This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Muhammad Sohail Danish*[1] Muhammad Akhtar Munir*[1] Syed Roshaan Ali Shah[2] Kartik Kuckreja[1]
Fahad Shahbaz Khan[1,3] Paolo Fraccaro[4] Alexandre Lacoste[5] Salman Khan[1,6]
[1]Mohamed bin Zayed University of Artificial Intelligence, [2]University College London, [3]Linköping University, Sweden
[4]IBM Research Europe, UK, [5]ServiceNow Research, [6]Australian National University

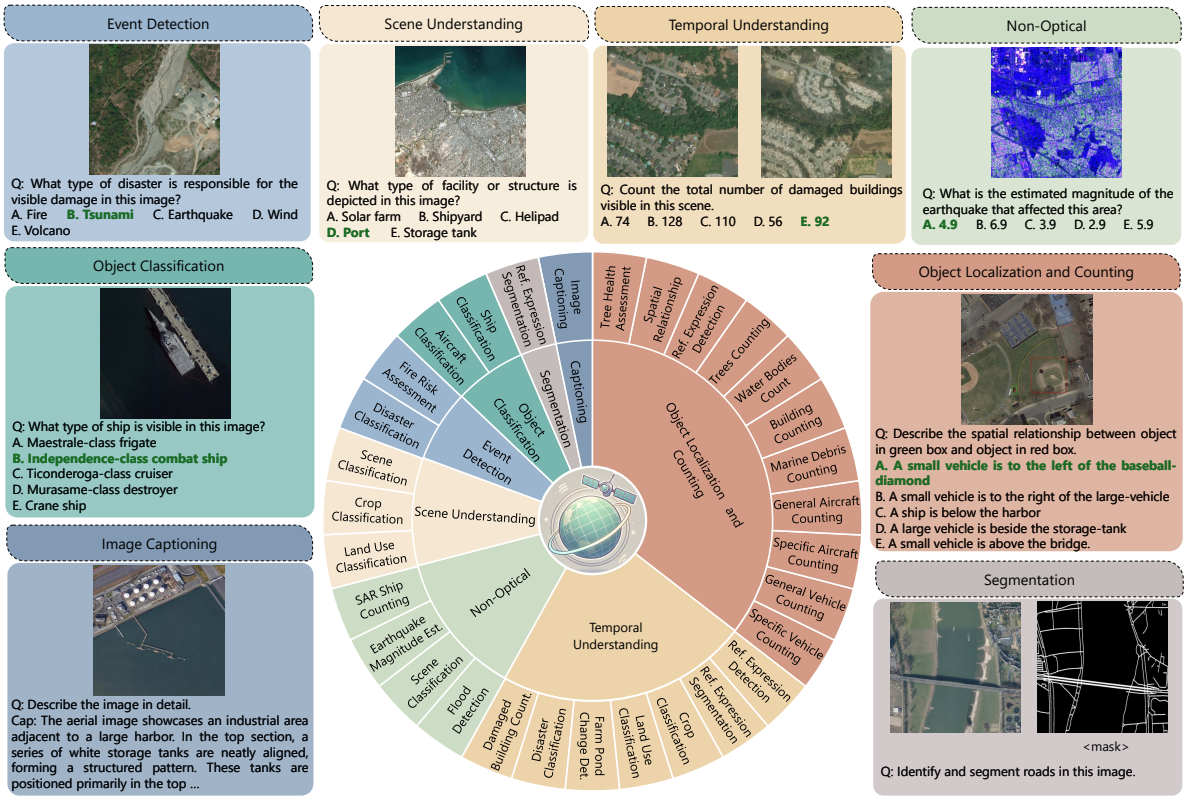


Figure 1. GEOBench-VLM comprehensively covers 31 fine-grained tasks categorized into 8 broad categories: scene and object classification, object detection, segmentation, captioning, event detection, non-optical and temporal understanding tasks.

Abstract

While numerous recent benchmarks focus on evaluating generic Vision-Language Models (VLMs), they do not effectively address the specific challenges of geospatial applications. Generic VLM benchmarks are not designed to handle the complexities of geospatial data, an essential component for applications such as environmental monitoring, urban planning, and disaster management. Key challenges in the geospatial domain include temporal change detection, large-scale object counting, tiny object detection, and understanding relationships between entities in remote sensing imagery. To bridge this gap, we present GEOBench-VLM, a comprehensive benchmark specifically designed to evaluate VLMs on geospatial tasks, including scene understanding, object counting, localization, fine-grained categorization, segmentation, and temporal analysis. Our benchmark features over 10,000 manually verified instructions and spanning diverse visual conditions, object types, and scales. We evaluate several state-of-the-art VLMs to assess performance on geospatial-specific challenges. The results indicate that although existing VLMs demonstrate potential, they face challenges when dealing with geospatial-specific tasks, highlighting the room for further improvements. Notably, the best-performing LLaVa-OneVision achieves only 41.7% accuracy on MCQs, slightly more than GPT-4o, which is approximately double

*Equally contributing first authors.

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 [13, 36, 50, 60, 66] and natural language processing [1, 4, 42, 43], enabling significant advancements in integrating visual and textual data. VLMs combine image and language understanding to handle tasks that involve visual and textual information [2, 8, 18, 35]. These models have demonstrated strong performance in tasks like image captioning [51], visual question answering [6], and object recognition [49], benefiting from training on large-scale visual-text datasets.

VLMs have the potential to enhance geospatial analysis in domains that rely on satellite and aerial imagery, such as urban planning, environmental assessment, and disaster management [14, 20, 27, 40, 45, 56]. By automating labor-intensive tasks such as land cover mapping, damage assessment, and object detection, VLMs can reduce manual effort while improving the efficiency and accuracy of geospatial analysis. However, despite their potential, VLMs remain limited in geospatial applications due to specific challenges inherent to geospatial data. This is due to the unique challenges posed by geospatial data, e.g., objects in satellite imagery can vary 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. The lack of geospatial-specific benchmarks hinders the objective assessment of VLMs, making it difficult to identify performance gaps and areas for improvement [17, 28, 57].

Existing benchmarks, such as SEED-Bench [17] and MMMU [57], have contributed to evaluating VLMs in general vision-language tasks such as visual question answering and scene understanding. However, these benchmarks do not adequately cover geospatial tasks requiring spatial reasoning and temporal analysis. On the other hand, VLEO benchmarks geospatial tasks, evaluating VLM performance in earth observation applications [59], but it lacks components for extended temporal analysis, non-optical data, and segmentation tasks. Another limitation is that VLEO primarily evaluates generic VLMs rather than models tailored for geospatial tasks, not considering specialized approaches. Although VLMs can improve the interpretation of data, the lack of a thorough evaluation highlights the need to assess both geospatial-specific and generic models on comprehensive geospatial benchmarks to effectively 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 integrates diverse datasets with varying visual conditions, resolutions, and object scales supplemented with both automated and manually verified annotations. It employs multiple-choice questions (MCQs) similar to [17], to ensure objective, scalable, and automated evaluation, reducing biases associated with open-ended responses. This structured approach also prevents hallucinations and addresses problems in understanding model responses for fair assessment. GEOBench-VLM covers key task categories, including *scene understanding, object counting, visual grounding, image captioning, temporal understanding, non-optical, referring segmentation, and relational reasoning* (Fig. 1). These tasks are essential for a wide range of applications, including but not limited to urban planning, monitoring deforestation, assessing environmental changes, and managing natural disaster response.

We evaluate several state-of-the-art VLMs using GEOBench-VLM to assess their performance in geospatial tasks. Our findings indicate that while generic VLMs, including commercial ones like GPT-4o, have potential, they struggle with geospatial tasks. We also include geospatial-specific VLMs in our evaluation to provide a comparative analysis of different models. Performance varies across geospatial tasks, with no single model excelling in all areas. LLaVA-OneVision [18] leads in object localization and counting; GPT-4o excels in object classification; Qwen2-VL [46] shows strength in event detection and interpreting non-optical imagery. Our comparative analysis 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 instructions.
2. We provide a detailed evaluation of ***13*** state-of-the-art VLMs, including generic (open and closed-source) and geospatial-specific VLMs, highlighting their capabilities and limitations in geospatial analysis.
3. We analyze performance across a range of tasks, including scene classification, counting, change detection, relationship prediction, referring expression detection, segmentation, image captioning, disaster detection, and temporal analysis, providing key insights that can help in improving VLMs for geospatial applications.

2. Benchmark Overview

VLM benchmarks cover tasks from general multimodal understanding to specialized geospatial applications. A de-

Benchmark	Domain	Modalities	Data Sources	Answer Type	Annotation Type	Human Verify	Year	RS Category
CulturalVQA[33]	General	O	Curated	FF	M	✗	2024	N/A
EXAMS-V[52]	General	O	Academic Exams	MCQ	M	-	2024	N/A
M4U[44]	General	O	Academic Exams	MCQ	M	✓	2024	N/A
MMMU[57]	General	O	Academic Exams	FF, MCQ	M	✓	2024	N/A
MME[63]	General	O	Various Open-Source	Yes/No	M	✓	2024	N/A
MMBench[28]	General	O	Various Open-Source	MCQ	A+M	✗	2024	N/A
MMStar[7]	General	O	Existing Benchmarks	MCQ	M	✓	2024	N/A
LAMM[53]	General	O, PC	Various Open-Source	FF	A+M	✗	2023	N/A
SEED-Bench[15]	General	O, V	Various Open-Source	MCQ	A+M	✓	2023	N/A
SEED-Bench2[17]	General	O, MI, V	Various Open-Source	MCQ	A+M	✓	2024	N/A
SEED-Bench-H[17]	General	O, MI, V	Public Data / Curated	MCQ	M	✓	2024	N/A
RSIEval[12]	RS	O, PAN	DOTA[48]	FF	M	✓	2023	6
LHRS-Bench[31]	RS	O	GE + OSM	SC	M	✓	2024	4
EarthGPT[62]	RS	O, IR, SAR	DRSD	FF, BBox	A+M	-	2024	5
SkyEyeGPT[58]	RS	O, V	DRSD	FF, MCQ	A+M	✓	2024	6
Fit-RSRC[29]	RS	O	DRSD	FF, MCQ	A+M	✓	2024	1
Fit-RSFG[29]	RS	O	DRSD	FF, MCQ	A+M	✗	2024	5
VRSBench[21]	RS	O	DOTA[48], DIOR[19]	FF, BBox	A+M	✓	2024	3
VLEO-Bench[59]	RS	O, (BT)	DRSD	FF, BBox, MCQ	A+M	✓	2024	6
EarthVQA[45]	RS	O	DRSD	FF	A+M	✗	2023	5
RemoteCount[23]	RS	O	DOTA[48]	SC	A+M	✓	2024	1
FineGrip[64]	RS	O	MAR20[47]	FF, Seg	A+M	✓	2024	5
GeoChat-Bench[14]	RS	O	DRSD	FF, BBox	A+M	✓	2023	6
GEOBench-VLM (Ours)	RS	O, MS, SAR, (BT, MT)	DRSD	MCQ, BBox, Seg	A+M	✓	2025	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).

tailed comparison is in Table 1.

Generic VLMs Benchmarks: Several benchmarks evaluate multimodal models across visual tasks, each with strengths and limitations. MMMU [57] evaluates models in multiple dimensions, incorporating diverse visual formats, and tests perceptual capabilities, but lacks geospatial tasks. SEED-Bench [17] focuses on spatial and temporal understanding with diverse datasets for complex multimodal contexts. SEED-Bench-2 [16] specializes in text-rich visual scenarios like charts and maps but interprets structured, generalized maps rather than complex geospatial data. Other benchmarks like MMBench [28], MM-Vet [55], and MMSTAR [7] evaluate spatial reasoning but do not address geospatial applications in earth observation.

Geospatial-specific: Few methods, along with benchmarks, including EarthVQA [45], LHRS [31], RS-LLaVA [3], EarthDial [40], and GeoChat [14], evaluate VLMs in remote sensing. RS-LLaVA [3] focuses on image captioning and VQA but lacks change detection and temporal analysis. LHRS [31] focuses on high-resolution remote sensing tasks with an RS-specific image-text dataset but lacks multi-temporal diversity. EarthVQA [45] supports complex relational reasoning with a multi-modal dataset, though its emphasis on non-temporal imagery restricts broader applicability. GeoChat [14] also evaluates geospatial tasks like region captioning and spatially grounded responses; however, it lacks diverse temporal datasets and segmentation. VLEO [59] is an earth observation benchmark; however, it is limited to non-remote sensing specialized methods and lacks the diversity of tasks. Specifically, VLEO does not include multi-temporal datasets, segmentation tasks, or non-optical imagery, restricting its effectiveness in evaluating geospatial VLM capabilities.

A comprehensive geospatial benchmark is essential for evaluating VLMs. GEOBench-VLM addresses this gap by integrating multi-temporal analysis, segmentation, and non-optical data, among other critical tasks.

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 outlined below.

Categories: The diverse geospatial tasks in GEOBench-VLM are carefully designed to evaluate the capabilities of VLMs. The 8 categories offer a comprehensive framework for assessing spatial reasoning, object classification, segmentation, temporal understanding, and more. These tasks are crucial for real-world applications like disaster response, urban planning, and environmental monitoring. See Fig. 2 for task illustrations. **1) Scene Understanding:** This category includes tasks like scene classification and land use classification, where models distinguish environments (e.g., airports, forests, bridges) and land types (e.g., agricultural, industrial, residential) respectively. Crop classification requires models to identify crops using vi-

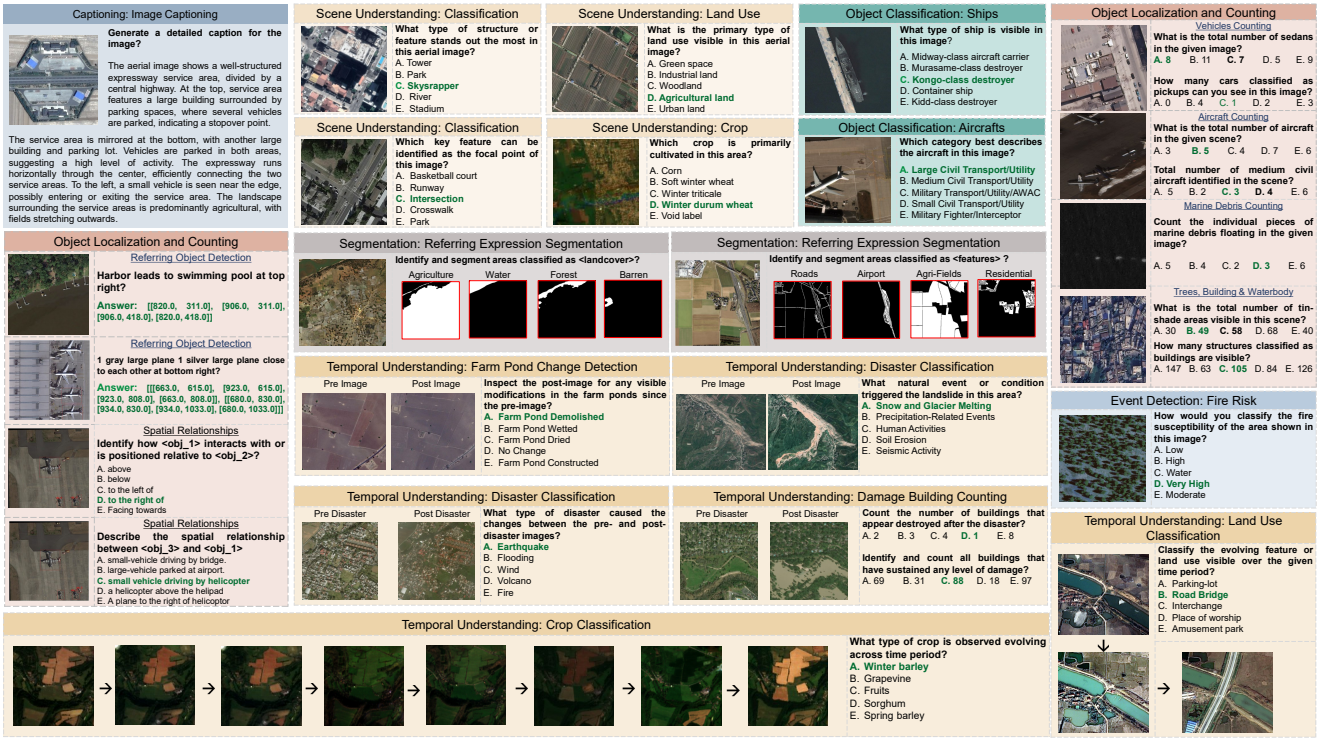


Figure 2. Comprehensive benchmark for VLMs in numerous geospatial tasks. This benchmark evaluates VLMs across eight core task categories, assessing their ability to interpret complex spatial data, classify scenes, identify and localize objects, detect events, generate captions, segment regions, analyze 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.

sual 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. **3) Object Localization and Counting:** 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 assessment. **4) Event Detection:** This focuses on fire risk detection and assesses fire hazards across forests, while disaster-type classification involves inferring potential disaster causes from post-event imagery. **5) Caption Generation:** This task evaluates image captioning, assessing a model's ability to generate descriptions that capture both overall scene context and specific object details. **6) Semantic Segmentation:** The task includes referring expression segmentation, where models generate binary masks for specified objects or regions (e.g., urban vs. non-urban areas). **7) Temporal Understanding:** This category includes tasks such as change detection and damage assessment to identify differences over time, such as post-disaster building conditions. Long temporal analysis includes tasks like crop classification. **8) Non-Optical:** This category focuses on non-optical imagery analysis, including ship detection, flood detection, and earthquake magnitude estimation, crucial for scenarios where optical imagery is insufficient.

Dataset pipeline: GEOBench-VLM integrates open datasets, and manual annotation aided with automated tools. For diversity, each task samples images from multiple datasets[1]. By combining multi-source datasets with structured question design, GEOBench-VLM enables scalable and high-quality evaluation of VLMs across geospatial tasks. Fig. 3 shows the overall pipeline.

For scene understanding tasks, including scene classification, land use classification, and crop type classification, we use classification datasets [9, 26, 30, 37, 38, 65]. 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 plausible alternatives. For counting tasks, we converted object detection data [11, 32, 39] 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 refer-

[1]For the complete list of datasets, see the supplementary material.

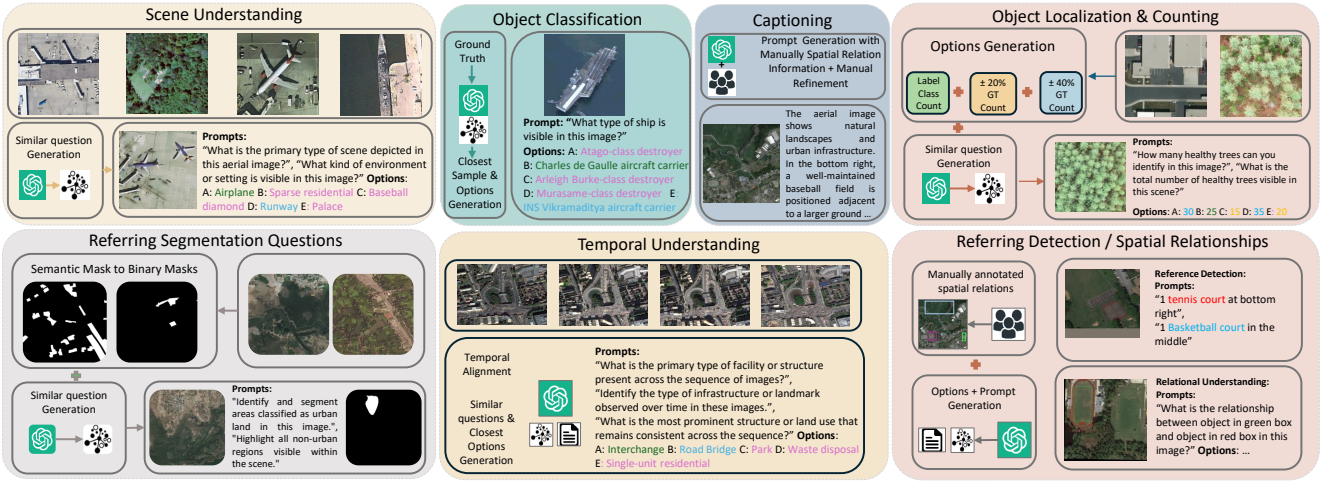


Figure 3. Data pipeline for the GEOBench-VLM: Our pipeline integrates diverse datasets, automated tools, and manual annotation. Tasks such as scene understanding, object classification, and non-optical analysis are based on classification datasets, while GPT-4o generates unique MCQs with five options: 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.

ring 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 [10, 41, 48], with cross-verification for consistency. Following our standard approach, five MCQs 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 incorporating object details and interactions. Each query is reviewed for clarity and complexity. For caption generation, we use GPT-4o with image, object attributes, and spatial relationships to create descriptions combining scene context and object details, which are manually refined for clarity, removing irrelevant or repetitive content.

4. VLMs Benchmarking

4.1. Selection of VLMs

Model Variability and Capabilities: We selected both generic and geospatial-specific VLMs, prioritizing recently developed models with advanced capabilities across various tasks. Generic VLMs include LLaVA-1.5 [25], LLaVA-NeXT [24], LLaVA-OneVision [18], Sphinx [22], Ferret [54], InternVL2 [8], and Qwen2-VL [46]. Also, we added GPT-4o, a closed-source, commercially available model, to compare its adaptability to remote sensing tasks with open-source models. These VLMs demonstrate robust performance in tasks such as scene understanding, and fine-grained visual classification. For geospatial-specific VLMs, we selected GeoChat [14], RS-LLaVA [3], SkySenseGPT [29], EarthDial [40], and LHRS-Bot-Nova [31], models tailored for satellite and aerial image interpretation.

Domain Specific Model Suitability: We consider VLMs based on their relevance to geospatial tasks. Despite being trained on satellite and aerial data, domain-specific models [3, 14, 29, 31, 40] may struggle with counting and spatial relationships due to dataset or architecture constraints. Meanwhile, we evaluate whether generic VLMs can effectively handle scene understanding, object detection, and visual reasoning in remote sensing. This selection ensures a focused study of VLM performance across diverse geospatial applications.

Open vs Closed Models: Open-source models like LLaVA [18, 24, 25], Qwen2-VL [46], GeoChat [14], and others provide transparency, aiding in understanding their strengths and limitations. Closed models like GPT-4o, despite their lack of transparency, demonstrate strong generalization and performance on complex tasks, benefiting from proprietary datasets and advanced architectures. Evaluating both ensures a comprehensive assessment of VLM capabilities.

4.2. Benchmarking Approaches

Task Setup & Complexity Levels: Our benchmark evaluates VLMs across varied geospatial task complexities, spanning basic (e.g., scene classification), intermediate (e.g., image captioning, object counting), and advanced tasks (e.g., risk assessment, spatial relations, change detection) that require spatial and temporal reasoning. The dataset includes annotations, temporal sequences, and segmentation data for a structured evaluation.

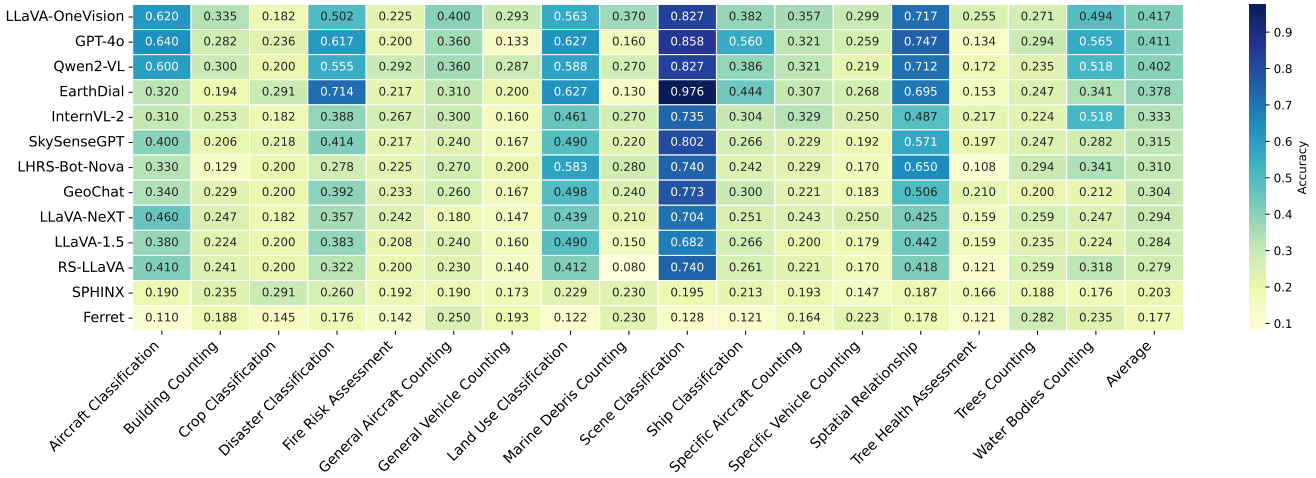


Figure 4. Performance summary of VLMs. LLaVA-OneVision achieves the average accuracy (41.7%), slightly outperforming GPT-4o, which is relatively better in building counting, and general aircraft counting. EarthDial demonstrates strong results in scene classification. The overall results highlight VLMs' varying strengths across geospatial tasks, with even the best models achieving accuracy only slightly above double the random guess.

Model	Event Det	Object Cls	Counting	Scene Und	Image Cap
GPT-4o [34]	0.4726	**0.5863**	0.3965	0.7114	0.6418
EarthDial [40]	**0.5418**	0.4039	0.3626	**0.7705**	0.5378
Qwen2-VL [46]	0.4640	0.4560	0.4019	0.6761	0.5895
LLaVA-OneVision [18]	0.4063	0.4593	**0.4377**	0.6636	0.6317
SkySenseGPT [29]	0.3458	0.3094	0.3119	0.6205	0.6416
InternVL-2 [8]	0.3458	0.3062	0.3280	0.5727	0.5968
GeoChat [14]	0.3372	0.3127	0.2922	0.6091	0.4395
LHRS-Bot-Nova [31]	0.2594	0.2704	0.3286	0.6330	0.6275
LLaVA-NeXT [24]	0.3170	0.3192	0.2737	0.5477	0.6293
LLaVA-1.5 [25]	0.3228	0.3029	0.2618	0.5625	0.6346
RS-LLaVA [3]	0.2795	0.3094	0.2534	0.5534	0.5604
SPHINX [22]	0.2363	0.2052	0.1860	0.2170	**0.6451**
Ferret[54]	0.1643	0.1173	0.1956	0.1261	0.5615

Table 2. VLMs accuracies across geospatial tasks. Evaluation includes event detection, object classification, counting, scene understanding, and image captioning. EarthDial performs better in event detection and scene understanding, demonstrating geospatial reasoning capabilities. GPT-4o performs best in object classification, while LLaVA-OneVision leads in counting. As evaluated by BERT score [61], Sphinx excels in image captioning.

Metrics: Our evaluation framework employs task-specific metrics to assess model performance. For MCQ-based tasks, we report accuracy as the evaluation metric. Precision evaluates referring expression detection, mIOU measures segmentation performance, and BERT score estimates image caption quality.

5. Evaluations and Results

In this section, we evaluate and discuss the results for each task in GEOBench-VLM as follows.

Scene Understanding: Scene understanding in remote sensing encompasses agriculture, urban planning, and environmental monitoring. Models struggle with crop classification, probably due to low-resolution imagery. GPT-4o and EarthDial [40] excel in land use, and scene classification (Fig. 4). Ferret [54] 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 achieves superior performance in object classification, probably benefiting from extensive training data and advanced model design that supports fine-grained recognition, while Ferret [54] performs lower in this task as it prioritizes region-based spatial localization over explicit object type classification. For detailed results, we refer to Fig. 4.

Object Localization & Counting: In remote sensing, this category involves locating and counting objects like water bodies, buildings, and trees. The task is challenging due to diverse scenes, varying object scales, and the need for high spatial resolution. Counting is framed as a classification task with multiple answer choices.

LLaVA-OneVision [18] performs relatively better in tasks such as *building/vehicle counting, specific vehicle/aircraft type counting, marine debris counting, and tree health assessment*, demonstrating the ability to capture fine details. Despite its remote sensing focus, RS-LLaVA [3] underperforms in *marine debris counting task*. LHRS-Bot-Nova [31] performs better in the *tree counting* task among geospatial models. In addition to Fig. 4 results, we present the precision for the referring expression detection task in Table 4. These results demonstrate that Sphinx achieves the relatively highest precision scores at *IoU: 0.25 & 0.50* compared to other models within this category while GPT-4o performs the worst. Among geospatial models, EarthDial [40] performs the best and is the second-best overall for this task. A lower IoU value captures the model's approximate localization ability and its effectiveness in interpreting

Model	Crop Cls	Damaged Bldg Cnt	Disaster Cls	Farm Pond CD	Land Use Cls
EarthDial[40]	**0.2182**	0.4362	0.5727	**0.2105**	**0.6623**
GPT-4o[34]	0.1818	**0.5667**	**0.6300**	0.1711	0.6525
LLaVA-OneVision[18]	0.1455	0.4810	0.4537	0.1842	0.5869
Qwen2-VL[46]	0.1091	0.5000	0.5991	0.1974	0.5967

Table 3. VLM performance on temporal geospatial tasks. Evaluation spans crop type classification, disaster type classification, farm pond change detection (CD), land use classification, and damaged building counting. EarthDial performs best in land use classification, while GPT-4o achieves better performance in disaster classification and damaged building counting. Qwen2-VL stands second in disaster classification.

task-specific instructions, while an IoU of 0.5 reflects the standard setting for this task.

Event Detection: Event detection involves identifying disasters crucial for environmental monitoring and risk management. Tasks such as fire risk assessment and disaster classification play a key role in mitigation and response efforts. Qwen2-VL [46] show potential in fire risk assessment, while EarthDial [40] performs better in disaster classification. We also report overall category results in Tab. 2.

Caption Generation: In this category, we evaluate image captioning using BERTScore[61], which measures the semantic similarity between generated captions and reference descriptions. Sphinx [22] achieves the highest BERTScore[61], followed closely by SkySenseGPT [29] and GPT-4o, indicating their strong ability to generate semantically rich and contextually relevant captions (Tab. 2). GeoChat [14], have the lowest scores, suggesting limitations in capturing fine-grained visual details. Our captions include both fine-grained and high-level scene descriptions. Since Sphinx is trained with detailed visual grounding, it outperforms other models, demonstrating its effectiveness in generating precise and contextually relevant captions.

Semantic Segmentation: Including referring segmentation in our benchmark enhances the evaluation of query-driven segmentation in remote sensing, crucial for land-use mapping and urban analysis. While no remote sensing-specific models currently support this task, non-specialized models like GlaMM [35] achieve a baseline mIoU of 0.1411. This integration helps assess adaptability of models to spatial queries in remote sensing.

Temporal Understanding: GEOBench-VLM also incorporates temporal tasks to analyze changes over time, an essential aspect of remote sensing. We evaluate GPT-4o, Qwen2-VL [46], EarthDial [40], and LLaVA-OneVision [18] on five tasks: crop classification, damaged building counting, disaster type classification, farm pond change detection, and land use classification (Table 3). Overall GPT-4o emerges as the top performer in disaster classification and damaged building counting, while Qwen2-VL trails GPT-4o. EarthDial performs best in land use classification. While temporal information has the potential to enhance performance, it is not yet fully exploited by existing models. The low accuracy in change detection and crop classification suggests that current VLMs struggle to leverage temporal dependencies effectively. Future improvements should focus on incorporating long-term temporal reasoning to better capture changes in geospatial data.

Model	Prec@0.5	Prec@0.25
Sphinx [22]	**0.3408**	**0.5289**
EarthDial [40]	0.2429	0.4139
GeoChat [14]	0.1151	0.2100
Ferret [54]	0.0943	0.2003
Qwen2-VL [46]	0.1518	0.2524
GPT-4o [34]	0.0087	0.0386
LHRS-Nova [31]	0.0930	0.2423
SkySenseGPT [29]	0.1082	0.3224

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

Model	EQME	LU Cls
RS-LLaVA [3]	0.0863	0.3123
GeoChat [14]	0.1403	0.3156
Qwen2-VL [46]	**0.2734**	0.2525
GPT-4o [34]	0.0827	**0.3256**

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

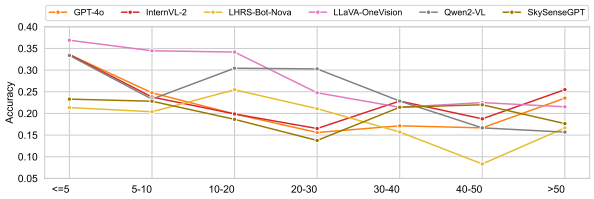


Figure 5. Object Density vs. Counting Accuracy. VLMs are evaluated on how well they maintain counting accuracy as the number of objects increases. LLaVA-OneVision shows better performance in less dense ranges, whereas Qwen2-VL, LHRS-Bot-Nova, and SkySenseGPT exhibit more significant performance drops at higher object densities.

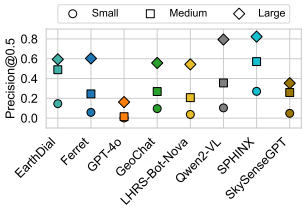


Figure 6. Performance comparison of different models on Referring Expression Detection across various object sizes.

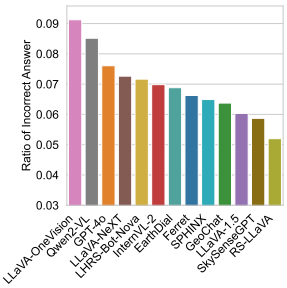


Figure 7. Percentage of incorrect answers falling within the 20% error range

Non-Optical: In the non-optical domain, we evaluate land use classification and earthquake impact estimation. Table 5 shows Qwen2-VL [46] scores highest in earthquake estimation, while GPT-4o ranks lowest. For land use classification, GPT-4o performs best, with GeoChat [14] and RS-LLaVA [3] performing comparably but struggling with earthquake magnitude estimation. While challenging, prior work [5] shows that seismic intensity can be inferred from SAR data via deformation patterns, making it a valuable proxy for assessing model grounding in non-optical data.

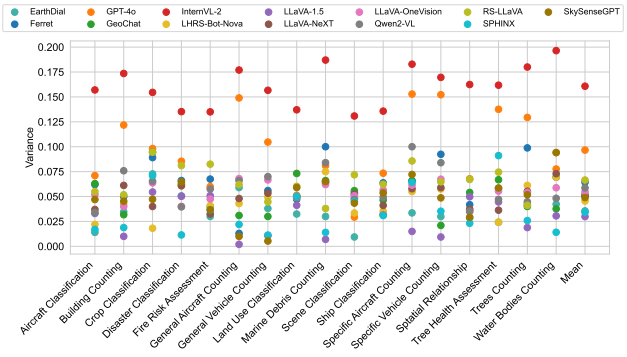


Figure 8. Accuracy Variance across different prompts.

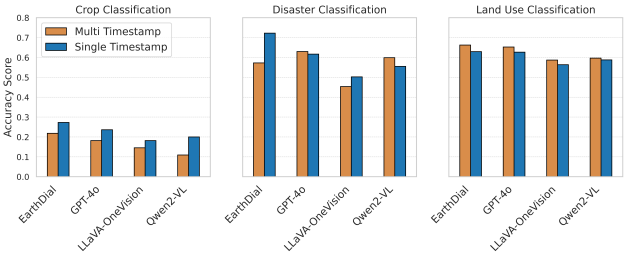


Figure 9. Single vs Multi-temporal.

6. Analysis

Counting Accuracy by Object Count Range: The object count ranges ($\leq$10 to $>$50) illustrate how well each model maintains counting accuracy from sparse to highly dense scenes (Fig. 5). LLaVA-OneVision [18] leads from low to moderate densities ($\leq$5 up to 20). InternVL2 [8] and GPT-4o perform best in the highest density ($>$50) category. Qwen2-VL [46] remains competitive in mid-density scenes but drops at higher object counts, and SkySenseGPT [29] usually exhibits the accuracy decline as density grows. These results emphasize the need for robust spatial alignment and fine-grained detection when counting objects in diverse density conditions.

Incorrect Answer Percentage by Option Range Distribution: Fig. 7 illustrates the ratio of incorrect responses when answer choices deviate from the ground truth within 20%. All models, including LLaVA-OneVision, Qwen2-VL, GPT-4o, and LHRS-Bot-Nova, exhibit that these models are sensitive to shifts in the distribution of possible options. Notably, no model demonstrates stable performance as the gap between the correct and alternative answers increases, suggesting weak numerical reasoning capabilities in counting tasks. This inconsistency highlights a critical limitation in VLMs when precise object count is required, making them prone to errors in complex geospatial scenes.

Prompt-Based Performance Variance: We find that models exhibit different sensitivity levels to prompt variations, affecting task accuracy. In Fig. 8, it is shown, GPT-4o and InternVL2 [8] exhibiting high variance across multiple tasks, suggesting a stronger sensitivity to prompt changes. Models such as Qwen2-VL [46] and RS-LLaVA [3] appear in the mid-range, while LLaVA-NeXT [24], EarthDial [40] and SkySenseGPT [29] generally show lower variance, indicating stability under varied prompt structures.

Single vs. Multi-Temporal Data: For the single vs. multi-temporal data analysis (Fig. 9), we evaluate GPT-4o, LLaVA-OneVision [18], EarthDial [40], and Qwen2-VL [46] across three tasks: crop, disaster, and land use classifications. In crop classification, multi-temporal data result in slightly lower performance across all models, suggesting that the temporal information introduces variability that the models struggle to effectively capture. In disaster classification, GPT-4o and Qwen2-VL benefit from multi-temporal data, whereas LLaVA-OneVision and EarthDial exhibit a decline. In land use classification, all models show improved performance with multi-temporal data, likely due to the stability of land use patterns over time, enabling better generalization for this task.

Impact of Object Size on Detection Performance: In referring expression detection across different object sizes (Fig. 6), we evaluate multiple models on small, medium, and large objects. Sphinx [22] and Qwen2-VL [46] demonstrate relatively high performance, particularly excelling in detecting large objects. GPT-4o, however, performs the worst, struggling significantly with all sized objects, suggesting limitations in fine-grained localization. Ferret [54], EarthDial [40], LHRS-Bot-Nova [31] and GeoChat [14] exhibit moderate performance on large objects. For medium objects, EarthDial [40] and Sphinx [22] perform overall better. These results highlight the challenges in detection, especially in medium and small objects.

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 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 13 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. 2

[2] 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

[3] 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, 8

[4] 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. 2

[5] Daniele Rege Cambrin, Isaac Corley, Paolo Garza, and Peyman Najafirad. Estimating earthquake magnitude in sentinel-1 imagery via ranking, 2024. 7

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain 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

[7] 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

[8] Zhe Chen, Jiannan Wu, Wenhai 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

[9] 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

[10] 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

[11] CSE499DeforestationSatellite. Deforestation-satellite-imagery dataset. https://universe.roboflow.com/cse499deforestationsatellite/deforestation-satellite-imagery-335n4. 4

[12] 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

[13] 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. 2

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] Fan Liu, Delong Chen, Zhangqingyun 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

[24] 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, 8

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5, 6

[26] 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

[27] 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

[28] 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

[29] 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, 5, 6, 7, 8

[30] 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

[31] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 440–457. Springer, 2024. 3, 5, 6, 7, 8

[32] 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. 4

[33] Shravan Nayak, Kanishk Jain, Rabiul 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

[34] OpenAI. Hello gpt-4o, 2024. 6, 7

[35] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman 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

[36] 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. 2

[37] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *ICCV*, 2021. 4

[38] 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

[39] Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. Rareplanes dataset, 2020. 4

[40] Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. *arXiv preprint arXiv:2412.15190*, 2024. 2, 3, 5, 6, 7, 8

[41] 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

[42] 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. 2

[43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 3, 5

[49] 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

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transform-

ers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 2

[51] 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

[52] 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

[53] 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

[54] 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, 8

[55] 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

[56] 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

[57] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi 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

[58] 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

[59] 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

[60] 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. 2

[61] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 6, 7

[62] 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

[63] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong 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

[64] 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

[65] 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

[66] 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. 2