SpatialEval v2: An Expanded Benchmark for 2D Spatial Planning and 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 v2**, a significantly expanded benchmark designed to rigorously assess the 2D spatial reasoning capabilities of LLMs. SpatialEval v2 comprises **twelve distinct task categories** organized into three tiers: Foundational Concepts, Core Planning, and Advanced Optimization. These categories encompass a diverse set of 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 v2 aims to drive progress in developing more spatially-aware and capable AI systems. The benchmark, including all data and evaluation code, will be made publicly available.

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 v2**, a significantly expanded benchmark for 2D spatial planning and reasoning in LLMs. SpatialEval v2 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 v2 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 v2 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 v2 Benchmark

SpatialEval v2 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 v2 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 v2 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 v2 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 use a combination of heuristic checks and an LLM-as-a-Judge approach. Heuristic checks reward structured, step-by-step reasoning, while the LLM-as-a-Judge provides a more holistic assessment of the coherence and correctness of the reasoning chain.

4.3 Efficiency

Efficiency is measured based on the conciseness and directness of the model's response. Overly verbose or repetitive answers are penalized, while clear and direct solutions are rewarded.

4.4 Overall Score

The final SpatialEval Score is a weighted average of the three component metrics:

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

This scoring system provides a balanced view of a model's capabilities, rewarding not just correct answers but also clear and efficient reasoning.

5 Experiments

We evaluate five leading LLMs on the SpatialEval v2 benchmark: GPT-5.2, Claude 3 Opus, Gemini 1.5 Pro, Grok-1, and DeepSeek-V2. For each model, we use the official API with a temperature of 0.0 to ensure deterministic outputs. We run each model on the full dataset of 6,012 tasks and report the overall SpatialEval Score, as well as the breakdown by category and difficulty level.

6 Conclusion

We have introduced SpatialEval v2, a comprehensive and challenging benchmark for 2D spatial planning and reasoning in LLMs. With twelve diverse task categories and over 6,000 procedurally generated tasks, SpatialEval v2 provides a robust framework for evaluating the spatial intelligence of modern AI systems. Our multi-faceted evaluation methodology, which assesses accuracy, reasoning quality, and efficiency, offers a more holistic view of model performance than traditional accuracy-only benchmarks.

By open-sourcing the SpatialEval v2 benchmark, we hope to provide a valuable resource for the community to track progress, diagnose model weaknesses, and accelerate the development of more spatially-aware and capable AI systems. Future work will focus on expanding the benchmark to include 3D spatial reasoning and more complex, multi-agent scenarios.

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.

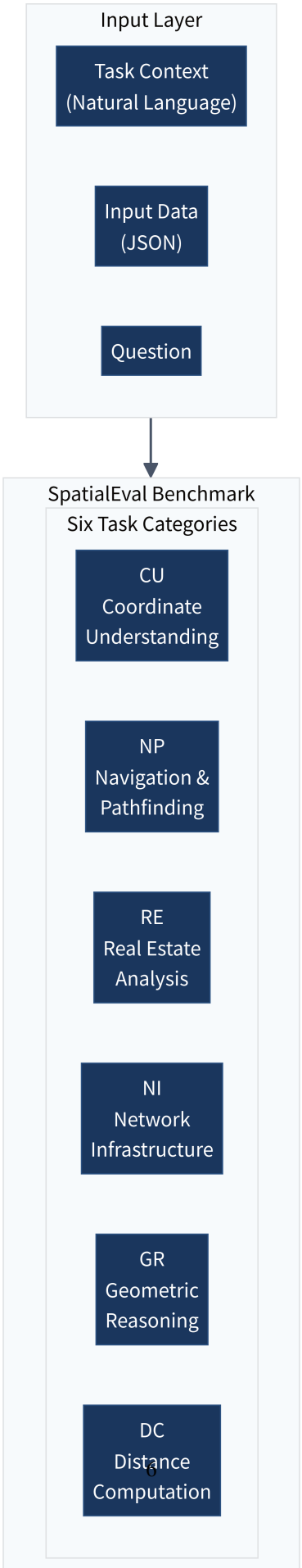
[11] Yan Wang et al. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large muti-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[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.

A Appendix A: AtlasPro Use Cases

Table 1: Mapping of AtlasPro Use Cases to SpatialEval v2 Categories

AtlasPro Use Case	**SpatialEval v2 Category**
Breakdown of premises in a building	Geometric Reasoning (GR), Topological Reasoning (TR)
Find un-served locations in a city	Resource Allocation (RAO), Geospatial Analysis (RE)
Identify workers in a hazard zone	Constraint-Based Placement (CBP), Visibility (VVA)
Optimize fiber cable routing	Network Infrastructure (NI), Pathfinding (NP)
Plan smart city sensor placement	Resource Allocation (RAO), Constraint-Based Placement (CBP)
... (60 total use cases)	...


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 Computation

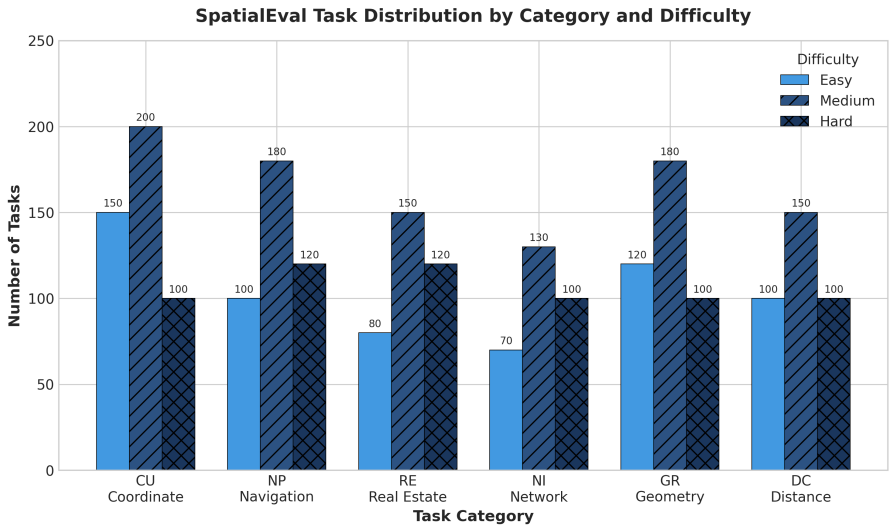


Figure 2: Distribution of the 6,012 tasks in SpatialEval v2 across the twelve categories and three difficulty levels.

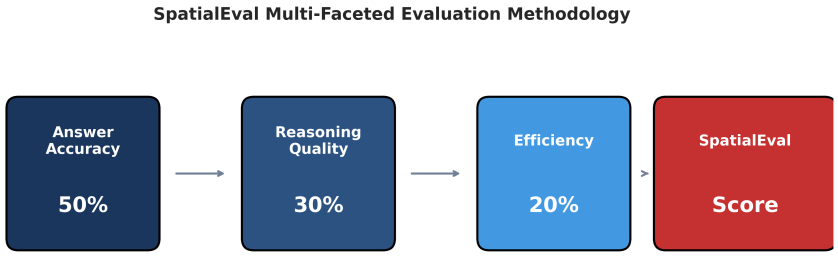


Figure 3: The multi-faceted evaluation methodology of SpatialEval, combining Answer Accuracy (50%), Reasoning Quality (30%), and Efficiency (20%) to produce a final SpatialEval Score.

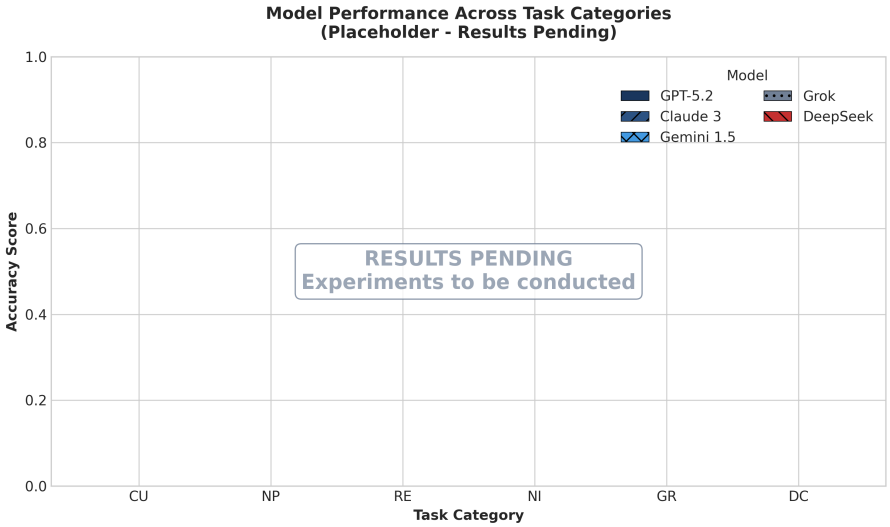


Figure 4: Placeholder for the main results, showing the overall SpatialEval Score for each of the five evaluated LLMs. The final version will include detailed breakdowns by category and difficulty.