Autonomous Spatial Intelligence: A Survey of Agentic AI Methods and Evaluation

Gloria Felicia
AtlasPro AI
gloria.felicia@atlaspro.ai

Nolan Bryant
AtlasPro AI
nolan.bryant@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, Xi et al., 2023, Guo et al., 2024], embodied AI [Wang et al., 2023, Driess et al., 2023, Xiao et al., 2023], and geospatial analysis [Janowicz et al., 2025, Mai

et al., 2024]. 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 over 300 papers.

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, Shah et al., 2023b], vision-language navigation [Anderson et al., 2018, Chen et al., 2019, Hong et al., 2020, Shah et al., 2023a], and exploration [Wang et al., 2023, Zhou et al., 2023a].

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, Peng et al., 2023], semantic segmentation [Dai et al., 2017, Takmaz et al., 2023], and spatial relationship understanding [Johnson et al., 2017, Suhr et al., 2019, Hudson and Manning, 2019, Armeni et al., 2019, Wald et al., 2020, Gu et al., 2024].

Manipulation. The ability to interact with and modify objects in the environment. This includes tasks like object rearrangement [Lin et al., 2022, Wu et al., 2023a], tool use [Schick et al., 2023, Liang et al., 2023], and assembly [Okamura et al., 2000].

Geospatial Analysis. The ability to reason about and analyze large-scale geographic data. This includes tasks like land use classification [Sumbul et al., 2019, Janowicz et al., 2025], change detection [Zhang et al., 2018], and urban planning [Zheng et al., 2014, Jin et al., 2023].

3 Core Components and Architectures 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] and Sumers et al. [2024]:

Memory. The ability to store and retrieve information from past experiences. This includes short-term memory for in-context learning [Brown et al., 2020] 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, Huang et al., 2023a].

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, Lin 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 [Brohan et al., 2022, 2023].

3.1 Agentic Architectures

Several prominent architectures have emerged to orchestrate these components:

ReAct (Reason+Act). This architecture [Yao et al., 2023b] interleaves reasoning traces with actions, allowing the agent to create, maintain, and adjust plans while interacting with an external environment. The reasoning traces enable the model to handle exceptions, and update its plan based on the outcomes of its actions.

Reflexion. This framework [Shinn et al., 2023] enhances agents with dynamic memory and self-reflection capabilities. After a task failure, the agent reflects on the feedback to identify the cause of the error and updates its internal memory to avoid repeating the same mistake in subsequent trials.

Tree of Thoughts (ToT). ToT [Yao et al., 2023a] generalizes over chain-of-thought by exploring multiple reasoning paths in a tree structure. This allows the agent to deliberately explore different lines of reasoning, self-evaluate its progress, and backtrack when necessary, making it more suitable for complex planning tasks.

Multi-Agent Systems. A growing area of research focuses on the collaboration of multiple agents to solve complex problems. Frameworks like AutoGen [Wu et al., 2023b], MetaGPT [Hong et al., 2024], and CAMEL [Li et al., 2023] enable sophisticated multi-agent conversations and workflows.

4 State-of-the-Art Methods and Industry Agents

4.1 Embodied AI and Vision-Language-Action (VLA) Models

Embodied AI focuses on creating agents that can learn and act in physical or simulated environments. A key development in this area is the rise of Vision-Language-Action (VLA) models, which are trained to map multimodal inputs (vision, language) directly to robotic actions. These models are at the forefront of creating general-purpose robots.

RT-2 (Robotics Transformer 2). Developed by Google DeepMind, RT-2 [Brohan et al., 2023] is a VLA model that leverages large-scale web data to learn general concepts about the world and transfer them to robot control. It demonstrates emergent capabilities, such as reasoning about novel objects and executing tasks it was not explicitly trained on.

PaLM-E. This 562-billion parameter embodied multimodal language model from Google [Driess et al., 2023] integrates continuous sensor data from robotic systems directly into a large language model. This allows it to ground language in real-world perception and perform a variety of robotic tasks without task-specific training.

Octo and OpenVLA. Recent open-source efforts like Octo [Team et al., 2024] and OpenVLA [Kim et al., 2024] are democratizing access to powerful VLA models, enabling broader research and development in the community.

4.2 Industry Agents for Spatial Planning

Several notable agents from industry showcase the application of these architectures to complex spatial planning tasks:

Voyager. Developed by NVIDIA, Voyager [Wang et al., 2023] is an LLM-powered embodied agent that excels at open-ended exploration and skill acquisition in Minecraft. It uses GPT-4 to generate a curriculum, write code for new skills, and store them in a skill library for long-term use.

SayCan. This Google project [Ahn et al., 2022] grounds language models in robotic affordances. It uses an LLM to determine high-level actions and a learned value function to assess the feasibility of those actions for a given robot, effectively bridging the gap between abstract reasoning and physical capability.

Code as Policies. This approach [Liang et al., 2023] uses LLMs to generate Python code that serves as a reactive policy for a robot. This allows for more complex and dynamic behaviors than direct action prediction and leverages the extensive knowledge of programming and logic embedded in LLMs.

5 A Brief Overview of Benchmarks

A comprehensive analysis of benchmarks for spatial AI agents will be the subject of a follow-up paper. However, a brief overview is necessary to contextualize the current state of evaluation. Existing benchmarks can be broadly categorized into:

- **Navigation Benchmarks:** R2R [Anderson et al., 2018], Habitat [Savva et al., 2019], ZSON [Majumdar et al., 2022], CoWs on Pasture [Gadre et al., 2023]
- **Manipulation Benchmarks:** ALFWorld [Shridhar et al., 2021], BEHAVIOR [Srivastava et al., 2021], TidyBot [Wu et al., 2023a]
- **Spatial Reasoning Benchmarks:** CLEVR [Johnson et al., 2017], GQA [Hudson and Manning, 2019], Open3DVQA [Zhang et al., 2025]
- **Integrated Agent Benchmarks:** AgentBench [Liu et al., 2023], EmbodiedBench [Yang et al., 2025], WebArena [Zhou et al., 2023b]

These benchmarks, while valuable, often focus on narrow aspects of spatial intelligence. A key challenge for the community is the development of more holistic evaluation frameworks that can assess the full range of agentic capabilities in complex, open-ended spatial tasks.

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, Kerr et al., 2023, Rosinol et al., 2020, Hughes et al., 2022].

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, Huang et al., 2023b].

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, Zhao et al., 2020].

World Models. A promising direction is the development of world models that can learn a comprehensive understanding of the world and predict future states [Feng et al., 2025, Ding et al., 2024, Zhen et al., 2024]. These models could enable more robust and generalizable agents.

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.

Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. *arXiv preprint arXiv:1910.02527*, 2019.

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.

Anthony Brohan, Noah Brown, Justice Carbajal, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

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.

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.

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.

Mingyu Ding et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 2024.

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.

Tuo Feng, Yixiao Wang, Jiaxin Chen, et al. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. *arXiv preprint arXiv:2203.10421*, 2023.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2024.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, et al. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

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.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, et al. Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2024.

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.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. *arXiv preprint arXiv:2210.05714*, 2023a.

Siyuan Huang, Zhengkai Jiang, Hao Dong, et al. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023b.

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.

Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*, 2022.

Krzysztof Janowicz et al. Geofm: how will geo-foundation models reshape spatial data science and geoai? *International Journal of Geographical Information Science*, 2025.

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.

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.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. *arXiv preprint arXiv:2303.09553*, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

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.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for mind exploration of large language model society. *arXiv preprint arXiv:2303.17760*, 2023.

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.

Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 2023.

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

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*, 2023.

Gengchen Mai, Weiming Huang, Jin Sun, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *ACM SIGSPATIAL Special*, 2024.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *arXiv preprint arXiv:2206.12403*, 2022.

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.

Allison M Okamura, Nils Smaby, and Mark R Cutkosky. An overview of dexterous manipulation. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, volume 1, pages 255–262. IEEE, 2000.

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.

Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. *arXiv preprint arXiv:2211.15654*, 2023.

Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. *arXiv preprint arXiv:1910.02490*, 2020.

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.

Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. *arXiv preprint arXiv:2207.04429*, 2023a.

Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv preprint arXiv:2306.14846*, 2023b.

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.

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

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 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.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2024.

Ayca Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023.

Octo Model Team, Dibya Ghosh, Homer Walke, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

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

Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. *arXiv preprint arXiv:2004.03967*, 2020.

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.

Jimmy Wu, Rika Antonova, Adam Kan, et al. Tidybot: Personalized robot assistance with large language models. *arXiv preprint arXiv:2305.05658*, 2023a.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023b.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Xing Xiao, Han Liu, Yinuo Li, and Dong Zhao. Robot learning in the era of foundation models: A survey. *arXiv preprint arXiv:2311.14379*, 2023.

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.

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.

Wei Zhang, Zheng Zhou, Zhen Zheng, et al. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *arXiv preprint arXiv:2503.11094*, 2025.

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

Wenyu Zhao, Jorge Pena Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744. IEEE, 2020.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, et al. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

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.

Kaiwen Zhou, Kaizhi Zheng, Connor Pratt, et al. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. *arXiv preprint arXiv:2301.13166*, 2023a.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023b.