datasets, coupled with the overlapping features of the objects that require advanced fine-grained recognition.

As shown in Fig. 14, models performed well on disaster assessment with relatively clear indicators, such as “fire” damage, where only Sphinx [27] and Ferret [61] failed. For the second image, depicting “flooding”, Ferret struggled.

Name	Task	Annotation Type	Sensor (Res)	Year	
AiRound[35]			RGB, Sentinel-2 (10m)	2020	
RESICS45[10]			RGB	2017	
PatternNet[72]	Scene Understanding, Object Classification	Class	RGB	2018	
MtSCCD[31]			RGB (1m)	2024	
FireRisk[45]			RGB (1m)	2023	
FGSCR[14]			RGB	2021	
FAIR1M[47]	Spatial Relation Classification, Referring Expression Detection, Captioning	Bounding Box	RGB (0.3–0.8m)	2021	
DIOR[11]			RGB	2020	
DOTA[55]			RGB (0.1–1m)	2021	
Forest Damage[2]			RGB	2021	
Deforestation[12]			RGB	2024	
COWC[37]	Counting	Bounding Box	RGB (15 cm)	2016	
NASA Marine Debris[44]			RGB (3m)	2024	
The RarePlanes Dataset[46]			RGB (0.3m)	2020	
fMoW[18]		Class	RGB (1m)	2018	
xBD[15]		Bounding Box, Instance Mask, Class	RGB (0.8m)	2019	
PASTIS[43]			Semantic Mask	MSI (10m)	2021
FPCD[50]			Semantic Mask	RGB (1m)	2022
GVLM[69]			Class	RGB (0.6m)	2023
DeepGlobe Land Cover[13]	Referring Expression Segmentation	Semantic Mask	RGB (0.5m)	2018	
GeoNRW[4]			RGB (1m)	2021	
So2Sat[74]	Non-Optical	Class	SAR, MSI (10m)	2020	
QuakeSet[41]		Number	SAR (10m)	2024	

Table 6. Comprehensive overview of geospatial datasets utilized for evaluating Vision-Language Models (VLMs) across diverse tasks, including Scene Understanding, Spatial Relation, Object Classification, Spatial Relation Classification, Referring Expression Detection, Captioning, Temporal Understanding, Referring Expression Segmentation, and Non-Optical tasks. The datasets are categorized by annotation types (e.g., class, bounding box, semantic mask) and sensor types (e.g., RGB imagery, Multispectral Imaging (MSI), Synthetic Aperture Radar (SAR)), highlighting their versatility for a wide range of geospatial applications.

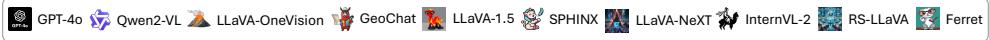
The third image depicts “tsunami” damage, characterized by disrupted layouts, scattered debris, and damaged buildings, which are often visually similar to flooding. Models may misclassify this due to overlapping features, insufficient tsunami-specific training data, and a lack of contextual reasoning. For the last image, only Qwen2VL identified the “seismic activity”, as others likely misclassified it due to overlapping features with “precipitation-related events”.

In Fig. 15, for the first image, models performed well because the objects are close to each other, easy to identify, and have minimal visual complexity. In the remaining images, models struggled because the objects were farther apart, making it harder to identify their spatial relationships. The cluttered environments and larger spatial gaps made it difficult for the models to accurately understand the relationships between the objects. Fig. 18 shows an aerial image alongside its ground truth caption and responses from

various models. The ground truth provides a detailed and accurate description of the scene, while the model generated captions vary in capturing key elements such as urban & natural features, pathways, and architectural structures. This comparison highlights differences in model responses for image captioning tasks.

S3. Multi-spectral

In this section, we compare how GPT-4o and Qwen2-VL [53] perform on Crop Type Classification and Land Use Classification tasks using RGB and multispectral (MS) data (Fig. 16). The models perform much better with RGB inputs because they are designed and trained specifically for RGB images. The accuracy drops significantly for multispectral data, especially in crop-type classification. To use MS data with these models, Sentinel-2 bands were com-



What type of facility or structure is depicted in this image? What type of facility or structure is depicted in this image? What is the primary type of land use visible in this aerial image? Which crop is primarily cultivated in this area?



- | | | | | | | | |
|----------------------------|----------------|---------------------|----------------------|------------|------------|---------------------|--------------------|
| A. Single-unit residential | B. Lighthouse | A. Crop field | B. Military facility | A. Airport | B. Statue | A. Winter triticale | B. Winter rapeseed |
| C. Road bridge | D. Interchange | C. Debris or rubble | D. Solar farm | C. Park | D. Stadium | C. Beet | D. Mixed cereal |
| E. Nuclear power plant | | E. Toll booth | | E. Tower | | E. Void label | |

	Interchange	✓		Solar farm	✓		Stadium	✓		Void label	✗
	Interchange	✓		Solar farm	✓		Stadium	✓		Mixed cereal	✓
	Interchange	✓		Solar farm	✓		Stadium	✓		Winter triticale	✗
	Interchange	✓		Solar farm	✓		Stadium	✓		Mixed cereal	✓
	Road bridge	✗		Solar farm	✓		Stadium	✓		Winter rapeseed	✗
	Interchange	✓		Solar farm	✓		Stadium	✓		Mixed cereal	✓
	Interchange	✓		Solar farm	✓		Stadium	✓		Winter triticale	✗
	Road bridge	✗		Solar farm	✓		Stadium	✓		Mixed cereal	✓
	Instruction not followed	✗		Instruction not followed	✗		Airport	✗		Winter triticale	✗

What is the level of fire risk depicted in this image? What type of facility or structure is depicted in this image? What is the primary type of scene depicted in this aerial image? What is the primary type of scene depicted in this aerial image?



- | | | | | | | | |
|-----------------|-------------|-----------------------------|-------------------|-------------------------------|-----------------|-----------------|------------------------|
| A. High | B. Low | A. Crop field | B. Dam | A. Ferry terminal | B. Oil well | A. Nursing home | B. Ferry terminal |
| C. Non-burnable | D. Very low | C. Lighthouse | D. Railway bridge | C. Storage tank | D. Tennis court | C. Harbor | D. Christmas tree farm |
| E. Very High | | E. Water treatment facility | | E. Wastewater treatment plant | | E. Tennis court | |

	Low	✓		Water treatment facility	✓		Wastewater treatment plant	✓		Harbor	✗
	Low	✓		Water treatment facility	✓		Wastewater treatment plant	✓		Harbor	✗
	Low	✓		Water treatment facility	✓		Storage tank	✗		Harbor	✗
	High	✗		Water treatment facility	✓		Storage tank	✗		Harbor	✗
	Low	✓		Water treatment facility	✓		Storage tank	✗		Harbor	✗
	Low	✓		Water treatment facility	✓		Storage tank	✗		Harbor	✗
	Non-burnable	✗		Water treatment facility	✓		Storage tank	✗		Harbor	✗
	Low	✓		Dam	✗		Storage tank	✗		Harbor	✗
	High	✗		Lighthouse	✗		Instruction not followed	✗		Harbor	✗

Figure 11. Scene Understanding: This illustrates model performance on geospatial scene understanding tasks, highlighting successes in clear contexts and challenges in ambiguous scenes. The results emphasize the importance of contextual reasoning and addressing overlapping visual cues for accurate classification.

bined into three channels sequentially to mimic RGB inputs. For land use classification, which depends more on spatial patterns than detailed spectral information, the drop in performance is smaller. These results show the need for improved methods to adapt MS data for such tasks.

S4. Effect of Temperature

In VLMs, increasing the temperature parameter during sampling introduces variability, allowing the model to explore a broader range of responses. Coupled with multiple runs

Model Comparison: Counting Tasks								
GPT-4o			Qwen2-VL			LLaVA-OneVision		
How many pickup trucks are visible in this image?	A. 0	B. 1	How many water bodies can you identify in this image?	A. 3	B. 2	How many buildings can you identify in this image?	A. 168	B. 140
C. 4	D. 3	E. 2	C. 0	D. 4	E. 1	C. 84	D. 196	E. 112
	1	✓		1	✓		84	✗
	1	✓		1	✓		196	✗
	1	✓		1	✓		196	✗
	3	✗		3	✗		196	✗
	0	✗		3	✗		168	✗
	4	✗		0	✗		168	✗
	2	✗		1	✓		168	✗
	3	✗		4	✗		168	✗
	0	✗		3	✗		168	✗
	4	✗		2	✗		168	✗
A. 69	B. 52	E. 86	C. 103	D. 120				
How many medium-sized civil transport or utility aircraft are visible in this image?	A. 5	B. 2	How many large civil transport or utility aircraft can you spot in this image?	A. 4	B. 2	How many trees show light damage in this image?	A. 30	B. 18
C. 6	D. 4	E. 3	C. 3	D. 5	E. 1	C. 24	D. 36	E. 42
	3	✓		3	✗		24	✗
	3	✓		4	✗		24	✗
	4	✗		5	✗		24	✗
	5	✗		4	✗		18	✗
	5	✗		5	✗		18	✗
	5	✗		3	✗		18	✗
	5	✗		2	✓		24	✗
	5	✗		5	✗		18	✗
	6	✗		3	✗		30	✓
	5	✗		4	✗		24	✗
A. 3	B. 4	E. 2	C. 0	D. 1				
How many pieces of marine debris are visible in this image?	A. 0	B. 4						
C. 0	D. 1							

Figure 12. Counting: The figure showcases model performance on counting tasks, where Qwen 2-VL, GPT-4o and LLaVA-One have better performance in identifying objects. Other models, such as Ferret, struggled with overestimation, highlighting challenges in object differentiation and spatial reasoning.

and argmax aggregation, this method could assess the stability of certain models. For this analysis, we fix a prompt for multiple runs for a fair comparison. Tasks with inherent ambiguity, like counting (e.g., water bodies and specific vehicle types) and spatial relation classification, showed some

improvement. However, tasks such as ship type classification and tree health assessment showed no improvement. Mixed results are observed in tree counting. These findings suggest that multiple runs may be effective for tasks with certain ambiguities (Fig. 19 & Fig. 20).

 GPT-4o	 Qwen2-VL	 LLaVA-OneVision	 GeoChat	 LLaVA-1.5	 SPHINX	 LLaVA-NeXT	 InternVL-2	 RS-LLaVA	 Ferret
What type of ship is visible in this image?		What type of aircraft is visible in this image?			What type of ship is visible in this image?			What type of ship is visible in this image?	
									
A. Atago-class destroyer	B. Type 45 destroyer	A. Military Bomber	B. Medium Civil Transport/Utility	A. Murasame-class destroyer	B. Kongo-class destroyer	A. Civil yacht	B. INS Vikramaditya carrier		
C. Mega yacht	D. Mega yacht	C. Military Trainer	D. Small Civil Transport/Utility	C. Civil yacht	D. Arleigh Burke-class destroyer	C. Atago-class destroyer	D. Garibaldi aircraft carrier		
E. Mistral-class amphibious assault ship		E. Large Civil Transport/Utility		E. Kitty Hawk-class aircraft carrier		E. Kitty Hawk-class aircraft carrier			
 Atago-class destroyer	✓	 Small Civil Transport/Utility	✓	 Kitty Hawk-class aircraft carrier	✗	 INS Vikramaditya carrier	✗		
 Atago-class destroyer	✓	 Small Civil Transport/Utility	✓	 Kitty Hawk-class aircraft carrier	✗	 Kitty Hawk-class aircraft carrier	✗		
 Atago-class destroyer	✓	 Small Civil Transport/Utility	✓	 Kitty Hawk-class aircraft carrier	✗	 INS Vikramaditya carrier	✗		
 Atago-class destroyer	✓	 Small Civil Transport/Utility	✓	 Kitty Hawk-class aircraft carrier	✗	 Kitty Hawk-class aircraft carrier	✗		
 Atago-class destroyer	✓	 Small Civil Transport/Utility	✓	 Kitty Hawk-class aircraft carrier	✗	 Kitty Hawk-class aircraft carrier	✗		
 Type 45 destroyer	✗	 Small Civil Transport/Utility	✓	 Arleigh Burke-class destroyer	✗	 INS Vikramaditya carrier	✗		
 Atago-class destroyer	✓	 Small Civil Transport/Utility	✓	 Kitty Hawk-class aircraft carrier	✗	 Kitty Hawk-class aircraft carrier	✗		
 Atago-class destroyer	✓	 Small Civil Transport/Utility	✓	 Arleigh Burke-class destroyer	✗	 Kitty Hawk-class aircraft carrier	✗		
 Instruction not followed	✗	 Small Civil Transport/Utility	✓	 Kitty Hawk-class aircraft carrier	✗	 Instruction not followed	✗		
 Atago-class destroyer	✓	 Instruction not followed	✗	 Kitty Hawk-class aircraft carrier	✗	 INS Vikramaditya carrier	✗		

Figure 13. Object Classification: The figure highlights model performance on object classification, showing success with familiar objects like the “atago-class destroyer” and “small civil transport/utility” aircraft. However, models struggled with rarer objects like the “murasame-class destroyer” and “garibaldi aircraft carrier” indicating a need for improvement on less common classes and fine-grained recognition.

S5. Bi-temporal vs. Multi-temporal

We compare bi-temporal and multi-temporal image classification performance for Crop Type and Land Use identification tasks (Fig. 17). Bi-temporal data slightly outperforms multi-temporal data for crop type classification, suggesting that two distant time points are sufficient to capture key temporal changes, while additional time points may introduce redundancy. For land use classification, there is no significant difference between bi-temporal and multi-temporal inputs, reflecting the task's support on static spatial patterns that remain consistent over time.

S6. Geographical Analysis

In this section, we detail the geographic distribution of benchmarking datasets used in the studied geospatial tasks. It categorizes datasets into global/diverse datasets and regional/localized datasets. Global datasets provide extensive coverage with samples from over 100 countries or diverse regions worldwide. On the other hand, regional and localized datasets, are tailored to specific tasks. The map in Fig. 21 highlights that our benchmark is well represented across the globe.

S7. Word Cloud

The breakdown in Fig. 22a leverages the word cloud as part of evaluating VLMs in geospatial tasks, with image captioning being one of the key areas of interest. The word cloud highlights terms commonly used in captions describing aerial or geospatial imagery. Words like “aerial”, “surrounding” and “residential” reflect spatial and contextual elements frequently addressed in such descriptions, while terms such as “harbor”, “ship”, “tennis court” and “greenery” represent specific features often observed in geospatial data. This provides a basis for understanding the capabilities and limitations of these models in capturing spatial relationships and identifying key features within geospatial tasks.

The word cloud in Fig. 22b shows the terms used in MCQs. Keywords such as “large vehicle”, “transport utility”, “harbor”, “bridge” and “small vehicle” emphasize important categories and features frequently mentioned in the questions. Additional terms like “aircraft carrier”, “runway”, “basketball court” and “helicopter” represent a blend of transportation, infrastructure, and activity-based elements often linked to geospatial data. The use of varied and domain-specific vocabulary ensures the MCQs encompass a broad spectrum of scenarios for testing model capabilities.

GPT-4o		Qwen2-VL		LLaVA-OneVision		GeoChat		LLaVA-1.5		SPHINX		LLaVA-NeXT		InternVL-2		RS-LLaVA		Ferret	
What type of disaster is responsible for the visible damage in this image?					What type of disaster is responsible for the visible damage in this image?					What type of disaster is responsible for the visible damage in this image?					What was the primary trigger for this landslide?				
	A. volcano		B. flooding		A. earthquake		B. tsunami		A. flooding		B. volcano		C. earthquake	D. wind		E. tsunami		A. Soil Erosion	B. Human Activities
	C. tsunami		D. fire		C. volcano		D. wind		C. earthquake		D. wind		E. flooding			E. tsunami		C. Snow and Glacier Melting	D. Seismic Activity
	E. earthquake																E. Precipitation-Related Events		
																		Precipitation-Related Events	
																		Seismic Activity	
																		Precipitation-Related Events	
																		Soil Erosion	
																		Human Activities	
																		Precipitation-Related Events	
																		Instruction not followed	
																		Instruction not followed	

Figure 14. Event Detection: Model performance on disaster assessment tasks, with success in scenarios like ‘fire’ and ‘flooding’ but challenges in ambiguous cases like ‘tsunami’ and ‘seismic activity’. Misclassifications highlight limitations in contextual reasoning and insufficient exposure on overlapping disaster features.

GPT-4o		Qwen2-VL		LLaVA-OneVision		GeoChat		LLaVA-1.5		SPHINX		LLaVA-NeXT		InternVL-2		RS-LLaVA		Ferret		
What is the relationship between object in green box and object in red box in this image?					What is the relationship between object in green box and object in red box in this image?					What is the relationship between object in green box and object in red box in this image?					What is the relationship between object in green box and object in red box in this image?					
	A. A large vehicle is aligned with the bridge.		B. A helicopter is positioned next to the helipad.		C. A overpass leads to a golffield.		D. A tennis court is beside the basketball court.		E. A runway connects to the airport.		A. A large vehicle is aligned with the bridge.		B. A plane is parked near the runway.		C. An a350 is aligned with the runway.		D. A large-vehicle is moving towards the storage-tank.	E. A large-vehicle is moving away from the a roundabout.	A. A small-vehicle is driving by the bridge.	
	C.		E.		B.		D.		C.		E.		B.		D.		C.			
					<img alt="Aerial view of a parking lot with a green box around the small vehicle and a red box															

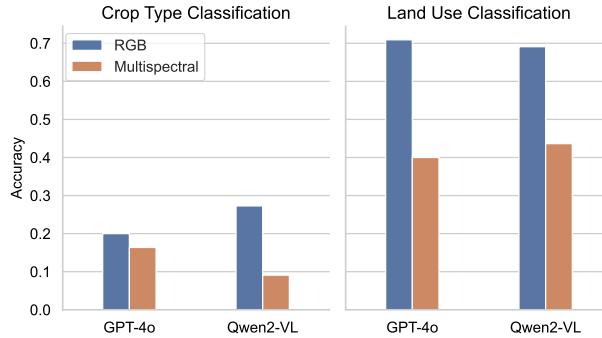


Figure 16. RGB and multispectral performance for Crop Type and Land Use Classification, showing a performance drop in multispectral accuracy.

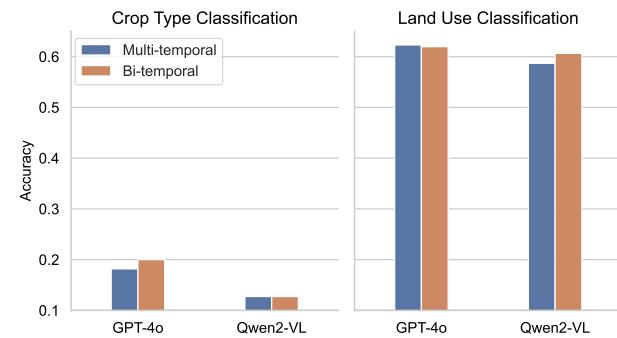


Figure 17. Bi-temporal and Multi-temporal performance, showing slight bi-temporal advantages for crop classification and similar results for static land use classification.



Figure 18. Image Captioning: Example response of different models.

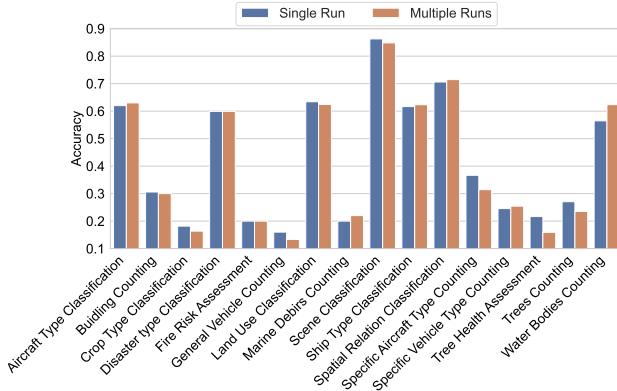


Figure 19. GPT4-o: Single run vs Multi runs

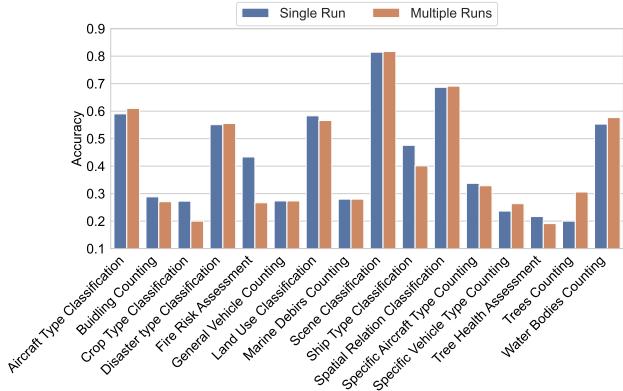


Figure 20. Qwen2-VL: Single run vs Multi runs

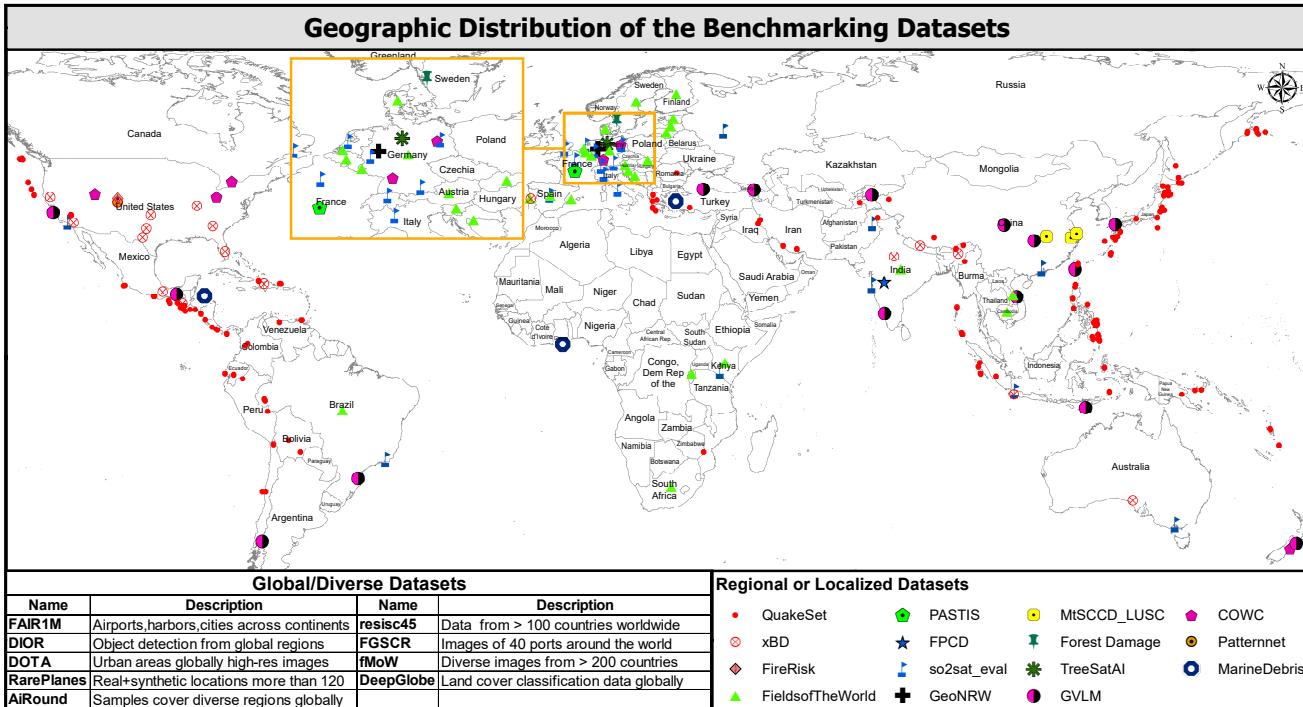
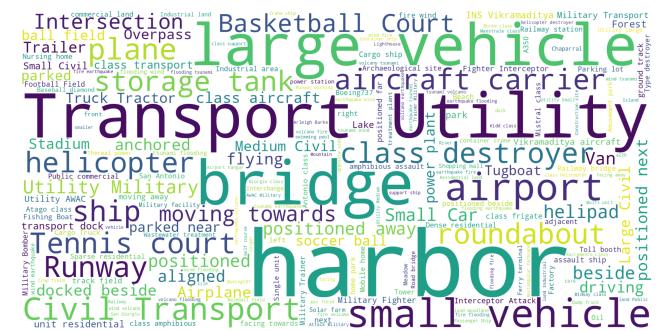


Figure 21. Figure shows the geographic distribution of benchmarking datasets, highlighting global coverage and regional specialization.



(a) Word cloud from captions used in geospatial image description tasks.



(b) Word cloud from MCQs designed for geospatial task evaluation.

Figure 22. Word clouds showcasing terms used in evaluating VLMs on geospatial tasks, with the first focusing on image captions and the second on MCQs.