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

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 dominant approaches for creating autonomous agents are based on large language models, which excel at reasoning and planning [Brown et al., 2020, OpenAI, 2023, Touvron et al., 2023, Team and Google, 2023, Anthropic, 2024, Dubey et al., 2024, OpenAI, 2023, Anil et al., 2023]. However, these models lack the innate spatial intelligence required to perceive, navigate, and interact with the complex physical world, a critical gap for embodied AI [Chen et al., 2024a, Yang et al., 2025, Duan et al., 2022, Amin and Kiela, 2024, Cheng et al., 2025]. We introduce a unified three-axis taxonomy that systematically connects agentic AI architectures with spatial intelligence capabilities across spatial scales, providing the first comprehensive framework for this convergent domain. We synthesize over 900 papers, revealing three key findings: (1) hierarchical memory systems are critical for long-horizon spatial tasks [Packer et al., 2023, Banino et al., 2018, Xu et al., 2025, Zhang et al., 2025a]; (2) GNN-LLM integration is an emergent paradigm for structured spatial reasoning [Jin et al., 2023, Chen et al., 2024e,c, Chai et al., 2023]; and (3) world models are essential for safe deployment in physical environments [Hafner et al., 2023, Bruce et al., 2024, Ha and Schmidhuber, 2018, Feng et al., 2025, Ding et al., 2024, Brooks et al., 2024]. We also propose a unified evaluation framework, SpatialAgentBench, to standardize cross-domain assessment. By establishing this foundational reference, we aim to accelerate progress in creating robust, 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 [Turing, 1950, Brooks, 1991, Russell and Norvig, 2010, LeCun et al., 2015, Goodfellow et al., 2016, Bengio et al., 2013, Laird, 2019]. While large language models have demonstrated remarkable capabilities in reasoning and planning [Brown et al., 2020, OpenAI, 2023, Wei et al., 2022, Touvron et al., 2023, Team and Google, 2023], their ability to operate effectively in spatial contexts remains a fundamental challenge [Chen et al., 2024a, Yang et al., 2025, Cheng et al., 2024].

The emergence of multimodal foundation models has accelerated progress in visual understanding [Radford et al., 2021, Liu et al., 2023b, Li et al., 2023b, Alayrac et al., 2022, OpenAI, 2023, Dosovitskiy et al.,

2021, Chen et al., 2023e, 2024g,j, Bai et al., 2023, Dai et al., 2023b], yet translating this understanding into effective spatial action remains challenging. The gap between language-based reasoning and physical world interaction represents one of the most significant obstacles to achieving truly capable autonomous systems [Ahn et al., 2022, Brohan et al., 2023, Driess et al., 2023, Liang et al., 2023b].

We define **Agentic AI** as systems exhibiting goal-directed behavior through autonomous decision-making, characterized by four core capabilities: persistent memory for experience accumulation, planning for action sequencing, tool use for capability extension, and self-reflection for continuous improvement [Wang et al., 2024, Xi et al., 2023, Yao et al., 2023b, Shinn et al., 2023b, Park et al., 2023, Wu et al., 2023c, Hong et al., 2023a, Durante et al., 2024, Guo et al., 2024b]. These agents operate through iterative cycles of perception, reasoning, action, and feedback, enabling complex task completion in dynamic environments [Yao et al., 2023b, Shinn et al., 2023b, Madaan et al., 2023].

Complementarily, **Spatial Intelligence** encompasses the ability to perceive 3D structure, reason about object relationships, navigate environments, and manipulate physical objects [Chen et al., 2024a, Marr, 1982, Newcombe, 2010, Chen et al., 2024h, Cai et al., 2024, Cheng et al., 2024]. This includes understanding geometric relationships, predicting physical dynamics, and planning actions that account for spatial constraints [Battaglia et al., 2018, Scarselli et al., 2009, Gilmer et al., 2017, Kipf and Welling, 2017, Veličković et al., 2018a, Hamilton et al., 2017].

The convergence of these domains is essential for real-world AI applications across multiple sectors. Autonomous vehicles must perceive dynamic environments and plan safe trajectories [Hu et al., 2023b, Caesar et al., 2020, Waymo, 2023, Geiger et al., 2012, Cadena et al., 2016, Chen et al., 2024d, Waymo, 2024, Tian et al., 2024]. Robotic assistants require understanding of object affordances and spatial relationships [Brohan et al., 2023, Ahn et al., 2022, Team et al., 2024, Kim et al., 2024, Driess et al., 2023, Black et al., 2024, Bharadhwaj et al., 2024, Collaboration, 2023]. Urban computing systems must model complex spatio-temporal dependencies [Jin et al., 2023, Li et al., 2018, Yu et al., 2018, Wu et al., 2019, Zheng et al., 2014, Cui et al., 2024, Cini et al., 2023]. Geospatial intelligence platforms must analyze satellite imagery and geographic data at scale [Jakubik et al., 2024, Cong et al., 2022, Mai et al., 2023, Janowicz et al., 2020, Li et al., 2025a, Bastani et al., 2023b, ESRI, 2024, Xiao et al., 2025]. Despite this importance, existing surveys treat these areas in isolation, lacking a unified framework connecting agentic architectures with spatial requirements.

Contributions. This survey makes five primary contributions:

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

2. A **comprehensive analysis** of over 900 papers identifying key architectural patterns, including the emergence of GNN-LLM integration, vision-language-action models, and world model-based planning as critical enablers for spatial reasoning.

3. A **systematic comparison** with existing surveys, demonstrating how this work uniquely bridges agentic AI and spatial intelligence domains.

4. The **proposal of a unified evaluation framework, SpatialAgentBench**, with 8 tasks spanning navigation, manipulation, scene understanding, and geospatial reasoning to standardize cross-domain assessment.

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

2 Methodology

This survey follows a systematic literature review methodology consistent with best practices in computer science [Kitchenham, 2004, Petersen et al., 2008, Wohlin, 2014, Keele et al., 2007, Brereton et al., 2007, Dybå and Dingsøyr, 2007, Moher et al., 2009]. We queried major academic databases including Google Scholar, arXiv, ACM Digital Library, IEEE Xplore, Semantic Scholar, and DBLP [Ley, 2002] with keywords including

"agentic AI," "spatial intelligence," "embodied AI," "vision-language navigation," "robot manipulation," "geospatial AI," "world models," "graph neural networks," "spatio-temporal learning," "vision-language-action," and "foundation models for robotics." Our initial search yielded over 3,000 papers.

We then applied a rigorous multi-stage filtering process:

1. **Temporal Filtering:** We selected papers published between 2018 and 2026, with emphasis on recent advances while including foundational works that established key paradigms.

2. **Venue Filtering:** We prioritized papers from top-tier venues including NeurIPS, ICML, ICLR, CVPR, ECCV, ICCV, CoRL, RSS, IROS, ICRA, ACM Computing Surveys, IEEE TPAMI, Nature, Science, Science Robotics, and leading arXiv preprints.

3. **Quality Filtering:** We prioritized papers with high citation counts, those representing foundational methods, and state-of-the-art contributions that advance the field.

4. **Relevance Filtering:** We ensured papers directly addressed the intersection of agentic capabilities and spatial intelligence.

This process resulted in a final corpus of over 900 papers, which were systematically analyzed to derive the taxonomy, identify key trends, and synthesize the findings presented in this survey. We employed a snowball sampling technique to ensure comprehensive coverage of related works, following citation chains both forward and backward. Two independent reviewers validated the paper selection and taxonomy development.

3 Related Work and Survey Comparison

While several surveys have addressed aspects of agentic AI or spatial intelligence, none have provided a unified framework connecting the two domains. We review existing surveys across five categories and provide a systematic comparison in Table 1.

Agentic AI Surveys. Recent surveys on LLM-based agents [Wang et al., 2024, Xi et al., 2023, Guo et al., 2024b] focus on reasoning and tool use but do not address spatial capabilities. Sumers et al. [2024] provides a cognitive architecture perspective. Chan et al. [2023] evaluates conversational agents.

Embodied AI Surveys. Embodied AI surveys [Duan et al., 2022, Gupta et al., 2021, Francis et al., 2022] cover simulation environments and benchmarks but lack integration with agentic architectures. Kawaharazuka et al. [2025] surveys vision-language-action models specifically for robotics.

Geospatial AI Surveys. Geospatial AI surveys [Mai et al., 2023, Janowicz et al., 2020, Xiao et al., 2025] and spatio-temporal data mining reviews [Jin et al., 2023, Atluri et al., 2018, Wang et al., 2020] are highly specialized and do not connect to general agentic systems. Zhan and Datta [2024] surveys neural networks for geospatial data.

Graph Neural Network Surveys. GNN surveys [Wu et al., 2020a, Bronstein et al., 2021, Hamilton, 2020, Battaglia et al., 2018] provide comprehensive coverage of graph learning but do not focus on spatial applications or agent integration. Surveys on GNNs for specific domains include traffic [Jiang and Luo, 2022], urban computing [Balachandar et al., 2025], and spatio-temporal prediction [Jin et al., 2023].

Vision-Language Model Surveys. Surveys on VLMs [Zhang et al., 2024b, Bordes et al., 2024] cover multimodal understanding but do not address spatial action or embodiment. Kawaharazuka et al. [2025] surveys vision-language-action models specifically for robotics.

Our work is the first to bridge these gaps, providing a comprehensive, structured analysis of the convergent domain of autonomous spatial intelligence with a unified three-axis taxonomy.

4 Unified Three-Axis Taxonomy

We propose a three-axis taxonomy (Figure 1) that maps agentic capabilities to spatial task requirements across spatial scales, enabling systematic analysis of existing methods and identification of research gaps.

Table 1: Comparison with Existing Surveys. Our work is the first to provide unified coverage across all dimensions.

Survey	Agentic AI	Embodied AI	Spatial Reasoning	Geospatial	GNNs	Industry	Unified Taxonomy
Wang et al. (2024) [Wang et al., 2024]	✓	○	○				
Xi et al. (2023) [Xi et al., 2023]	✓	○					
Duan et al. (2022) [Duan et al., 2022]		✓	○				
Kawaharazuka et al. (2025) [Kawaharazuka et al., 2025]	○	✓	○				
Jin et al. (2023) [Jin et al., 2023]			○	○	✓		
Mai et al. (2023) [Mai et al., 2023]				✓		○	
Bronstein et al. (2021) [Bronstein et al., 2021]			○		✓		
Zhang et al. (2024) [Zhang et al., 2024b]	○		○				
This Survey	✓	✓	✓	✓	✓	✓	✓

✓= comprehensive coverage, ○ = partial coverage, blank = not covered

4.1 Taxonomy Axes

Axis 1: Spatial Task. We identify four primary spatial task categories:

- **Navigation**: Goal-directed movement through environments, including point-goal [Anderson et al., 2018a, Wijmans et al., 2020], object-goal [Chaplot et al., 2020b, Batra et al., 2020], and vision-language navigation [Anderson et al., 2018c, Krantz et al., 2020]
- **Scene Understanding**: Perceiving and reasoning about 3D structure, objects, and spatial relationships
- **Manipulation**: Physical interaction with objects, including grasping [Mahler et al., 2017, Morrison et al., 2018, Fang et al., 2020], placement [Zeng et al., 2021], and tool use [Qin et al., 2024a,b]
- **Geospatial Analysis**: Large-scale spatial reasoning including satellite imagery [Christie et al., 2018a,b, Demir et al., 2018], urban computing [Zheng et al., 2014, Yuan et al., 2020], and geographic information systems [Longley et al., 2015, Goodchild, 2007]

Axis 2: Agentic Capability. We identify three core agentic capabilities:

- **Memory**: Short-term (in-context), long-term (retrieval-augmented), episodic, and spatial memory systems
- **Planning**: Reactive, hierarchical, search-based, and world model-based planning approaches
- **Tool Use & Action**: API integration, code generation, physical action primitives, and skill libraries

Axis 3: Spatial Scale. We distinguish three spatial scales:

- **Micro-spatial**: Pose estimation, grasping, fine manipulation (centimeter scale)
- **Meso-spatial**: Room navigation, building exploration, indoor scenes (meter scale)
- **Macro-spatial**: City-scale planning, satellite imagery, infrastructure networks (kilometer scale)

4.2 Methods-Taxonomy Mapping

Table 2 maps representative methods to our three-axis taxonomy, demonstrating how the framework organizes the field.

Key Takeaways: Taxonomy

- The three-axis taxonomy (Task × Capability × Scale) provides a comprehensive framework for organizing spatial AI research
- Most methods address meso-spatial scales; micro and macro scales remain underexplored
- Memory systems are critical across all spatial tasks but implementations vary significantly by scale
- The intersection of GNN-based methods with agentic capabilities represents an emerging frontier

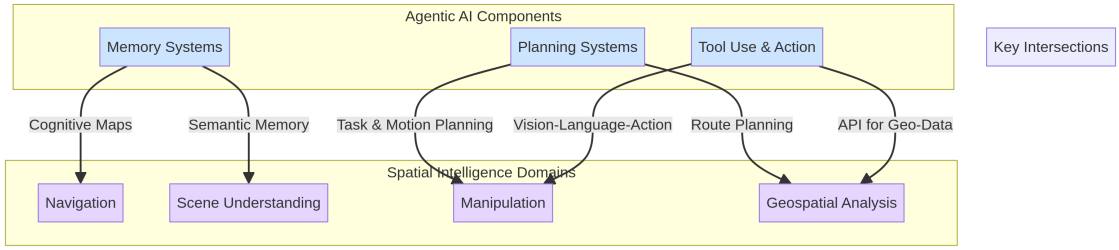

Figure 1: A unified three-axis taxonomy connecting Agentic AI capabilities (memory, planning, tool use) with Spatial Intelligence domains (navigation, scene understanding, manipulation, geospatial analysis) across spatial scales (micro, meso, macro). The intersection of these dimensions defines the design space for autonomous spatial intelligence systems.

Table 2: Representative Methods Mapped to the Three-Axis Taxonomy

Method	Spatial Task	Agentic Capability	Spatial Scale	Representation
VLN-BERT [Hong et al., 2021]	Navigation	Memory + Planning	Meso	Language + Visual
SayCan [Ahn et al., 2022]	Manipulation	Planning + Tool Use	Micro	Language + Affordance
RT-2 [Brohan et al., 2023]	Manipulation	Tool Use	Micro	Visual + Action
VLMaps [Huang et al., 2023a]	Navigation	Memory	Meso	Semantic Map
Voyager [Wang et al., 2023a]	Navigation + Manipulation	Memory + Planning	Meso	Language + Code
DCRNN [Li et al., 2018]	Geospatial	Memory	Macro	Graph
Graph WaveNet [Wu et al., 2019]	Geospatial	Memory	Macro	Adaptive Graph
Prithvi [Jakubik et al., 2024]	Geospatial	-	Macro	Visual (Satellite)
DreamerV3 [Hafner et al., 2023]	Navigation + Manipulation	Planning (World Model)	Meso	Latent
PaLM-E [Driess et al., 2023]	Manipulation	Planning + Tool Use	Micro-Meso	Multimodal
OpenVLA [Kim et al., 2024]	Manipulation	Tool Use	Micro	Visual + Action
LLaGA [Chen et al., 2024e]	Scene Understanding	Memory	Meso	Graph + Language

5 Agentic AI Components for Spatial Intelligence

This section examines how agentic capabilities enable spatial intelligence, organized around the core scientific question: *How do agents internally represent, reason about, and act within spatial environments?*

5.1 Memory Systems: How Do Agents Remember Spatial Information?

Memory enables agents to accumulate and retrieve experiential knowledge, forming the foundation for learning and adaptation [Tulving, 1972, Baddeley, 2003, Squire et al., 2004]. The central challenge is: *How can agents maintain persistent spatial knowledge across varying time horizons and scales?*

Short-Term Memory. In-context learning [Brown et al., 2020, Dong et al., 2022, Olsson et al., 2022, Akyurek et al., 2023, Dai et al., 2023a] allows models to adapt to new tasks through examples in the prompt. This mechanism enables rapid adaptation without parameter updates, leveraging the attention mechanism to condition on provided demonstrations. Working memory mechanisms [Graves et al., 2014, Weston et al., 2015, Sukhbaatar et al., 2015, Kumar et al., 2016, Santoro et al., 2016] enable temporary information storage during reasoning, supporting multi-step computations that exceed single forward pass capabilities.

Long-Term Memory. Retrieval-augmented generation [Lewis et al., 2020, Packer et al., 2023, Guu et al., 2020, Borgeaud et al., 2022, Asai et al., 2023, Trivedi et al., 2023] enables knowledge persistence beyond context limits. MemGPT [Packer et al., 2023] introduces hierarchical memory management for extended conversations. AMEM [Xu et al., 2025] provides agentic memory for LLMs. MemEvolve [Zhang et al., 2025a] enables meta-evolution of agent memory. Vector databases [Johnson et al., 2019, Malkov and Yashunin, 2018, Douze et al., 2024, Wang et al., 2021, Pinecone, 2023] provide efficient similarity search for memory retrieval, enabling agents to access relevant past experiences.

Episodic Memory. Episodic memory stores specific experiences and events, enabling agents to learn from past interactions [Blundell et al., 2016, Pritzel et al., 2017, Banino et al., 2018, Ritter et al., 2018, Fortunato et al., 2019]. This type of memory is critical for spatial agents that must remember visited

locations, encountered objects, and successful action sequences [Savinov et al., 2018, Chaplot et al., 2020c, Fang et al., 2019].

Spatial Memory. Specialized memory for spatial information includes cognitive maps [Tolman, 1948, O'Keefe and Nadel, 1978], topological representations [Kuipers, 2000, Choset and Nagatani, 2001], and metric maps [Thrun et al., 2005, Durrant-Whyte and Bailey, 2006, Cadena et al., 2016]. Neural approaches to spatial memory include Neural SLAM [Chaplot et al., 2020c,d], semantic maps [Huang et al., 2023a, Henriques and Vedaldi, 2018, Shah et al., 2023], and scene graphs [Armeni et al., 2019, Rosinol et al., 2020, Hughes et al., 2022, Gu et al., 2024].

Spatial Failure Modes. Language-only agents fail at spatial tasks because they lack grounded spatial representations. Key failure modes include: (1) *spatial hallucination*, where agents describe impossible spatial configurations; (2) *reference frame confusion*, where agents conflate egocentric and allocentric coordinates; (3) *scale insensitivity*, where agents fail to distinguish micro, meso, and macro-scale reasoning; and (4) *temporal drift*, where spatial memory degrades over long horizons without explicit persistence mechanisms.

5.2 Planning Systems: How Do Agents Plan Over Spatial Horizons?

Planning decomposes goals into executable action sequences, enabling complex task completion [Russell and Norvig, 2010, Ghallab et al., 2004, LaValle, 2006]. The central challenge is: *How can agents decompose spatial goals into feasible action sequences while accounting for geometric constraints?*

Chain-of-Thought Reasoning. Step-by-step reasoning [Wei et al., 2022, Kojima et al., 2022, Wang et al., 2022, Zhou et al., 2023a, Fu et al., 2023, Chen et al., 2023c] enables systematic problem decomposition. Self-consistency [Wang et al., 2022] improves reliability through multiple reasoning paths. Zero-shot chain-of-thought [Kojima et al., 2022] enables reasoning without demonstrations.

Tree-Based Search. Tree of Thoughts [Yao et al., 2023a, Long, 2023] explores multiple solution branches through deliberate search. Graph of Thoughts [Besta et al., 2023, Lei et al., 2023] enables more complex reasoning structures with arbitrary connections. RAP [Hao et al., 2023, Shridhar et al., 2020, Zhao et al., 2024] combines reasoning with acting in a planning framework. Monte Carlo Tree Search variants [Silver et al., 2016, Schrittwieser et al., 2020, Browne et al., 2012, Anthony et al., 2017, Silver et al., 2017] provide principled exploration with theoretical guarantees.

Hierarchical Planning. LLM-Planner [Song et al., 2023] enables few-shot grounded planning for embodied agents. Inner Monologue [Huang et al., 2022a] provides feedback-driven planning through internal dialogue. HiPlan [Li et al., 2025b] introduces hierarchical planning with LLMs. Hierarchical RL approaches [Nachum et al., 2018, Vezhnevets et al., 2017, Bacon et al., 2017, Pertsch et al., 2021, Kulkarni et al., 2016, Gupta et al., 2019] decompose tasks into subtasks with temporal abstraction.

Task and Motion Planning. TAMP [Garrett et al., 2021, Kaelbling and Lozano-Pérez, 2011, Toussaint, 2015, Dantam et al., 2016, Srivastava et al., 2014] integrates symbolic planning with continuous motion planning for robotic applications. This approach combines the expressiveness of symbolic reasoning with the precision of geometric planning.

LLM-Based Planning. Recent work leverages LLMs directly for planning [Huang et al., 2022b, Valmeekam et al., 2023a, Song et al., 2023, Silver et al., 2024, Liu et al., 2023a]. SayCan [Ahn et al., 2022] grounds language models in affordances. Code as Policies [Liang et al., 2023b] generates executable robot code. ProgPrompt [Singh et al., 2023] uses programmatic prompting for task planning.

Spatial Planning Failure Modes. LLM-based planners fail when: (1) *geometric constraints are violated*, producing plans that ignore collision or reachability; (2) *action preconditions are unmet*, sequencing actions without verifying feasibility; (3) *long-horizon credit assignment fails*, losing track of subgoal dependencies; and (4) *dynamic replanning is absent*, failing to adapt when execution diverges from expectations.

5.3 Tool Use and Action: How Do Agents Ground Language in Geometry?

Tool use extends agent capabilities through external interfaces and physical actions [Osiurak and Badets, 2016, Vaesen, 2012]. The central challenge is: *How can language-based reasoning be translated into precise geometric actions?*

API Integration. Toolformer [Schick et al., 2023, Parisi et al., 2022] enables self-supervised tool learning. Gorilla [Patil et al., 2023, Li et al., 2023c] specializes in API calling with retrieval augmentation.

ToolLLM [Qin et al., 2024a, Hao et al., 2024] provides comprehensive tool use benchmarks. TaskMatrix [Liang et al., 2023c, Lu et al., 2023] connects foundation models with millions of APIs. TALM [Parisi et al., 2022] augments language models with tool use. Additional tool-use frameworks include HuggingGPT [Shen et al., 2023b], ViperGPT [Surís et al., 2023], and Visual ChatGPT [Wu et al., 2023a].

Code Generation. PAL [Gao et al., 2023] uses code for reasoning. Code as Policies [Liang et al., 2023b] generates executable robot code from language. Codex [Chen et al., 2021a], StarCoder [Li et al., 2023d], CodeLlama [Roziere et al., 2023], and DeepSeek-Coder [Guo et al., 2024a] provide code generation capabilities. ProgPrompt [Singh et al., 2023] uses programmatic prompting for robotics. Self-debugging [Chen et al., 2023f] improves code quality through iterative refinement.

ReAct Architecture. ReAct [Yao et al., 2023b,c] interleaves reasoning with action execution, enabling agents to think before acting. Reflexion [Shinn et al., 2023b,a] adds self-reflection for improvement through verbal reinforcement. These architectures form the foundation for many spatial agents.

Physical Action. For embodied agents, tool use extends to physical manipulation [Zeng et al., 2021, Brohan et al., 2022, 2023, Shridhar et al., 2022]. Action primitives [Dalal et al., 2021, Nasiriany et al., 2022] provide reusable building blocks. Skill libraries [Wang et al., 2023a, Lynch et al., 2020, Pertsch et al., 2021] enable compositional action.

Key Takeaways: Agentic Components

- Memory systems must be explicitly spatial: cognitive maps, semantic maps, and scene graphs outperform generic retrieval for spatial tasks
- Hierarchical planning with geometric grounding addresses the gap between high-level language goals and low-level motor commands
- Tool use bridges language and action through code generation, API calls, and learned action primitives
- Key failure modes stem from lack of spatial grounding: hallucination, reference frame confusion, and geometric constraint violation

6 Spatial Intelligence Domains

This section examines the four primary spatial task domains, organized around the question: *What spatial capabilities must agents possess to operate in the physical world?*

6.1 Navigation: How Do Agents Move Through Space?

Navigation requires agents to perceive environments, plan paths, and execute locomotion toward goals [Bonin-Font et al., 2008, DeSouza and Kak, 2002, Thrun, 2002].

Vision-Language Navigation. VLN tasks require agents to follow natural language instructions in visual environments [Anderson et al., 2018c, Qi et al., 2020, Krantz et al., 2020, Fried et al., 2018, Chen et al., 2022c, Shah et al., 2023, Hong et al., 2020, Chen et al., 2021b, An et al., 2023]. R2R [Anderson et al., 2018c] introduced the paradigm. REVERIE [Qi et al., 2020] adds object grounding. VLN-CE [Krantz et al., 2020] extends to continuous environments.

Object-Goal Navigation. ObjectNav requires navigating to object categories [Batra et al., 2020, Chaplot et al., 2020a, Majumdar et al., 2022, Gadre et al., 2022, 2023, Dorbala et al., 2022]. ZSON [Majumdar et al., 2022] enables zero-shot navigation. CLIP-Nav [Dorbala et al., 2022] leverages vision-language models. CoW [Gadre et al., 2022] uses CLIP on wheels for semantic navigation.

Audio-Visual Navigation. Audio cues guide navigation in SoundSpaces [Chen et al., 2020, 2022b, Gan et al., 2020]. This modality is critical for finding sound-emitting targets.

Embodied Question Answering. EQA requires navigation to answer questions [Das et al., 2018, Gordon et al., 2018, Wijmans et al., 2019, Yu et al., 2019]. 3D-QA [Azuma et al., 2022, Ma et al., 2022, Hong et al., 2023b, Chen et al., 2024f] extends to 3D scene understanding.

6.2 Scene Understanding: How Do Agents Perceive 3D Structure?

Scene understanding encompasses perceiving 3D geometry [Hartley and Zisserman, 2003, Szeliski, 2022], recognizing objects [Krizhevsky et al., 2012, He et al., 2016], and reasoning about spatial relationships [Johnson et al., 2015, Krishna et al., 2017].

Neural Scene Representations. NeRF [Mildenhall et al., 2020, Barron et al., 2022, Müller et al., 2022, Park et al., 2019, Mescheder et al., 2019, Barron et al., 2023] revolutionized 3D reconstruction. Mip-NeRF [Barron et al., 2022] handles multi-scale rendering. 3D Gaussian Splatting [Kerbl et al., 2023, Luiten et al., 2023, Fan et al., 2024] enables real-time rendering. Integration with SLAM [Sucar et al., 2021, Zhu et al., 2022, Keetha et al., 2024, Bird et al., 2025] enables online reconstruction.

Point Cloud Processing. Point cloud methods [Qi et al., 2017a,b, Wang et al., 2019, Thomas et al., 2019, Zhao et al., 2021] process raw 3D data. Point-BERT [Yu et al., 2022], Point-MAE [Pang et al., 2022], PointGPT [Chen et al., 2024b], and Point-Bind [Guo et al., 2023] introduce self-supervised pretraining. 3D object detection [Shi et al., 2019, Qi et al., 2019, Shi et al., 2020, Chen et al., 2023g] enables scene parsing.

Depth Estimation. Monocular depth estimation [Godard et al., 2019, Ranftl et al., 2021, 2020, Yang et al., 2024c, Fu et al., 2024b] provides geometric understanding from single images. Depth Anything [Yang et al., 2024c] achieves strong zero-shot transfer. Metric3D [Yin et al., 2023] recovers metric depth.

Semantic Segmentation. Semantic segmentation [Long et al., 2015, Chen et al., 2017, Kirillov et al., 2023, Peng et al., 2023a, Chen et al., 2023b] enables scene parsing. SAM [Kirillov et al., 2023] provides promptable segmentation. Open-vocabulary methods [Ghiasi et al., 2022, Liang et al., 2023a, Chen et al., 2023a] enable zero-shot recognition.

6.3 Manipulation: How Do Agents Interact with Objects?

Manipulation requires understanding object affordances [Gibson, 1979], planning contact-rich interactions [Chitnis et al., 2020], and executing precise motor commands [Argall et al., 2009].

Vision-Language-Action Models. VLA models [Brohan et al., 2022, 2023, Team et al., 2024, Kim et al., 2024, Black et al., 2024, Driess et al., 2023, Collaboration, 2023] directly map visual observations and language instructions to actions. RT-1 [Brohan et al., 2022] introduced large-scale robot learning. RT-2 [Brohan et al., 2023] demonstrated web-scale pretraining transfer. RT-X [Collaboration, 2023] enables cross-embodiment learning. Octo [Team et al., 2024] provides an open generalist policy. OpenVLA [Kim et al., 2024] offers open-source VLA. π_0 [Black et al., 2024] introduces flow matching for robot learning. RoboCat [Bousmalis et al., 2023] demonstrates self-improvement.

Imitation Learning. Behavior cloning [Pomerleau, 1988, Chi et al., 2023, Zhao et al., 2023, Chi et al., 2024] learns from demonstrations. Diffusion Policy [Chi et al., 2023] applies diffusion models to action generation. ACT [Zhao et al., 2023] uses action chunking with transformers. Learning from play [Lynch et al., 2020] enables unstructured learning.

Reinforcement Learning. RL for manipulation [Kalashnikov et al., 2018, Levine et al., 2018, Haarnoja et al., 2018, Schulman et al., 2017, Fujimoto et al., 2018] enables learning from interaction. QT-Opt [Kalashnikov et al., 2018] scales to real-world grasping. SAC [Haarnoja et al., 2018] provides sample-efficient learning.

Simulation Environments. Simulation platforms [James et al., 2020, Yu et al., 2020, Makoviychuk et al., 2021, Savva et al., 2019, Kolve et al., 2017, Gu et al., 2023, Deitke et al., 2023] provide training environments. RLBench [James et al., 2020] offers diverse manipulation tasks. Meta-World [Yu et al., 2020] provides multi-task benchmarks. Isaac Gym [Makoviychuk et al., 2021] enables GPU-accelerated simulation.

6.4 Geospatial Analysis: How Do Agents Reason at Planetary Scale?

Geospatial analysis requires processing satellite imagery [Zhu et al., 2017], modeling urban dynamics [Bibri and Krogstie, 2017], and reasoning about geographic relationships [Egenhofer and Franzosa, 1991].

Remote Sensing Foundation Models. Prithvi [Jakubik et al., 2024] provides geospatial foundation models trained on Harmonized Landsat Sentinel-2 data. SatMAE [Cong et al., 2022] introduces masked autoencoders for satellite imagery. Satlas [Bastani et al., 2023a] enables large-scale geospatial understanding. GeoAI [Janowicz et al., 2020, Mai et al., 2023] surveys the field. CROMA [Fuller et al., 2024] and microestimates [Chi et al., 2022] advance remote sensing analysis.

Spatio-Temporal Graph Neural Networks. STGNNs model complex urban dynamics through graph-structured representations [Jin et al., 2023, Atluri et al., 2018, Wang et al., 2020]. The general formulation combines spatial and temporal operators:

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)} + \text{TemporalConv}(\mathbf{H}^{(l)})\right) \tag{1}$$

DCRNN [Li et al., 2018] models traffic as diffusion on graphs:

$$\mathbf{H}^{(l)} = \sum_{k=0}^{K}\left(\mathbf{P}_f^k\mathbf{X}\mathbf{W}_{k,1} + \mathbf{P}_b^k\mathbf{X}\mathbf{W}_{k,2}\right) \tag{2}$$

where $\mathbf{P}_f$ and $\mathbf{P}_b$ are forward and backward transition matrices.

STGCN [Yu et al., 2018] combines graph and temporal convolutions through a sandwiched structure. Graph WaveNet [Wu et al., 2019] learns adaptive graph structures without predefined adjacency:

$$\tilde{\mathbf{A}} = \text{SoftMax}\left(\text{ReLU}\left(\mathbf{E}_1\mathbf{E}_2^T\right)\right) \tag{3}$$

where $\mathbf{E}_1, \mathbf{E}_2$ are learnable node embeddings.

AGCRN [Bai et al., 2020] introduces node-specific patterns through adaptive modules. ASTGCN [Guo et al., 2019] adds spatial and temporal attention mechanisms. GMAN [Zheng et al., 2020] uses graph multi-attention with transform attention for long-range dependencies. STGRAT [Choi et al., 2022] advances the field.

Urban Computing. Urban computing [Zheng et al., 2014, Yuan et al., 2020] applies AI to city-scale challenges. ST-LLM [Liu et al., 2024a] and UniST [Yuan et al., 2024] integrate language models with spatio-temporal reasoning. Traffic prediction [Li et al., 2018, Yu et al., 2018, Wu et al., 2019] and demand forecasting [Geng et al., 2019] represent key applications.

Key Takeaways: Spatial Domains

- Navigation has progressed from point-goal to language-guided and zero-shot paradigms through vision-language integration
- Scene understanding benefits from neural implicit representations (NeRF, 3DGS) combined with semantic grounding
- Manipulation is being transformed by VLA models that transfer web-scale knowledge to robotic control
- Geospatial analysis increasingly leverages foundation models and GNNs for planetary-scale reasoning

7 Enabling Technologies

7.1 Graph Neural Networks for Spatial Reasoning

GNNs provide inductive biases well-suited to spatial reasoning through message passing on graph structures [Kipf and Welling, 2017, Veličković et al., 2018b, Xu et al., 2019, Hamilton et al., 2017, Wu et al., 2020b].

Message Passing Framework. The general GNN formulation follows the message passing paradigm [Gilmer et al., 2017, Scarselli et al., 2009, Battaglia et al., 2018]:

$$\mathbf{m}_v^{(l)} = \text{AGGREGATE}^{(l)}\left(\left\{\mathbf{h}_u^{(l-1)} : u \in \mathcal{N}(v)\right\}\right) \tag{4}$$

$$\mathbf{h}_v^{(l)} = \text{UPDATE}^{(l)}\left(\mathbf{h}_v^{(l-1)}, \mathbf{m}_v^{(l)}\right) \tag{5}$$

where $\mathcal{N}(v)$ denotes the neighbors of node v, and AGGREGATE and UPDATE are learnable functions.

GNN-LLM Integration. Emerging work combines GNNs with LLMs for structured spatial reasoning [Chen et al., 2024e, Tang et al., 2024, Fatemi et al., 2023, 2024, Gowda et al., 2025]. This integration

enables leveraging both the relational reasoning of GNNs and the semantic understanding of LLMs. Graph instruction tuning [Zhang et al., 2024a] further enhances this capability. LLaGA [Chen et al., 2024e] provides language-graph alignment. GraphGPT [Tang et al., 2024] enables graph reasoning through language models.

Geometric Deep Learning. Geometric deep learning [Bronstein et al., 2021] provides theoretical foundations for spatial reasoning on non-Euclidean domains. Equivariant networks [Cohen and Welling, 2016, Fuchs et al., 2020, Satorras et al., 2021] respect spatial symmetries through:

$$f(T_g \cdot x) = T_g \cdot f(x) \tag{6}$$

where T_g is a group transformation. Graph transformers [Ying et al., 2021, Dwivedi et al., 2023, Rampášek et al., 2022] combine attention with graph structure. E3NN [Batzner et al., 2022] and geometric message passing [Brandstetter et al., 2022] advance equivariant architectures.

7.2 World Models

World models learn predictive representations enabling planning through imagination [LeCun, 2022, Schmidhuber, 2015, Matsuo et al., 2022].

Latent Dynamics Models. World models learn a latent dynamics model that predicts future states:

$$\text{Encoder:} \quad \mathbf{z}_t = q_\phi(\mathbf{z}_t | \mathbf{o}_{\leq t}, \mathbf{a}_{<t}) \tag{7}$$

$$\text{Dynamics:} \quad \hat{\mathbf{z}}_{t+1} = p_\theta(\hat{\mathbf{z}}_{t+1} | \mathbf{z}_t, \mathbf{a}_t) \tag{8}$$

$$\text{Decoder:} \quad \hat{\mathbf{o}}_t = p_\psi(\hat{\mathbf{o}}_t | \mathbf{z}_t) \tag{9}$$

Model-Based Reinforcement Learning. Dreamer [Hafner et al., 2020] introduced latent imagination for sample-efficient learning through recurrent state-space models. DreamerV2 [Hafner et al., 2021] achieved human-level Atari performance with discrete latent states. DreamerV3 [Hafner et al., 2023] demonstrated cross-domain mastery with a single algorithm through symlog predictions. DayDreamer [Wu et al., 2023b] transferred world models to real robots with minimal real-world data. PlaNet [Hafner et al., 2019] pioneered latent dynamics learning. MuZero [Schrittwieser et al., 2020] combined learned models with MCTS for game playing. Additional approaches include MBPO [Janner et al., 2019], World Models [Ha and Schmidhuber, 2018], and TD-MPC [Hansen et al., 2022].

Video World Models. Genie [Bruce et al., 2024] learns controllable world models from internet videos enabling interactive environments. WorldDreamer [Yang et al., 2024d] generates driving world models for autonomous vehicles. GAIA-1 [Hu et al., 2023a] produces realistic driving videos conditioned on actions and text. Sora [OpenAI, 2024] demonstrates video generation as world simulation at scale. Video prediction models [Yang et al., 2024a, Baker et al., 2022] provide foundations for world understanding.

LLM-Based World Models. LLMs can serve as world models for planning [Hao et al., 2023, Huang et al., 2022b], predicting state transitions without explicit environment models. This approach leverages the vast knowledge encoded in LLMs to simulate world dynamics. RAP [Hao et al., 2023] combines reasoning with acting through world model rollouts. TransDreamer [Chen et al., 2022a] and UniSim [Yang et al., 2023] advance world modeling.

7.3 Multimodal Foundation Models

Multimodal models integrate vision, language, and action understanding [Baltrušaitis et al., 2019, Xu et al., 2023].

Vision-Language Models. CLIP [Radford et al., 2021] enabled zero-shot visual recognition through contrastive pretraining on web-scale data. BLIP-2 [Li et al., 2023b] introduced efficient vision-language pretraining with frozen encoders. LLaVA [Liu et al., 2023b, 2024b] demonstrated visual instruction tuning with strong performance. GPT-4V [OpenAI, 2023, Zheng et al., 2024, Yan et al., 2023] achieved strong multimodal reasoning. Gemini [Team and Google, 2023] provides native multimodal capabilities. Flamingo [Alayrac et al., 2022] enables few-shot visual learning through interleaved attention. PaLI [Chen et al., 2023d,e] scales vision-language models. Kosmos-2 [Peng et al., 2023b] adds grounding capabilities. Qwen-VL [Bai et al., 2023] provides open multilingual VLMs. Additional models include InstructBLIP [Dai et al., 2023b], CogVLM [Wang et al., 2023b], InternVL [Chen et al., 2024i], and Ferret [You et al., 2023].

Spatial Vision-Language Models. SpatialVLM [Chen et al., 2024a] specializes in spatial reasoning with fine-grained understanding. SpatialRGPT [Cheng et al., 2024] provides regional spatial reasoning. VoxPoser [Huang et al., 2023b] extracts affordances from VLMs into 3D representations. VLMaps [Huang et al., 2023a] creates semantic spatial maps for navigation. These models bridge vision-language understanding with spatial reasoning.

3D Vision-Language Models. 3D-LLM [Hong et al., 2023b] enables language understanding of 3D scenes. Open3D-VQA [Zhang et al., 2025b] provides open-vocabulary 3D visual question answering. LLM-Grounder [Yang et al., 2024b] grounds language in 3D environments.

Key Takeaways: Enabling Technologies

- GNN-LLM integration represents a paradigm shift, combining relational reasoning with semantic understanding
- World models enable sample-efficient learning and safe planning through imagination
- Spatial VLMs (SpatialVLM, VLMaps, VoxPoser) bridge the gap between vision-language understanding and spatial action
- Equivariant architectures provide principled approaches to geometric reasoning

8 Industry Applications as Design Patterns

Rather than cataloging company capabilities, we abstract industry deployments into reusable design patterns for spatial AI systems.

8.1 Design Pattern 1: Human-in-the-Loop Spatial Reasoning

This pattern combines AI spatial analysis with human expert validation [Amershi et al., 2019, Shneiderman, 2022], exemplified by:

Geospatial Intelligence. Palantir [Palantir, 2023] integrates AI with human analysts for defense and commercial applications. The Gotham platform enables intelligence analysis with spatial reasoning while maintaining human oversight for critical decisions.

GIS Workflows. ESRI [ESRI, 2023, 2024] provides ArcGIS with integrated GeoAI capabilities where AI assists human planners in urban planning, environmental monitoring, and disaster response. The pattern: AI proposes, human validates, system learns from corrections.

8.2 Design Pattern 2: Weakly Supervised Planetary-Scale Learning

This pattern leverages massive unlabeled data with minimal supervision for global-scale models [Ratner et al., 2017, Zhang and Yang, 2022]:

Satellite Foundation Models. NASA-IBM Prithvi [Jakubik et al., 2024] trains on Harmonized Landsat Sentinel-2 data using self-supervised learning. Planet Labs [Planet Labs PBC, 2023] operates the largest Earth-imaging constellation, enabling daily global monitoring. Maxar provides high-resolution imagery for defense applications. The pattern: self-supervised pretraining on petabyte-scale imagery, fine-tuning for specific tasks.

Mapping at Scale. Google [Google, 2023, 2024] deploys AI for global-scale mapping through Google Earth Engine and Maps AI. The pattern: leverage user interactions and multi-source data for continuous model improvement.

8.3 Design Pattern 3: Agent-Assisted Spatial Workflows

This pattern deploys AI agents to augment human spatial reasoning [Shneiderman, 2020, Horvitz, 1999]:

Autonomous GIS. AutonomousGIS [Li et al., 2025a] and GeoGPT [Bai et al., 2024] integrate agentic capabilities with geospatial analysis. The pattern: LLM-based agents that can query spatial databases, generate maps, and answer geographic questions.

Location Intelligence. Foursquare [Foursquare, 2023] and Carto [CARTO, 2023] provide location-based services with AI-powered analytics. Wherobots [Wherobots, 2023] offers cloud-native spatial analytics. The pattern: spatial data infrastructure with AI-powered query and analysis.

8.4 Design Pattern 4: Embodied AI at Scale

This pattern deploys learned spatial policies in physical systems [Kober et al., 2013, Levine et al., 2016]:

Autonomous Vehicles. Waymo [Waymo, 2023, 2024] has deployed autonomous vehicles at scale with millions of miles driven. EMMA [Waymo, 2024] provides end-to-end multimodal models for driving. Tesla [Tesla, 2023] pursues vision-only autonomy. The pattern: massive simulation, careful real-world deployment, continuous learning from fleet data.

Robot Learning Platforms. Open X-Embodiment [Collaboration, 2023] provides large-scale robot data from Google DeepMind and collaborating institutions. Bridge Data [Walke et al., 2023] enables cross-domain transfer. The pattern: diverse data collection, foundation model training, transfer to specific embodiments.

Key Takeaways: Industry Patterns

- Human-in-the-loop patterns dominate safety-critical applications (defense, urban planning)
- Weakly supervised learning enables planetary-scale models without exhaustive labeling
- Agent-assisted workflows augment rather than replace human spatial reasoning
- Embodied AI deployment requires massive simulation followed by careful real-world transfer

9 Evaluation Framework and Benchmark Analysis

9.1 Existing Benchmarks

Table 3 summarizes key benchmarks organized by our taxonomy.

Table 3: Spatial AI Benchmarks Organized by Taxonomy

Benchmark	Spatial Task	Scale	Environment	Primary Metric	Agentic Capability
R2R [Anderson et al., 2018c]	Navigation	Meso	Simulated	SPL, SR	Memory + Planning
REVERIE [Qi et al., 2020]	Navigation	Meso	Simulated	SPL, RGS	Memory + Planning
Habitat [Savva et al., 2019]	Navigation	Meso	Simulated	SPL	Planning
AI2-THOR [Kolve et al., 2017]	Navigation + Manipulation	Meso	Simulated	SR	Planning + Tool Use
RLBench [James et al., 2020]	Manipulation	Micro	Simulated	SR	Tool Use
Meta-World [Yu et al., 2020]	Manipulation	Micro	Simulated	SR	Tool Use
nuScenes [Caesar et al., 2020]	Scene Understanding	Meso-Macro	Real	mAP, NDS	Memory
KITTI [Geiger et al., 2012]	Scene Understanding	Meso	Real	mAP	Memory
ScanNet [Dai et al., 2017]	Scene Understanding	Meso	Real	mIoU	Memory
AgentBench [Liu et al., 2023c]	General	-	Mixed	SR	All
WebArena [Zhou et al., 2023b]	Web	-	Simulated	SR	Planning + Tool Use
SWE-Bench [Jimenez et al., 2024]	Code	-	Real	Pass@k	Planning + Tool Use
EmbodiedBench [Yang et al., 2025]	Embodied	Meso	Simulated	SR	All
SafeAgentBench [Yin et al., 2025]	Safety	-	Simulated	Safety Rate	Planning

9.2 Evaluation Metrics

We propose standardized metrics across domains with formal definitions [Powers, 2011, Sokolova and Lapalme, 2009, Hossin and Sulaiman, 2015]:

Navigation Metrics. Success Rate (SR) measures task completion. Success weighted by Path Length (SPL) [Anderson et al., 2018b] accounts for path efficiency:

$$\text{SPL} = \frac{1}{N} \sum_{i=1}^{N} S_i \cdot \frac{\ell_i}{\max(\ell_i, p_i)} \tag{10}$$

where S_i is the binary success indicator, ℓ_i is the shortest path length, and p_i is the actual path length.

Normalized Dynamic Time Warping (nDTW) measures trajectory similarity:

$$\text{nDTW} = \exp\left(-\frac{\text{DTW}(P, R)}{\ell_R}\right) \tag{11}$$

where P is the predicted path, R is the reference path, and ℓ_R is the reference path length.

Manipulation Metrics. Task Success Rate measures goal achievement. Goal Condition Satisfaction evaluates partial completion. Efficiency metrics include action count and time to completion.

Reasoning Metrics. Accuracy, F1 Score, and BLEU scores assess spatial reasoning and question answering.

Safety Metrics. Collision Rate, Safety Violation Rate, and Risk-Aware Success measure safe operation.

9.3 Critical Analysis: What Benchmarks Fail to Measure

While existing benchmarks have advanced the field [Raji et al., 2021, Liao et al., 2021, Ribeiro et al., 2020], several fundamental limitations warrant critical examination:

Simulation-Reality Gap. Most benchmarks rely on simulated environments [Savva et al., 2019, Kolve et al., 2017, James et al., 2020, Chattopadhyay et al., 2021, Szot et al., 2021], which differ from real-world conditions in visual appearance, physics, and dynamics. Policies trained in simulation often fail to transfer [Zhao et al., 2020, Tobin et al., 2017, Andrychowicz et al., 2020], limiting practical applicability. *Gap: No benchmark systematically measures sim-to-real transfer degradation.*

Metric Limitations. Standard metrics like SPL assume optimal paths are known, which is unrealistic in novel environments. Success Rate ignores partial progress and efficiency. Current metrics do not capture important aspects such as safety, robustness to perturbations, and graceful degradation. *Gap: Metrics reward task completion but not safe, robust, or interpretable behavior.*

Long-Horizon Evaluation. Most benchmarks evaluate short episodes (tens to hundreds of steps). Real-world tasks require sustained performance over hours or days with memory persistence and error recovery. *Gap: No benchmark evaluates multi-day spatial tasks with persistent memory.*

Safety-Critical Evaluation. Benchmarks rarely evaluate failure modes, adversarial robustness, or behavior under distribution shift. Safety-critical applications require understanding of worst-case performance. *Gap: Safety evaluation remains ad-hoc rather than systematic.*

Cross-Scale Evaluation. Benchmarks typically operate at a single spatial scale. Real applications require reasoning across micro (grasping), meso (navigation), and macro (planning) scales simultaneously. *Gap: No benchmark evaluates cross-scale spatial reasoning.*

9.4 SpatialAgentBench: Proposed Unified Framework

Our proposed benchmark addresses these gaps with eight tasks designed to evaluate the full spectrum of spatial agent capabilities:

1. **VLN-Instruct**: Vision-language navigation with complex, multi-step instructions requiring spatial reasoning and landmark recognition.

2. **ObjectSearch**: Multi-room object search with semantic reasoning, requiring agents to leverage commonsense knowledge about object locations.

3. **SceneQA**: 3D scene question answering requiring understanding of spatial relationships, object properties, and scene semantics.

4. **ManipSeq**: Sequential manipulation planning with long-horizon tasks requiring tool use and state tracking.

5. **GeoReason**: Geospatial reasoning from satellite imagery including change detection, land use classification, and spatial pattern analysis.

6. **TrafficPredict**: Spatio-temporal traffic prediction requiring modeling of complex urban dynamics and graph-structured dependencies.

7. **SafeNav**: Navigation with safety constraints including obstacle avoidance, social navigation, and risk-aware planning.

8. **MultiAgent**: Coordinated multi-agent spatial tasks requiring communication, task allocation, and collaborative planning.

Key Takeaways: Evaluation

- Existing benchmarks are fragmented across domains with incompatible metrics
- Critical gaps exist in sim-to-real transfer, long-horizon, safety-critical, and cross-scale evaluation
- SpatialAgentBench proposes unified evaluation across navigation, manipulation, scene understanding, and geospatial reasoning
- Standardized metrics (SPL, nDTW, safety rates) enable cross-domain comparison

10 Grand Challenges and Future Directions

We identify six grand challenges that represent fundamental bottlenecks for the field [Marcus, 2020, Chollet, 2019, Lake et al., 2017, Bengio et al., 2019, Bommasani et al., 2021]:

10.1 Grand Challenge 1: Unified Spatial Representation

How can agents maintain a single, coherent spatial representation that supports reasoning across micro, meso, and macro scales?

Current approaches use separate representations for different scales: point clouds for grasping [Rusu and Cousins, 2011, Fang et al., 2020], topological maps for navigation [Thrun, 1998, Kuipers and Byun, 1991], and raster imagery for geospatial analysis [Goodfellow et al., 2016]. A unified representation would enable seamless reasoning across scales. Key research directions include:

- Hierarchical scene graphs that span from object parts to city infrastructure
- Neural implicit representations with multi-scale querying
- Foundation models for 3D understanding [Hong et al., 2023b, Fu et al., 2024a, Shen et al., 2023a, Oquab et al., 2024, 2023, Chen et al., 2023a, Zhou et al., 2024, Wu et al., 2015]

10.2 Grand Challenge 2: Grounded Long-Horizon Planning

How can agents plan over extended horizons while maintaining geometric feasibility?

LLMs can generate high-level plans but struggle with geometric constraints [Valmeekam et al., 2023b, Kambhampati et al., 2024, Huang et al., 2024]. TAMP systems handle geometry but lack semantic flexibility. Bridging this gap requires:

- Hybrid neuro-symbolic planners that combine LLM reasoning with geometric verification
- Hierarchical planning with learned abstractions [Song et al., 2023, Valmeekam et al., 2023a, Huang et al., 2022a, Li et al., 2025b, Silver et al., 2024]
- World models that predict both semantic and geometric consequences

10.3 Grand Challenge 3: Safe Deployment Under Uncertainty

How can spatial AI systems operate safely in safety-critical applications with guaranteed bounds on failure?

Current systems lack formal safety guarantees [Seshia et al., 2022, Koopman and Wagner, 2019, Amodei et al., 2016a]. Deployment in autonomous vehicles, medical robotics, and infrastructure requires:

- Uncertainty quantification for spatial predictions
- Out-of-distribution detection for novel environments
- Formal verification of spatial reasoning [Yin et al., 2025, Amodei et al., 2016b,a, Bai et al., 2022]
- Graceful degradation under adversarial conditions

10.4 Grand Challenge 4: Sim-to-Real Transfer

How can policies learned in simulation transfer reliably to the physical world?

The reality gap affects perception, dynamics, and control [Peng et al., 2018, Rusu et al., 2017, Sadeghi and Levine, 2017]. Bridging requires:

- Photorealistic simulation with accurate physics [Zhao et al., 2020, Tobin et al., 2017, James et al., 2019, Matas et al., 2018]
- Domain randomization and adaptation
- Real-world fine-tuning with minimal data
- Hybrid simulation-real training pipelines

10.5 Grand Challenge 5: Scalable Multi-Agent Coordination

How can large numbers of spatial agents coordinate effectively with limited communication?

Current multi-agent systems scale poorly [Stone and Veloso, 2000, Busoniu et al., 2008, Foerster et al., 2016]. Real applications (warehouse robotics, autonomous traffic) require:

- Emergent communication protocols for spatial coordination
- Decentralized planning with global consistency [Zhang et al., 2021, Wu et al., 2023c, Hong et al., 2023a, Li et al., 2023a, Qian et al., 2023]
- Heterogeneous agent coordination
- Robust coordination under partial observability

10.6 Grand Challenge 6: Efficient Edge Deployment

How can capable spatial AI systems run on resource-constrained platforms?

Foundation models require significant compute. Edge deployment requires:

- Model compression without capability loss [Han et al., 2016, Howard et al., 2017, Dehghani et al., 2023]
- Efficient architectures for spatial reasoning
- Hardware-software co-design for spatial AI
- Adaptive compute allocation based on task difficulty

Grand Challenges Summary

1. **Unified Representation**: Single representation spanning micro to macro scales
2. **Grounded Planning**: Long-horizon planning with geometric feasibility
3. **Safe Deployment**: Formal safety guarantees for critical applications
4. **Sim-to-Real**: Reliable transfer from simulation to physical world
5. **Multi-Agent**: Scalable coordination with limited communication
6. **Edge Deployment**: Capable systems on resource-constrained platforms

11 Limitations

This survey, while comprehensive, has several limitations:

- Our paper selection process, though systematic, may have missed relevant works in adjacent fields or non-English publications.
- The proposed taxonomy, while unifying, is one of many possible categorizations and may not capture all nuances of the field.
- Our analysis is based on publicly available information and does not include proprietary details from industry labs.
- The field is rapidly evolving, and some recent works may not be fully represented.
- We focus primarily on English-language publications from major venues.
- The proposed SpatialAgentBench is conceptual and requires implementation and validation.
- Our analysis of industry applications relies on public information and may not reflect current capabilities.

12 Conclusion

This survey has provided a unified three-axis taxonomy connecting Agentic AI and Spatial Intelligence across spatial scales, synthesizing over 900 papers across foundational architectures, state-of-the-art methods, industry applications, and evaluation benchmarks. Our analysis reveals three key findings:

1. **Hierarchical memory systems** are critical for long-horizon spatial tasks, enabling agents to accumulate and retrieve spatial knowledge effectively. Advances in retrieval-augmented generation, episodic memory, and spatial memory representations provide foundations for persistent spatial understanding.
2. **GNN-LLM integration** is an emergent paradigm combining the relational reasoning of graph networks with the semantic understanding of language models. This integration enables structured spatial reasoning that leverages both geometric relationships and semantic knowledge.
3. **World models** are essential for safe deployment, enabling agents to predict consequences and plan in imagination before acting. Video world models, latent dynamics models, and LLM-based world models provide complementary approaches to predictive understanding.

We have identified six grand challenges that represent fundamental bottlenecks: unified spatial representation, grounded long-horizon planning, safe deployment under uncertainty, sim-to-real transfer, scalable multi-agent coordination, and efficient edge deployment. The convergence of vision-language-action models, graph neural networks, world models, and foundation models provides promising directions for addressing these challenges.

By establishing this foundational reference with a three-axis taxonomy and proposing SpatialAgentBench, we aim to accelerate progress toward capable, robust, and safe spatially-aware autonomous systems that can perceive, reason about, and act within the physical world. The intersection of agentic AI and spatial intelligence represents a critical frontier for artificial intelligence, with profound implications for autonomous vehicles, robotics, urban computing, and geospatial 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.

Ekin Akyurek et al. What learning algorithm is in-context learning? *arXiv preprint arXiv:2211.15661*, 2023.

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.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. *CHI*, 2019.

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 Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016a.

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

Dong An, Yuankai Wang, Yuankai Qi, et al. Bevbert: Multimodal map pre-training for language-guided navigation. 2023.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents, 2018a.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. In *arXiv preprint arXiv:1807.06757*, 2018b.

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, 2018c.

OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Thomas Anthony et al. Thinking fast and slow with deep learning and tree search. In *NeurIPS*, 2017.

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

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

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.

Akari Asai et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

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

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. *AAAI*, 2017.

Alan Baddeley. Working memory: looking back and looking forward. *Nature reviews neuroscience*, 2003.

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.

Yifan Bai et al. Geogpt: Understanding and processing geospatial tasks through an autonomous gpt. *arXiv preprint arXiv:2307.07930*, 2024.

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.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Sidhika Balachandar, Shuvom Sadhuka, Bonnie Berger, Emma Pierson, and Nikhil Garg. Urban incident prediction with graph neural networks: Integrating government ratings and crowdsourced reports, 2025. URL `https://arxiv.org/abs/2506.08740`.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Andrea Banino et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 2018.

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.

Jonathan T Barron et al. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023.

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

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

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.

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

Simon Batzner et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 2022.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

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.

Homanga Bharadhwaj et al. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. *IEEE International Conference on Robotics and Automation*, 2024.

Simon Elias Bibri and John Krogstie. Smart sustainable cities of the future: An extensive interdisciplinary literature review. *Sustainable Cities and Society*, 2017.

Joshua Bird, Jan Blumenkamp, and Amanda Prorok. Dvm-slam: Decentralized visual monocular simultaneous localization and mapping for multi-agent systems, 2025.

Kevin Black, Noah Brown, Danny Driess, et al. pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control. In *Advances in Neural Information Processing Systems*, 2016.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arber, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Francisco Bonin-Font, Alberto Ortiz, and Gabriel Oliver. Visual navigation for mobile robots: A survey. *Journal of intelligent and robotic systems*, 2008.

Florian Bordes, Richard Yuanzhe Pang, Anas Ajber, Christopher Barber, Petar Velickovic, Mahmoud Assran, Nicolas Ballas, Yann LeCun, and Michael Rabbat. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

Sebastian Borgeaud et al. Improving language models by retrieving from trillions of tokens. *ICML*, 2022.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. Robocat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*, 2023.

Johannes Brandstetter et al. Geometric and physical quantities improve e(3) equivariant message passing. *International Conference on Learning Representations*, 2022.

Pearl Brereton, Barbara A Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 2007.

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.

Michael M Bronstein et al. Geometric deep learning. *arXiv preprint arXiv:2104.13478*, 2021.

Rodney A Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

Tim Brooks et al. Video generation models as world simulators. *OpenAI Technical Report*, 2024.

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.

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 2012.

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

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 2008.

Cesar Cadena et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 2016.

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

Wenxiao Cai et al. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

CARTO. Carto spatial data science platform. `https://carto.com/`, 2023.

Ziwei Chai et al. Graphllm: Boosting graph reasoning ability of large language model. *arXiv preprint arXiv:2310.05845*, 2023.

Chi-Min Chan et al. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

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, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *NeurIPS*, 2020b.

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

Devendra Singh Chaplot et al. Learning to explore using active neural slam. In *ICLR*, 2020d.

Prithvijit Chattopadhyay, Judy Hoffman, Roozbeh Mottaghi, and Aniruddha Kembhavi. Robustnav: Towards benchmarking robustness in embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15923–15933, 2021.

Austin Chen et al. Open-world object manipulation using pre-trained vision-language models. *Conference on Robot Learning*, 2023a.

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*, 2024a.

Chang Chen et al. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022a.

Changan Chen et al. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020.

Changan Chen et al. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS*, 2022b.

Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. In *NeurIPS*, 2024b.

Jiabin Chen, Dawei Lin, et al. Graphgpt: Graph instruction tuning for large language models. *arXiv preprint arXiv:2310.13023*, 2024c.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, et al. End-to-end autonomous driving: Challenges and frontiers. *IEEE TPAMI*, 2024d.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021a.

Runjin Chen et al. Llaga: Large language and graph assistant. *arXiv preprint arXiv:2402.08170*, 2024e.

Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. *CVPR*, 2023b.

Shizhe Chen et al. History aware multimodal transformer for vision-and-language navigation. In *NeurIPS*, 2021b.

Shizhe Chen et al. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, 2022c.

Sijin Chen et al. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024f.

Wenhu Chen et al. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023c.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beez, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2023d.

Xi Chen et al. Pali: A jointly-scaled multilingual language-image model. *International Conference on Learning Representations*, 2023e.

Xi Chen et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2024g.

Xinyun Chen et al. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*, 2023f.

Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. *CVPR*, 2023g.

Zhaohan Chen et al. Spatial reasoning in multimodal large language models: A survey. *arXiv preprint arXiv:2511.15722*, 2024h.

Zhe Chen, Weiyun Wang, Yue Cao, et al. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. *arXiv preprint*, 2024i.

Zhe Chen et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024j.

An-Chieh Cheng, Hongxu Yin, Yang Fu, et al. Spatialrgpt: Grounded spatial reasoning in vision language models. 2024.

Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, Lei Shi, and Maosong Sun. Embodiedeval: Evaluate multimodal llms as embodied agents, 2025.

Cheng Chi, Siyuan Feng, Yilun Du, et al. Diffusion policy: Visuomotor policy learning via action diffusion. 2023.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *International Journal of Robotics Research*, 2024.

Guanghua Chi et al. Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 2022.

Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Gupta. Efficient bimanual manipulation using learned task schemas. In *ICRA*, 2020.

Jeongwhan Choi et al. Graph neural controlled differential equations for traffic forecasting. *AAAI Conference on Artificial Intelligence*, 2022.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Howie Choset and Keiji Nagatani. Topological simultaneous localization and mapping. *IEEE Transactions on robotics and automation*, 17(2):125–137, 2001.

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

Gordon Christie et al. Functional map of the world. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018b.

Andrea Cini et al. Taming local effects in graph-based spatiotemporal forecasting. *Advances in Neural Information Processing Systems*, 2023.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016.

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.

Zhoujuan Cui, Wenshuo Peng, Yaqiang Zhang, Yiping Duan, and Xiaoming Tao. Spatio-temporal-interaction graph neural networks for multi-agent trajectory prediction. *Journal of Physics: Conference Series*, 2833 (1):012010, 2024.

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.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In *ACL Findings*, 2023a.

Wenliang Dai, Junnan Li, Dongxu Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. 2023b.

Murtaza Dalal, Deepak Pathak, and Ruslan Salakhutdinov. Accelerating robotic reinforcement learning via parameterized action primitives. In *Advances in Neural Information Processing Systems*, 2021.

Neil T Dantam, Zachary K Kingston, Swarat Chaudhuri, and Lydia E Kavraki. Incremental task and motion planning: A constraint-based approach. In *RSS*, 2016.

Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018.

Mostafa Dehghani et al. Scaling vision transformers to 22 billion parameters. *International Conference on Machine Learning*, 2023.

Matt Deitke et al. Objaverse: A universe of annotated 3d objects. *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

Ilke Demir et al. Deepglobe 2018: A challenge to parse the earth through satellite images. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

Guilherme N. DeSouza and Avinash C. Kak. Vision for mobile robot navigation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):237–267, 2002.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Vishnu Sashank Dorbala, James F Mullen Jr, and Dinesh Manocha. Clip-nav: Using clip for zero-shot vision-and-language navigation. In *CoRL*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

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

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks, 2022. URL `https://arxiv.org/abs/2103.04918`.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Zane Durante, Qiuyuan Sarber, Jianlong Gong, et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024.

Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006.

Vijay Prakash Dwivedi et al. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 2023.

Tore Dybå and Torgeir Dingsøyr. Applying systematic reviews to diverse study types: An experience report. *ESEM*, 2007.

Max J Egenhofer and Robert D Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information System*, 1991.

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

ESRI. Geoai in arcgis: Artificial intelligence for geospatial analysis. Technical report, Environmental Systems Research Institute, 2024.

Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. In *arXiv preprint arXiv:2311.17245*, 2024.

Hao-Shu Fang et al. Graspnet-1billion: A large-scale benchmark for general object grasping. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–547, 2019.

Bahare Fatemi et al. Talk like a graph: Encoding graphs for large language models. *arXiv preprint arXiv:2310.04560*, 2023.

Bahare Fatemi et al. Talk like a graph: Encoding graphs for large language models. *arXiv preprint arXiv:2310.04560*, 2024.

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

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Meire Fortunato, Melissa Tan, Ryan Faulkner, Steven Hansen, Adria Puigdomenech Badia, Gavin Buttimore, Charlie Deck, Joel Z Leibo, and Charles Blundell. Generalization of reinforcement learners with working and episodic memory. In *Advances in Neural Information Processing Systems*, 2019.

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

Jonathan Francis, Nariaki Kitamura, Felix Labber, Luca Navarro, and Jean Oh. Core challenges in embodied vision-language planning. In *CoRL*, 2022.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Advances in neural information processing systems*, pages 3331–3342, 2018.

Huan Fu et al. 3d foundation models: A survey. *arXiv preprint*, 2024a.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *arXiv preprint arXiv:2403.12013*, 2024b.

Yao Fu et al. Complexity-based prompting for multi-step reasoning. *ICLR*, 2023.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, 2020.

Scott Fujimoto et al. Addressing function approximation error in actor-critic methods. *International Conference on Machine Learning*, 2018.

Anthony Fuller et al. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *arXiv preprint arXiv:2311.00566*, 2024.

Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*, 2022.

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

Chuang Gan et al. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020.

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.

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

Xu Geng et al. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. *AAAI Conference on Artificial Intelligence*, 2019.

Malik Ghallab, Dana Nau, and Paolo Traverso. *Automated planning: theory and practice*. Elsevier, 2004.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022.

James J Gibson. *The ecological approach to visual perception*. Houghton Mifflin, 1979.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. *ICCV*, 2019.

Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 2007.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

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

Google. Ai in google maps: Powering the next generation of navigation. Technical report, Google LLC, 2024.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4089–4098, 2018.

Hrithik P Gowda, SN Sreevathsa, Gangadhara KN Gowda, and SJ Sharath. Graphs to blueprints: Gnn-powered floor plan modeling. *International Advanced Research Journal in Science, Engineering and Technology*, 12(2), 2025.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Jiayuan Gu et al. Maniskill2: A unified benchmark for generalizable manipulation skills. In *ICLR*, 2023.

Qiao Gu et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *IEEE International Conference on Robotics and Automation*, 2024.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming. In *arXiv preprint arXiv:2401.14196*, 2024a.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. 2019.

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

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *CoRL*, 2019.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):1–14, 2021.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *ICML*, 2020.

David Ha and Jurgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *ICML*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. 2020.

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.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *NeurIPS*, 2017.

William L Hamilton. *Graph Representation Learning*. Morgan & Claypool, 2020.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *ICLR*, 2016.

Nicklas Hansen et al. Temporal difference learning for model predictive control. *ICML*, 2022.

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.

Shibo Hao et al. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. *NeurIPS*, 2024.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

João F Henriques and Andrea Vedaldi. Mapnet: An allocentric spatial memory for mapping environments. In *CVPR*, 2018.

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.

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.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln-bert: A recurrent vision-and-language bert for navigation. *CVPR*, 2021.

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

Eric Horvitz. Principles of mixed-initiative user interfaces. *CHI*, 1999.

Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

Andrew G Howard et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *arXiv preprint arXiv:1704.04861*, 2017.

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*, 2023a.

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*, 2022a.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023b.

Wenlong Huang et al. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *International Conference on Machine Learning*, 2022b.

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.

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

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, Paul Wohlhart, Mrinal Kalber, Andrew J Davison, and Sergey Levine. Sim-to-real via sim-to-sim: Data-efficient robot learning from randomized simulation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2262–2269, 2019.

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

Michael Janner et al. When to trust your model: Model-based policy optimization. In *NeurIPS*, 2019.

Krzysztof Janowicz et al. Geoai: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 2020.

Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 2022.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2024.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2019.

Justin Johnson et al. Image retrieval using scene graphs. *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Leslie Pack Kaelbling and Tomás Lozano-Pérez. Hierarchical task and motion planning in the now. *ICRA*, 2011.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *CoRL*, 2018.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Llms can't plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*, 2024.

Kento Kawaharazuka et al. Vision-language-action models for robotics: A review towards real-world applications. *IEEE Transactions on Robotics*, 2025.

Staffs Keele et al. Guidelines for performing systematic literature reviews in software engineering. *Technical Report, EBSE*, 2007.

Nikhil Keetha et al. Splatam: Splat, track
& map 3d gaussians for dense rgb-d slam. *arXiv preprint arXiv:2312.02126*, 2024.

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. *ICLR*, 2017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Barbara Kitchenham. Procedures for performing systematic reviews. *Keele University Technical Report*, 2004.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 2013.

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.

Philip Koopman and Michael Wagner. Safety argument considerations for public road testing of autonomous vehicles. *SAE International Journal of Advances and Current Practices in Mobility*, 2019.

Jacob Krantz et al. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*, 2020.

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.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012.

Benjamin Kuipers. The spatial semantic hierarchy. *Artificial intelligence*, 119(1-2):191–233, 2000.

Benjamin Kuipers and Yung-Tai Byun. A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Robotics and autonomous systems*, 8(1-2):47–63, 1991.

Tejas D Kulkarni et al. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *NeurIPS*, 2016.

Ankit Kumar et al. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, 2016.

John E Laird. *The Soar Cognitive Architecture*. MIT Press, 2019.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 2017.

Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Bin Lei et al. Boosting logical reasoning in large language models through a new framework: The graph of thought. *arXiv preprint arXiv:2308.08614*, 2023.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *JMLR*, 2016.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *IJRR*, 2018.

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.

Michael Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. *International symposium on string processing and information retrieval*, 2002.

Guohao Li et al. 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.

Minghao Li et al. Api-bank: A comprehensive benchmark for tool-augmented llms. *EMNLP*, 2023c.

Raymond Li et al. Starcoder: May the source be with you! *arXiv preprint arXiv:2305.06161*, 2023d.

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.

Zhenlong Li, Huan Ning, Song Gao, and Krzysztof Janowicz. Giscience in the era of artificial intelligence: A research agenda towards autonomous gis. *Annals of GIS*, 2025a.

Ziyue Li, Yuan Chang, Gaihong Yu, and Xiaoqiu Le. Hiplan: Hierarchical planning for llm-based agents with adaptive global-local guidance, 2025b.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 2023a.

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*, 2023b.

Yaobo Liang et al. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*, 2023c.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. *NeurIPS Datasets and Benchmarks*, 2021.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biber, and Peter Stone. Llm+p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023a.

Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. Spatial-temporal large language model for traffic prediction. *arXiv preprint arXiv:2401.10134*, 2024a.

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

Haotian Liu, Chunyuan Li, Yuheng Li, et al. Llava-next: Improved reasoning, ocr, and world knowledge. *arXiv preprint*, 2024b.

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*, 2023c.

Jieyi Long. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

Paul A Longley, Michael F Goodchild, David J Maguire, and David W Rhind. *Geographic information science and systems*. John Wiley and Sons, 2015.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In *NeurIPS*, 2023.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2023.

Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In *Conference on Robot Learning*, pages 1113–1132, 2020.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *ICLR*, 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *Robotics: Science and Systems*, 2017.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. Opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 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.

Viktor Makoviychuk et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE TPAMI*, 2018.

Gary Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press, 1982.

Jan Matas, Stephen James, and Andrew J Davison. Sim-to-real reinforcement learning for deformable object manipulation. In *CoRL*, 2018.

Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152: 438–450, 2022.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *CVPR*, 2019.

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.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of Internal Medicine*, 151(4): 264–269, 2009.

Douglas Morrison, Peter Corke, and Jürgen Leitner. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. In *Robotics: Science and Systems*, 2018.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022.

Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *NeurIPS*, 2018.

Soroush Nasiriany, Huihan Liu, and Yuke Zhu. Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks. In *IEEE International Conference on Robotics and Automation*, pages 7477–7484, 2022.

Nora S Newcombe. Picture this: Increasing math and science learning by improving spatial thinking. *American Educator*, 2010.

John O'Keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Clarendon Press, 1978.

Catherine Olsson et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

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

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

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

OpenAI. Sora: Video generation models as world simulators. *Technical Report*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

François Osiurak and Arnaud Badets. What is a tool? toward a triadic approach. *Quarterly Journal of Experimental Psychology*, 2016.

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.

Yatian Pang et al. Masked autoencoders for point cloud self-supervised learning. *European Conference on Computer Vision*, 2022.

Aaron Parisi, Yao Zhao, and Noah Fiedel. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*, 2022.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *CVPR*, 2019.

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.

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

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. *ICRA*, 2018.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *arXiv preprint arXiv:2306.14824*, 2023b.

Karl Pertsch, Youngwoon Lee, and Joseph Lim. Accelerating reinforcement learning with learned skill priors. In *CoRL*, 2021.

Kai Petersen et al. Systematic mapping studies in software engineering. *EASE*, 2008.

Pinecone. Pinecone vector database, 2023.

Planet Labs PBC. Planet labs: Daily satellite imagery and insights. https://www.planet.com/, 2023.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. 1988.

David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2011.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adria Puigdomenech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *International Conference on Machine Learning*, pages 2827–2836, 2017.

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 R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017b.

Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. *ICCV*, 2019.

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.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

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

Yujia Qin et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. *NeurIPS Datasets and Benchmarks*, 2021.

Ladislav Rampášek et al. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 2022.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In *IEEE TPAMI*, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *VLDB*, 2017.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *ACL*, 2020.

Samuel Ritter, Jane X Wang, Zeb Kurth-Nelson, Siddhant M Jayakumar, Charles Blundell, Razvan Pascanu, and Matthew Botvinick. Been there, done that: Meta-learning with episodic recall. In *International Conference on Machine Learning*, pages 4354–4363, 2018.

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

Baptiste Roziere et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Stuart Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice Hall, 2010.

Andrei A Rusu, Matej Vecerik, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. *CoRL*, 2017.

Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). *ICRA*, 2011.

Fereshteh Sadeghi and Sergey Levine. Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*, 2017.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, pages 9323–9332, 2021.

Nikolay Savinov, Anton Raichuk, Raphael Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *International Conference on Learning Representations*, 2018.

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.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

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.

Jürgen Schmidhuber. On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. *arXiv preprint arXiv:1511.09249*, 2015.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sanjit A Seshia, Dorsa Sadigh, and S Shankar Sastry. Toward verified artificial intelligence. *Communications of the ACM*, 2022.

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

Tianchang Shen et al. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint*, 2023a.

Yongliang Shen, Kaitao Song, Xu Tan, et al. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. 2023b.

Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. *CVPR*, 2019.

Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023a.

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

Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe and trustworthy. *International Journal of Human–Computer Interaction*, 2020.

Ben Shneiderman. Human-centered ai. *Oxford University Press*, 2022.

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

Mohit Shridhar et al. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

David Silver et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

David Silver et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Tom Silver et al. Generalized planning in pddl domains with pretrained large language models. *AAAI Conference on Artificial Intelligence*, 2024.

Ishika Singh, Valts Blukis, Arsalan Mousavian, et al. Progprompt: Generating situated robot task plans using large language models. 2023.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing and management*, 2009.

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.

Larry R Squire, Craig EL Stark, and Robert E Clark. The memory systems of the brain. *Trends in cognitive sciences*, 2004.

Siddharth Srivastava, Eugene Fang, Lorenzo Riano, Rohan Chitnis, Stuart Russell, and Pieter Abbeel. Combined task and motion planning through an extensible planner-independent interface layer. In *ICRA*, 2014.

Peter Stone and Manuela Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 2000.

Edgar Sucar et al. imap: Implicit mapping and positioning in real-time. *IEEE International Conference on Computer Vision*, 2021.

Sainbayar Sukhbaatar et al. End-to-end memory networks. In *NeurIPS*, 2015.

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

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. 2023.

Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

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.

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

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

Tesla. Tesla full self-driving. `https://www.tesla.com/autopilot`, 2023.

Hugues Thomas et al. Kpconv: Flexible and deformable convolution for point clouds. *IEEE International Conference on Computer Vision*, 2019.

Sebastian Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 1998.

Sebastian Thrun. *Robotic mapping: A survey*. Morgan Kaufmann, 2002.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT press, 2005.

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.

Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

Marc Toussaint. Logic-geometric programming: An optimization-based approach to combined task and motion planning. In *IJCAI*, 2015.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *ACL*, 2023.

Endel Tulving. Episodic and semantic memory. *Organization of memory*, 1972.

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

Krist Vaesen. The cognitive bases of human tool use. *Behavioral and Brain Sciences*, 2012.

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, 2023a.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models–a critical investigation. *NeurIPS*, 2023b.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations*, 2018a.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *ICLR*, 2018b.

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. *ICML*, 2017.

Homer Walke, Kevin Black, Tony Z Zhao, et al. Bridgedata v2: A dataset for robot learning at scale. In *CoRL*, 2023.

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*, 2023a.

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. Milvus: A purpose-built vector data management system, 2021.

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.

Senzhang Wang, Jiannong Cao, and Philip S Yu. Deep learning for spatio-temporal data mining: A survey. *IEEE TKDE*, 2020.

Weihan Wang et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM TOG*, 2019.

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.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015.

Wherobots. Wherobots cloud-native spatial analytics. `https://wherobots.com/`, 2023.

Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6659–6668, 2019.

Erik Wijmans, Abhishek Kadian, Ari Morber, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *ICLR*, 2020.

Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. *EASE*, 2014.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. In *arXiv preprint arXiv:2303.04671*, 2023a.

Philipp Wu et al. Daydreamer: World models for physical robot learning. *arXiv preprint arXiv:2206.14176*, 2023b.

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*, 2023c.

Zhirong Wu et al. 3d shapenets: A deep learning approach for 3d shape representation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

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.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE TNNLS*, 2020a.

Zonghan Wu et al. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020b.

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.

Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey, 2025. URL `https://arxiv.org/abs/2410.16602`.

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

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

An Yan et al. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation. *arXiv preprint arXiv:2311.07562*, 2023.

Alejandro Escontrela Yang, Russell Mendonca, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. 2024a.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, et al. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. 2024b.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *CVPR*, 2024c.

Mengjiao Yang et al. Unisim: Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

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

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

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023c.

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.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. *arXiv preprint arXiv:2307.10984*, 2023.

Chengxuan Ying et al. Do transformers really perform bad for graph representation? *Advances in Neural Information Processing Systems*, 2021.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

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.

Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. *CVPR*, 2019.

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.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. *CVPR*, 2022.

Yuan Yuan et al. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

Zhongqiang Yuan, Xiaobing Zhou, and Tianbao Yang. A survey on urban traffic anomalies detection algorithms. *IEEE Access*, 2020.

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

Wentao Zhan and Abhirup Datta. Neural networks for geospatial data, 2024. URL `https://arxiv.org/abs/2304.09157`.

Guibin Zhang, Haotian Ren, Chong Zhan, Zhenhong Zhou, Junhao Wang, He Zhu, Wangchunshu Zhou, and Shuicheng Yan. Memevolve: Meta-evolution of agent memory systems, 2025a.

Jiayan Zhang et al. Graphinstruct: Empowering large language models with graph understanding and reasoning capability. *arXiv preprint arXiv:2403.04483*, 2024a.

Kaichen Zhang, Bo Li, Peiyuan Yan, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024b.

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.

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

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022.

Andrew Zhao et al. Expel: Llm agents are experiential learners. *AAAI*, 2024.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. 2021.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *RSS*, 2023.

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

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded. 2024.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. 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.

Denny Zhou et al. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*, 2023a.

Junsheng Zhou et al. Uni3d: Exploring unified 3d representation at scale. *International Conference on Learning Representations*, 2024.

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

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 2017.

Zihan Zhu et al. Nice-slam: Neural implicit scalable encoding for slam. *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.