Autonomous Spatial Intelligence: A Comprehensive Technical Report on Agentic AI Methods, Architectures, 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 technical report 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 over 500 papers spanning foundational agentic architectures [Yao et al., 2023b, Shinn et al., 2023, Yao et al., 2023a, Wei et al., 2022], embodied AI systems [Wang et al., 2023, Driess et al., 2023, Brohan et al., 2023, Ahn et al., 2022], vision-language-action models [**?**Kim et al., 2024, Liu et al., 2023a, Alayrac et al., 2022], graph neural networks for spatial reasoning [Kipf and Welling, 2017, Velickovic et al., 2018, Hamilton et al., 2017, **?**], world models [Hafner et al., 2023, 2021, **?**, **?**], and geospatial foundation models [Jakubik et al., 2024, **?**, Cong et al., 2022, **?**]. Through comprehensive analysis of state-of-the-art methods, industry applications from Palantir, ESRI, Foursquare, Google, and Waymo, and evaluation benchmarks, we provide a foundational reference for researchers and practitioners. 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 [Wang et al., 2024, Xi et al., 2023, **?**]. The foundational work on large language models [**???**Touvron et al., 2023, Team and Google, 2023, Anthropic, 2024] has enabled a new generation of AI agents that can reason, plan, and execute complex tasks through natural language interfaces [Wei et al., 2022, **?**, **?**].

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 [Chen et al., 2024, Yang et al., 2025, Thompson et al., 2025]. The convergence of these two domains is essential for developing AI systems that can effectively and safely navigate real-world complexities, from autonomous vehicles [?Caesar et al., 2020, ?, Hu et al., 2023, ?] and robotic assistants [Brohan et al., 2023, Driess et al., 2023, ?] to large-scale urban planning [Zheng et al., 2014, Jin et al., 2023, Li et al., 2018] and disaster response systems [??Bastani et al., 2023].

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, ?, Patil et al., 2023, Qin et al., 2024], embodied AI [Wang et al., 2023, Driess et al., 2023, ?, Amin and Kiela, 2024], multimodal large language models [Liu et al., 2023a, Alayrac et al., 2022, Li et al., 2023b, Bai et al., 2023, ?], graph neural networks [?Kipf and Welling, 2017, Velickovic et al., 2018, Hamilton et al., 2017, Xu et al., 2019, Battaglia et al., 2018], spatio-temporal prediction [Jin et al., 2023, Li et al., 2018, Yu et al., 2018, Wu et al., 2019, ?], world models [Hafner et al., 2023, 2021, ?, ?, ?], 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 technical report 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:

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 over 500 papers covering state-of-the-art methods, evaluation benchmarks, and real-world industry 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.

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, each with distinct challenges and methodological approaches.

2.1 Navigation

Navigation encompasses the ability to plan and execute paths in physical or simulated environments. This domain has seen remarkable progress through vision-language navigation (VLN) [Anderson et al., 2018, Chen et al., 2019, Hong et al., 2020, ?, Ku et al., 2020], which requires agents to follow natural language instructions in realistic environments. The Room-to-Room (R2R) benchmark [Anderson et al., 2018] established a foundational evaluation framework, while subsequent work has extended to continuous environments [?], outdoor settings [Chen et al., 2019], and cross-lingual scenarios [Yan et al., 2020].

Point-to-point navigation has been advanced through the Habitat platform [Savva et al., 2019, Szot et al., 2021, ?], which provides high-fidelity simulation environments for training and evaluating embodied agents. Object-goal navigation [?Chaplot et al., 2020a, ?] requires agents to navigate to specific object categories, while image-goal navigation [?] uses visual targets. Zero-shot object navigation (ZSON) [Majumdar et al., 2022, ?] leverages vision-language models to navigate to novel objects without task-specific training.

Semantic mapping approaches [Gupta et al., 2019, Chaplot et al., 2020b, Huang et al., 2023] build spatial representations that enable more efficient navigation. VLMaps [Huang et al., 2023] creates open-vocabulary 3D semantic maps by fusing CLIP features with depth information, enabling natural language queries about spatial locations. Recent work on visual navigation transformers [Shah et al., 2023, ?] has demonstrated impressive generalization across diverse environments through large-scale pretraining.

2.2 Scene Understanding

Scene understanding involves perceiving and reasoning about the objects, relationships, and context of 3D environments. This domain spans multiple levels of abstraction, from low-level perception to high-level semantic reasoning.

3D Reconstruction and Representation. Neural Radiance Fields (NeRF) [Mildenhall et al., 2020, ?, Barron et al., 2022] have revolutionized novel view synthesis by representing scenes as continuous volumetric functions. More recently, 3D Gaussian Splatting [Kerbl et al., 2023] has emerged as a faster alternative with explicit scene representations. These representations enable agents to build detailed mental models of their environments.

3D Object Detection and Segmentation. Point cloud processing through PointNet [Qi et al., 2017] and PointNet++ [?] established foundational architectures for 3D understanding. Subsequent work has developed more sophisticated approaches including Point Transformers [??], voxel-based methods [??], and hybrid approaches. Indoor scene understanding has been advanced through datasets like ScanNet [Dai et al., 2017], Matterport3D [Chang et al., 2017], and S3DIS [?].

Scene Graphs. Scene graph generation [Xu et al., 2017, ?, ?] provides structured representations of objects and their relationships, enabling higher-level reasoning about spatial configurations. Visual Genome [?] established a large-scale dataset for this task, while recent work has explored 3D scene graphs [Armeni et al., 2019, ?] for more complete environmental understanding.

Spatial Reasoning Benchmarks. CLEVR [Johnson et al., 2017] introduced compositional visual reasoning, while GQA [Hudson and Manning, 2019] extended this to real-world images. NLVR2 [Suhr et al., 2019] focuses on grounded language understanding, and SpatialVLM [Chen et al., 2024] specifically targets spatial reasoning in vision-language models. Recent benchmarks like REM [Thompson et al., 2025] and EmbodiedBench [Yang et al., 2025] evaluate spatial reasoning in embodied contexts.

2.3 Manipulation

Manipulation encompasses the ability to interact with and modify objects in the environment. This domain is critical for robotic applications and requires tight integration of perception, planning, and control.

Robotic Manipulation. Transporter Networks [Zeng et al., 2021] introduced a spatial action representation for pick-and-place tasks. CLIPort [Shridhar et al., 2022] combined this with CLIP for language-conditioned manipulation. More recent work has developed general-purpose manipulation policies through large-scale imitation learning [Brohan et al., 2022, 2023, ?, Kim et al., 2024].

6D Pose Estimation. Accurate object pose estimation is fundamental for manipulation. PoseCNN [?] established a baseline approach, while recent work has developed foundation models for pose estimation [??] that generalize to novel objects without retraining.

Task and Motion Planning. Integrating high-level task planning with low-level motion planning remains a key challenge [??]. LLM-based planners [Song et al., 2023, ?, Huang et al., 2022] have shown promise in generating task plans from natural language, while approaches like SayCan [Ahn et al., 2022] ground these plans in robotic affordances.

Dexterous Manipulation. Learning dexterous manipulation skills, particularly for multi-fingered hands, has been advanced through simulation [Akkaya et al., 2019, ?] and real-world learning [?]. TidyBot [Wu et al., 2023a] demonstrated household tidying through LLM-guided manipulation.

2.4 Geospatial Analysis

Geospatial analysis involves reasoning about large-scale geographic data, from satellite imagery to urban sensor networks. This domain has seen rapid advancement through foundation models and graph neural networks.

Remote Sensing Foundation Models. Prithvi [Jakubik et al., 2024] introduced a geospatial foundation model pretrained on NASA's Harmonized Landsat Sentinel-2 data. SatMAE [Cong et al., 2022] and SatCLIP [?] developed self-supervised approaches for satellite imagery. Scale-MAE [?] addressed the multi-scale nature of remote sensing data. These models enable transfer learning across diverse geospatial tasks including land use classification [Sumbul et al., 2019, ?], change detection [Zhang et al., 2018], and building damage assessment [?].

Spatio-Temporal Graph Networks. Traffic forecasting has been revolutionized by graph neural networks that model spatial dependencies between sensors. DCRNN [Li et al., 2018] introduced diffusion convolution for traffic prediction, while STGCN [Yu et al., 2018] combined graph convolution with temporal convolution. Graph WaveNet [Wu et al., 2019] learned adaptive adjacency matrices, and AGCRN [?] introduced attention mechanisms. These approaches have been extended to broader urban computing applications [Jin et al., 2023, Atluri et al., 2018].

Urban Computing. Smart city applications leverage spatial AI for traffic management [Li et al., 2018, Yu et al., 2018], crime prediction [?], air quality monitoring, and urban planning [Zheng et al., 2014]. The integration of multiple data sources—sensors, social media, satellite imagery—enables comprehensive urban intelligence [??].

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] and subsequent analyses [Xi et al., 2023, ?, Huang et al., 2024].

3.1 Memory Systems

Memory enables agents to store and retrieve information from past experiences, supporting both short-term reasoning and long-term knowledge accumulation.

Short-Term Memory. In-context learning [??] allows agents to adapt to new tasks through examples provided in the prompt. Chain-of-thought prompting [Wei et al., 2022, ?] enables step-by-step reasoning within a single context window. Self-consistency [?] improves reasoning by sampling multiple reasoning paths.

Long-Term Memory. Retrieval-augmented generation (RAG) [Lewis et al., 2020, ?] extends agent knowledge through external retrieval. Generative Agents [Park et al., 2023] demonstrated emergent social behaviors through memory streams and reflection. MemGPT [?] introduced hierarchical memory management for extended conversations. Recent work on agentic memory [Xu et al., 2025] explores more sophisticated memory architectures.

Spatial Memory. For embodied agents, spatial memory is critical for navigation and manipulation. Cognitive mapping approaches [Gupta et al., 2019, Chaplot et al., 2020b] build metric maps of environments. Semantic mapping [Huang et al., 2023, ?] adds language-grounded understanding to spatial representations.

3.2 Planning Systems

Planning enables agents to decompose high-level goals into executable action sequences. This capability is essential for complex spatial tasks that require multi-step reasoning.

Chain-of-Thought Planning. CoT prompting [Wei et al., 2022] elicits step-by-step reasoning from language models. Zero-shot CoT [?] demonstrated that simple prompts like "Let's think step by step" can improve reasoning. Self-consistency [?] aggregates multiple reasoning paths for more robust planning.

Tree-Based Planning. Tree of Thoughts [Yao et al., 2023a] generalizes CoT by exploring multiple reasoning paths in a tree structure, enabling deliberate search and backtracking. Graph of Thoughts [?] further extends this to arbitrary graph structures. These approaches are particularly valuable for complex spatial planning tasks.

Iterative Refinement. Reflexion [Shinn et al., 2023] enables agents to learn from failures through verbal self-reflection. Self-Refine [?] iteratively improves outputs through self-feedback. These approaches are critical for robust planning in uncertain environments.

Hierarchical Planning. LLM-Planner [Song et al., 2023] decomposes high-level goals into subgoals for embodied agents. Inner Monologue [Huang et al., 2022] uses language as an interface between planning and perception. RAP [Hao et al., 2023] treats planning as reasoning with world models.

Classical Planning Integration. Recent work has explored combining LLMs with classical planners [???] to leverage the complementary strengths of neural and symbolic approaches.

3.3 Tool Use and Action

Tool use extends agent capabilities through external APIs, code execution, and physical actuators.

API and Tool Integration. Toolformer [Schick et al., 2023] trained language models to decide when and how to use tools. MRKL [Karpas et al., 2022] proposed a modular architecture combining LLMs with specialized modules. Gorilla [Patil et al., 2023] and ToolLLM [Qin et al., 2024] scaled tool use to thousands of APIs. ART [?] automates multi-step reasoning and tool use.

Code Generation. Program-aided language models [?] use code as an intermediate representation for reasoning. Code as Policies [Liang et al., 2023] generates executable robot policies as Python code. This approach enables more complex and dynamic behaviors than direct action prediction.

ReAct Architecture. ReAct [Yao et al., 2023b] interleaves reasoning traces with actions, creating a synergistic loop between thinking and acting. This architecture has become foundational for agentic systems, enabling agents to create, maintain, and adjust plans while interacting with environments.

3.4 Multi-Agent Systems

Multi-agent architectures enable collaboration and specialization among multiple AI agents.

Collaborative Frameworks. AutoGen [Wu et al., 2023b] provides a framework for building multi-agent conversations. CAMEL [Li et al., 2023a] explores role-playing for cooperative task completion. MetaGPT [Hong et al., 2023] assigns different roles (architect, engineer, etc.) to agents for software development.

Multi-Agent Coordination. Research on multi-agent reinforcement learning [**??**Yuan et al., 2023] provides foundations for coordinated behavior. Multi-agent geosimulation [?] applies these concepts to spatial domains.

4 State-of-the-Art Methods

4.1 Vision-Language-Action Models

Vision-Language-Action (VLA) models represent a paradigm shift in robotics, directly mapping visual observations and language instructions to robot actions through end-to-end learning.

Proprietary VLA Models. RT-1 [Brohan et al., 2022] demonstrated that transformer-based policies trained on large-scale robot data can generalize across tasks. RT-2 [Brohan et al., 2023] extended this by co-training on web-scale vision-language data, enabling emergent capabilities like reasoning about novel objects. PaLM-E [Driess et al., 2023], a 562B parameter model, integrates continuous sensor data directly into a language model for embodied reasoning.

Open-Source VLA Models. Octo [?] provides an open-source generalist robot policy trained on the Open X-Embodiment dataset. OpenVLA [Kim et al., 2024] offers a 7B parameter open-source alternative with strong performance. These models democratize access to VLA capabilities and enable community-driven research.

Multimodal Foundation Models. LLaVA [Liu et al., 2023a, ?] pioneered visual instruction tuning for multimodal understanding. Flamingo [Alayrac et al., 2022] introduced few-shot learning for vision-language tasks. BLIP-2 [Li et al., 2023b] efficiently bootstraps vision-language pretraining. Qwen-VL [Bai et al., 2023, ?] and InternVL [?] provide strong open-source alternatives. GPT-4V [OpenAI, 2023b,a] and Gemini [Team and Google, 2023] represent the frontier of proprietary multimodal capabilities.

4.2 Embodied AI Agents

Embodied AI agents operate in physical or simulated environments, requiring tight integration of perception, reasoning, and action.

Open-Ended Exploration. Voyager [Wang et al., 2023] demonstrated open-ended exploration in Minecraft through LLM-driven curriculum learning and skill library construction. MineDojo [Fan et al., 2022] provides a benchmark suite for open-ended embodied agents. DEPS [?] uses language descriptions to enable efficient exploration.

Grounded Language Agents. SayCan [Ahn et al., 2022] grounds language models in robotic affordances by combining LLM planning with learned value functions. Code as Policies [Liang et al., 2023] generates executable robot code from language instructions. LLM-Planner [Song et al., 2023] enables few-shot grounded planning for embodied agents.

Simulation Environments. Habitat [Savva et al., 2019, Szot et al., 2021, ?] provides high-fidelity simulation for embodied AI research. iGibson [Shen et al., 2021, ?] offers interactive environments with realistic physics. AI2-THOR [Kolve et al., 2017] enables research on interactive visual AI. Gibson [Xia et al., 2018] provides real-world scanned environments.

4.3 Graph Neural Networks for Spatial Intelligence

Graph Neural Networks (GNNs) provide powerful tools for modeling spatial relationships and dependencies.

Foundational Architectures. Graph Convolutional Networks (GCN) [Kipf and Welling, 2017] introduced spectral convolution on graphs. Graph Attention Networks (GAT) [Velickovic et al., 2018] added attention mechanisms for adaptive aggregation. GraphSAGE [Hamilton et al., 2017] enabled inductive learning on large graphs. Graph Isomorphism Networks (GIN) [Xu et al., 2019] provided theoretical analysis of GNN expressiveness.

Geometric GNNs. Geometric deep learning [Han et al., 2024, ?] extends GNNs to handle geometric data with equivariance properties. E(n) Equivariant GNNs [?] preserve Euclidean symmetries. These approaches are critical for molecular modeling, protein structure prediction, and physical simulation.

Spatio-Temporal GNNs. Traffic forecasting has driven innovation in spatio-temporal graph learning. DCRNN [Li et al., 2018] models traffic as diffusion on a graph. STGCN [Yu et al., 2018] combines graph and temporal convolutions. Graph WaveNet [Wu et al., 2019] learns adaptive graph structures. AGCRN [?] introduces node-specific patterns. These methods have been surveyed comprehensively [Jin et al., 2023, Atluri et al., 2018].

GNN + LLM Integration. Recent work explores combining GNNs with LLMs for enhanced reasoning. GraphGPT [?] aligns graph encoders with language models. LLM-GNN [??] uses LLMs to enhance graph learning. GNN-RAG [?] combines graph retrieval with language generation. This integration holds promise for spatial reasoning tasks that require both structural and semantic understanding.

4.4 World Models

World models learn predictive representations of environments, enabling planning through imagination.

Model-Based Reinforcement Learning. Dreamer [?] introduced latent imagination for model-based RL. DreamerV2 [Hafner et al., 2021] achieved human-level performance on Atari through discrete world models. DreamerV3 [Hafner et al., 2023] demonstrated mastery across diverse domains with a single algorithm. DayDreamer [?] transferred world models to real robots.

Video Prediction Models. Video prediction provides a form of world modeling through pixel-space forecasting. Genie [Bruce et al., 2024] learns controllable world models from internet videos. Sora [?] demonstrates impressive video generation capabilities. WorldDreamer [?] generates world models for autonomous driving.

World Models for Autonomous Driving. GAIA-1 [?] generates realistic driving videos conditioned on actions. UniSim [?] provides a unified simulator for real-world interaction. DriveWorld [?] learns structured world models for driving. These approaches enable scalable training of autonomous driving systems.

LLM-Based World Models. Recent work explores using LLMs as world models for planning [Hao et al., 2023, ?]. LLMs can predict state transitions and outcomes, enabling model-based planning without explicit environment models.

4.5 Autonomous Driving Systems

Autonomous driving represents a critical application domain for spatial AI, requiring integration of perception, prediction, and planning.

End-to-End Driving. UniAD [Hu et al., 2023] unifies perception, prediction, and planning in a single model. VAD [?] vectorizes scene representation for efficient planning. DriveVLM [?] integrates vision-language models for driving. EMMA [Waymo, 2024] from Waymo demonstrates end-to-end multimodal driving.

BEV Perception. Bird's-eye-view (BEV) representations have become standard for autonomous driving perception. LSS [?] introduced lift-splat-shoot for BEV generation. BEVFormer [??] uses transformers for BEV feature extraction. These representations enable unified perception across multiple cameras.

Datasets and Benchmarks. nuScenes [Caesar et al., 2020] provides a large-scale multimodal dataset. Waymo Open Dataset [?] offers high-quality sensor data. Argoverse 2 [?] includes HD maps and diverse scenarios. KITTI [Geiger et al., 2012] remains a foundational benchmark.

5 Industry Applications

The convergence of agentic AI and spatial intelligence has enabled transformative applications across industries.

5.1 Geospatial Intelligence Platforms

Palantir. Palantir Technologies [???] has pioneered the integration of AI with geospatial analysis for government and commercial applications. Their platforms enable analysis of satellite imagery, sensor data, and geographic information for defense, logistics, and urban planning applications.

ESRI. ESRI [?] provides the ArcGIS platform, which has increasingly integrated AI capabilities for geospatial analysis. Their GeoAI tools enable automated feature extraction, land use classification, and spatial pattern recognition. Recent integration of foundation models [Jakubik et al., 2024] enables more sophisticated analysis.

Google Earth and Maps. Google [?] has deployed AI extensively for mapping, navigation, and location-based services. Their systems process satellite imagery at global scale, enable real-time traffic prediction, and power location-based recommendations.

5.2 Location Intelligence

Foursquare. Foursquare [??] provides location intelligence through analysis of movement patterns, points of interest, and spatial behavior. Their platforms enable businesses to understand customer behavior, optimize site selection, and target marketing based on location.

Smart City Applications. Urban computing [Zheng et al., 2014, ?] leverages spatial AI for traffic management, public safety, resource optimization, and urban planning. Cities worldwide are deploying AI-powered systems for real-time monitoring and decision support.

5.3 Autonomous Vehicles

Waymo. Waymo [?Waymo, 2024] has deployed autonomous vehicles at scale, demonstrating the viability of spatial AI for real-world transportation. Their systems integrate perception, prediction, and planning for safe navigation in complex urban environments.

Tesla. Tesla [?] has pursued a vision-based approach to autonomous driving, leveraging large-scale data collection from their vehicle fleet. Their systems demonstrate the potential for scalable spatial AI through fleet learning.

5.4 Enterprise Spatial AI

The integration of spatial AI with enterprise data systems enables new applications in business intelligence and decision support.

Data Integration. Combining spatial data with enterprise systems like Snowflake, SAP, and Salesforce enables location-aware business analytics. This integration supports applications like sales territory optimization, supply chain planning, and customer segmentation based on geographic patterns.

Automated GIS Analysis. AI agents can automate complex GIS workflows that previously required teams of specialists. This includes automated feature extraction, change detection, and spatial pattern analysis at scale.

Real-Time Sensor Analytics. Processing millions of sensor data points in real-time enables applications like predictive maintenance, environmental monitoring, and smart infrastructure management.

6 Evaluation Benchmarks

Comprehensive evaluation is essential for measuring progress in spatial AI. We categorize existing benchmarks by their focus areas.

6.1 Navigation Benchmarks

Vision-language navigation benchmarks include R2R [Anderson et al., 2018], RxR [Ku et al., 2020], and REVERIE [Qi et al., 2020]. Object-goal navigation is evaluated through Habitat ObjectNav [?] and SOON [Zhu et al., 2021]. Continuous navigation benchmarks [?] extend discrete graph-based evaluation.

6.2 Manipulation Benchmarks

ALFWorld [Shridhar et al., 2021] provides text-based household tasks. BEHAVIOR [Srivastava et al., 2021] offers realistic household activities. RLBench [James et al., 2020] provides diverse manipulation tasks. Meta-World [Yu et al., 2020] enables multi-task evaluation.

6.3 Spatial Reasoning Benchmarks

CLEVR [Johnson et al., 2017] tests compositional visual reasoning. GQA [Hudson and Manning, 2019] evaluates real-world visual reasoning. SpatialVLM [Chen et al., 2024] specifically targets spatial reasoning. REM [Thompson et al., 2025] evaluates embodied spatial reasoning in MLLMs.

6.4 Integrated Agent Benchmarks

AgentBench [Liu et al., 2023b] provides comprehensive LLM agent evaluation. WebArena [Zhou et al., 2023] tests web-based agent capabilities. OSWorld [?] evaluates computer use agents. EmbodiedBench [Yang et al., 2025] comprehensively evaluates embodied MLLMs. SafeAgentBench [Yin et al., 2025] focuses on safe task planning.

6.5 Geospatial Benchmarks

BigEarthNet [Sumbul et al., 2019] provides multi-label land use classification. fMoW [?] tests temporal reasoning in satellite imagery. xBD [?] evaluates building damage assessment. SpaceNet [?] focuses on building and road extraction.

7 Open Challenges and Future Directions

Despite significant progress, several fundamental challenges remain for spatial AI agents.

7.1 Robust Spatial Representation

Developing representations that capture the complexity of 3D environments and generalize across different scenes remains challenging [Mildenhall et al., 2020, Kerbl et al., 2023, Dai et al., 2017, Chang et al., 2017]. Current approaches often struggle with novel viewpoints, lighting conditions, and scene compositions. Foundation models for 3D understanding [?Xu et al., 2024] represent promising directions.

7.2 Long-Horizon Planning

Creating agents that can plan over extended time horizons and decompose complex spatial tasks into manageable sub-goals is essential for real-world applications [Song et al., 2023, ?, Hao et al., 2023, ?]. Current LLM-based planners often struggle with tasks requiring many sequential steps or complex spatial reasoning.

7.3 Safe and Reliable Operation

Ensuring that agents operate safely, especially in safety-critical applications, is paramount [Yin et al., 2025, Bai et al., 2022, ?, ?, Amodei et al., 2016]. This includes robust handling of uncertainty, graceful degradation under distribution shift, and alignment with human values.

7.4 Sim-to-Real Transfer

Bridging the gap between simulation and the real world remains a key challenge for deploying embodied agents [Zhao et al., 2020, Tobin et al., 2017, ?, Savva et al., 2019, Shen et al., 2021]. Domain randomization, system identification, and real-world fine-tuning are active research areas.

7.5 Multi-Modal Integration

Effectively integrating information across modalities—vision, language, audio, touch, proprioception—is essential for robust spatial intelligence. Current approaches often struggle to leverage complementary information across modalities.

7.6 Scalable Data Collection

Training capable spatial AI agents requires large-scale, diverse data. Approaches like Open X-Embodiment [?] demonstrate the value of data sharing, but scaling data collection for embodied AI remains challenging.

8 Conclusion

This technical report has provided a comprehensive overview of the intersection of Agentic AI and Spatial Intelligence, reviewing over 500 papers spanning foundational architectures, state-of-the-art methods, industry applications, and evaluation benchmarks. We have proposed a unified taxonomy connecting agentic components (memory, planning, tool use) with spatial intelligence domains (navigation, scene understanding, manipulation, geospatial analysis).

The convergence of large language models, vision-language models, graph neural networks, and world models is enabling a new generation of spatially-aware autonomous agents. Industry applications from Palantir, ESRI, Foursquare, Google, Waymo, and others demonstrate the transformative potential of these technologies.

Key challenges remain in robust spatial representation, long-horizon planning, safe operation, and sim-to-real transfer. Addressing these challenges will require continued collaboration across the AI, robotics, and geospatial communities.

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

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.

N. Amin and D. Kiela. Embodied language learning: Opportunities, challenges, and future directions. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

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.

Anthropic. Claude 3 model card. *Anthropic Technical Report*, 2024.

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.

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

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. 2023.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

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.

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

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

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

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

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.

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.

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.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. 2012.

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.

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

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

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

Jiaqi Han, Jiacheng Cen, Liming Wu, Zongzhao Li, Xiangzhe Kong, Rui Jiao, Ziyang Yu, Tingyang Xu, Fandi Wu, Zihe Wang, et al. A survey of geometric graph neural networks: Data structures, models and applications. *arXiv preprint arXiv:2403.00485*, 2024.

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*, 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.

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

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.

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.

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

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.

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.

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.

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

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*, 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*, 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.

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.

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

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

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.

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.

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

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. 2020.

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

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

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.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. 2022.

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.

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

Gemini Team and Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

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.

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.

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

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.

Waymo. Waymo safety report: Building the world's most experienced driver. Technical report, Waymo LLC, 2024.

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.

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.

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.

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.

Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. *arXiv preprint arXiv:1808.10654*, 2018.

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

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2019.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents, 2025.

Zhiyuan Xu et al. A survey on robotics with foundation models: Toward embodied ai. *arXiv preprint arXiv:2402.02385*, 2024.

An Yan, Xin Eric Wang, Jiangtao Feng, Lei Li, and William Yang Wang. Cross-lingual vision-language navigation, 2020.

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.

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. 2020.

Lei Yuan, Ziqian Zhang, Lihe Li, Cong Guan, and Yang Yu. A survey of progress on cooperative multi-agent reinforcement learning in open environment, 2023.

Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, et al. Transporter networks: Rearranging the visual world for robotic manipulation. 2021.

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.

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.

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*, 2023.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenarios for object navigation with natural language instructions. 2021.