OmniEarth-Bench: Towards Holistic Evaluation of Earth's Six Spheres and Cross-Spheres Interactions with Multimodal Observational Earth Data

Fengxiang Wang[1,2], Mingshuo Chen[3], Xuming He[4], Yi-Fan Zhang, Feng Liu[2], Zijie Guo[2], Zhenghao Hu[5], Jiong Wang[2] Jingyi Xu[2], Zhangrui Li[2], Fenghua Ling[2], Ben Fei[2], Weijia Li[5], Long Lan,[1] Wenjing Yang[1]*, Wenlong Zhang[2]*, Lei Bai[2]

[1] NUDT [2]Shanghai AI Lab [3] BUPT [4] ZJU [5] SYSU [6] FDU

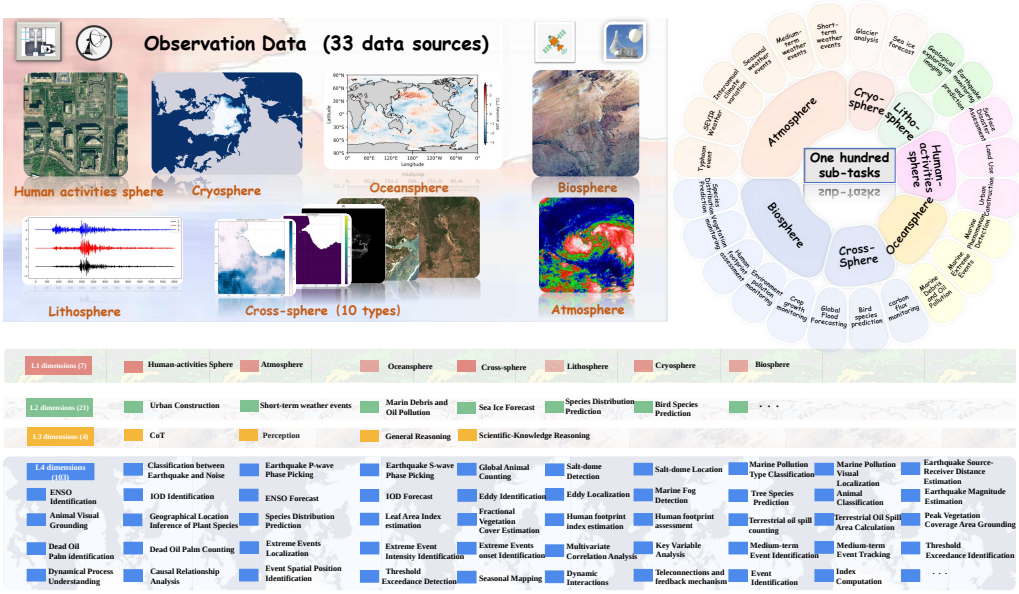

Figure 1: **Overview of OmniEarth-Bench.** Our benchmark spans six Earth science spheres and cross-sphere scenarios, encompassing 100 sub-tasks derived from 33 sensor types.

Abstract

Existing benchmarks for Earth science multimodal learning exhibit critical limitations in systematic coverage of geosystem components and cross-sphere interactions, often constrained to isolated subsystems (only in Human-activities sphere or atmosphere) with limited evaluation dimensions ($\leq$ 16 tasks). To address these gaps, we introduce **OmniEarth-Bench**, the first comprehensive multimodal benchmark spanning all six Earth science spheres (atmosphere, lithosphere, Oceansphere, cryosphere, biosphere and Human-activities sphere) and cross-spheres with one hundred expert-curated evaluation dimensions. Leveraging observational data from satellite sensors and in-situ measurements, OmniEarth-Bench integrates 29,779 annotations across four tiers: perception, general reasoning, expert-knowledge deductive reasoning and chain-of-thought (CoT) reasoning. This involves the efforts of 2–5 experts per sphere to establish authoritative evaluation dimensions and curate relevant observational datasets, 40 crowd-sourcing annotators to assist experts for annotations, and finally, OmniEarth-Bench is validated via hybrid expert-crowd workflows to reduce label ambiguity. Experiments on 9 state-of-the-art MLLMs

*Corresponding authors

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

reveal that even the most advanced models struggle with our benchmarks, where none of them reach 35% accuracy. Especially, in some cross-spheres tasks, the performance of leading models like GPT-4o drops to 0.0%. OmniEarth-Bench sets a new standard for geosystem-aware AI, advancing both scientific discovery and practical applications in environmental monitoring and disaster prediction. The dataset, source code, and trained models were released at OmniEarth-Bench.

1 Introduction

Earth scientists address critical environmental and societal challenges through modeling Earth's interconnected systems [1]: the atmosphere, lithosphere, hydrosphere, cryosphere, biosphere, and human activities [2]. By analyzing cross-system interactions, researchers derive impactful findings such as flood prediction [3], a complex task requiring multi-domain expertise (e.g., atmospheric precipitation, biospheric soil moisture, and lithospheric runoff). These discoveries are systematically validated in high-impact journals including Nature and Science [4, 5, 6, 7, 8, 9].

Existing MLLMs (e.g., GPT-4o [10], Gemini [11] and Claude [12]) excel at considerable tasks and have motivated benchmarks that explicitly test core skills. These benchmarks span diverse evaluation dimensions and explicitly include: Visual understanding [13, 14], Vision–language alignment [15, 16], Long-context modeling [17, 18], Chain-of-Thought (CoT) reasoning [19, 20], Scientific knowledge reasoning [17, 21] and so on [22, 23, 24]. In Earth science, existing multimodal benchmarks often focus on visual question answering using remote sensing data, covering a variety of satellite observation modalities and resolutions [25, 26, 27]. However, these existing benchmarks mainly focus on the human-activities sphere, with few or no multimodal benchmarks for other spheres. Moreover, while the semantic information in the observation data of the human-activities sphere is well-defined (e.g., buildings, roads and ships), other Earth systems lack precise scientific information formulation. This presents a new challenge: **How to establish scientific information definitions across multi-sphere Earth observations for effectively evaluating multimodal models?**

To address this challenge, we introduce OmniEarth-Bench to evaluate the scientific information processing capabilities of multimodal models across six Earth science spheres and cross-sphere scenarios. Considering the professional expertise required for analyzing Earth observation data, we have established four tasks: basic perception tasks, general reasoning tasks, specialized scientific reasoning tasks, and specialized scientific CoT reasoning tasks. The basic perception tasks are designed to assess the model's ability to perceive and recognize fundamental features and patterns in the Earth observation data. The general reasoning tasks evaluate the ability to draw logical conclusions based on the perceived information. The specialized scientific reasoning tasks aim to assess the ability to interpret scientific knowledge related to observational data. The specialized scientific CoT reasoning tasks evaluate the ability to perform step-by-step analysis of the observation data and derive accurate conclusions based on scientific knowledge.

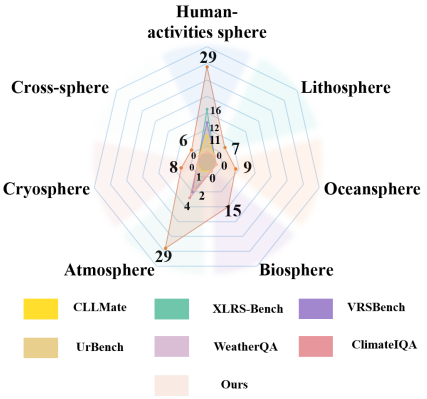

Figure 2: **Dimensions Categories of L4 dimensions.** Our benchmark spans 6 spheres and cross-sphere, across 100 typical subtasks (L4 dimensions).

Fig. 1 shows the typical examples across 6 spheres and cross-spheres. We engaged 2–5 experts (PhD holders or candidates) per sphere to identify representative real-world tasks, establish authoritative evaluation dimensions, and curate relevant observational datasets (either existing datasets or original data sourced from satellites like MODIS [28]). After defining these dimensions, we enlisted 40 crowd-sourcing annotators (undergraduate and master's students, 5–10 per sphere) to assist experts in annotation, followed by rigorous cross-validation to ensure quality. Ultimately, OmniEarth-Bench comprises 100 sub-dimensions (L-4 tasks) across seven categories (atmosphere, lithosphere, Oceansphere, cryosphere, biosphere, Human-activities sphere, and cross-sphere). As illustrated in Fig. 2, OmniEarth-Bench substantially surpasses existing benchmarks in comprehensiveness and coverage. Tab. 2 summarizes its quantitative and qualitative advantages. The key contributions are:

- **Comprehensive Evaluation Across All Six Spheres.** OmniEarth-Bench is the first benchmark to extensively cover all Earth science spheres, offering 58 practical and comprehensive evaluation dimensions that significantly surpass prior benchmarks.
- **Pioneering Cross-Sphere Evaluation Dimensions.** To address complex real-world scenarios, OmniEarth-Bench introduces cross-sphere evaluation capabilities for societally important tasks such as disaster prediction and ecological forecasting.
- **CoT-Based Reasoning Evaluations in Earth Science.** OmniEarth-Bench establishes, for the first time, CoT-based evaluations tailored for complex Earth science reasoning tasks, addressing scenarios where previous benchmarks showed near-zero accuracy, and explores how CoT strategies might enhance reasoning capabilities in the Earth domain.

2 Related Work

Earth Multimodal Benchmark. Recent advancements in large multimodal models (MLLMs) have accelerated progress in Earth sciences [29, 30], leading to the development of several evaluation benchmarks [25, 26, 31, 32]. Current benchmarks primarily target the Human-activities sphere and atmosphere. In the Human-activities sphere, remote sensing-based benchmarks include RSIEval [33], featuring 100 human-annotated captions and 936 VQA pairs; VRSBench [25], containing 29,614 images, 52,472 object references, and 123,221 QA pairs; and XLRS-Bench, which offers the largest dataset to date with an average resolution of 8500×8500. Atmospheric benchmarks include WeatherQA [31], designed specifically to evaluate severe weather predictions in two dimensions; ClimateIQA [32], built from climate reanalysis data for extreme weather event detection across four question types; and CLLMate [34], focused on weather and climate event forecasting using numerical meteorological data and textual event descriptions. However, these benchmarks exhibit notable limitations: 1) They typically address isolated spheres, neglecting cross-sphere interactions essential to real-world Earth science challenges. 2) They offer limited evaluation dimensions, with atmospheric benchmarks assessing fewer than four question types, and even the most extensive Human-activities sphere benchmark covering only 16 dimensions. Overall, comprehensive benchmarks addressing all six spheres and evaluating cross-sphere capabilities are still lacking in Earth sciences.

General Multimodal Benchmark. Large-scale vision-language models (VLMs) have shown great promise in multimodal tasks such as scene understanding and visual sentiment analysis, prompting the development of diverse benchmarks to quantitatively assess their capabilities. However, earlier benchmarks mostly targeted specific domains with limited evaluation tasks (*e.g.*, visual grounding [35, 36] or visual question answering (VQA) [37, 38, 39, 40, 41]). Recent efforts aim for more comprehensive assessments: MME [15] evaluates 14 perceptual and cognitive tasks; MMBench [13] offers over 3,000 questions spanning 20 skill dimensions like object localization and social reasoning; Seed-Bench [16] scales up further with 19,000 questions; MMT-Bench [24] integrates real-world scenarios like autonomous driving; and MME-Realworld [18] includes five real-world contexts with high-resolution imagery. Multimodal benchmarks focusing on scientific disciplines have also emerged. HLE [42] covers numerous academic disciplines with 2,500 questions; MMMU-Pro [43] evaluates multidisciplinary visual-textual integration skills at scale. Recently, multimodal chain-of-thought (CoT) benchmarks were developed: MME-CoT [19] includes 1,130 questions annotated with 3,865 reasoning steps; and ZeroBench [20] provides 100 handpicked questions and 334 simpler sub-questions. Despite these advancements, two critical limitations remain: 1) Earth sciences have been largely neglected, with only SuperGPQA featuring a minimal number (100) of geophysics-related textual questions, and multimodal CoT benchmarks lacking Earth science content entirely. 2) Existing benchmarks overlook the importance of observational data, a distinctive strength of Earth sciences (e.g., satellite imagery, climate data grids, seismic signals). In summary, current general-domain benchmarks fail to sufficiently evaluate multimodal models in Earth sciences, particularly concerning observational data and CoT reasoning scenarios.

3 OmniEarth-Bench

OmniEarth-Bench stands out from existing multimodal understanding benchmarks with three key features: i) It is the first benchmark based on Earth observational data to comprehensively cover all six Earth spheres, with evaluation dimensions grounded in real-world needs and rigorously validated by domain experts. ii) It firstly introduces the cross-sphere evaluation dimensions in geoscience, enabling

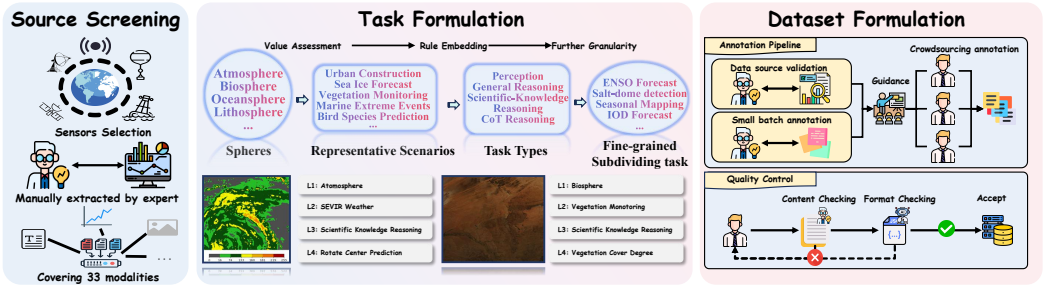

Figure 3: **Pipeline of OmniEarth-Bench.** Our pipeline comprises 4 stages—Source Screening, Task Formulation, Expert Annotation, and Quality Control—all led by experts. The first two stages are exclusively conducted by experts, while crowdsourcing annotators assist in the latter two stages.

MLLMs to be tested on realistic, interdisciplinary Earth science cross-sphere tasks. iii) It firstly establishes the Chain-of-Thought (CoT) reasoning benchmark for geoscience, using expert-reviewed human annotations and cross-validation to assess CoT effectiveness in complex scientific reasoning.

3.1 Pipeline of Benchmark

Source Screening. Our Benchmark comprises not only publicly available open-source datasets but also a significant portion of data manually extracted by experts from satellite imagery and raw observational sources. For example, Vegetation Monitoring uses satellite imagery from MODIS and expert-curated data from the Global Land Surface Satellite (GLASS), including Leaf Area Index, Fractional Vegetation Cover and Peak Vegetation Coverage Area. Moreover, for the Eddy data in oceansphere, the chlorophyll (CHL) data used in this study were obtained by applying the OCI empirical algorithm to Level-2 data acquired by the Geostationary Ocean Color Imager I (GOCI) aboard the Oceanography and Meteorology Satellite (COMS). After careful selection and integration, we compiled a comprehensive dataset covering 33 different data modalities across all Earth spheres. Tab.1 is a summary of the data sources used for each Earth sphere, with detailed data organization and construction procedures presented in the appendix.

Table 1: **Data source of different spheres**, including open-source datasets, satellite websites and other observation data sources. We only exhibit the L1 and L2 dimensions.

L1 dimensons	L2 deminesons	Data Source	Annotations Volume
Cross-sphere	Global Flood Forecasting	GFF [3]	873
	Bird Species Prediction	SatBird [44]	2,253
	Carbon Flux Monitoring	CarbonSense [45]	330
Human-activities sphere	Urban Construction	UBCv1 [46], BHdataset [47]	3,161
	Land Use	WHU-OHS [48]	2,990
	Surface Disaster Assessment	XView [49]	3,851
Biosphere	Species Distribution Prediction	TreeSatAI [50], Penguin [51] OAM-TCD [52], TaxaBench [53]	2,819
	Vegetation Monitoring	GLASS [54], MODIS [28]	900
	Environmental Pollution Monitoring	ROSID [55]	246
	Human Footprint Assessment	HFP [56], MODIS [28]	600
	Crop Growth Monitoring	MOPAD [57]	1,656
Atmosphere	SEVIR Weather	SEVIR [58]	893
	Typhoon Events	DigitalTyphoon [59]	5,082
	Short-term meteorological events	ERA5 [60]	140
	Medium-term meteorological events	ERA5 [60]	160
	Seasonal meteorological events	ERA5 [60]	60
	Interannual climate change	ERA5 [60]	60
Lithosphere	Earthquake monitoring and prediction	STRAD [61]	1,500
	Geological exploration imaging	TGS-Salt [62]	631
Oceansphere	Marine Debris and Oil Pollution	MADOS [63]	221
	Marine Extreme Events	ERASSTv5 [64]	583
	Marine Phenomenon Detection	COMS [65], M4Fog [66]	570
Cryosphere	Sea ice forecast	G02202 (SIC) [67], NSIDC-0079 [68] PIOMAS [69],GIOMAS [70]	200
	Glacier analysis	CryoSat-2 [71] IceBridge [72], ICESat-2 [73]	30

Task Formulation. As shown in Fig.3, OmniEarth-Bench defines tasks across four hierarchical levels (L1–L4): L1 covers the seven domains based on established geophysical spheres: atmosphere, lithosphere, oceansphere, cryosphere, biosphere, Human-activities sphere and cross-sphere. L2 includes expert-approved, representative scenarios within each sphere, selected based on their scientific and practical value (e.g., earthquake prediction). Tab.1 illustrates representative scenarios covered by the L1 and L2 levels. Detailed descriptions of the L3 and L4 dimensions for each sphere are provided in the appendix. L3 comprises four core abilities: Perception, General Reasoning, Scientific-Knowledge Reasoning and CoT Reasoning. Perception and General Reasoning align with previous works such as MMBench [13] and XLRS-Bench [26], where Perception focuses on sensory inputs and Reasoning

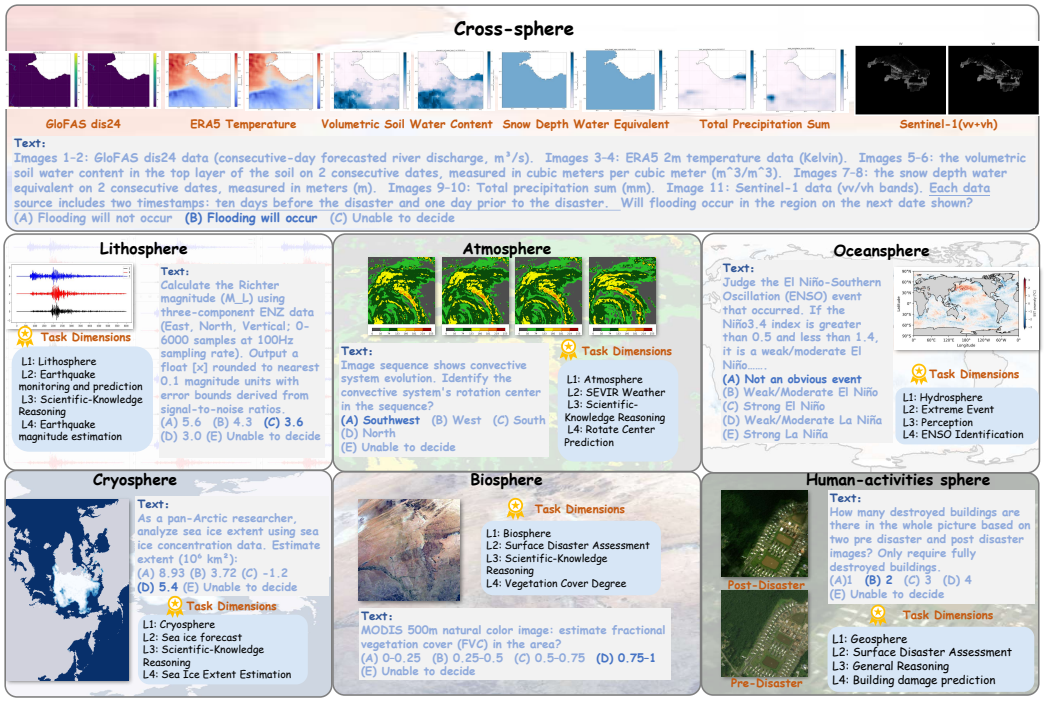

Figure 4: **Examples of OmniEarth-Bench.** OmniEarth-Bench comprises 100 unique L4 tasks, each with distinct questions, answers, and images. Spanning diverse data sources, timeframes, and natural variables, all tasks are jointly defined by domain experts across spheres.

on inference. Scientific-Knowledge Reasoning addresses complex reasoning tasks requiring deep domain expertise in Earth sciences. CoT Reasoning evaluates the effectiveness of chain-of-thought processes within Earth science scenarios. L4 provides further granularity by subdividing tasks based on the L1–L3 dimensions. Each L4 category is verified by domain experts to ensure practical relevance. Examples include fractional vegetation cover estimation in the biosphere and earthquake magnitude estimation in the lithosphere. Achieving robust general intelligence in Earth sciences requires MLLMs to perform effectively across all hierarchical levels. OmniEarth-Bench provides the first comprehensive framework designed for such an evaluation.

Expert Annotations. For each of the six Earth spheres, we enlisted 2–5 domain experts (Ph.D. holders or candidates) and 5–10 crowd-sourcing annotators (undergraduate and master's students). (1) For each sphere, evaluation dimensions were collaboratively defined by domain experts and MLLM specialists, ensuring high practical value and complexity. Cross-sphere tasks involved experts from multiple domains. This approach addresses the limitations observed when crowd-sourcing annotators proposed overly simplistic tasks—for example, "Estimated Maximum Precipitation Level" in atmosphere, which GPT-4o solved with 97.7% accuracy. Expert-led design ensures meaningful evaluation. (2) Experts were also responsible for defining data sources. Attempts to delegate this to annotators led to issues such as low sample difficulty and data scarcity. For complex tasks, annotators struggled with downloading and aligning data (e.g., MODIS and GLASS from NASA). Thus, experts curated and organized datasets, with annotators assisting.

Quality Control. To ensure data integrity and task relevance, the quality control process involved two main steps. Cross-Validation: Annotator outputs were systematically compared against expert-provided annotation examples. Any discrepancies were flagged and reviewed by domain experts to ensure annotation correctness, especially for complex tasks involving multi-source or challenging data. Final Quality Assessment: MLLM specialists conducted thorough reviews to confirm that annotations adhered to expert standards and maintained consistency across all tasks and Earth spheres. High-quality annotations were approved and incorporated into the dataset, while annotations that did not meet quality standards underwent iterative refinement through a feedback loop involving annotators and expert supervision. This cyclical process ensured continuous improvement and maintained the overall reliability of the dataset.

Table 2: **Comparison between existing vision-language benchmarks and our benchmark.** ✗ represents semi-automated, *i.e.*, machine generation followed by human verification.

Dataset	Spheres	Cross-Sphere	Observation Data	Data Source Volume	VQA and Visual Grounding			CoT	
					Volume	Dimensions Volume	Expert Annotation	Volume	Average key step annotation
ScienceQA [17]	⊗	⊗	⊗	-	21,000	127	⊗	⊗	⊗
Seed-Bench [16]	⊗	⊗	⊗	-	19,242	12	⊗	⊗	⊗
MME [15]	⊗	⊗	⊗	-	2,374	14	✓	⊗	⊗
MMBench [13]	⊗	⊗	⊗	-	3,217	20	✓	⊗	⊗
MME-Realworld [18]	⊗	⊗	⊗	-	29,429	43	✓	⊗	⊗
ZeroBench [20]	⊗	⊗	⊗	-	⊗	⊗	⊗	100	⊗
MME-CoT [19]	⊗	⊗	⊗	-	⊗	⊗	⊗	1,130	3.2
VRSBench [25]	Human-activities sphere	⊗	✓	2	175,703	12	✗	⊗	⊗
XLRS-Bench [26]	Human-activities sphere	⊗	✓	6	45,008	16	✓	⊗	⊗
RSIEval [33]	Human-activities sphere	⊗	✓	1	933	1	⊗	⊗	⊗
UrBench [27]	Human-activities sphere	⊗	✓	6	11,600	11	⊗	⊗	⊗
WeatherQA [31]	Atmosphere	⊗	✓	1	8,000	2	⊗	⊗	⊗
ClimateIQA [34]	Atmosphere	⊗	✓	2	254,040	4	⊗	⊗	⊗
CLLMate [34]	Atmosphere	⊗	✓	2	7,747	1	⊗	⊗	⊗
OmniEarth-Bench	6 Spheres	✓	✓	33	29,779	100	✓	610	5.8

3.2 Task Dimensions

OmniEarth-Bench defines tasks across four hierarchical levels (L1–L4), comprising 7 L1 dimensions, 23 L2 dimensions, 4 L3 dimensions, and 103 expert-defined L4 subtasks with real-world applicability. One representative L4 subtask from each L1 sphere is illustrated in Fig 4. Detailed descriptions of the L3 and L4 dimensions are provided in the appendix.

Cross-sphere. Cross-sphere tasks in Earth science carry high practical and societal importance [4, 5, 6, 9]. To evaluate MLLMs, we select three representative L2 scenarios from socially impactful applications, including *Global Flood Forecasting (L2), Bird Species Prediction (L2) and Carbon Flux Monitoring (L2)*. Due to their reliance on expert knowledge and complex reasoning, all are categorized as Scientific-Knowledge Reasoning (L3). Their L4 dimensions are collaboratively defined by experts from the relevant spheres. Despite the complexity of cross-sphere scenarios, we successfully collaborated with domain experts to construct **6 high-value subtasks (L4 dimensions) .**

Lithosphere. We firstly construct an MLLM benchmark for the lithosphere based on observational data, comprising **7 practical subtasks (L4 dimensions) .** We define two representative L2 scenarios within the lithosphere: *Seismic Monitoring and Prediction (L2) and Geophysical Exploration (L2).* Seismic monitoring and prediction [74], a critical domain in geosciences, aims to uncover Earth's internal dynamics and earthquake nucleation mechanisms, forming a theoretical basis for early warning and disaster mitigation. Geophysical exploration imaging [75], by analyzing subsurface responses to physical fields such as seismic waves, electromagnetic fields, and gravity/magnetic anomalies, enables high-resolution geological modeling essential for understanding subsurface structures, hydrocarbon exploration, and geological hazard assessment.

Human-activities sphere. The Human-activities sphere leverages remote sensing and mapping technologies across three key scenarios: *Urban Construction (L2), Land Use (L2), and Surface Disaster Assessment (L2).* Urban construction supports planning and socio-economic analysis; land use classification underpins environmental monitoring and resource management; and disaster assessment enables rapid post-event response and risk mitigation. OmniEarth-Bench spans all four L3 capability dimensions in the Human-activities sphere—Perception, General Reasoning, Scientific-Knowledge Reasoning, and CoT Reasoning—with **29 subtasks (L4 dimensions)**, surpassing all existing benchmarks in this domain [25, 26].

Atmosphere. The atmosphere is a key domain in Earth sciences with high practical value and extensive research interest [76, 77]. While existing benchmarks target specific atmospheric sub-scenarios [31, 32, 34], they lack comprehensive domain-wide coverage. OmniEarth-Bench addresses this gap by defining evaluation dimensions across six representative scenarios using data from ERA5 [60], SEVIR [58], and Typhoon [59] datasets: *Short-term Weather Events (L2), Medium-term Weather Events (L2), Seasonal Weather Events (L2), Interannual Climate Variation (L2), Typhoon Event (L2), and SEVIR Weather (L2).* For example, the *Typhoon Event* dimension serves as a flagship benchmark for atmospheric machine learning, supporting operational hazard forecasting and advancing research on tropical cyclone intensity and structure. These six scenarios (L2) span **30 expert-designed subtasks (L4 dimensions)** with strong real-world relevance, substantially surpassing existing atmospheric benchmarks. Full task details are provided in the appendix.

Oceansphere. We build a multi-layer MLLM benchmark for the oceansphere based primarily on satellite and analysis data products, featuring **9 practical L4 subtasks**. This domain includes three representative L2 scenarios: *Marine Oil Spills and Debris Monitoring (L2), Extreme Oceanic*

Events Warning (L2), and Ocean Phenomena Detection (L2). The Marine Oil Spills and Debris Monitoring [78] scenario uses multi-source remote sensing and in situ water quality data to track the spatial distribution and temporal dynamics of oil contamination and floating debris, supporting environmental management and emergency response. The Extreme Oceanic Events Warning [79, 80] scenario targets the detection and prediction of major climate modes such as El Niño–Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD), aiming to mitigate their societal and economic impacts. The Ocean Phenomena Detection scenario [81, 82] involves identifying ocean features like eddies and marine fog, which are key for maritime safety and ecological studies.

Cryosphere. We conduct a MLLM benchmark for the cryosphere primarily based on sea ice reanalysis data, glacial imagery, and graphical plots, incorporating **8 practical L4 subtasks**. We identify two representative L2 scenarios within the cryosphere: *Sea Ice Forecasting (L2) and Glacier Analysis (L2)*. SSea ice forecasting focuses on predicting the dynamic changes of sea ice in polar regions. Arctic sea ice is crucial for understanding global climate change [83, 84]. Its continuous decline over the last few decades has made sea ice forecasting significant for navigating through the Arctic Ocean during melting seasons. Moreover, the loss of the Antarctic sea ice would greatly impact the global sea level. Glacier analysis [85, 86], aims to study the glacial movements and changes of glaciers over time.

Biosphere. We present a biosphere-focused MLLM benchmark built on observational data and retrieval products, featuring **16 practical L4 subtasks**. It includes four representative L2 scenarios: Vegetation Monitoring (L2), Human Footprint Assessment (L2), Environmental Pollution Monitoring (L2), Species Distribution Prediction (L2) and Crop Growth Monitoring (L2). Vegetation Monitoring [87] evaluates plant and ecosystem health to support function assessment, carbon accounting, and climate response. Human Footprint Assessment [88] quantifies human impact on nature, informing sustainability and biodiversity strategies. Environmental Pollution Monitoring [89] identifies pollution events and their extent, guiding environmental policy and mitigation. Species Distribution is a key concern in the biosphere, as it guides biodiversity conservation and supports modeling species range shifts under climate and land-use change. Crop Growth Monitoring [90] assesses crop health for precision agriculture and sustainable farming.

3.3 Statistics and Analyses

Overview Statistics. OmniEarth-Bench includes 100 expert-defined, high-value evaluation dimensions and 29,779 samples annotated by both experts and crowdsourced contributors. As shown in Tab. 2, it offers clear advantages over existing benchmarks. Uniquely built on observational Earth science data—rather than exam-style datasets—OmniEarth-Bench spans all six spheres and cross-sphere scenarios. Consistency metrics are reported in Tab. 3, with additional details and dimension-specific indicators provided in the appendix.

Table 3: **Main statistics in OmniEarth-Bench**

Statistic	Number
Total questions	29,779
- Cross-sphere	3,456
- Human-activities sphere	9,362
- Biosphere	6,221
- Atmosphere	6,395
- Lithosphere	2,131
- Oceansphere	1,374
- Cryosphere	230
Multiple-choice questions	27,082
Visual grounding questions	2,697
Single-image questions	24,108
Multi-image questions	5,671
Maximum question length	213
Average question length	48.2
CoT	
- Total key step annotation	3,473
- Average key step annotation	5.8
- Average key step length	14.8
- Maximum question length	101
- Average question length	50.2

Observational Data vs. Exam-questions Data. Unlike subject-based benchmarks like ScienceQA [17], which rely on exam questions or online learning problems followed by manual filtering, our approach takes a fundamentally different path. While such methods could theoretically span all six Earth spheres, they face two key limitations: (1) Benchmarks like ScienceQA focus on scientific inquiry rather than practical geoscience applications, limiting their real-world relevance. (2) Their evaluation dimensions are constrained by a bottom-up design—questions are derived from existing image-text pairs in question banks or papers, then filtered and categorized. In contrast, OmniEarth-Bench follows a top-down strategy: domain experts first define evaluation dimensions based on real-world geoscience challenges and data availability, then curate corresponding data. This ensures each task is both meaningful and grounded in practical utility.

Table 4: **Experimental results on each sphere of VQA tasks, with models ranked by average performance.** '*Avg*' represents the average accuracy across sub-tasks. Proprietary models are highlighted in gray. '*Experts*' means evaluation results of 100 examples in each sphere by experts. We mark the highest score of each metric in red, and second highest underlined.

Method	Speres (L1 dimensions)							Avg.
	Cross-sphere	Atmosphere	Lithosphere	Oceansphere	Cryosphere	Biosphere	Human-activities sphere	
Experts	90	96	91	95	93	97	95	93.4
Closed-source MLLMs								
Claude-3.7-Sonnet [12]	30.68	24.72	28.15	23.12	54.46	31.21	11.18	29.07
Gemini-2.0 [11]	16.93	20.83	38.94	16.94	58.52	20.83	23.74	28.1
GPT-4o [10]	0.04	9.64	12.8	13.35	37.48	1.97	2.76	11.15
Open-source MLLMs								
InternVL3-72b [97]	19.19	33.98	23.39	20.22	74.56	31.99	29.46	33.26
InternVL3-7b [97]	42.85	30.1	37.47	20.28	49.27	28.74	23.18	33.13
LLaVA-Onevision-7b [92]	19.26	33.69	28.72	24.54	46.4	37.31	30.62	31.51
Internlm-Xcomposer2.5-7b [98]	19.78	17.45	28.88	21.06	40.04	30.67	24.76	26.09
Qwen2.5-VL-7B [99]	9.85	9.25	18.65	13.95	17.85	10.94	6.23	12.39
Qwen2.5-VL-72B [99]	3.92	4.82	22.43	16.27	5.88	14.91	8.63	10.98

Human Annotations vs. GPT Annotations. All annotations are finished by experts and crowdsourcing annotators. Unlike MMBench [13], we did not use tools like GPT-4o [10]. It was driven by two key reasons: (1) GPT-4o cannot generate VQA data requiring deep domain expertise. Tasks under the Scientific-Knowledge Reasoning (L3) demand substantial background knowledge and must be constructed collaboratively by experts. (2) Although GPT-4o can generate samples for general perception or simple reasoning tasks, expert evaluation found the data to be low quality and insufficiently challenging. For example, in visual grounding task, GPT-4o only detects highly salient structures, failing to support our goal of testing MLLMs on locating diverse buildings across complex scenes. As a result, all OmniEarth-Bench data was exclusively created by experts and annotators.

4 Experiment

Experimental Setup. The MLLMs evaluated on OmniEarth-Bench are grouped into two categories: (a) open-source VLMs, including Qwen2.5-VL [91], LLava-Onevision [92], InterVL3 [93] and InternLM-XComposer-2.5 [94]; (b) closed-source VLMs, such as GPT-4o [10], Gemini-2.0 [11] and Claude 3.7 Sonnet [12] All models were evaluated using LMMs-Eval [95, 96]. Following MMBench [13] and MME-Realworld [18] methods, in the VQA task, we manually created 5 options for each question: one correct answer, three distractors and one special answer (unable to answer). We evaluated the accuracy and reported of L-1 dimension for the VQA task, with L-3 and L-4 results available in the appendix. All scores in Tables 4 are reported as percentages (%). For the Grounding task, we used precision, assessing accuracy based on the intersection between predicted and ground truth bounding boxes, with predictions deemed correct if IoU exceeds a threshold (0.5 and 0.7).

All MLLMs exhibit suboptimal performance across all 7 domains. As illustrated in Tab. 4 and Tab. 5, nearly all MLLMs achieve accuracy rates below 40%, significantly underperforming relative to their success on traditional perception or reasoning benchmarks [13, 100, 41]. Several factors likely contribute to this challenge. First, current multimodal large models are typically trained without domain-specific Earth science data, which impedes their ability to comprehend related queries. Second, many Earth science problems are inherently complex, particularly cross-domain prediction tasks that demand in-depth, specialized knowledge, which existing LLMs or MLLMs may not possess. Finally, OmniEarth-Bench provides high-resolution, intricate imagery, and the task of interpreting such complex visuals presents unique obstacles for MLLMs. This underscores the pressing need for specialized models or advanced post-training techniques to effectively address these challenges.

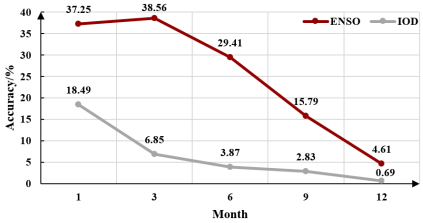

Figure 5: **GPT-4o performance on ENSO and IOD prediction with different lead months (previous).**

Time-sensitive task. The Earth's seven spheres encompass numerous temporally correlated tasks. ENSO, a key climate mode influencing global weather extremes via teleconnections [101], has seen improved forecasts through domain-specific AI models [79, 102]. As shown in Fig. 5, prediction accuracy declines with longer lead times, echoing the limitations of specialized models. However, performance of MLLMs still lags behind tailored models. Performance drops further for Indian Ocean Dipole (IOD) predictions, aligning with challenges faced by existing methods [103, 80].

Table 5: **Visual grounding performance on OmniEarth-Bench.** 3 spheres have the groudning task.

Benchmark	Method	GPT-4o	Gemini-2.0	Claude-3.7-Sonnet	Qwen2.5-VL	LLaVA-OneVision	InternVL3 8B	InternVL3 78B
Human-activities sphere	Acc@0.5	0.02	0.03	0	0.59	0.2	0	2.36
	Acc@0.7	0	0	0	0	0	0	0.2
Lithosphere	Acc@0.5	0.08	0.13	0.02	5.3	0	8.94	4.3
	Acc@0.7	0	0.04	0	0.33	0	1.66	0.33
Oceansphere	Acc@0.5	0.12	0.34	0.2	1.81	1.51	6.63	13.86
	Acc@0.7	0.01	0.06	0.07	0	0	0.6	3.61

Limited Gains from Scaling Model Size on Earth-Science Tasks. We evaluate two sizes of the InterVL3 model and find that the 72B InterVL3 does not provide a significant advantage over the 7B model in our benchmarks, with performance even declining in some evaluation metrics. This contrasts with the substantial improvements observed in general-domain tasks. This performance bottleneck likely stems from the lack of Earth-science-specific knowledge, rather than a limitation in model capacity. Even large MLLMs struggle to reason about unfamiliar scientific concepts without targeted training on domain-specific data. These findings highlight the importance of prioritizing the integration of domain-specific knowledge in future Earth-science MLLM development, rather than merely increasing model size.

Impact of Model Safety on Results. In Tab. 4, we observe that Qwen2.5-VL and GPT-4o perform very poorly, even falling below the level of random guessing. However, this does not mean that these two models have the worst perceptual and science-related abilities. We observe that these models tend to refuse to answer when they are uncertain, whereas InternVL3 and LLaVA-Onevision randomly guess an answer. This safety mechanism in the models leads to the poor performance of Qwen2.5-VL and GPT-4o. Detailed refusal rate statistics for each model are provided in the appendix—for instance, QwenVL2.5-VL-72B refused to answer 18,258 questions.

CoT performance. Following the MME-CoT [19], building upon our annotated key steps in Earth Observation data, we leverage two interpretable metrics to evaluate the CoT correctness: recall and precision. The two metrics respectively attend to the two aspects of the CoT correctness: informativeness and accuracy. Finally, we calculate the F1 score as the metric of CoT quality. As shown in Tab.6, InternVL3 outperformed Qwen2.5-VL and LaVA-OneVision with the highest F1 score. Larger open-source variants showed superior performance, underscoring the scalability of model size.

Table 6: **CoT performance of OmniEarth-Bench**

Models	LLaVA-OneVision-7B	Qwen2.5-VL-7B	InternVL3-8B	InternVL3-78B
Precision	89.83	92.72	94.02	94.74
Recall	23.41	29.12	34.47	35.5
F1	37.14	44.32	50.45	51.65

Expert Validation. Although OmniEarth-Bench's evaluation dimensions and data sources were curated by experts, we further validated its quality through expert upper-bound assessments. We randomly sampled 100 questions from each sphere and invited independent experts—unaffiliated with the annotation team—to answer them. As shown in Tab. 4, expert accuracy consistently exceeded 90%, confirming the benchmark's reliability. Occasional errors arose mainly in tasks requiring precise calculations or counting.

5 Conclusion

We have introduced OmniEarth-Bench, a comprehensive multimodal Earth science benchmark that encompasses all six spheres of the Earth system (atmosphere, lithosphere, Oceansphere, cryosphere, biosphere, and human-activities) along with their cross-sphere interactions. This benchmark introduces 100 expert-curated evaluation dimensions and four hierarchical levels of reasoning (perception, general reasoning, expert-knowledge reasoning, and chain-of-thought reasoning), representing a novel and rigorous evaluation design for geoscientific MLLMs. Our results show that even state-of-the-art MLLMs (e.g., Claude) struggle with OmniEarth-Bench; none of the tested models surpassed 35% accuracy. This stark performance gap underscores the benchmark's difficulty and exposes fundamental limitations in current models' geoscientific understanding. The significance of OmniEarth-Bench lies in its breadth and depth, providing a unified challenge that pushes beyond existing capabilities and highlighting the need for MLLMs that integrate visual perception with expert domain knowledge and advanced reasoning. We anticipate that OmniEarth-Bench will serve as a catalyst for future research in geoscientific AI, guiding the development of models capable of expert-level analysis across Earth's spheres and enabling advanced applications in environmental monitoring, climate science, and Earth system management. **Limitations.** Due to the high cost and difficulty of data acquisition, some spheres currently include only eight evaluation dimensions. We plan to expand these by partnering with relevant institutions and companies to obtain more data.

References

[1] Karianne J Bergen, Paul A Johnson, Maarten V de Hoop, and Gregory C Beroza. Machine learning for data-driven discovery in solid earth geoscience. *Science*, 363(6433):eaau0323, 2019.

[2] James Super. Geoscientists excluded. *Nature Geoscience*, 16(3):194–194, 2023.

[3] Brandon Victor, Mathilde Letard, Peter Naylor, Karim Douch, Nicolas Longépé, Zhen He, and Patrick Ebel. Off to new shores: A dataset & benchmark for (near-) coastal flood inundation forecasting. *Advances in Neural Information Processing Systems*, 37:114797–114811, 2024.

[4] Francesca Di Giuseppe, Joe McNorton, Anna Lombardi, and Fredrik Wetterhall. Global data-driven prediction of fire activity. *Nature Communications*, 16(1):2918, 2025.

[5] Benjamin J Wallis and Anna E Hogg. Satellite data shows antarctic peninsula glaciers flow faster in summer. *Nature Geoscience*, 16(3):196–197, 2023.

[6] Gustau Camps-Valls, Miguel-Ángel Fernández-Torres, Kai-Hendrik Cohrs, Adrian Höhl, Andrea Castelletti, Aytac Pacal, Claire Robin, Francesco Martinuzzi, Ioannis Papoutsis, Ioannis Prapas, et al. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature Communications*, 16(1):1919, 2025.

[7] Jesús Aguirre-Gutiérrez, Sami W Rifai, Xiongjie Deng, Hans Ter Steege, Eleanor Thomson, Jose Javier Corral-Rivas, Aretha Franklin Guimaraes, Sandra Muller, Joice Klipel, Sophie Fauset, et al. Canopy functional trait variation across earth's tropical forests. *Nature*, pages 1–8, 2025.

[8] François Keck, Tianna Peller, Roman Alther, Cécilia Barouillet, Rosetta Blackman, Eric Capo, Teofana Chonova, Marjorie Couton, Lena Fehlinger, Dominik Kirschner, et al. The global human impact on biodiversity. *Nature*, pages 1–6, 2025.

[9] Qiang Dai, Jingxuan Zhu, Guonian Lv, Latif Kalin, Yuanzhi Yao, Jun Zhang, and Dawei Han. Radar remote sensing reveals potential underestimation of rainfall erosivity at the global scale. *Science advances*, 9(32):eadg5551, 2023.

[10] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o, 2024.

[11] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[12] Anthropic. Anthropic ai, 2023.

[13] 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, 2024.

[14] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[16] 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.

[17] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

[18] 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.

[19] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

[20] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

[21] Pengfei Zhou, Fanrui Zhang, Xiaopeng Peng, Zhaopan Xu, Jiaxin Ai, Yansheng Qiu, Chuanhao Li, Zhen Li, Ming Li, Yukang Feng, et al. Mdk12-bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models. *arXiv preprint arXiv:2504.05782*, 2025.

[22] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[23] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: a benchmark for vision-language instruction following inspired by real-world use. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 26898–26922, 2023.

[24] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. In *International Conference on Machine Learning*, pages 57116–57198. PMLR, 2024.

[25] 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.

[26] Fengxiang Wang, Hongzhen Wang, Mingshuo Chen, Di Wang, Yulin Wang, Zonghao Guo, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? *arXiv preprint arXiv:2503.23771*, 2025.

[27] Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10):10707–10715, 2025.

[28] Eric Vermote. Mod09a1 modis/terra surface reflectance 8-day l3 global 500m sin grid v006, 2015.

[29] 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.

[30] 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.

[31] Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*, 2024.

[32] Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Zixuan Yuan, Bing Zhu, and Junwei Liang. Vision-language models meet meteorology: Developing models for extreme weather events detection with heatmaps. *arXiv preprint arXiv:2406.09838*, 2024.

[33] 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.

[34] Haobo Li, Zhaowei Wang, Jiachen Wang, Alexis Kai Hon Lau, and Huamin Qu. Cllmate: A multimodal llm for weather and climate events forecasting. *arXiv preprint arXiv:2409.19058*, 2024.

[35] Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 404–412, 2022.

[36] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[37] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[38] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[39] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[40] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[41] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[42] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[43] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[44] Mélisande Teng, Amna Elmustafa, Benjamin Akera, Yoshua Bengio, Hager Radi, Hugo Larochelle, and David Rolnick. Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data. *Advances in Neural Information Processing Systems*, 36:75925–75950, 2023.

[45] Matthew Fortier, Mats L Richter, Oliver Sonnentag, and Chris Pal. Carbonsense: A multimodal dataset and baseline for carbon flux modelling. *arXiv preprint arXiv:2406.04940*, 2024.

[46] Xingliang Huang, Kaiqiang Chen, Deke Tang, Chenglong Liu, Libo Ren, Zheng Sun, Ronny Hänsch, Michael Schmitt, Xian Sun, Hai Huang, et al. Urban building classification (ubc) v2—a benchmark for global building detection and fine-grained classification from satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

[47] Yinxia Cao and Qihao Weng. A deep learning-based super-resolution method for building height estimation at 2.5 m spatial resolution in the northern hemisphere. *Remote Sensing of Environment*, 310:114241, 2024.

[48] Jiayi Li, Xin Huang, and Lilin Tu. Whu-ohs: A benchmark dataset for large-scale hersepctral image classification. *International Journal of Applied Earth Observation and Geoinformation*, 113:103022, 2022.

[49] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 10–17, 2019.

[50] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024.

[51] Yifei Qian, Grant RW Humphries, Philip N Trathan, Andrew Lowther, and Carl R Donovan. Counting animals in aerial images with a density map estimation model. *Ecology and Evolution*, 13(4):e9903, 2023.

[52] Josh Veitch-Michaelis, Andrew Cottam, Daniella Schweizer, Eben Broadbent, David Dao, Ce Zhang, Angelica Almeyda Zambrano, and Simeon Max. Oam-tcd: A globally diverse dataset of high-resolution tree cover maps. *Advances in Neural Information Processing Systems*, 37:49749–49767, 2024.

[53] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1765–1774. IEEE, 2025.

[54] Shunlin Liang, Jie Cheng, Kun Jia, Bo Jiang, Qiang Liu, Zhiqiang Xiao, Yunjun Yao, Wenping Yuan, Xiaotong Zhang, Xiang Zhao, et al. The global land surface satellite (glass) product suite. *Bulletin of the American Meteorological Society*, 102(2):E323–E337, 2021.

[55] Daniyar B Nurseitov, Galymzhan Abdimanap, Abdelrahman Abdallah, Gulshat Sagatdinova, Larissa Balakay, Tatyana Dedova, Nurkuisa Rametov, and Anel Alimova. Rosid: Remote sensing satellite data for oil spill detection on land. *Engineered Science*, 32:1348, 2024.

[56] Eric W Sanderson, Malanding Jaiteh, Marc A Levy, Kent H Redford, Antoinette V Wannebo, and Gillian Woolmer. The human footprint and the last of the wild: the human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. *BioScience*, 52(10):891–904, 2002.

[57] Juepeng Zheng, Haohuan Fu, Weijia Li, Wenzhao Wu, Le Yu, Shuai Yuan, Wai Yuk William Tao, Tan Kian Pang, and Kasturi Devi Kanniah. Growing status observation for oil palm trees using unmanned aerial vehicle (uav) images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:95–121, 2021.

[58] Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in Neural Information Processing Systems*, volume 33, pages 22009–22019, 2020.

[59] Asanobu Kitamoto, Jared Hwang, Bastien Vuillod, Lucas Gautier, Yingtao Tian, and Tarin Clanuwat. Digital typhoon: Long-term satellite image dataset for the spatio-temporal modeling of tropical cyclones. *Advances in Neural Information Processing Systems*, 36:40623–40636, 2023.

[60] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

[61] S Mostafa Mousavi, Yixiao Sheng, Weiqiang Zhu, and Gregory C Beroza. Stanford earthquake dataset (stead): A global data set of seismic signals for ai. *IEEE Access*, 7:179464–179476, 2019.

[62] S Kainkaryam, C Ong, S Sen, and A Sharma. Crowdsourcing salt model building: Kaggle-tgs salt identification challenge. In *81st EAGE Conference and Exhibition 2019*, pages 1–5. European Association of Geoscientists & Engineers, 2019.

[63] Katerina Kikaki, Ioannis Kakogeorgiou, Ibrahim Hoteit, and Konstantinos Karantzalos. Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024.

[64] Boyin Huang, Peter W Thorne, Viva F Banzon, Tim Boyer, Gennady Chepurin, Jay H Lawrimore, Matthew J Menne, Thomas M Smith, Russell S Vose, and Huai-Min Zhang. Noaa extended reconstructed sea surface temperature (ersst), version 5. *(No Title)*, 2017.

[65] Yan Wang, Ge Chen, Jie Yang, Zhipeng Gui, and Dehua Peng. A submesoscale eddy identification dataset in the northwest pacific ocean derived from goci i chlorophyll a data based on deep learning. *Earth System Science Data*, 16(12):5737–5752, 2024.

[66] Mengqiu Xu, Ming Wu, Kaixin Chen, Yixiang Huang, Mingrui Xu, Yujia Yang, Yiqing Feng, Yiying Guo, Bin Huang, Dongliang Chang, et al. M4fog: A global multi-regional, multi-modal, and multi-stage dataset for marine fog detection and forecasting to bridge ocean and atmosphere. *arXiv preprint arXiv:2406.13317*, 2024.

[67] W.N. Meier, F. Fetterer, A.K. Windnagel, and J.S. Stewart. Noaa/nsidc climate data record of passive microwave sea ice concentration. (g02202, version 4). *National Snow and Ice Data Center*, 2021.

[68] Josefino Comiso. Bootstrap sea ice concentrations from nimbus-7 smmr and dmsp ssm/i-ssmis, version 4, 2023.

[69] Axel Schweiger, Ronald Lindsay, Jinlun Zhang, Michael Steele, and H Stern. Uncertainty in modeled arctic sea ice volume. *Journal of Geophysical Research*, 116(C00D06):1–15, 2011.

[70] Jinlun Zhang and D.A. Rothrock. Modeling global sea ice with a thickness and enthalpy distribution model in generalized curvilinear coordinates. *Monthly Weather Review*, 131(5):681–697, 2003.

[71] V. Helm, A. Humbert, and H. Miller. Elevation and elevation change of greenland and antarctica derived from cryosat-2. *The Cryosphere*, 8(4):1539–1559, 2014.

[72] Michael Studinger. Icebridge atm l2 icessn elevation, slope, and roughness, version 2, 2014.

[73] Ben Smith, Susheel Adusumilli, Be??ta Csathó, Denis Felikson, Helen Fricker, Alex Gardner, Nick Holschuh, Jeff Lee, Johan Nilsson, Fernando Paolo, Matthew Siegfried, Tyler Sutterley, and The ICESat-2 Science Team. Atlas/icesat-2 l3a land ice height, version 6, 2023.

[74] Richard M Allen and Diego Melgar. Earthquake early warning: Advances, scientific challenges, and societal needs. *Annual Review of Earth and Planetary Sciences*, 47(1):361–388, 2019.

[75] Siwei Yu and Jianwei Ma. Deep learning for geophysics: Current and future trends. *Reviews of Geophysics*, 59(3):e2021RG000742, 2021.

[76] Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*, 2024.

[77] Jason Stock, Kyle Hilburn, Imme Ebert-Uphoff, and Charles Anderson. Srvit: Vision transformers for estimating radar reflectivity from satellite observations at scale. *arXiv preprint arXiv:2406.16955*, 2024.

[78] Rami Al-Ruzouq, Mohamed Barakat A Gibril, Abdallah Shanableh, Abubakir Kais, Osman Hamed, Saeed Al-Mansoori, and Mohamad Ali Khalil. Sensors, features, and machine learning for oil spill detection and monitoring: A review. *Remote Sensing*, 12(20):3338, 2020.

[79] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year enso forecasts. *Nature*, 573(7775):568–572, 2019.

[80] Fenghua Ling, Jing-Jia Luo, Yue Li, Tao Tang, Lei Bai, Wanli Ouyang, and Toshio Yamagata. Multi-task machine learning improves multi-seasonal prediction of the indian ocean dipole. *Nature Communications*, 13(1):7681, 2022.

[81] Zijun Duo, Wenke Wang, and Huizan Wang. Oceanic mesoscale eddy detection method based on deep learning. *Remote Sensing*, 11(16):1921, 2019.

[82] Darko Koračin, Clive E Dorman, John M Lewis, James G Hudson, Eric M Wilcox, and Alicia Torregrosa. Marine fog: A review. *Atmospheric research*, 143:142–175, 2014.

[83] Dagmar Budikova. Role of arctic sea ice in global atmospheric circulation: A review. *Global and Planetary Change*, 68(3):149–163, 2009.

[84] W. Zhou, L.R. Leung, and J. Lu. Steady threefold arctic amplification of externally forced warming masked by natural variability. *Nat. Geosci.*, 17:508–515, 2024.

[85] S.A. Khan, Y. Choi, M. Morlighem, et al. Extensive inland thinning and speed-up of northeast greenland ice stream. *Nature*, 611:727–732, 2022.

[86] T.R. Chudley, I.M. Howat, M.D. King, et al. Increased crevassing across accelerating greenland ice sheet margins. *Nat. Geosci.*, 18:148–153, 2025.

[87] Thomas W Crowther, Henry B Glick, Kristofer R Covey, Charlie Bettigole, Daniel S Maynard, Stephen M Thomas, Jeffrey R Smith, Gregor Hintler, Marlyse C Duguid, Giuseppe Amatulli, et al. Mapping tree density at a global scale. *Nature*, 525(7568):201–205, 2015.

[88] Oscar Venter, Eric W Sanderson, Ainhoa Magrach, James R Allan, Jutta Beher, Kendall R Jones, Hugh P Possingham, William F Laurance, Peter Wood, Balázs M Fekete, et al. Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nature communications*, 7(1):12558, 2016.

[89] Daniyar B Nurseitov, Galymzhan Abdimanap, Abdelrahman Abdallah, Gulshat Sagatdinova, Larissa Balakay, Tatyana Dedova, Nurkuisa Rametov, and Anel Alimova. Rosid: Remote sensing satellite data for oil spill detection on land. *Engineered Science*, 32:1348, 2024.

[90] Juepeng Zheng, Haohuan Fu, Weijia Li, Wenzhao Wu, Le Yu, Shuai Yuan, Wai Yuk William Tao, Tan Kian Pang, and Kasturi Devi Kanniah. Growing status observation for oil palm trees using unmanned aerial vehicle (uav) images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:95–121, 2021.

[91] 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.

[92] 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.

[93] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[94] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.

[95] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024.

[96] Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimodal models. https://github.com/EvolvingLMMs-Lab/lmms-eval, March 2024.

[97] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[98] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[99] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[100] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[101] Axel Timmermann, Soon-Il An, Jong-Seong Kug, Fei-Fei Jin, Wenju Cai, Antonietta Capotondi, Kim M Cobb, Matthieu Lengaigne, Michael J McPhaden, Malte F Stuecker, et al. El niño–southern oscillation complexity. *Nature*, 559(7715):535–545, 2018.

[102] Zijie Guo, Pumeng Lyu, Fenghua Ling, Lei Bai, Jing-Jia Luo, Niklas Boers, Toshio Yamagata, Takeshi Izumo, Sophie Cravatte, Antonietta Capotondi, et al. Data-driven global ocean modeling for seasonal to decadal prediction. *arXiv preprint arXiv:2405.15412*, 2024.

[103] Jun Liu, Youmin Tang, Yanling Wu, Tang Li, Qiang Wang, and Dake Chen. Forecasting the indian ocean dipole with deep learning techniques. *Geophysical Research Letters*, 48(20):e2021GL094407, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction accurately reflect the scope of our dataset and our contributions to the field of Earth sciences.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We openly acknowledge our limitations. Due to the high cost and difficulty of data acquisition, some spheres currently include only eight evaluation dimensions. We plan to expand these by partnering with relevant institutions and companies to obtain more data.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer:[Yes]

Justification: We conducted thorough analyses of the proposed dataset, each supported by complete evidence.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our experimental results are reproducible under the described setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided a direct link to access the dataset.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Our results can be reproduced by following the provided experimental guidelines and dataset usage instructions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All experiments were conducted three times using standard settings, and the results were averaged.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All our experiments are evaluation-based and follow the VLMEval framework referenced in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We strictly adhered to ethical guidelines and ensured the preservation of anonymity.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed the significance of conducting comprehensive evaluations across the entire Earth science domain.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our paper does not involve data with a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All our data was obtained through publicly accessible sources, downloaded legitimately, and processed and annotated by our team. The proposed OmniEarth-Bench follows a highly permissive license (CC-BY 4.0), enabling broad use for various evaluations.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have documented the dataset and made it publicly available, along with detailed usage instructions.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer:[Yes]

    Justification: We provided generous compensation to crowdsourcing annotators and offered detailed explanations of their collaboration with domain experts.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Our annotation process was thoroughly explained to crowdsourcing annotators in advance. With their informed consent and approval from relevant organizations, we proceeded with the annotation work, providing substantial compensation.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We only used LLMs to assist with writing, complying with the LLM policy. For benchmark evaluation, MLLMs were employed as assessment tools, following a widely accepted evaluation protocol in this field.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.