Autonomous Spatial Intelligence: A Survey of Agentic AI Methods for Physical World Understanding and Interaction

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 represents a transformative frontier in creating machines capable of autonomous operation in physical environments. This survey provides the first unified taxonomy systematically connecting agentic AI architectures—encompassing memory, planning, and tool use—with spatial intelligence capabilities spanning navigation, scene understanding, manipulation, and geospatial analysis. We synthesize over 300 papers across foundational agentic frameworks [Yao et al., 2023b, Shinn et al., 2023, Wang et al., 2024b], vision-language-action models [Brohan et al., 2023, Team et al., 2024, Kim et al., 2024], graph neural networks for spatial reasoning [Kipf and Welling, 2017, Velickovic et al., 2018, Wu et al., 2019a], world models [Hafner et al., 2023, Hu et al., 2023a], and geospatial foundation models [Jakubik et al., 2024, Cong et al., 2022]. Our analysis reveals three key findings: (1) the critical role of hierarchical memory systems in enabling long-horizon spatial tasks, (2) the emergence of GNN-LLM integration as a powerful paradigm for structured spatial reasoning, and (3) the growing importance of world models for safe deployment in physical environments. We present a comprehensive evaluation framework and identify open challenges including robust spatial representation, sim-to-real transfer, and multi-agent coordination. This survey establishes a foundational reference for advancing spatially-aware autonomous systems.

1 Introduction

The pursuit of artificial general intelligence increasingly centers on creating agents that can perceive, reason about, and act within physical environments [McCarthy et al., 1955, Turing, 1950]. While large language models have demonstrated remarkable capabilities in reasoning and planning [Brown et al., 2020, OpenAI, 2023, Wei et al., 2022], their ability to operate effectively in spatial contexts remains a fundamental challenge [Chen et al., 2024, Yang et al., 2025].

We define **Agentic AI** as systems exhibiting goal-directed behavior through autonomous decision-making, characterized by three core capabilities: persistent memory for experience accumulation, planning for action sequencing, and tool use for capability extension [Wang et al., 2024b, Xi et al., 2023, Weng, 2023].

Complementarily, **Spatial Intelligence** encompasses the ability to perceive 3D structure, reason about object relationships, navigate environments, and manipulate physical objects [Chen et al., 2024, Thompson et al., 2025].

The convergence of these domains is essential for real-world AI applications. Autonomous vehicles must perceive dynamic environments and plan safe trajectories [Hu et al., 2023b, Caesar et al., 2020]. Robotic assistants require understanding of object affordances and spatial relationships [Brohan et al., 2023, Ahn et al., 2022]. Urban computing systems must model complex spatio-temporal dependencies [Jin et al., 2023, Li et al., 2018]. Despite this importance, existing surveys treat these areas in isolation, lacking a unified framework connecting agentic architectures with spatial requirements.

Contributions. This survey makes three primary contributions:

1. A **unified taxonomy** connecting agentic AI components (memory, planning, tool use) with spatial intelligence domains (navigation, scene understanding, manipulation, geospatial analysis), providing a structured framework for interdisciplinary research.

2. A **comprehensive analysis** of over 300 papers identifying key architectural patterns, including the emergence of GNN-LLM integration and world model-based planning as critical enablers for spatial reasoning.

3. A **forward-looking roadmap** identifying open challenges and research directions for developing robust, safe, and capable spatially-aware autonomous systems.

2 Unified Taxonomy

We propose a two-dimensional taxonomy (Figure **??**) that maps agentic capabilities to spatial task requirements, enabling systematic analysis of existing methods and identification of research gaps.

2.1 Agentic AI Components

Memory Systems. Memory enables agents to accumulate and retrieve experiential knowledge. Short-term memory through in-context learning [Brown et al., 2020] supports immediate reasoning, while long-term memory via retrieval-augmented generation [Lewis et al., 2020, Packer et al., 2023] enables knowledge persistence. For spatial tasks, cognitive mapping [Gupta et al., 2019b, Chaplot et al., 2020b] and semantic spatial memory [Huang et al., 2023] are critical for navigation and scene understanding.

Planning Systems. Planning decomposes goals into executable action sequences. Chain-of-thought reasoning [Wei et al., 2022, Kojima et al., 2022] enables step-by-step problem solving. Tree-based search [Yao et al., 2023a, Besta et al., 2023] explores multiple solution paths. Hierarchical planning [Song et al., 2023, Huang et al., 2022] bridges high-level goals with low-level actions. For spatial domains, planning must account for geometric constraints, physical dynamics, and uncertainty.

Tool Use and Action. Tool use extends agent capabilities through external interfaces. API integration [Schick et al., 2023, Patil et al., 2023, Qin et al., 2024] enables access to specialized functions. Code generation [Gao et al., 2023, Liang et al., 2023] provides flexible action specification. The ReAct architecture [Yao et al., 2023b] interleaves reasoning with action execution, forming the foundation for many spatial agents.

2.2 Spatial Intelligence Domains

Navigation. Navigation requires path planning and execution in physical or simulated environments. Vision-language navigation [Anderson et al., 2018, Ku et al., 2020, Qi et al., 2020] follows natural language instructions. Object-goal navigation [Batra et al., 2020, Chaplot et al., 2020a] locates target object categories. Zero-shot approaches [Majumdar et al., 2022, Gadre et al., 2022] leverage vision-language models for novel object navigation.

Scene Understanding. Scene understanding encompasses 3D perception and semantic reasoning. Neural radiance fields [Mildenhall et al., 2020, Barron et al., 2022] and 3D Gaussian splatting [Kerbl et al., 2023] enable novel view synthesis. Point cloud processing [Qi et al., 2017a,b] supports 3D object detection. Scene graphs [Xu et al., 2017, Krishna et al., 2017, Armeni et al., 2019] represent object relationships for higher-level reasoning.

Manipulation. Manipulation involves physical interaction with objects. Vision-language-action models [Brohan et al., 2022, 2023, Team et al., 2024, Kim et al., 2024] directly map observations to robot actions. Task and motion planning [Garrett et al., 2021, Ahn et al., 2022] integrates high-level reasoning with low-level control. Dexterous manipulation [Akkaya et al., 2019, Chen et al., 2022] addresses complex hand-object interactions.

Geospatial Analysis. Geospatial analysis reasons about large-scale geographic data. Remote sensing foundation models [Jakubik et al., 2024, Cong et al., 2022, Bastani et al., 2023] enable transfer learning across satellite imagery tasks. Spatio-temporal graph networks [Li et al., 2018, Yu et al., 2018, Wu et al., 2019b, Bai et al., 2020] model urban dynamics for traffic prediction and city planning.

3 State-of-the-Art Methods

3.1 Vision-Language-Action Models

VLA models represent a paradigm shift in robotics, directly mapping multimodal inputs to actions through end-to-end learning.

Proprietary Models. RT-1 [Brohan et al., 2022] demonstrated transformer-based policies trained on large-scale robot data. RT-2 [Brohan et al., 2023] co-trained on web-scale vision-language data, enabling emergent reasoning about novel objects. PaLM-E [Driess et al., 2023] integrated continuous sensor data into a 562B parameter language model for embodied reasoning.

Open-Source Models. Octo [Team et al., 2024] provides a generalist robot policy trained on the Open X-Embodiment dataset [Collaboration, 2023]. OpenVLA [Kim et al., 2024] offers a 7B parameter alternative with competitive performance. These models democratize VLA research and enable community-driven advancement.

Multimodal Foundations. LLaVA [Liu et al., 2023] pioneered visual instruction tuning. Flamingo [Alayrac et al., 2022] introduced few-shot multimodal learning. BLIP-2 [Li et al., 2023b] efficiently bootstraps vision-language pretraining. Qwen-VL [Bai et al., 2023, Wang et al., 2024c] and GPT-4V [OpenAI, 2023] represent frontier multimodal capabilities.

3.2 Graph Neural Networks for Spatial Reasoning

GNNs provide powerful tools for modeling spatial relationships and dependencies, with emerging integration with language models.

Foundational Architectures. GCN [Kipf and Welling, 2017] introduced spectral graph convolution. GAT [Velickovic et al., 2018] added attention mechanisms. GraphSAGE [Hamilton et al., 2017] enabled inductive learning. GIN [Xu et al., 2019] provided theoretical expressiveness analysis. These architectures form the basis for spatial graph learning.

Spatio-Temporal Networks. DCRNN [Li et al., 2018] models traffic as graph diffusion. STGCN [Yu et al., 2018] combines graph and temporal convolutions. Graph WaveNet [Wu et al., 2019b] learns adaptive graph structures. AGCRN [Bai et al., 2020] introduces node-specific patterns. Comprehensive surveys [Jin et al., 2023, Atluri et al., 2018] detail these advances.

GNN-LLM Integration. Recent work explores combining GNNs with LLMs for enhanced reasoning. GraphGPT [Tang et al., 2024] aligns graph encoders with language models. GNN-RAG [Wang et al., 2024a] combines graph retrieval with language generation. This integration holds significant promise for spatial reasoning requiring both structural and semantic understanding.

3.3 World Models

World models learn predictive representations enabling planning through imagination, critical for safe deployment in physical environments.

Model-Based RL. Dreamer [Hafner et al., 2019] introduced latent imagination. DreamerV2 [Hafner et al., 2021] achieved human-level Atari performance. DreamerV3 [Hafner et al., 2023] demonstrated cross-domain mastery. DayDreamer [Wu et al., 2023a] transferred world models to real robots.

Video World Models. Genie [Bruce et al., 2024] learns controllable world models from internet videos. WorldDreamer [Yang et al., 2024] generates driving world models. GAIA-1 [Hu et al., 2023a] produces realistic driving videos conditioned on actions.

LLM-Based World Models. LLMs can serve as world models for planning [Hao et al., 2023, Guan et al., 2023], predicting state transitions without explicit environment models.

3.4 Embodied AI Agents

Open-Ended Exploration. Voyager [Wang et al., 2023] demonstrated open-ended exploration in Minecraft through LLM-driven curriculum learning. MineDojo [Fan et al., 2022] provides benchmarks for open-ended embodied agents.

Grounded Language Agents. SayCan [Ahn et al., 2022] grounds language models in robotic affordances. Code as Policies [Liang et al., 2023] generates executable robot code. LLM-Planner [Song et al., 2023] enables few-shot grounded planning.

Simulation Platforms. Habitat [Savva et al., 2019, Szot et al., 2021, Puig et al., 2024] provides high-fidelity embodied AI simulation. iGibson [Shen et al., 2021, Li et al., 2021] offers interactive environments. AI2-THOR [Kolve et al., 2017] enables interactive visual AI research.

4 Industry Applications

4.1 Geospatial Intelligence

Palantir [Palantir, 2023, Bailey, 2021] integrates AI with geospatial analysis for defense and commercial applications. **ESRI** [ESRI, 2023] provides ArcGIS with integrated GeoAI capabilities. **Google** [Google, 2023] deploys AI for global-scale mapping and navigation.

4.2 Location Intelligence

Foursquare [Foursquare, 2023] provides location intelligence through movement pattern analysis. Smart city applications [Zheng et al., 2014, Allam and Dhunny, 2020] leverage spatial AI for traffic management and urban planning.

4.3 Autonomous Vehicles

Waymo [Waymo, 2023, 2024] has deployed autonomous vehicles at scale. End-to-end approaches including UniAD [Hu et al., 2023b], VAD [Jiang et al., 2023], and DriveVLM [Tian et al., 2024] unify perception, prediction, and planning.

5 Evaluation Framework

5.1 Navigation Benchmarks

R2R [Anderson et al., 2018], RxR [Ku et al., 2020], REVERIE [Qi et al., 2020] for vision-language navigation. Habitat ObjectNav [Batra et al., 2020] and SOON [Zhu et al., 2021] for object-goal navigation.

5.2 Manipulation Benchmarks

RLBench [James et al., 2020], Meta-World [Yu et al., 2020], BEHAVIOR [Srivastava et al., 2021, Li et al., 2023a] for robotic manipulation evaluation.

5.3 Spatial Reasoning Benchmarks

CLEVR [Johnson et al., 2017], GQA [Hudson and Manning, 2019], SpatialVLM [Chen et al., 2024], REM [Thompson et al., 2025], EmbodiedBench [Yang et al., 2025] for spatial reasoning evaluation.

5.4 Geospatial Benchmarks

BigEarthNet [Sumbul et al., 2019], fMoW [Christie et al., 2018], xBD [Gupta et al., 2019a], SpaceNet [Van Etten et al., 2018] for remote sensing evaluation.

6 Open Challenges and Future Directions

Robust Spatial Representation. Developing representations that generalize across scenes, viewpoints, and conditions remains challenging [Mildenhall et al., 2020, Kerbl et al., 2023]. Foundation models for 3D understanding [Hong et al., 2023b] represent promising directions.

Long-Horizon Planning. Creating agents that plan over extended horizons and decompose complex spatial tasks is essential [Song et al., 2023, Valmeekam et al., 2023]. Integration of neural and symbolic planning approaches shows promise.

Safe and Reliable Operation. Ensuring safe operation in safety-critical applications is paramount [Yin et al., 2025, Amodei et al., 2016, Bai et al., 2022]. Robust uncertainty handling and alignment with human values are critical.

Sim-to-Real Transfer. Bridging simulation and reality remains challenging [Zhao et al., 2020, Tobin et al., 2017]. Domain randomization and real-world fine-tuning are active research areas.

Multi-Agent Coordination. Scaling to multi-agent systems for complex spatial tasks requires advances in coordination and communication [Zhang et al., 2021, Wu et al., 2023b, Hong et al., 2023a].

7 Conclusion

This survey has provided a unified taxonomy connecting Agentic AI and Spatial Intelligence, synthesizing over 300 papers across foundational architectures, state-of-the-art methods, industry applications, and evaluation benchmarks. Our analysis reveals the critical importance of hierarchical memory, GNN-LLM integration, and world models for spatial reasoning. Key challenges remain in robust representation, long-horizon planning, and safe deployment. By establishing this foundational reference, we aim to accelerate progress toward capable, robust, and safe spatially-aware autonomous systems.

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.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

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.

Zaheer Allam and Zaynah A Dhunny. On big data, artificial intelligence and smart cities. *Cities*, 89:80–91, 2020.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mane. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

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.

Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys*, 2018.

Jinze Bai, Shuai Bai, Shusheng Yang, et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Lei Bai et al. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems*, 2020.

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.

Jonathan Bailey. Palantir technologies: Building the operating system for the modern enterprise. Industry Report, 2021.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022.

Favyen Bastani et al. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *arXiv preprint arXiv:2211.15660*, 2023.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. In *arXiv preprint arXiv:2006.13171*, 2020.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajber, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2023.

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.

Jake Bruce, Michael Dennis, Ashley Edwards, et al. Genie: Generative interactive environments. In *ICML*, 2024.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, et al. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *NeurIPS*, 2020a.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *ICLR*, 2020b.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024.

Tao Chen et al. A system for general in-hand object re-orientation. *CoRL*, 2022.

Gordon Christie et al. Functional map of the world. *CVPR*, 2018.

Open X-Embodiment Collaboration. Open x-embodiment, 2023.

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.

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.

ESRI. Esri arcgis: The mapping and analytics platform. `https://www.esri.com`, 2023.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.

Foursquare. Foursquare location intelligence. `https://foursquare.com`, 2023.

Samir Yitzhak Gadre et al. Clip on wheels. *arXiv preprint arXiv:2203.10421*, 2022.

Luyu Gao et al. Pal: Program-aided language models. *International Conference on Machine Learning*, 2023.

Caelan Reed Garrett et al. Integrated task and motion planning. *Annual Review of Control*, 2021.

Google. Google maps platform. `https://cloud.google.com/maps-platform`, 2023.

Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36, 2023.

Ritwik Gupta et al. xbd: A dataset for assessing building damage. *arXiv preprint arXiv:1911.09296*, 2019a.

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, 2019b.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *ICLR*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Danijar Hafner et al. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 2017.

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, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023a.

Yining Hong et al. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 2023b.

Anthony Hu et al. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023a.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, et al. Planning-oriented autonomous driving. In *CVPR*, 2023b.

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

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

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.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark. *IEEE Robotics and Automation Letters*, 2020.

Bo Jiang et al. Vad: Vectorized scene representation for efficient autonomous driving. *IEEE International Conference on Computer Vision*, 2023.

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.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuhler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 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.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2017.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal of Computer Vision*, volume 123, pages 32–73, 2017.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *EMNLP*, 2020.

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.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *Conference on Robot Learning*, pages 455–465, 2021.

Chengshu Li, Ruohan Zhang, Josiah Wong, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *CoRL*, 2023a.

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*, 2023b.

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.

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.

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

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.

John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence. *AI Magazine*, 27(4):12–12, 1955.

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-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Palantir. Palantir technologies. `https://www.palantir.com`, 2023.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Dhruv Batra, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots. In *International Conference on Learning Representations*, 2024.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, 2017b.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, 2020.

Yujia Qin, Shihao Liang, Yining Ye, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *ICLR*, 2024.

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.

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.

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.

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.

Andrew Szot, Alexander Clegg, Eric Undersander, et al. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, 2021.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. *arXiv preprint arXiv:2310.13023*, 2024.

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

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

Xiaoyu Tian et al. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Josh Tobin, Rocky Fong, Alex Ray, John Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30. IEEE, 2017.

Alan M Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models–a critical investigation. *Advances in Neural Information Processing Systems*, 36, 2023.

Adam Van Etten et al. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2018.

Costas Wang et al. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*, 2024a.

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*, 2024b.

Peng Wang et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024c.

Waymo. Waymo: The world's most experienced driver. `https://waymo.com`, 2023.

Waymo. Introducing Waymoś Research on an End-to-End Multimodal Model for Autonomous Driving. Waymo Blog, October 2024. URL `https://waymo.com/blog/2024/10/introducing-emma`.

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.

Lilian Weng. Llm powered autonomous agents. *Lil'Log*, 2023. `https://lilianweng.github.io/posts/2023-06-23-agent/`.

Philipp Wu et al. Daydreamer: World models for physical robot learning. *arXiv preprint arXiv:2206.14176*, 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.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1): 4–24, 2019a.

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, 2019b.

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.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 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.

Xiaofeng Yang et al. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024.

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.

Sheng Yin, Xianghe Xiong, Wenhao Huang, et al. Safeagentbench: A benchmark for safe task planning of embodied llm agents. *arXiv preprint arXiv:2412.13178*, 2025.

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.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, et al. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

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.

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. Soon: Scenarios for object navigation with natural language instructions. In *CVPR*, 2021.