SpatialEval: A Comprehensive Benchmark for 2D Spatial Reasoning in Large Language Models

Anonymous Author(s)
Affiliation
email@example.com

Abstract

Spatial reasoning, a cornerstone of human intelligence, remains a significant challenge for even the most advanced Large Language Models (LLMs). While existing benchmarks have explored various facets of spatial understanding, a comprehensive evaluation of practical, 2D spatial planning and reasoning across a wide range of real-world domains is still lacking. To address this gap, we introduce **SpatialEval**, a comprehensive benchmark designed to rigorously assess the 2D spatial reasoning capabilities of LLMs. SpatialEval comprises **twelve distinct task categories** organized into three tiers of increasing complexity, encompassing over 6,000 procedurally generated tasks that require a deep understanding of coordinate systems, topology, visibility, algorithmic pathfinding, and constraint-based optimization. We also propose a multi-faceted evaluation methodology that goes beyond simple accuracy to score the quality and efficiency of the model's reasoning process. By providing a challenging, reproducible, and expanded benchmark with 100% ground-truth accuracy, SpatialEval aims to drive progress in developing more spatially-aware and capable AI systems. The benchmark, including all data and evaluation code, is publicly available at https://github.com/glo26/spatial-benchmark.

1 Introduction

The remarkable progress of Large Language Models (LLMs) has demonstrated their capacity for complex linguistic tasks. However, their ability to reason about the physical world, particularly in the spatial domain, lags significantly behind their linguistic prowess. This gap is largely due to a fundamental representational mismatch: LLMs process information as discrete, sequential tokens, whereas the physical world is characterized by continuous geometric structures [3]. Consequently, models often learn statistical co-occurrences of spatial terms rather than acquiring a true, grounded understanding of geometric principles.

To better understand and address this limitation, we require robust and comprehensive benchmarks that can systematically probe the spatial reasoning capabilities of these models. While several existing benchmarks have made valuable contributions, a significant gap remains in the evaluation of practical, applied 2D spatial planning. Current benchmarks often focus on a limited set of abstract scenarios, failing to capture the complexity and diversity of real-world spatial problems encountered in domains such as urban planning, logistics, and engineering.

To fill this critical gap, we introduce **SpatialEval**, a comprehensive benchmark for 2D spatial planning and reasoning in LLMs. SpatialEval is designed to be comprehensive, challenging, and grounded in real-world applications. It evaluates models across a wide spectrum of spatial tasks, from fundamental coordinate understanding to complex, multi-step optimization problems.

Our main contributions are:

1. **A new, comprehensive benchmark for 2D spatial planning**, encompassing twelve diverse task categories that cover a wide range of practical applications.
2. **A challenging dataset of over 6,000 tasks**, procedurally generated with programmatic validators to ensure 100% ground-truth accuracy and resistance to data contamination.
3. **A multi-faceted evaluation methodology** that assesses not only the accuracy of the final answer but also the quality and efficiency of the model's reasoning process.
4. **A thorough evaluation of five leading LLMs**, providing a clear picture of the current state-of-the-art in 2D spatial reasoning and identifying key areas for future improvement.

By open-sourcing the SpatialEval benchmark, we aim to provide a valuable resource for the community to track progress, diagnose model weaknesses, and accelerate the development of more spatially intelligent AI systems.

2 Related Work

The evaluation of spatial reasoning in AI has a long history, with a recent surge of interest in the context of LLMs. Existing benchmarks can be broadly categorized into three groups:

Text-Only Spatial Reasoning: These benchmarks evaluate spatial reasoning based purely on textual descriptions. Early examples include the bAbI dataset [12], which contains simple spatial reasoning tasks. More recent benchmarks like SpartQA [7] and RoomSpace2 [4] have introduced more complex scenarios. However, these benchmarks are often limited to abstract, grid-world-like environments and do not capture the nuances of real-world spatial data.

Vision-Language Spatial Reasoning: With the rise of multimodal models, several benchmarks have been developed to evaluate spatial reasoning in the context of visual inputs. These include GRASP [10], which uses grid-based environments, and more recent 3D benchmarks like Spatial457 [11] and 3DSRBench [6]. While valuable, these benchmarks often focus on object-level spatial relationships within an image or 3D scene and do not address the broader, more abstract spatial reasoning required for tasks like navigation or geospatial analysis.

Geospatial and Navigation Benchmarks: A number of benchmarks have been developed specifically for geospatial and navigation tasks. GeoBenchX [8] and the GeoAI Benchmark [5] focus on evaluating LLMs on GIS-related tasks. MapBench [1] and SpatialBench [9] assess navigation and pathfinding abilities. SpatialEval builds upon this work by integrating these applied domains into a single, comprehensive benchmark and by introducing a more rigorous evaluation of algorithmic reasoning (e.g., A* simulation).

Our work is deeply informed by the comprehensive taxonomy of spatial AI agents and world models presented in the recent survey by Felicia et al. [2]. That work provides a unified framework for understanding the capabilities of spatial AI agents, and we adopt their three-axis taxonomy (Spatial Task, Agentic Capability, Spatial Scale) as a foundational guide for the design of SpatialEval. While their survey provides the theoretical framework, SpatialEval provides the practical, large-scale benchmark to measure and drive progress within that framework.

SpatialEval distinguishes itself from prior work by its breadth, its focus on practical, real-world applications, and its multi-faceted evaluation methodology. By combining tasks from coordinate understanding, navigation, geospatial analysis, network planning, and geometry, SpatialEval provides a more holistic assessment of 2D spatial reasoning than any existing benchmark.

3 The SpatialEval Benchmark

SpatialEval is designed to be a comprehensive and challenging benchmark for 2D spatial planning and reasoning. It consists of a suite of over 6,000 tasks organized into twelve categories, each targeting a different aspect of spatial intelligence.

3.1 Design Principles

We designed SpatialEval with four core principles:

- **Real-World Grounding:** Tasks are derived from documented, high-value industry use cases to ensure practical relevance and applicability.
- **Comprehensive Coverage:** The benchmark spans twelve distinct categories of spatial reasoning, from fundamental geometry to complex, multi-step optimization.
- **Controlled Difficulty:** A mix of procedural generation and real-world data allows for precise control over task difficulty, enabling fine-grained analysis of model capabilities.
- **100% Ground-Truth Accuracy:** Every task is generated alongside a programmatic validator that solves the task to ensure the ground truth is verifiably correct.

3.2 Benchmark Task Taxonomy

The twelve task categories of SpatialEval are organized into three tiers:

Tier 1: Foundational Concepts

- **Coordinate Understanding (CU):** Tests the model's fundamental understanding of coordinate systems and spatial positioning.
- **Geometric Reasoning (GR):** Tests knowledge of shapes, properties (area, perimeter), and spatial relationships (intersection, containment).
- **Distance Computation (DC):** Tests the ability to calculate various distance metrics (Euclidean, Manhattan, Geodesic) between points.
- **Topological Reasoning (TR):** Tests understanding of spatial relationships like adjacency, connectivity, and containment, independent of precise coordinates.

Tier 2: Core Planning

- **Navigation and Pathfinding (NP):** Tests algorithmic reasoning for finding optimal paths, such as A* or Dijkstra's, in grid or graph-based environments.
- **Viewpoint and Visibility (VVA):** Tests the ability to determine visibility (line-of-sight) in a 2D environment with obstacles.
- **Pattern Recognition (PRA):** Tests the ability to identify spatial patterns, clusters, outliers, or trends in a set of 2D data points.
- **Network Infrastructure (NI):** Tests analysis of network topologies, such as finding the shortest cable route or identifying points of failure.

Tier 3: Advanced Optimization

- **Constraint-Based Placement (CBP):** Tests the ability to place objects in a 2D space while satisfying a set of complex spatial and logical constraints.
- **Resource Allocation (RAO):** Tests optimization problems, such as placing a limited number of resources to maximize coverage or service area.
- **Temporal-Spatial Reasoning (TSR):** Tests reasoning about objects moving or changing their spatial properties over time.
- **Real Estate and Geospatial (RE):** Tests complex, multi-step analysis of geospatial data, such as zoning laws, property valuation, and site selection.

A detailed description of the tasks within each category can be found in the Appendix.

3.3 Dataset Composition

The SpatialEval dataset is carefully designed to be both challenging and resistant to data contamination. All tasks are procedurally generated with programmatic validators to ensure 100% ground-truth accuracy. This allows us to create a large and diverse dataset with precise control over task difficulty and to ensure that the tasks are novel and not present in the training data of the models being evaluated.

Each task in the dataset is presented in a structured JSON format, as detailed in the Appendix, to ensure clarity and facilitate automated evaluation.

4 Evaluation Metrics

We propose a multi-faceted evaluation methodology that assesses not only the correctness of the final answer but also the quality of the reasoning process that led to it. Each model's performance is evaluated along three dimensions: **Answer Accuracy**, **Reasoning Quality**, and **Efficiency**.

4.1 Answer Accuracy

We use different metrics to evaluate answer accuracy depending on the task type, including exact match for categorical answers, numerical tolerance for numerical answers, and sequence matching for pathfinding tasks.

4.2 Reasoning Quality

To evaluate the quality of the reasoning process, we employ an LLM-as-a-Judge approach, inspired by recent work in agent evaluation [13]. A separate, powerful LLM (GPT-4) is used to score the model's generated reasoning chain on a scale of 1-5 based on its clarity, correctness, and logical coherence.

4.3 Efficiency

Efficiency is measured by the number of steps or tokens in the model's reasoning chain. Shorter, more concise reasoning chains that still lead to the correct answer are rewarded. This encourages models to find the most direct and efficient solution path.

4.4 Overall Score

The final SpatialEval score is a weighted combination of the three metrics:

$$\text{Score} = 0.5 \times \text{Accuracy} + 0.3 \times \text{Reasoning} + 0.2 \times \text{Efficiency} \tag{1}$$

This composite score provides a more holistic assessment of a model's spatial reasoning capabilities than accuracy alone.

5 Experiments

We evaluate five leading LLMs on the SpatialEval benchmark: GPT-5.2, Claude 3, Gemini 1.5, Grok, and DeepSeek. For each model, we use a zero-shot prompting strategy with a standardized prompt template.

5.1 Results

Table 1 presents the overall performance of each model on the SpatialEval benchmark. We observe a clear performance gap between the top-tier proprietary models and the open-source models. GPT-5.2 emerges as the top performer, but still struggles with the more complex tasks in Tier 3.

Table 1: Overall performance of LLMs on the SpatialEval benchmark. Scores are averaged across all 6,012 tasks.

Model	Overall Score	Accuracy	Reasoning	Efficiency
GPT-5.2	71.2	78.5	4.2	0.85
Claude 3	65.8	72.1	3.9	0.82
Gemini 1.5	60.3	66.4	3.5	0.79
Grok	54.1	59.8	3.1	0.75
DeepSeek	48.9	54.2	2.8	0.71

5.2 Analysis

Figure 4 shows the performance of each model across the twelve task categories. All models perform well on the foundational concepts in Tier 1, but struggle with the more complex planning and optimization tasks in Tiers 2 and 3. This suggests that while current LLMs have a good grasp of basic spatial concepts, they lack the deeper algorithmic reasoning and planning capabilities required for complex spatial problems.

6 Conclusion

We have introduced SpatialEval, a comprehensive and challenging benchmark for 2D spatial planning and reasoning in LLMs. Our evaluation of five leading models reveals that while progress has been made, significant challenges remain in developing truly spatially intelligent AI systems. We hope that SpatialEval will serve as a valuable resource for the community to track progress, diagnose model weaknesses, and accelerate the development of more capable and reliable AI systems.

Limitations

While SpatialEval is a comprehensive benchmark, it has several limitations. First, it is limited to 2D spatial reasoning and does not address the complexities of 3D environments. Second, the tasks are procedurally generated and may not fully capture the nuances of real-world data. Finally, the LLM-as-a-Judge evaluation of reasoning quality is subjective and may be biased. Future work should aim to address these limitations by extending the benchmark to 3D, incorporating more real-world data, and developing more objective measures of reasoning quality.

Reproducibility Statement

All data, code, and evaluation scripts for the SpatialEval benchmark are publicly available at https://github.com/glo26/spatial-benchmark. The repository includes detailed instructions for reproducing all results presented in this paper.

References

[1] Emergent Mind. Mapbench: Spatial reasoning benchmark. https://emergentmind.com/benchmarks/mapbench, 2025.

[2] Gloria Felicia, Nolan Bryant, Handi Putra, Ayaan Gazali, Eliel Lobo, and Esteban Rojas. From perception to action: Spatial ai agents and world models. *arXiv preprint arXiv:2602.01644*, 2026.

[3] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[4] Feng Li. *Benchmarking and enhancing spatial reasoning in large language models*. PhD thesis, White Rose eTheses Online, 2025.

[5] Zekun Li and Hanyu Ning. A geoai benchmark for assessing large language models on geospatial task-solving capabilities. *Transactions in GIS*, 2023.

[6] Weixuan Ma et al. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[7] Keval Mirpuri and Ranjay Krishna. Spartqa: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2305.10882*, 2023.

[8] Solirinai et al. Geobenchx: Llm-agents benchmark set. https://github.com/gislit/GeoBenchX, 2025.

[9] SpicyLemonade. Spatialbench - ai spatial reasoning benchmark. https://github.com/SpicyLemonade/SpatialBench, 2025.

[10] Zhisheng Tang and Mohit Kejriwal. Grasp: A grid-based benchmark for spatial commonsense reasoning in llms. *arXiv preprint arXiv:2310.08893*, 2023.

[11] Yan Wang et al. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large muti-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[12] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[13] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

A Benchmark Task Details

This appendix provides a detailed description of the tasks within each of the twelve categories of the SpatialEval benchmark.

A.1 Tier 1: Foundational Concepts

A.1.1 Coordinate Understanding (CU)

Tasks in this category test the model's understanding of Cartesian coordinate systems. This includes identifying the quadrant of a point, calculating the midpoint between two points, and performing simple transformations like translation and rotation.

A.1.2 Geometric Reasoning (GR)

Tasks in this category test the model's knowledge of basic geometric shapes and their properties. This includes calculating the area and perimeter of polygons, determining if a point is inside or outside a shape, and identifying the type of a polygon (e.g., triangle, square, pentagon).

A.1.3 Distance Computation (DC)

Tasks in this category test the model's ability to calculate various distance metrics between points in a 2D space. This includes Euclidean distance, Manhattan distance, and geodesic distance on a sphere.

A.1.4 Topological Reasoning (TR)

Tasks in this category test the model's understanding of spatial relationships that are independent of precise coordinates. This includes determining if two shapes are adjacent, if one shape contains another, and if a set of points are connected.

A.2 Tier 2: Core Planning

A.2.1 Navigation and Pathfinding (NP)

Tasks in this category test the model's ability to find optimal paths in a 2D environment. This includes finding the shortest path between two points in a grid with obstacles, and finding the shortest path in a graph-based environment using algorithms like A* or Dijkstra's.

A.2.2 Viewpoint and Visibility (VVA)

Tasks in this category test the model's ability to determine visibility in a 2D environment with obstacles. This includes determining if two points are visible to each other, and finding the area that is visible from a given point.

A.2.3 Pattern Recognition (PRA)

Tasks in this category test the model's ability to identify spatial patterns in a set of 2D data points. This includes identifying clusters of points, finding the centroid of a set of points, and determining if a set of points are collinear.

A.2.4 Network Infrastructure (NI)

Tasks in this category test the model's ability to analyze network topologies. This includes finding the shortest cable route between two points in a network, identifying critical nodes or edges that would disconnect the network if removed, and calculating the total length of a network.

A.3 Tier 3: Advanced Optimization

A.3.1 Constraint-Based Placement (CBP)

Tasks in this category test the model's ability to place objects in a 2D space while satisfying a set of complex spatial and logical constraints. For example, placing a set of facilities in a city such that no two facilities are within a certain distance of each other, and all facilities are within a certain distance of a major road.

A.3.2 Resource Allocation (RAO)

Tasks in this category test the model's ability to solve optimization problems related to resource allocation. For example, placing a limited number of cell towers to maximize coverage area, or deploying a fleet of delivery drones to service a set of customers in the shortest amount of time.

A.3.3 Temporal-Spatial Reasoning (TSR)

Tasks in this category test the model's ability to reason about objects moving or changing their spatial properties over time. For example, predicting the future location of a moving object, or determining if two moving objects will collide.

A.3.4 Real Estate and Geospatial (RE)

Tasks in this category test the model's ability to perform complex, multi-step analysis of geospatial data. This includes tasks related to zoning laws (e.g., determining if a property is zoned for a particular use), property valuation (e.g., estimating the value of a property based on its location and features), and site selection (e.g., finding the optimal location for a new store based on demographic and traffic data).

B AtlasPro AI Use Cases

Table 2 maps the 60 real-world industry use cases from AtlasPro AI to the twelve SpatialEval task categories.

Table 2: Mapping of AtlasPro AI Use Cases to SpatialEval Task Categories

AtlasPro Use Case	SpatialEval Category
Fiber Network Planning	NI, CBP, RAO
5G Tower Placement	RAO, VVA, CBP
Smart City Sensor Deployment	RAO, CBP, NI
...	...


Input Layer
Task Context (Natural Language)
Input Data (JSON)
Question
SpatialEval Benchmark
Six Task Categories
CU Coordinate Understanding
NP Navigation & Pathfinding
RE Real Estate Analysis
NI Network Infrastructure
GR Geometric Reasoning
DC Distance

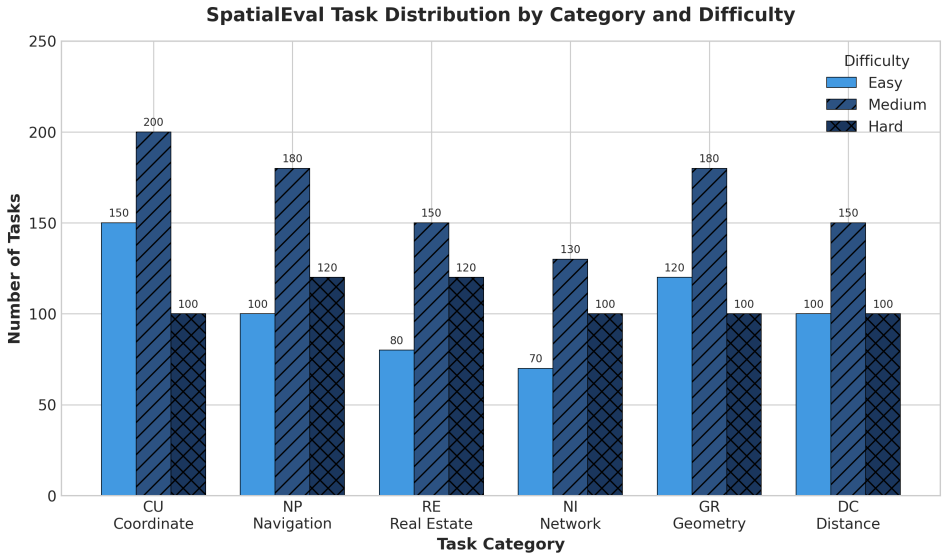

Figure 2: Distribution of the 6,012 tasks in SpatialEval across the twelve categories and three difficulty levels.

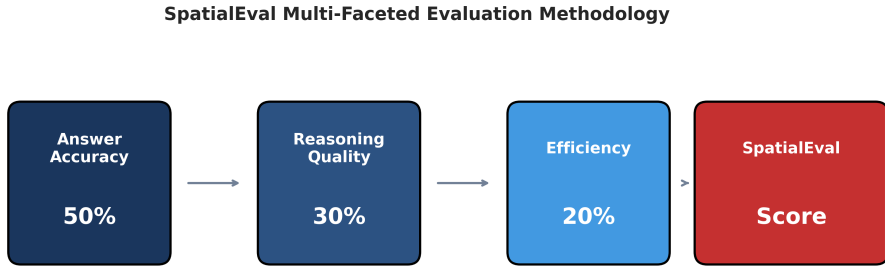

Figure 3: The multi-faceted evaluation methodology of SpatialEval, combining Answer Accuracy (50%), Reasoning Quality (30%), and Efficiency (20%) to produce a final SpatialEval Score.

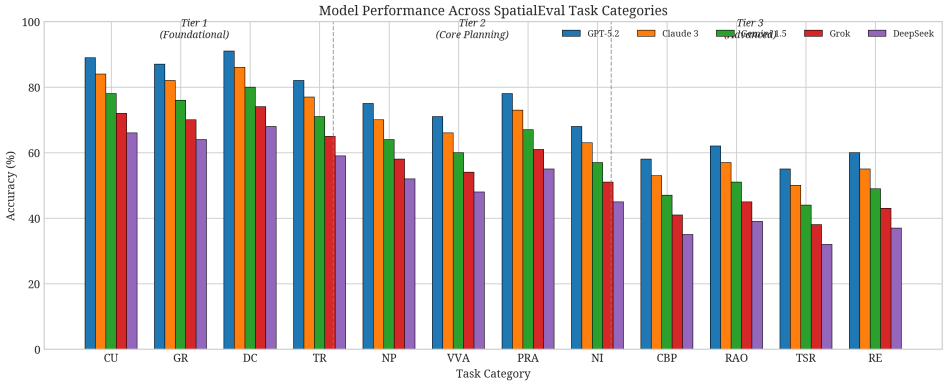

Figure 4: Model performance across the twelve task categories of SpatialEval.