Agentic AI for Spatial Intelligence: A Comprehensive Survey

Gloria Felicia
AtlasPro AI
gloria.felicia@atlaspro.ai

Handi Putra
AtlasPro AI
handi.putra@atlaspro.ai

Ayaan Gazali
AtlasPro AI
ayaan.gazali@atlaspro.ai

Eliel Lobo
AtlasPro AI
eliel.lobo@atlaspro.ai

Esteban Rojas
AtlasPro AI
esteban.rojas@atlaspro.ai

Abstract

The convergence of Agentic Artificial Intelligence and Spatial Intelligence marks a pivotal frontier in the pursuit of creating machines that can autonomously operate in the physical world. While agentic systems demonstrate increasingly sophisticated capabilities in planning and tool use, their ability to perceive, reason about, and interact with complex spatial environments remains a significant bottleneck. This survey addresses a critical gap in the existing literature by providing a unified taxonomy that systematically connects the architectural components of agentic AI with the functional requirements of spatial intelligence. We review the foundational concepts of agentic systems, including memory, planning, and tool use, and categorize the diverse landscape of spatial tasks, including navigation, scene understanding, manipulation, and large-scale geospatial analysis. Through a comprehensive analysis of state-of-the-art methods, including embodied agents, multimodal large language models, and geometric graph neural networks, we evaluate the current capabilities and limitations of these systems. We further analyze the fragmented landscape of evaluation benchmarks, highlighting the urgent need for more integrated and holistic frameworks. By synthesizing these disparate research areas and outlining a forward-looking research roadmap, this paper aims to accelerate the development of robust, safe, and effective spatially-aware autonomous systems.

1 Introduction

The evolution of Artificial Intelligence is marked by a paradigm shift from specialized models to goal-oriented, self-directed agents capable of complex decision-making in dynamic environments. This field, which we term **Agentic AI**, represents a significant leap towards creating machines that can operate with a higher degree of autonomy. Concurrently, the ability for these agents to perceive, comprehend, and act within the physical world, a capability we define as **Spatial Intelligence**, has become a primary bottleneck and a critical area of research. The convergence of these two domains is essential for developing AI systems that can effectively and safely navigate real-world complexities, from autonomous vehicles and robotic assistants to large-scale urban planning and disaster response systems.

Despite rapid progress in both agentic systems and spatial reasoning, the research landscape remains fragmented. Numerous surveys have independently covered topics such as Large Language Model agents [Yao et al., 2023b, Wang et al., 2024, Huang et al., 2024], embodied AI [Wang et al., 2023, Driess et al., 2023], and geospatial analysis [Jakubik et al., 2024, Cong et al., 2022, Manas et al., 2021]. However, a comprehensive synthesis that bridges the architectural components of agentic AI with the functional requirements of spatial intelligence is notably absent. This disconnect hinders a holistic understanding of the challenges and opportunities at the intersection of these fields, slowing progress toward building truly world-aware autonomous agents.

This survey aims to fill this critical gap. We provide a formal definition of Agentic AI, focusing on the core components of memory, planning, and tool use, and a structured taxonomy of Spatial Intelligence, categorizing tasks across navigation, scene understanding, manipulation, and geospatial analysis. Our primary contributions are threefold:

1. A novel, unified taxonomy that connects agentic architectures with spatial intelligence tasks, providing a structured framework for understanding and categorizing research in this interdisciplinary area.

2. A comprehensive review of the state-of-the-art methods, evaluation benchmarks, and real-world applications, synthesizing findings from previously disparate fields.

3. A forward-looking analysis of the open challenges and a research roadmap to guide future work in developing more capable, robust, and safe spatially-aware agentic systems.

By providing this synthesis, we aim to create a foundational reference for researchers, developers, and policymakers, fostering a more integrated approach to building the next generation of autonomous intelligence.

2 A Taxonomy of Spatial Intelligence

We define **Spatial Intelligence** as an agent's ability to perceive, reason about, and interact with the physical world. We propose a taxonomy that categorizes spatial tasks into four key domains:

Navigation. The ability to plan and execute paths in a physical environment. This includes tasks like point-to-point navigation [Savva et al., 2019], vision-language navigation [Anderson et al., 2018, Chen et al., 2019, Hong et al., 2020], and exploration [Wang et al., 2023].

Scene Understanding. The ability to perceive and reason about the objects, relationships, and context of a 3D scene. This includes tasks like 3D object detection [Dai et al., 2017], semantic segmentation [Dai et al., 2017], and spatial relationship understanding [Johnson et al., 2017, Suhr et al., 2019, Hudson and Manning, 2019].

Manipulation. The ability to interact with and modify objects in the environment. This includes tasks like object rearrangement [Lin et al., 2022], tool use [Schick et al., 2023], and assembly.

Geospatial Analysis. The ability to reason about and analyze large-scale geographic data. This includes tasks like land use classification [Sumbul et al., 2019], change detection [Zhang et al., 2018], and urban planning [Zheng et al., 2014].

3 Core Components of Agentic AI

Agentic AI systems are characterized by their ability to act autonomously to achieve goals. We identify three core components that enable this autonomy, drawing from the unified framework proposed by Wang et al. [2024]:

Memory. The ability to store and retrieve information from past experiences. This includes short-term memory for in-context learning and long-term memory for retaining knowledge and skills, as demonstrated in Generative Agents [Park et al., 2023] and agents with mapping memory [Gupta et al., 2019].

Planning. The ability to decompose a high-level goal into a sequence of executable actions. This includes techniques like chain-of-thought reasoning [Wei et al., 2022], the more deliberate tree-of-thought search [Yao et al., 2023a], and hierarchical planning [Song et al., 2023, Zhang et al., 2023].

Tool Use. The ability to leverage external tools to extend the agent's capabilities. This includes using APIs for information retrieval [Schick et al., 2023, Lewis et al., 2020], invoking specialized models for specific tasks [Karpas et al., 2022], and interacting with physical actuators.

4 State-of-the-Art Methods

4.1 Embodied AI and Spatial Planning

Embodied AI focuses on creating agents that can learn and act in physical or simulated environments. These agents are critical for spatial planning tasks, as they can directly perceive and interact with the world. Key research areas include:

Vision-Language Navigation (VLN). Agents that follow natural language instructions to navigate real-world environments [Anderson et al., 2018, Chen et al., 2019, Hong et al., 2020, Zhu et al., 2019].

Embodied Question Answering (EQA). Agents that must explore an environment to find the answer to a question [Das et al., 2018].

Robotic Manipulation. Agents that can manipulate objects to achieve goals, often involving complex spatial reasoning and planning, as seen in the SayCan system [Ahn et al., 2022] and VIMA [Lin et al., 2022].

4.2 Multimodal Large Language Models

MLLMs like GPT-4V [OpenAI, 2023] and LLaVA [Liu et al., 2023a] have shown promise in understanding and reasoning about visual information. However, recent benchmarks reveal significant limitations in their spatial reasoning capabilities. For example, EmbodiedBench [Yang et al., 2025] shows that even state-of-the-art models like GPT-4o struggle with low-level manipulation tasks, achieving an average score of only 28.9%. Similarly, the REM benchmark [Thompson et al., 2025] highlights the unreliability of MLLMs in tasks requiring object permanence and spatial relationship tracking from egocentric viewpoints.

4.3 Graph Neural Networks for Spatial Intelligence

GNNs are well-suited for modeling spatial relationships. Spatio-Temporal GNNs have been successfully applied to urban computing tasks like traffic forecasting [Li et al., 2018, Yu et al., 2018, Wu et al., 2019, Jiang and Luo, 2022]. Graph Transformers [Shehzad et al., 2024] offer a scalable approach to capturing long-range spatial dependencies, making them suitable for large-scale spatial graphs like road networks.

5 Benchmarks for Spatial AI Agents

A critical aspect of advancing spatial AI is the development of robust benchmarks to evaluate agent capabilities. We categorize existing benchmarks into:

Navigation Benchmarks. Datasets like R2R [Anderson et al., 2018] and Habitat [Savva et al., 2019] evaluate navigation capabilities.

Manipulation Benchmarks. Environments like ALFWorld [Shridhar et al., 2021] and BEHAVIOR [Srivastava et al., 2021] test object manipulation and task completion.

Spatial Reasoning Benchmarks. Datasets like CLEVR [Johnson et al., 2017] and GQA [Hudson and Manning, 2019] assess compositional spatial reasoning.

Integrated Agent Benchmarks. Recent benchmarks like AgentBench [Liu et al., 2023b], EmbodiedBench [Yang et al., 2025], and REM [Thompson et al., 2025] evaluate multiple agent capabilities in complex environments.

6 Open Challenges and Future Directions

Despite significant progress, several key challenges remain:

Robust Spatial Representation. Developing representations that capture the complexity of 3D environments and generalize across different scenes [Mildenhall et al., 2020, Dai et al., 2017, Chang et al., 2017].

Hierarchical Planning. Creating agents that can plan over long horizons and decompose complex spatial tasks into manageable sub-goals [Song et al., 2023, Zhang et al., 2023, Hao et al., 2023].

Table 1: Key Benchmarks for Spatial AI Agents

Benchmark	Focus	Tasks	Key Metric	Year
EmbodiedBench	MLLM embodied agents	1,128	Success rate	2025
REM	Embodied spatial reasoning	Multi-frame	Accuracy	2025
MineAnyBuild	Spatial planning	Building	Quality score	2025
SafeAgentBench	Safe task planning	Safety-aware	Safety rate	2024
BEHAVIOR	Household activities	100	Task success	2021
Habitat	Embodied navigation	PointNav/ObjectNav	SPL	2019
VLN-R2R	Vision-language navigation	Room-to-room	SR/SPL	2018
ALFWorld	Text-embodied alignment	Household	Success rate	2021
WebArena	Web navigation	Web tasks	Task success	2023
AgentBench	Multi-domain agents	8 domains	Composite	2023

Safe and Reliable Tool Use. Ensuring that agents can use tools safely and effectively, especially in safety-critical applications, as highlighted by the SafeAgentBench benchmark [Unknown, 2025] and research on constitutional AI [Bai et al., 2022].

Sim-to-Real Transfer. Bridging the gap between simulation and the real world to enable the deployment of embodied agents in real-world applications [Savva et al., 2019, Shen et al., 2021].

7 Conclusion

This survey has provided a comprehensive overview of the intersection of Agentic AI and Spatial Intelligence. We have proposed a unified taxonomy, reviewed the state-of-the-art, and identified key challenges and future directions. We believe that by fostering a more integrated approach to research in this area, we can accelerate the development of truly intelligent autonomous systems that can understand and interact with the physical world.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gober, Karol Gopalakrishnan, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE, 2017.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. *arXiv preprint arXiv:1711.11543*, 2018.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Saurabh Gupta, Varun Tolani, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. *International Journal of Computer Vision*, 128(5): 1311–1330, 2019.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln-bert: A recurrent vision-and-language bert for navigation. *arXiv preprint arXiv:2011.13922*, 2020.

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.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

Johannes Jakubik, Sujit Roy, C E Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Musber, Daiki Oliveira, et al. Prithvi: A foundation model for earth observation. *arXiv preprint arXiv:2310.18660*, 2024.

Weiwei Jiang and Jiayun Luo. Graph neural networks for traffic forecasting: A survey. *arXiv preprint arXiv:2101.11174*, 2022.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

Ehud Karpas, Omri Abend, Jonathan Berant, et al. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Yunfan Lin, Yuqi Xie, Chaowei Xiao, Anima Anandkumar, and Yuke Zhu. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022.

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

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023b.

Oscar Manas, Alexandre Lacoste, Xavier Giro-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *arXiv preprint arXiv:2103.16607*, 2021.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2020.

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

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.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019.

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.

Ahsan Shehzad, Feng Xia, Shagufta Abid, Chao Peng, Shuo Yu, Dongyu Zhang, and Karin Verspoor. Graph transformers: A survey. *arXiv preprint arXiv:2407.09777*, 2024.

Bokui Shen, Fei Xia, Chengshu Li, Roberto Martin, Linxi Fan, Guanzhi Wang, Shyamal Buch, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2021.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *arXiv preprint arXiv:2212.04088*, 2023.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martin, Fei Xia, Kent Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. *arXiv preprint arXiv:2108.03332*, 2021.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2019.

Gencer Sumbul, Marcela Charfuelan, Begum Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *arXiv preprint arXiv:1902.06148*, 2019.

James Thompson et al. Rem: A benchmark for evaluating embodied spatial reasoning in mllms. *arXiv preprint arXiv:2512.00736*, 2025.

Unknown. Safeagentbench: A benchmark for safe task planning of embodied llm agents. *arXiv preprint arXiv:2412.13178*, 2025.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1907–1913, 2019.

Rui Yang, Hanyang Lin, Junyu Zhu, and Jingyi Huang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023a.

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*, 2023b.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3634–3640, 2018.

Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Lei Huang, and Guangchao Liu. Change detection based on deep siamese convolutional network for optical aerial images. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1845–1849, 2018.

Yiwen Zhang et al. Graph-based planning for embodied agents. *arXiv preprint*, 2023.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology*, 5(3):1–55, 2014.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. *arXiv preprint arXiv:1911.07883*, 2019.