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

Gloria Felicia
AtlasPro AI
gloria.felicia@atlaspro.ai

Nolan Bryant
AtlasPro AI
nolan.bryant@atlaspro.ai

Handi Putra
AtlasPro AI
handi.putra@atlaspro.ai

Ayaan Gazali
AtlasPro AI
ayaan.gazali@atlaspro.ai

Eliel Lobo
AtlasPro AI
eliel.lobo@atlaspro.ai

Esteban Rojas
AtlasPro AI
esteban.rojas@atlaspro.ai

Abstract

The convergence of Agentic Artificial Intelligence and Spatial Intelligence marks a pivotal frontier in the pursuit of creating machines that can autonomously operate in the physical world. While agentic systems demonstrate increasingly sophisticated capabilities in planning and tool use, their ability to perceive, reason about, and interact with complex spatial environments remains a significant bottleneck. This survey addresses a critical gap in the existing literature by providing a unified taxonomy that systematically connects the architectural components of agentic AI with the functional requirements of spatial intelligence. We review over 1,000 papers spanning foundational agentic architectures [????], embodied AI systems [????], vision-language-action models [????], graph neural networks for spatial reasoning [????], world models [????], and geospatial foundation models [????]. 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 [???]. The foundational work on large language models [??????] has enabled a new generation of AI agents that can reason, plan, and execute complex tasks through natural language interfaces [???].

Concurrently, the ability for these agents to perceive, comprehend, and act within the physical world, a capability we define as **Spatial Intelligence**, has become a primary bottleneck and a critical area of research [???]. The convergence of these two domains is essential for developing AI systems that can effectively and safely navigate real-world complexities, from autonomous vehicles [?????] and robotic assistants [???] to large-scale urban planning [???] and disaster response systems [???].

Despite rapid progress in both agentic systems and spatial reasoning, the research landscape remains fragmented. Numerous surveys have independently covered topics such as Large Language Model agents [??????], embodied AI [????], multimodal large language models [?????], graph neural networks [??????],

spatio-temporal prediction [?????], world models [?????], and geospatial analysis [?????]. However, a comprehensive synthesis that bridges the architectural components of agentic AI with the functional requirements of spatial intelligence is notably absent. This disconnect hinders a holistic understanding of the challenges and opportunities at the intersection of these fields, slowing progress toward building truly world-aware autonomous agents.

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

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 1,000 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) [?????], which requires agents to follow natural language instructions in realistic environments. The Room-to-Room (R2R) benchmark [?] established a foundational evaluation framework, while subsequent work has extended to continuous environments [?], outdoor settings [?], and cross-lingual scenarios [?].

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

Semantic mapping approaches [???] build spatial representations that enable more efficient navigation. VLMaps [?] 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 [??] 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) [???] have revolutionized novel view synthesis by representing scenes as continuous volumetric functions. More recently, 3D Gaussian Splatting [?] 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 [?] 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 [?], Matterport3D [?], and S3DIS [?].

Scene Graphs. Scene graph generation [???] 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 [??] for more complete environmental understanding.

Spatial Reasoning Benchmarks. CLEVR [?] introduced compositional visual reasoning, while GQA [?] extended this to real-world images. NLVR2 [?] focuses on grounded language understanding, and SpatialVLM [?] specifically targets spatial reasoning in vision-language models. Recent benchmarks like REM [?] and EmbodiedBench [?] 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 [?] introduced a spatial action representation for pick-and-place tasks. CLIPort [?] combined this with CLIP for language-conditioned manipulation. More recent work has developed general-purpose manipulation policies through large-scale imitation learning [????].

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 [???] have shown promise in generating task plans from natural language, while approaches like SayCan [?] ground these plans in robotic affordances.

Dexterous Manipulation. Learning dexterous manipulation skills, particularly for multi-fingered hands, has been advanced through simulation [??] and real-world learning [?]. TidyBot [?] 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 [?] introduced a geospatial foundation model pre-trained on NASA's Harmonized Landsat Sentinel-2 data. SatMAE [?] 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 [??], change detection [?], and building damage assessment [?].

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

Urban Computing. Smart city applications leverage spatial AI for traffic management [??], crime prediction [?], air quality monitoring, and urban planning [?]. 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 ? and subsequent analyses [???].

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 [??] 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) [??] extends agent knowledge through external retrieval. Generative Agents [?] demonstrated emergent social behaviors through memory streams and reflection. MemGPT [?] introduced hierarchical memory management for extended conversations. Recent work on agentic memory [?] explores more sophisticated memory architectures.

Spatial Memory. For embodied agents, spatial memory is critical for navigation and manipulation. Cognitive mapping approaches [??] build metric maps of environments. Semantic mapping [??] 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 [?] 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 [?] 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 [?] 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 [?] decomposes high-level goals into subgoals for embodied agents. Inner Monologue [?] uses language as an interface between planning and perception. RAP [?] 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 [?] trained language models to decide when and how to use tools. MRKL [?] proposed a modular architecture combining LLMs with specialized modules. Gorilla [?] and ToolLLM [?] 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 [?] generates executable robot policies as Python code. This approach enables more complex and dynamic behaviors than direct action prediction.

ReAct Architecture. ReAct [?] 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 [?] provides a framework for building multi-agent conversations. CAMEL [?] explores role-playing for cooperative task completion. MetaGPT [?] assigns different roles (architect, engineer, etc.) to agents for software development.

Multi-Agent Coordination. Research on multi-agent reinforcement learning [???] 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 [?] demonstrated that transformer-based policies trained on large-scale robot data can generalize across tasks. RT-2 [?] extended this by co-training on web-scale vision-language data, enabling emergent capabilities like reasoning about novel objects. PaLM-E [?], 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 [?] 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 [??] pioneered visual instruction tuning for multimodal understanding. Flamingo [?] introduced few-shot learning for vision-language tasks. BLIP-2 [?] efficiently bootstraps vision-language pretraining. Qwen-VL [??] and InternVL [?] provide strong open-source alternatives. GPT-4V [??] and Gemini [?] 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 [?] demonstrated open-ended exploration in Minecraft through LLM-driven curriculum learning and skill library construction. MineDojo [?] provides a benchmark suite for open-ended embodied agents. DEPS [?] uses language descriptions to enable efficient exploration.

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

Simulation Environments. Habitat [???] provides high-fidelity simulation for embodied AI research. iGibson [??] offers interactive environments with realistic physics. AI2-THOR [?] enables research on interactive visual AI. Gibson [?] 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) [?] introduced spectral convolution on graphs. Graph Attention Networks (GAT) [?] added attention mechanisms for adaptive aggregation. GraphSAGE [?] enabled inductive learning on large graphs. Graph Isomorphism Networks (GIN) [?] provided theoretical analysis of GNN expressiveness.

Geometric GNNs. Geometric deep learning [??] 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 [?] models traffic as diffusion on a graph. STGCN [?] combines graph and temporal convolutions. Graph WaveNet [?] learns adaptive graph structures. AGCRN [?] introduces node-specific patterns. These methods have been surveyed comprehensively [??].

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 [?] achieved human-level performance on Atari through discrete world models. DreamerV3 [?] 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 [?] 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 [??]. 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 [?] unifies perception, prediction, and planning in a single model. VAD [?] vectorizes scene representation for efficient planning. DriveVLM [?] integrates vision-language models for driving. EMMA [?] 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 [?] provides a large-scale multimodal dataset. Waymo Open Dataset [?] offers high-quality sensor data. Argoverse 2 [?] includes HD maps and diverse scenarios. KITTI [?] 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 [?] 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 [??] 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 [??] 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 [?], RxR [?], and REVERIE [?]. Object-goal navigation is evaluated through Habitat ObjectNav [?] and SOON [?]. Continuous navigation benchmarks [?] extend discrete graph-based evaluation.

6.2 Manipulation Benchmarks

ALFWorld [?] provides text-based household tasks. BEHAVIOR [?] offers realistic household activities. RLBench [?] provides diverse manipulation tasks. Meta-World [?] enables multi-task evaluation.

6.3 Spatial Reasoning Benchmarks

CLEVR [?] tests compositional visual reasoning. GQA [?] evaluates real-world visual reasoning. SpatialVLM [?] specifically targets spatial reasoning. REM [?] evaluates embodied spatial reasoning in MLLMs.

6.4 Integrated Agent Benchmarks

AgentBench [?] provides comprehensive LLM agent evaluation. WebArena [?] tests web-based agent capabilities. OSWorld [?] evaluates computer use agents. EmbodiedBench [?] comprehensively evaluates embodied MLLMs. SafeAgentBench [?] focuses on safe task planning.

6.5 Geospatial Benchmarks

BigEarthNet [?] 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 [????]. Current approaches often struggle with novel viewpoints, lighting conditions, and scene compositions. Foundation models for 3D understanding [??] 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 [????]. 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 [?????]. 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 [?????]. 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 survey has provided a comprehensive overview of the intersection of Agentic AI and Spatial Intelligence, reviewing over 1,000 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.