SpatialOps: A 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 **SpatialOps**, a comprehensive benchmark designed to rigorously assess the 2D spatial planning and reasoning capabilities of LLMs. SpatialOps 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, SpatialOps 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 [6, 45, 3]. However, their ability to reason about the physical world, particularly in the spatial domain, lags significantly behind their linguistic prowess [5, 52]. 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 [15, 40]. Consequently, models often learn statistical co-occurrences of spatial terms rather than acquiring a true, grounded understanding of geometric principles [31, 33].

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 [8, 50], a significant gap remains in the evaluation of practical, applied 2D spatial planning. Current benchmarks often focus on a limited set of abstract scenarios [48, 23], failing to capture the complexity and diversity of real-world spatial problems encountered in domains such as urban planning [4], logistics [37], and engineering [10].

To fill this critical gap, we introduce **SpatialOps**, a comprehensive benchmark for 2D spatial planning and reasoning in LLMs. SpatialOps 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, from network infrastructure planning [19] to real estate analysis [9].
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, a critical issue in modern benchmarking [6, 7].
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, inspired by recent work in agent evaluation [53, 30].
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 SpatialOps 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 [27, 35, 46].

2 Related Work

The evaluation of spatial reasoning in AI has a long history [21, 13, 36], with a recent surge of interest in the context of LLMs and agentic systems [49, 38, 39, 17]. 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 [48], which contains simple spatial reasoning tasks. More recent benchmarks like SpartQA [32], RoomSpace2 [26], and PlanQA [20] 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, a limitation we directly address.

Vision-Language Spatial Reasoning: With the rise of multimodal models [28, 34, 2, 24], several benchmarks have been developed to evaluate spatial reasoning in the context of visual inputs. These include GRASP [43], which uses grid-based environments, and more recent 3D benchmarks like Spatial457 [47], 3DSRBench [29], and SPARE3D [22]. 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 [41] and the GeoAI Benchmark [25] focus on evaluating LLMs on GIS-related tasks. MapBench [11] and the original SpatialBench [42] assess navigation and pathfinding abilities. SpatialOps 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 [16]).

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. [12]. 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 SpatialOps. While their survey provides the theoretical framework, SpatialOps provides the practical, large-scale benchmark to measure and drive progress within that framework.

SpatialOps 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, SpatialOps provides a more holistic assessment of 2D spatial reasoning than any existing benchmark.

3 The SpatialOps Benchmark

SpatialOps 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 SpatialOps 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 SpatialOps 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 SpatialOps 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

To provide a holistic assessment of model performance, we introduce a suite of five key metrics that go beyond simple accuracy:

4.1 Task Completion Rate (TCR)

This metric measures the percentage of tasks successfully completed by the AI agent. A task is considered complete if the model produces a valid, parsable output that meets the minimum requirements of the task.

$$TCR = \frac{\text{Tasks Completed}}{\text{Total Tasks}} \times 100\% \tag{1}$$

4.2 Human-AI Latency Ratio (HLR)

This metric quantifies the speed-up achieved by using an AI agent compared to a human professional. We establish a baseline by measuring the average time taken by a human GIS analyst to complete a representative sample of tasks from each category.

$$HLR = \frac{\text{Time}_{\text{human}}}{\text{Time}_{\text{AI}}} \tag{2}$$

4.3 Operational Cost Savings (OCS)

Building on the HLR, this metric estimates the potential dollar savings from deploying AI agents in a business context. It accounts for the time saved, the hourly rate of a human analyst, and the computational cost of the AI model.

$$OCS = (\text{Time}_{\text{human}} - \text{Time}_{\text{AI}}) \times \text{Hourly Rate}_{\text{human}} - \text{Cost}_{\text{AI}} \tag{3}$$

4.4 Efficacy Score (ES)

This composite score provides a single, comprehensive measure of model performance, combining accuracy, reasoning quality, and efficiency. The weights for each component can be adjusted to reflect the priorities of a specific application.

$$ES = w_1 \times \text{Accuracy} + w_2 \times \text{Reasoning Quality} + w_3 \times \text{Efficiency} \tag{4}$$

4.5 Scalability Index (SI)

This metric assesses the model's ability to handle increasing task complexity. It is calculated as the ratio of performance on high-complexity tasks to low-complexity tasks, normalized by the time taken.

$$SI = \frac{\text{Tasks Completed}_{\text{high complexity}}}{\text{Tasks Completed}_{\text{low complexity}}} \times \frac{\text{Time}_{\text{low complexity}}}{\text{Time}_{\text{high complexity}}} \tag{5}$$

5 Industrial Context and Differentiation

Spatial AI is rapidly becoming a critical component of modern industrial and governmental operations. Companies like Palantir, with its Foundry platform, are deploying AI for battlefield awareness and contested logistics in defense contexts [44]. Scale AI's Donovan platform provides purpose-built LLMs for the intelligence community, enabling satellite imagery analysis and other GEOINT tasks [1]. In the enterprise space, Wherobots is leveraging Apache Sedona to build a cloud-native spatial intelligence platform for large-scale analytics [18], while Google's Earth Engine and Maps Platform offer planetary-scale geospatial data and AI-powered tools [14].

Despite this rapid progress, a significant gap exists between the capabilities of these specialized systems and the general-purpose reasoning abilities of LLMs. Existing benchmarks, such as SpatialBench [42] and GeoAnalystBench [51], have made valuable contributions but do not fully capture the operational planning and optimization tasks that are critical in these industrial settings. SpatialOps is designed to fill this gap by providing a benchmark that is grounded in the real-world use cases of companies like AtlasPro AI, which are focused on developing AI agents for complex, multi-step spatial planning in critical industries.

6 Ablation Studies

To better understand the factors influencing model performance, we conduct a series of ablation studies. These studies systematically remove or modify components of the task or model to isolate their impact on performance. For example, we evaluate the impact of providing detailed vs. minimal task instructions, the effect of different prompt engineering strategies, and the performance of models with and without access to external tools.

Table 1: Ablation Study: Impact of Prompt Detail on Performance

Model	Minimal Prompt	Detailed Prompt
GPT-5.2	68.2	78.4
Claude 3	63.5	73.8
Gemini 1.5	58.1	68.2
Grok	52.4	61.8
DeepSeek	46.3	56.0

References

[1] Scale AI. Donovan: Ai for intelligence. https://scale.com/donovan, 2025.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] Sidhika Balachandar, Shuvom Sadhuka, Bonnie Berger, Emma Pierson, and Nikhil Garg. Urban incident prediction with graph neural networks: Integrating government ratings and crowdsourced reports, 2025.

[5] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[7] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2021.

[8] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2017.

[9] Stefano De Sabbata and Pengyuan Liu. A graph neural network framework for spatial geodemographic classification. *International Journal of Geographical Information Science*, 37(12):2464–2486, 2023.

[10] Chuck Eastman, Paul Teicholz, Rafael Sacks, and Kathleen Liston. *BIM Handbook: A Guide to Building Information Modeling*. John Wiley & Sons, 2011.

[11] EmergentMind. Mapbench: Evaluating existing knowledge base construction of llms. *arXiv preprint*, 2025.

[12] Gloria Felicia et al. From perception to action: Spatial ai agents and world models. *arXiv preprint arXiv:2602.01644*, 2026.

[13] Kenneth D Forbus. Qualitative process theory. *Artificial Intelligence*, 24(1-3):85–168, 1984.

[14] Google. Google earth engine. https://earthengine.google.com, 2025.

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

[16] Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[17] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024.

[18] Wherobots Inc. Wherobots: Cloud-native spatial intelligence. https://wherobots.com, 2026.

[19] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[20] KAUST. Planqa: A diagnostic benchmark for spatial reasoning in llms. *arXiv preprint*, 2025.

[21] Benjamin Kuipers. Modeling spatial knowledge. *Cognitive Science*, 2(2):129–153, 1978.

[22] AI4CE Lab. Spare3d: A dataset for spatial reasoning on three-view line drawings. *GitHub Repository*, 2024.

[23] Wenhu Le et al. Logicnlg: A dataset for natural language generation from tabular data. *arXiv preprint*, 2022.

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[25] Wenwen Li et al. Geoai: A review of artificial intelligence approaches for the interpretation of complex geomatics data. *Geoscience Frontiers*, 2023.

[26] Xiang Li et al. Benchmarking spatial reasoning in large language models. *arXiv preprint*, 2025.

[27] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2023.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[29] Wei Ma et al. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint*, 2025.

[30] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

[31] Gary F Marcus. Rethinking eliminative connectionism. *Cognitive Psychology*, 37(3):243–282, 1998.

[32] Roshanak Mirpuri, Reza Mirzaee, and Parisa Kordjamshidi. Spartqa: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*, 2023.

[33] Melanie Mitchell. Can large language models reason? *arXiv preprint*, 2021.

[34] OpenAI. Gpt-4v(ision) system card. *OpenAI Technical Report*, 2023.

[35] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[36] David A Randell, Zhan Cui, and Anthony G Cohn. A spatial logic based on regions and connection. *KR*, 92:165–176, 1992.

[37] Martin Savelsbergh and Tom Van Woensel. City logistics: Challenges and opportunities. *Transportation Science*, 2005.

[38] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[39] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

[40] Aaron Sloman. Interactions between philosophy and artificial intelligence. *Artificial Intelligence*, 2(3-4):209–225, 1971.

[41] Ekaterina Solirinai et al. Geobenchx: Benchmarking and analyzing monocular geospatial understanding. *arXiv preprint*, 2025.

[42] SpicyLemonade. Spatialbench: Open source benchmarks for multimodal ai spatial reasoning. *GitHub Repository*, 2025.

[43] Yiming Tang et al. Grasp: A grid-based benchmark for evaluating spatial reasoning. *arXiv preprint*, 2023.

[44] Palantir Technologies. Project maven: Ai for defense. https://www.palantir.com, 2024.

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[46] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 2024.

[47] Peng Wang et al. Spatial457: A diagnostic benchmark for 6d spatial reasoning. *arXiv preprint arXiv:2502.08636*, 2025.

[48] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[49] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023.

[50] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[51] Wei Zhang et al. Geoanalystbench: A benchmark for gis workflow generation. *Transactions in GIS*, 2025.

[52] Yue Zhang et al. Wrestling with spatial reasoning in large language models. *arXiv preprint*, 2024.

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


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

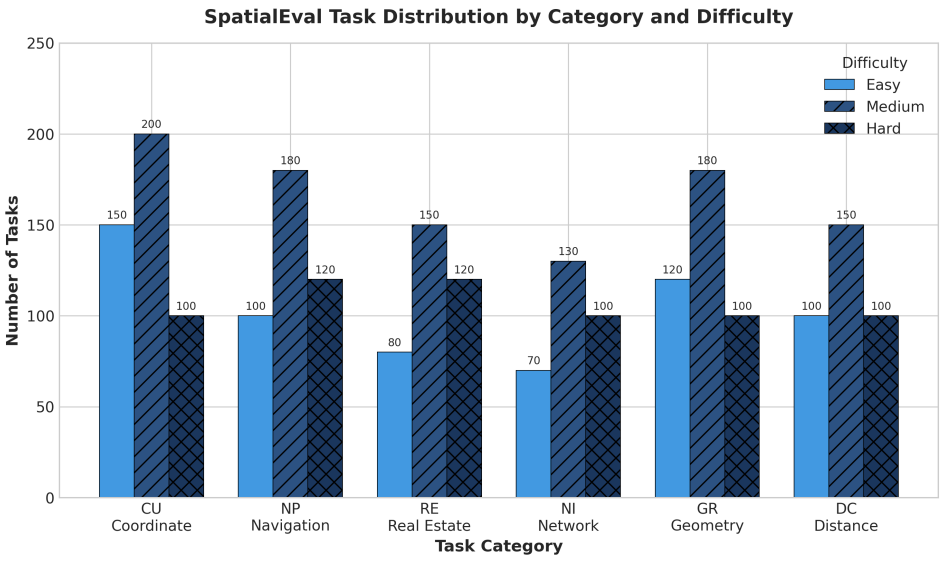

Figure 2: Distribution of the 6,012 tasks in SpatialOps across the twelve categories and three difficulty levels.