EARTHSPATIALBENCH: BENCHMARKING SPATIAL REASONING CAPABILITIES OF MULTIMODAL LLMS ON EARTH IMAGERY

Zelin Xu[1], **Yupu Zhang**[1], **Saugat Adhikari**[2], **Saiful Islam**[1], **Tingsong Xiao**[1], **Zibo Liu**[1], **Shigang Chen**[1], **Da Yan**[2], **Zhe Jiang**[1]
[1]Department of Computer & Information Science & Engineering, University of Florida
[2]Department of Computer Science, Indiana University Bloomington
{zelin.xu, zhe.jiang}@ufl.edu

ABSTRACT

Benchmarking spatial reasoning in multimodal large language models (MLLMs) has attracted growing interest in computer vision due to its importance for embodied AI and other agentic systems that require precise interaction with the physical world. However, spatial reasoning on Earth imagery has lagged behind, as it uniquely involves grounding objects in georeferenced images and quantitatively reasoning about distances, directions, and topological relations using both visual cues and vector geometry coordinates (e.g., 2D bounding boxes, polylines, and polygons). Existing benchmarks for Earth imagery primarily focus on 2D spatial grounding, image captioning, and coarse spatial relations (e.g., simple directional or proximity cues). They lack support for quantitative direction and distance reasoning, systematic topological relations, and complex object geometries beyond bounding boxes. To fill this gap, we propose **EarthSpatialBench**, a comprehensive benchmark for evaluating spatial reasoning in MLLMs on Earth imagery. The benchmark contains over 325K question–answer pairs spanning: (1) qualitative and quantitative reasoning about spatial distance and direction; (2) systematic topological relations; (3) both single-object and object-pair queries and compositional aggregate group queries; and (4) object references expressed via textual descriptions, visual overlays, and explicit geometric coordinates, including 2D bounding boxes, polylines, and polygons. We conducted extensive experiments on both open-source and proprietary models to identify limitations in the spatial reasoning of MLLMs.

1 INTRODUCTION

Benchmarking the spatial reasoning capabilities of multimodal large language models (MLLMs) (Achiam et al., 2023; Liu et al., 2023; Bai et al., 2023; Chen et al., 2024b) has attracted growing interest in computer vision (Yang et al., 2025; Deng et al., 2025; Cheng et al., 2024; Chen et al., 2024a; Ma et al., 2025). Spatial reasoning involves analyzing and interpreting the positions, orientations, and spatial relationships of objects based on 2D photos. It is crucial for embodied AI and other agentic systems that require precise interaction with the physical world. A large number of benchmarking datasets have been developed for natural images (Deng et al., 2025; Chen et al., 2024a; Cheng et al., 2024; Shiri et al., 2024; Cai et al., 2025; Zhang et al., 2025; Li et al., 2024a).

However, benchmarking spatial reasoning on Earth imagery has lagged behind. Earth imagery, spanning satellite, aerial, and drone observations, is pervasive. Benchmarking and developing spatial reasoning capabilities of MLLMs on Earth imagery has the potential to solve many grand societal challenges, such as natural disaster response, urban planning, precision agriculture, and ecological monitoring (Stallings, 2003; Shi et al., 2020; Shekhar et al., 2015; Shekhar & Xiong, 2007). For example, during a deadly flood disaster, MLLMs can assist first responders in quickly locating the impacted communities, estimating the economic loss, and planning rescue operations (e.g., counting damaged houses within certain neighborhood blocks, and locating them within a distance range from major roads).

Table 1: A comparison of existing benchmark datasets for MLLMs on Earth imagery. (✓ "Fully Benchmarked"; △ "Partially Benchmarked"; × "Not Benchmarked")

Datasets	Number of QA Pairs	Geometry Type	Spatial Reasoning		
			Distance	Direction	Topological
EarthVQA (Wang et al., 2024)	209K	BBox	×	×	×
GeoChatSet (Kuckreja et al., 2024)	318K	BBox	×	×	×
VRSBench (Li et al., 2024b)	123K	BBox	×	×	×
SkyEyeGPT (Zhan et al., 2025)	968K	BBox	×	×	×
XLRS-Bench (Wang et al., 2025a)	46K	BBox	×	△	×
GeoBench-VLM (Danish et al., 2025)	10K	BBox	×	△	×
DisasterM3 (Wang et al., 2025b)	123K	BBox	×	△	×
EarthSpatialBench	**325K**	**BBox, Polygon, Polyline**	✓	✓	✓

Spatial reasoning on Earth imagery poses several unique challenges. First, compared with natural imagery (on which existing MLLMs are largely pretrained), Earth imagery is often noisy with denser objects (e.g., tens of houses) in a bird's-eye view (e.g., houses are tiny with only roofs visible). Second, an object in Earth imagery can be referred to in multiple ways, including both textual descriptions of visual cues (e.g., object types and spatial context) and vector geometric coordinates (e.g., 2D bounding boxes, polylines, and polygons). Third, spatial reasoning involves the estimation of quantitative distance and direction (azimuth angle), as well as complex topological relations between objects in rich geometries (e.g., counting houses within a certain distance of a river, locating houses within a park polygon).

Table 1 summarizes the existing MLLM benchmarking datasets on Earth imagery. These datasets mainly focus on spatial grounding (e.g., locating 2D BBoxes of objects on Earth imagery). A few datasets (Wang et al., 2025a; Danish et al., 2025; Wang et al., 2025b) include coarse direction queries (e.g., left of, above, beside), and sometimes the reference to input objects is based on drawing boxes on the original Earth imagery. Almost none include quantitative direction (azimuth angle), distance metrics, and spatial topological relations. In addition, these datasets do not incorporate rich object geometry (polyline, polygon).

To fill this gap, we propose **EarthSpatialBench**, a comprehensive benchmark for evaluating spatial reasoning in MLLMs on Earth imagery. The benchmark contains over 325K question–answer pairs spanning: (1) qualitative and quantitative reasoning about spatial distance and direction; (2) systematic topological relations; (3) both single-object and object-pair queries and compositional aggregate group queries based on distance, direction, or topological relation; and (4) object references expressed via both textual descriptions, visual overlays, and explicit geometry coordinates, including 2D bounding boxes, polylines, and polygons. We conducted evaluations of multiple MLLMs on our datasets to identify limitations in the spatial reasoning.

2 EARTHSPATIALBENCH

To guide the spatial reasoning evaluation of MLLMs on Earth imagery, as shown in Figure 1, EarthSpatialBench is designed around three core research questions:

- **(RQ1) Spatial Relation Understanding:** How well can MLLMs conduct spatial reasoning based on different types of spatial relations on Earth imagery (e.g., distance, direction, and topological relations)?
- **(RQ2) Reasoning with Different Geometric Representations:** When performing spatial reasoning, how effectively do MLLMs handle different reference expressions of objects, including textual description (object types, spatial context), visual overlays, and explicit geometric coordinates of objects? How do they perform on objects in different geometric types, such as 2D bounding boxes, polygons, and polylines?
- **(RQ3) Task Format and Reasoning Difficulty:** How do MLLMs perform across different question formats, including qualitative questions (e.g., discrete choices, yes/no) versus quantitative (Azimuth degree, Euclidean distance), single-object pair (exact distance or direction between two objects) versus composite group aggregate questions (e.g., objects within a certain distance or direction range from another object), and counting versus localization questions.

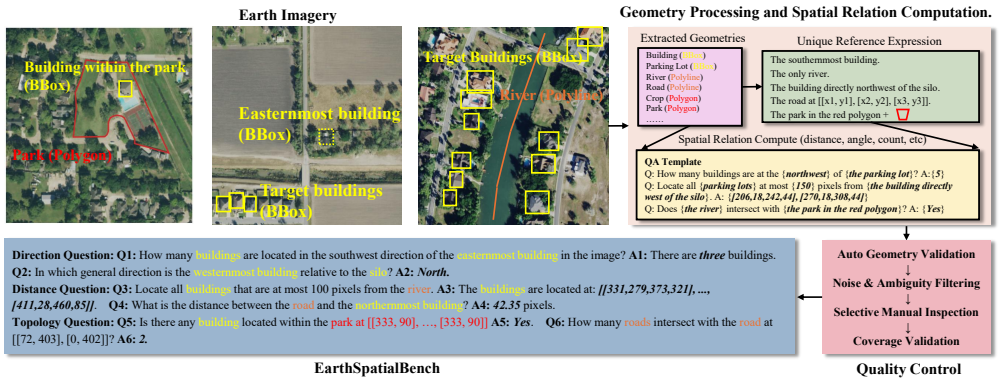


Figure 1: Overview of EarthSpatialBench. The benchmark spans diverse spatial relations evaluated over multiple geometry types, object representation modalities, and question formats, enabling comprehensive assessment of spatial reasoning. High-resolution Earth imagery and geometric annotations are collected from SatlasPretrain. Spatial relations (distance, direction, topology) are computed using standard GIS procedures and combined with curated question templates to generate diverse reasoning tasks. A multi-stage quality control process ensures geometric accuracy, semantic clarity, and reliable ground-truth annotations.

These research questions allow us to systematically characterize the strengths and limitations of current MLLMs in geospatial reasoning on Earth imagery data. Accordingly, EarthSpatialBench is organized around orthogonal dimensions that collectively capture the breadth of spatial reasoning required for Earth imagery.

2.1 Evaluation Dimensions

Spatial Relations. We categorize spatial relations into three main types: distance, direction, and topological relations. Together, they capture the common spatial relationships involved in geospatial reasoning on Earth imagery.

- **Distance.** Distance relations describe how far objects or regions are from one another under a metric. They are central to proximity analysis and buffer-based operations. EarthSpatialBench includes tasks that require estimating the distance between two objects, determining whether two objects fall within a specified distance threshold, and counting or identifying objects within a defined buffer.
- **Direction.** Direction relations encode the relative orientation of objects or regions, expressed either as angular measurements or as coarse directional categories (for example, the eight compass directions). Such reasoning is important for navigation, spatial alignment, and understanding scene layout. EarthSpatialBench includes queries that involve estimating angles between objects, assigning objects to directional bins, and counting or locating objects within particular directional sectors.
- **Topological relations.** Topological relations describe how geometric entities connect, overlap, or are embedded within one another, independent of distance or orientation. We focus on two common cases in geospatial analysis: within (containment) and intersect. These tasks involve determining whether one region lies entirely inside another, identifying intersections between geometries, and locating objects that satisfy these relationships.

Together, distance, direction, and topology provide a principled decomposition of spatial reasoning that goes well beyond conventional perception benchmarks.

Geometry Types. Spatial relations in Earth imagery are expressed over geometric entities rather than only over discrete object instances. EarthSpatialBench explicitly models three primary geometric types: bounding boxes, polygons, and polylines, which together cover a broad set of common geospatial elements. We exclude point geometries because they are typically not visible at Earth imagery resolution. Although a bounding box is formally a simple polygon, we treat it as a separate category because it serves as a stable, perceptually grounded representation for small objects (e.g.,

buildings), while polygons preserve the true shapes of larger region-level entities (e.g., parks, airports) where geometry is essential for reasoning. By systematically combining these geometric types (e.g., polyline–polygon, polyline–bounding box), EarthSpatialBench covers a rich spectrum of real-world spatial configurations and allows EarthSpatialBench to evaluate both object-level and region-level spatial reasoning.

- **Bounding boxes.** Used for small, visually identifiable objects such as buildings and facilities. For example, *"Locate the building (bounding box) closest to the railway (polyline)."*
- **Polygons.** Represent areal regions such as parks, agricultural fields, and lakes. For example, *"How many river segments (polyline) intersect this park (polygon)?"*
- **Polylines.** Capture linear features such as roads, railways, and rivers. For example, *"Does this road segment (polyline) intersect the river (polyline)?"*

Object Representations. Objects in Earth imagery can be referenced through multiple representation modalities. EarthSpatialBench supports three complementary object representation methods: textual descriptions, visual overlays, and explicit geometric coordinates.

- **Textual descriptions.** Objects are specified through semantic attributes and spatial context, such as object types, functional roles, and surrounding features. For example, *unique object references* (e.g., the only hospital in the image), *extreme position references* (e.g., the northernmost building), and *spatial context references* (e.g., the building north of a unique parking lot). This format evaluates whether models can ground diverse linguistic cues in complex geospatial scenes.
- **Visual overlays.** Target objects are highlighted directly on the image using visual markers or masks. Visual overlays facilitate explicit visual grounding and reduce ambiguity in object referencing.
- **Geometric coordinates.** Objects are represented through precise spatial specifications in the form of bounding boxes, polylines, or polygons. This format enables geometry-aware reasoning and tests whether models can leverage structured spatial inputs.

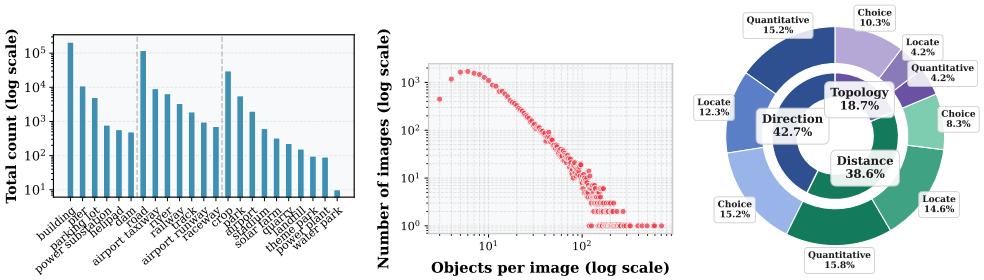
Question Formats. Beyond the spatial relation and geometry involved, the formulation of a question determines the type of reasoning an MLLM must perform. EarthSpatialBench includes three complementary question formats: choice-based, quantitative, and localization, each probing a different aspect of spatial reasoning.

- **Choice-based questions.** These require selecting from a discrete set of options, such as Yes or No, or one of the cardinal directions. Choice-based formats emphasize categorical spatial decisions and are less sensitive to small numerical deviations, making them reliable for robust comparison across models.
- **Quantitative questions.** These require explicit numerical reasoning, including counting objects, estimating distances, or estimating angles. Quantitative tasks test whether MLLMs can integrate spatial scale, extent, and relative geometry rather than relying solely on visual recognition.
- **Localization questions.** These focus on spatial grounding by asking the model to identify or locate specific objects or regions that satisfy a spatial constraint. Localization tasks assess whether an MLLM can connect abstract relational reasoning back to concrete image regions, which is essential for practical geospatial analysis.

2.2 Dataset Construction

Data Source. We collect source data from SatlasPretrain (Bastani et al., 2023), a large-scale publicly available remote sensing dataset containing high-resolution imagery paired with geometric object annotations. We first retain only images that contain geometric objects with visually discernible footprints, focusing primarily on buildings and other man-made structures. To ensure the presence of non-trivial spatial relationships, we further require that each image contains at least two distinct semantic categories beyond generic building footprints (e.g., river, road, park, dam). This avoids degenerate cases where only a single object type appears and no meaningful geometric relations can be formed.

Geometry Processing and Spatial Relation Computation. SatlasPretrain provides polygon and polyline labels for a variety of object classes. We preserve polygons for medium- and large-scale regions such as stadiums or airports. For small-scale objects such as buildings and parking lots whose polygon shapes may be too fine-grained for consistent visual interpretation, we convert their polygonal annotations into axis-aligned bounding boxes. For every pair of geometry instances in an



(a) (b) (c)

Figure 2: (a) Distribution of all 23 geometric object types in EarthSpatialBench (log scale). The dataset includes 6 small-object bounding-box classes (building, dam, parking lot, pier, power substation, helipad), 7 polyline classes (airport runway, airport taxiway, raceway, railway, river, road, track), and 10 large-region polygon classes (airport, crop, landfill, park, quarry, solar farm, water park, theme park, stadium, power plant). (b) Distribution of the total number of annotated objects per image on a log–log scale, showing the wide variation in scene complexity across images. (c) Distribution of question counts across spatial relation categories (distance, direction, topology) and question formats (choice-based, quantitative, localization).

image, we compute spatial relationships following standard GIS procedures. Distance is computed using Euclidean metrics for geometries (the nearest distance for each point in each object), and buffer queries use a consistent metric scale derived from georeferenced pixel resolution. Direction is computed using the azimuth angle between object centroids and discretized into 8-direction bins. Topology is computed using the DE-9IM model, the standard formalism for topological relations in computational geometry. We combine computed spatial relations with curated question templates for each reasoning category (distance, direction, topology). Templates cover choice-style, quantitative, and localization tasks. The final dataset includes multiple question types per relation, with both positive and negative exemplars to avoid answer biases. Because all answers are derived from exact geometry computations, EarthSpatialBench provides unambiguous ground-truth labels.

Quality Control. To ensure that EarthSpatialBench provides reliable spatial reasoning tasks, we implement a multi-stage quality assurance pipeline that combines automated checks with targeted manual review. We first perform automated checks to confirm that all polygons, polylines, and bounding boxes fall within image bounds and have non-zero area or length. Objects that are too small, noisy, or visually ambiguous (such as chimneys, wells, mineshafts, or flagpoles) are removed. We also exclude geometry pairs whose spatial relations are unstable or ambiguous, including cases where boundaries are nearly touching and produce inconsistent topological results under minor geometric variation. We then conduct manual review on a selected subset of samples that require closer inspection. These include (i) images containing very large geometric regions, (ii) scenes with high object density (for example, more than 100 buildings), (iii) objects intersecting image boundaries, and (iv) polygons with highly complex shapes or a large number of vertices. Observations from this review are used to refine question templates, remove problematic examples, and adjust geometric thresholds. To ensure comprehensive coverage, we also verify that each spatial relation type includes diverse examples across geometry types and object categories. Through this multi-stage verification process, EarthSpatialBench maintains high geometric fidelity, semantic clarity, and visual interpretability, enabling robust evaluation of spatial reasoning in Earth imagery.

2.3 Dataset Statistics

EarthSpatialBench is constructed to provide large-scale, geometry-rich supervision for evaluating spatial reasoning in multimodal models. Table 2 summarizes the key dataset properties. The benchmark contains 21K optical RGB images sourced from SatlasPretrain, each captured at approximately 1 m ground resolution and standardized to 512 × 512 pixels. On average, each image contains 19.3 annotated geometric entities, including 10.7 bounding-box objects, 6.7 polylines, and 1.9 polygonal regions, reflecting the heterogeneous composition of real-world Earth imagery.

Table 2: Overview of EarthSpatialBench.

Property	**Value**
Source dataset	SatlasPretrain (Bastani et al., 2023)
Sensor type	Optical RGB
Image resolution	1m
Image size	512 × 512
Total images	21K
Total QA pairs	325K
Bbox objects	6 classes
Polyline objects	7 classes
Polygon objects	10 classes
Average objects per image	19
Average bboxes per image	11
Average polylines per image	7
Average polygons per image	2

Figure 2 provides additional insights into dataset composition. The distribution of geometry types indicates substantial variation in the spatial structures present across scenes, ranging from dense built environments to large-region land-cover polygons. The wide spread in object counts per image demonstrates that the benchmark captures both simple and highly complex spatial layouts, which is essential for evaluating model robustness under varying scene densities. The distribution of question types further confirms that the benchmark spans a broad range of reasoning demands, ensuring that no single relation type or question format dominates the evaluation. These statistics show that EarthSpatialBench offers a large, diverse, and geometrically complex foundation for systematically evaluating MLLMs' spatial reasoning capabilities in Earth imagery.

Table 3: Performance on **Distance** reasoning tasks.

Model	Exact Distance		Count		Localization		Binary Cls.	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	F1 ↑	IoU ↑	Acc ↑	F1 ↑
Qwen3VL-I-30B	127.93	168.31	1.28	2.06	0.25	0.18	0.58	0.56
Qwen3VL-I-30B-CoT	124.78	169.16	1.19	1.41	0.16	0.16	0.53	0.34
Qwen3VL-T-30B	114.87	158.92	1.02	1.77	0.25	0.19	0.71	0.48
Qwen3VL-T-30B-CoT	121.17	158.39	1.33	1.68	**0.26**	**0.20**	0.56	0.34
InternVL3.5-38B	144.21	189.73	**0.89**	**1.19**	0.23	0.19	0.59	0.54
InternVL3.5-38B-CoT	131.51	174.07	1.21	1.47	0.16	0.13	0.60	0.53
Claude-3.5-Sonnet	146.51	191.15	1.47	2.21	0.12	0.05	0.60	0.59
Claude-3.5-Sonnet-CoT	115.86	155.57	1.24	1.63	0.07	0.02	0.69	0.66
Gemini-2.5-Pro	119.44	161.97	1.90	3.27	0.23	0.12	0.74	0.74
Gemini-2.5-Pro-CoT	97.82	134.22	1.65	1.77	0.24	0.10	0.72	0.71
GPT-5	**93.91**	150.61	1.73	2.85	0.18	0.07	**0.86**	**0.86**
GPT-5-CoT	100.23	**129.66**	1.15	1.62	0.12	0.04	0.70	0.67

Table 4: Performance on **Direction** reasoning tasks.

Model	Exact Angle		Count		Localization		8-Direction Cls.		Binary Cls.	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	F1 ↑	IoU ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Qwen3VL-I-30B	117.20	145.86	2.88	6.37	**0.24**	**0.16**	0.35	0.31	0.67	0.67
Qwen3VL-I-30B-CoT	80.31	130.91	1.74	5.21	0.04	0.05	0.34	0.34	0.57	0.38
Qwen3VL-T-30B	117.98	140.94	1.64	**2.90**	0.21	0.13	0.26	0.26	0.71	0.71
Qwen3VL-T-30B-CoT	**54.99**	96.28	2.24	6.37	0.16	0.11	0.45	0.38	0.70	0.46
InternVL3.5-38B	126.31	158.25	1.55	3.14	0.16	0.11	0.30	0.29	0.64	0.64
InternVL3.5-38B-CoT	71.06	113.33	1.47	3.30	0.16	0.10	0.25	0.24	0.54	0.50
Claude-3.5-Sonnet	130.18	154.41	1.61	3.10	0.07	0.03	0.36	0.31	0.67	0.67
Claude-3.5-Sonnet-CoT	65.68	112.24	1.74	3.32	0.10	0.03	0.45	0.41	0.70	0.68
Gemini-2.5-Pro	134.47	159.88	2.02	3.91	0.09	0.03	0.37	0.31	0.67	0.67
Gemini-2.5-Pro-CoT	61.20	97.58	1.69	3.71	0.07	0.03	0.48	0.46	0.66	0.62
GPT-5	119.26	142.71	1.76	3.34	0.12	0.05	0.43	0.41	0.70	0.70
GPT-5-CoT	55.66	**95.78**	**1.46**	2.97	0.11	0.04	**0.51**	**0.49**	**0.73**	**0.72**

Table 5: Performance on **Topological** reasoning tasks.

Model	Count				Localization				Binary Cls.			
	Intersect		Within		Intersect		Within		Intersect		Within	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	F1 ↑	IoU ↑	F1 ↑	IoU ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Qwen3VL-I-30B	1.14	**1.82**	**2.45**	**4.30**	0.24	0.19	0.13	0.15	0.68	0.63	0.90	0.71
Qwen3VL-I-30B-CoT	1.10	1.98	2.83	6.55	0.19	0.17	0.15	0.13	0.71	0.69	0.94	0.77
Qwen3VL-T-30B	1.07	1.88	2.79	6.54	0.21	0.18	**0.21**	0.20	0.73	0.72	0.91	0.68
Qwen3VL-T-30B-CoT	1.07	2.09	3.06	6.84	0.19	0.18	0.12	0.07	0.72	0.72	0.94	0.68
InternVL3.5-38B	1.17	2.00	3.23	6.86	0.15	0.13	0.04	0.09	0.60	0.58	0.92	0.65
InternVL3.5-38B-CoT	1.26	2.07	3.27	6.87	0.16	0.19	0.02	**0.42**	0.72	0.70	0.93	0.59
Claude-3.5-Sonnet	1.23	1.97	2.83	6.49	0.14	0.15	0.13	0.06	0.61	0.56	0.89	0.68
Claude-3.5-Sonnet-CoT	1.06	1.91	3.25	7.09	0.22	0.19	0.12	0.07	0.73	0.71	0.95	0.70
Gemini-2.5-Pro	**1.05**	1.93	2.94	6.90	0.12	0.06	0.01	0.00	0.77	0.77	**0.98**	0.91
Gemini-2.5-Pro-CoT	1.09	2.01	3.02	7.05	0.25	0.18	0.01	0.00	0.80	0.79	0.97	**0.93**
GPT-5	1.34	2.09	2.60	6.25	0.09	0.12	**0.21**	0.12	0.75	0.73	0.96	0.85
GPT-5-CoT	1.24	2.05	3.17	6.91	**0.26**	**0.27**	0.19	0.11	**0.82**	**0.81**	0.97	0.88

3 EXPERIMENT

3.1 EXPERIMENTAL SETUP

Evaluated MLLMs. We evaluated a representative set of both open-source and proprietary multimodal LLMs. The open-source models include Qwen3VL-30B-Instruct and Qwen3VL-30B-Thinking (Qwen team, 2025) and InternVL3.5-38B (Wang et al., 2025c). The proprietary models include GPT-5, Gemini 2.5 Pro, and Claude 3.5 Sonnet.

Evaluation Protocol. We adopt task-specific metrics aligned with the three question formats in EarthSpatialBench. For quantitative questions (e.g., counting, distance and angle estimation), we report mean absolute error (MAE) and root mean squared error (RMSE) with respect to the ground truth. For choice-based questions (e.g., Yes/No or directional classification), we report overall accuracy and macro-averaged F1 score. For localization questions, which require identifying 2D bounding boxes satisfying spatial constraints, we evaluate intersection-over-union (IoU) and Macro F1 score based on the overlap between predicted and ground-truth boxes, using an IoU threshold of 0.1.

Computational Resources. All open-source models are evaluated on a compute node equipped with 8 NVIDIA B200 GPUs. Proprietary models are accessed through their respective APIs using default inference configurations.

3.2 MAIN RESULTS

Tables 3, 4, and 5 summarize the performance of multiple MLLMs with chain-of-thought (CoT) prompting on distance, direction, and topological reasoning tasks. Overall, we observe substantial performance gaps across reasoning types, indicating that current models exhibit uneven capabilities in numerical estimation, spatial grounding, and geometric reasoning.

For quantitative tasks, GPT-5 achieves the lowest error in distance estimation, while Qwen3VL-Thinking and its CoT variant demonstrate strong performance on angle regression, suggesting superior numerical reasoning ability. For localization tasks, Qwen3VL variants generally outperform other models, whereas proprietary models exhibit comparatively weaker performance on distance- and direction-related localization. For classification tasks, large proprietary models achieve the best performance, reflecting strong symbolic and relational reasoning capabilities. CoT prompting consistently improves classification and regression performance for most models, particularly on topological reasoning tasks, but often yields limited or negative gains for localization, suggesting that explicit reasoning traces do not directly translate to improved spatial grounding.

Finally, strong reasoning performance does not necessarily imply robust visual grounding. For example, Claude and Gemini attain competitive classification accuracy but perform poorly on localization metrics, revealing a clear decoupling between high-level spatial reasoning and low-level visual grounding. These results indicate that, despite recent advances, multimodal LLMs remain

limited in jointly reasoning over precise geometry, numerical relations, and visual evidence in Earth observation scenarios.

3.3 GROUNDING AS A PREREQUISITE FOR SPATIAL REASONING

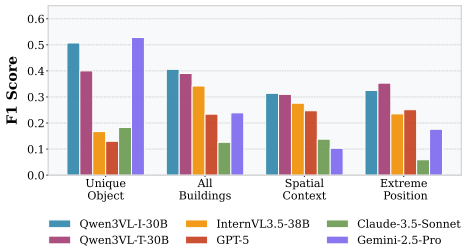
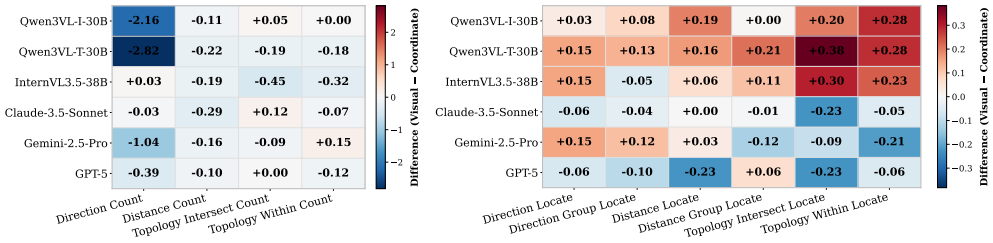


Figure 3: Performance of MLLMs on four grounding variants.

Most spatial reasoning tasks depend on accurate object localization, as geometric relations computed using GIS procedures require precise spatial inputs. We therefore evaluate grounding as a prerequisite for reliable spatial reasoning. Figure 3 reports F1 scores on four grounding variants that isolate object identification and localization difficulty. We observe substantial model-dependent variation, with Qwen3VL variants achieving the strongest and most consistent performance, while GPT-5 and Claude exhibit marked degradation on exhaustive grounding (*all buildings*). Gemini 2.5 Pro performs well on *unique object* grounding but degrades on context-dependent and extreme cases, indicating limited generalization to ambiguous scenarios. Across models, *extreme position* objects remain particularly challenging, highlighting rare-attribute grounding as a major bottleneck. Overall, these results demonstrate that strong spatial reasoning is tightly coupled with robust low-level grounding, and that joint grounding and reasoning in complex Earth imagery remains an open challenge.

3.4 COMPARISONS ON DIFFERENT GEOMETRIC REPRESENTATIONS



(a) (b)

Figure 4: Performance differences between visual-overlay-based and coordinate-based object references. (a) Change in MAE for counting. (b) Change in F1 score for localization.

Figure 4 compares model performance when spatial references are provided as visual overlays versus explicit coordinates. Overall, visual overlays improve localization performance, while their impact on counting and regression tasks is more heterogeneous. Across models, we observe a clear divergence between grounding-oriented and reasoning-oriented behaviors. Qwen3VL models consistently leverage visual overlays to improve localization and counting, demonstrating strong vision–language alignment. In contrast, GPT-5 achieves competitive performance under coordinate-based descriptions but fails to effectively exploit visual overlays, revealing a preference for text-centric reasoning. These results highlight substantial differences in how current MLLMs internalize spatial information and suggest that future systems must better unify visual grounding and geometric reasoning for robust Earth observation analysis.

Figure 5 further examines the impact of object representation on direction estimation by comparing coordinate-based and language-based references. When explicit coordinates are provided for both

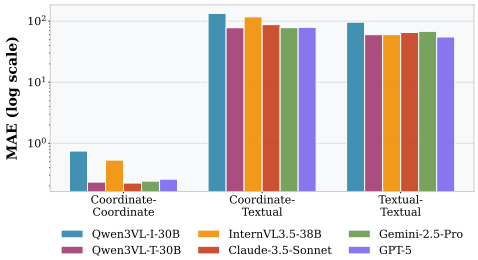


Figure 5: MAE of angle estimation under different object representation settings.

objects (*coordinate–coordinate*), all models achieve substantially lower MAE, indicating that they can reliably perform angle computation through direct numerical reasoning. In contrast, performance degrades markedly when objects are specified using unique textual descriptions, highlighting the difficulty of grounding linguistic references to precise spatial locations. Notably, the *coordinate–textual* setting yields even higher errors than *textual–textual*, suggesting that mixing symbolic coordinates with ambiguous language descriptions may introduce additional grounding inconsistencies. These results emphasize that accurate spatial reasoning depends critically on reliable object grounding, and that partial access to geometric information is insufficient when linguistic grounding remains uncertain.

3.5 Comparisons on Different Geometric Types

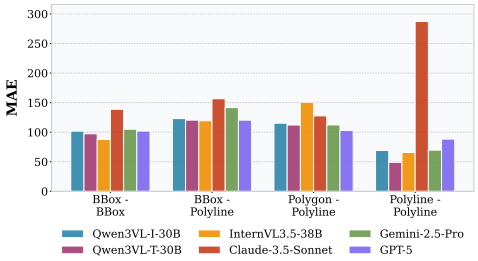


Figure 6: MAE of distance estimation across different geometry type combinations.

Figure 6 reports the MAE of distance estimation across different geometry type combinations. We observe substantial performance variations induced purely by geometric representation, with heterogeneous pairs (BBox–Polyline and Polygon–Polyline) consistently yielding higher errors than homogeneous cases. Notably, polyline–polyline queries achieve lower errors than box–box queries for most models, suggesting that linear structures such as roads and coastlines provide clearer geometric cues than coarse bounding boxes. In contrast, polygon–polyline settings remain among the most challenging, reflecting difficulties in reasoning over irregular region–curve interactions. From a model perspective, Qwen3VL variants exhibit relatively stable performance across geometry types, whereas GPT-5 and Claude show pronounced degradation on polyline-based queries. These results highlight geometric representation as a critical yet underexplored factor and motivate the incorporation of explicit geometric reasoning mechanisms in future models.

4 Conclusion and Future Work

We introduced EarthSpatialBench, a comprehensive benchmark for evaluating spatial reasoning in multimodal large language models on Earth imagery. By formalizing distance, direction, and topology relations and incorporating diverse geometric entities and question formats, it enables systematic assessment of spatial intelligence in real-world geospatial settings. Our experiments reveal persistent limitations of state-of-the-art MLLMs. Future work will extend the benchmark to multitemporal and multimodal inputs (e.g., DEM, SAR, and LiDAR) to further advance geometry-grounded multimodal reasoning in geospatial AI.

REFERENCES

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16772–16782, 2023.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498. IEEE, 2025.

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 (CVPR)*, pp. 14455–14465, June 2024a.

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*, pp. 24185–24198, 2024b.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Xingguang Wei, Tianshuo Yang, Min Dou, Tong He, Wenqi Shao, Kaipeng Zhang, Yi Wang, Botian Shi, Yanting Zhang, Jifeng Dai, Yu Qiao, Hongjie Zhang, and Wenhai Wang. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.XXXXX*, 2025.

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*, pp. 27831–27840, 2024.

Junnan Li, Zhenfei Fang, Peiyuan Sun, Xinyu Chen, Yifei Zhang, Peng Hu, Hanlin Qiu, Xinlong Wang, Xiang Li, Jianping Shi, and Jifeng Dai. Cambrian-1: A fully open, vision-centric exploration of multimodal large language models. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024a.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934, October 2025.

Qwen team. Qwen3-vl: Sharper vision, deeper thought, broader action, 2025.

Shashi Shekhar and Hui Xiong. *Encyclopedia of GIS*. Springer Science & Business Media, 2007.

Shashi Shekhar, Steven K Feiner, and Walid G Aref. Spatial computing. *Communications of the ACM*, 59(1):72–81, 2015.

Peijun Shi, Tao Ye, Ying Wang, Tao Zhou, Wei Xu, Juan Du, Jing'ai Wang, Ning Li, Chongfu Huang, Lianyou Liu, et al. Disaster risk science: A geographical perspective and a research framework. *International Journal of Disaster Risk Science*, 11(4):426–440, 2020.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21440–21455, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1195.

Robert A Stallings. *Methods of disaster research*. Xlibris Corporation, 2003.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, 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? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14325–14336, 2025a.

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*, volume 38, pp. 5481–5489, 2024.

Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, et al. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. *arXiv preprint arXiv:2505.21089*, 2025b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025.

Weichen Zhang, Zile Zhou, Xin Zeng, Xuchen Liu, Jianjie Fang, Chen Gao, Yong Li, Jinqiang Cui, Xinlei Chen, and Xiao-Ping Zhang. Open3d-vqa: A benchmark for comprehensive spatial reasoning with multimodal large language models in open space. *arXiv preprint arXiv:2503.11094*, 2025.