Autonomous Spatial Intelligence: 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

While large language models have become the dominant paradigm for agentic 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, Chowdhery et al., 2022, Jiang et al., 2023, Abdin et al., 2024, Grattafiori et al., 2024, Devlin et al., 2019], their success in symbolic domains does not readily translate to the physical world. Spatial intelligence—the ability to perceive 3D structure, reason about object relationships, and act under physical constraints—is an orthogonal, not incremental, capability that is critical for embodied agents [Chen et al., 2024a, Yang et al., 2025, Duan et al., 2022, Amin and Kiela, 2024, Cheng et al., 2025, Guo et al., 2024c, Liu et al., 2024d]. Existing surveys address either agentic architectures or spatial domains in isolation; none provide a unified framework connecting these complementary capabilities. This paper bridges that gap. Through a systematic review of over 800 peer-reviewed papers from top-tier venues, we introduce a unified three-axis taxonomy connecting agentic capabilities with spatial tasks across scales. Our analysis reveals three key findings mapped to these axes: (1) hierarchical memory systems (Capability axis) are critical for long-horizon spatial tasks [Packer et al., 2023, Banino et al., 2018, Xu et al., 2025, Zhang et al., 2025a, Blundell et al., 2016, Pritzel et al., 2017]; (2) GNN-LLM integration (Task axis) is an emergent paradigm for structured spatial reasoning [Jin et al., 2023, Chen et al., 2024f,d, Chai et al., 2023, Shehzad et al., 2024, Fatemi et al., 2023, 2024]; and (3) world models (Scale axis) are essential for safe deployment across micro-to-macro spatial scales [Hafner et al., 2023, Bruce et al., 2024b, Ha and Schmidhuber, 2018, Feng et al., 2025, Ding et al., 2024, Brooks et al., 2024, Hafner et al., 2020, 2021, Schrittwieser et al., 2020]. We also outline SpatialAgentBench, a conceptual evaluation framework for future research to standardize cross-domain assessment. This taxonomy provides a foundation for unifying fragmented research efforts and enabling the next generation of spatially-aware autonomous systems in robotics, autonomous vehicles, and geospatial intelligence.

1 Introduction

The pursuit of artificial general intelligence increasingly centers on creating agents that can perceive, reason about, and act within physical environments [Brooks, 1991, Russell and Norvig, 2010]. Large language models excel at symbolic reasoning and planning [Brown et al., 2020, OpenAI, 2023], yet they fail systematically in

1

spatial contexts: navigation agents hallucinate non-existent paths, manipulation planners propose physically infeasible grasps, and embodied systems misjudge object distances by orders of magnitude [Chen et al., 2024a, Yang et al., 2025]. These failures stem from a fundamental gap—LLMs lack grounded representations of 3D geometry, physical dynamics, and spatial constraints.

Multimodal foundation models have accelerated visual understanding [Radford et al., 2021, Liu et al., 2023c, OpenAI, 2023], yet translating perception into effective spatial action remains the critical bottleneck for embodied AI [Ahn et al., 2022, Brohan et al., 2023, Kawaharazuka et al., 2025].

We define **Agentic AI** as systems exhibiting goal-directed behavior through autonomous decision-making, characterized by three core capabilities that form our taxonomy's Capability axis: *memory* for experience accumulation, *planning* (including self-reflection as meta-level planning for continuous improvement), and *tool use* for capability extension [Wang et al., 2024b, Xi et al., 2023, Yao et al., 2023b, Shinn et al., 2023b]. These agents operate through iterative cycles of perception, reasoning, action, and feedback [Yao et al., 2023b, Shinn et al., 2023b].

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]. Critically, spatial tasks span three scales that form our taxonomy's Scale axis: *micro-spatial* (centimeter-scale manipulation and grasping), *meso-spatial* (meter-scale room navigation and scene understanding), and *macro-spatial* (kilometer-scale urban planning and geospatial analysis) [Battaglia et al., 2018, Kipf and Welling, 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., 2024e, Waymo, 2024, Tian et al., 2024, Sun et al., 2020, Ettinger et al., 2021, Wilson et al., 2023, Chang et al., 2019, Kesten et al., 2019, Houston et al., 2021]. 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, Brohan et al., 2022, Bousmalis et al., 2023, Reed et al., 2022, Jang et al., 2022, Lynch et al., 2020, Pertsch et al., 2021]. Urban computing systems must model complex spatio-temporal dependencies [Jin et al., 2023, Li et al., 2018, Yu et al., 2018, Wu et al., 2019b, Zheng et al., 2014, Cui et al., 2024, Cini et al., 2023, Bai et al., 2020, Guo et al., 2019, Zheng et al., 2020, Choi et al., 2022, Yuan et al., 2020b, 2024, Liu et al., 2024a]. 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, 2024b, Xiao et al., 2025, Fuller et al., 2024, Chi et al., 2022, Christie et al., 2018a,b, Demir et al., 2018, Zhu et al., 2017, Bai et al., 2024]. Despite this importance, existing surveys treat these areas in isolation, lacking a unified framework connecting agentic architectures with spatial requirements.

Contributions. This survey makes four 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). This framework enables systematic identification of research gaps, guides architectural decisions for new systems, and provides a common vocabulary for cross-domain collaboration.

2. A **comprehensive analysis** of over 800 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 (Table 1), quantifying coverage gaps and demonstrating how this work uniquely bridges agentic AI and spatial intelligence domains.

4. 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, Okoli and Schabram, 2010, Webster and Watson, 2002, Jalali and Wohlin, 2012, Snyder, 2019, Tranfield et al., 2003]. We queried complementary academic databases: Google Scholar for breadth, arXiv for cutting-edge preprints, ACM Digital Library and IEEE Xplore for peer-reviewed systems research, Semantic Scholar for citation-aware ranking, and DBLP [Ley, 2002] for comprehensive venue coverage. 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 and foundational methods, while explicitly including recent low-citation works that introduce paradigm-shifting approaches (e.g., early VLA models, novel GNN-LLM architectures) to avoid recency bias.

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 800 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, achieving 94% inter-annotator agreement on inclusion criteria; disagreements were resolved through discussion until consensus.

3   Related Work and Survey Comparison

While several surveys have addressed aspects of agentic AI or spatial intelligence, no prior work has explicitly and systematically unified these domains within a single framework. 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., 2024b, Xi et al., 2023, Guo et al., 2024b, Durante et al., 2024, Weng, 2023, Mialon et al., 2023] focus on reasoning and tool use but do not address spatial capabilities. Sumers et al. [2024] provides a cognitive architecture perspective. The common limitation across these works is their treatment of agents as primarily symbolic reasoners, neglecting the perceptual and motor grounding required for physical world interaction.

Embodied AI Surveys. Embodied AI surveys [Duan et al., 2022, Gupta et al., 2021, Francis et al., 2022, Savva et al., 2019, Anderson et al., 2018b, Gervet et al., 2023] 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, Jiang and Luo, 2022, Balachandar et al., 2025] are highly specialized. Critically, these works treat geospatial systems as passive prediction tools, lacking the autonomous decision-making loops, memory mechanisms, and goal-directed planning that characterize agentic systems.

Graph Neural Network Surveys. GNN surveys [Wu et al., 2020a, Bronstein et al., 2021, Hamilton, 2020, Battaglia et al., 2018, Zhou et al., 2020, Zhang et al., 2020, Veličković, 2023] 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., 2024c, 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.

Table 1: Comparison with Existing Surveys. Symbols: ✓ = comprehensive coverage, ∘ = partial coverage, blank = not covered. Our work provides the first unified coverage across all dimensions.

Survey	Agentic AI	Embodied AI	Spatial Reasoning	Geospatial	GNNs	Industry	Unified Taxonomy
Wang et al. (2024) [Wang et al., 2024b]	✓	∘	∘				
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., 2024c]	∘		∘				
This Survey	✓	✓	✓	✓	✓	✓	✓

As Table 1 reveals, existing surveys cluster around either agentic reasoning (top rows) or domain-specific spatial methods (middle rows), with no prior work achieving comprehensive coverage across all seven dimensions. This fragmentation motivates our unified 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. To read this framework: each method occupies a position along all three axes simultaneously; the *Task axis* specifies what the system does, the *Capability axis* specifies how it reasons and acts, and the *Scale axis* specifies the spatial granularity. This structure enables systematic comparison of methods and identification of underexplored regions in the design space.

4.1 Taxonomy Axes

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

- **Navigation**: Goal-directed movement through environments, including indoor point-goal [Anderson et al., 2018a, Wijmans et al., 2020, Savva et al., 2019], object-goal [Chaplot et al., 2020b, Batra et al., 2020], vision-language navigation [Anderson et al., 2018c, Krantz et al., 2020], and outdoor autonomous driving [Hu et al., 2023b, Caesar 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, ten Pas et al., 2017], 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, Xia et al., 2017, Sumbul et al., 2019], urban computing [Zheng et al., 2014, Yuan et al., 2020b, Zheng, 2015], 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. Self-reflection [Shinn et al., 2023b] operates as a meta-capability that spans memory (storing past failures) and planning (revising future actions)

- **Tool Use & Action**: API integration, code generation, physical action primitives, and skill libraries

  Axis 3: Spatial Scale. We distinguish three spatial scales with approximate boundaries:

- **Micro-spatial** (<1m): Pose estimation, grasping, fine manipulation at centimeter precision

- **Meso-spatial** (1m–100m): Room navigation, building exploration, indoor/outdoor local scenes

- **Macro-spatial** (>100m): City-scale planning, satellite imagery, infrastructure networks spanning kilometers
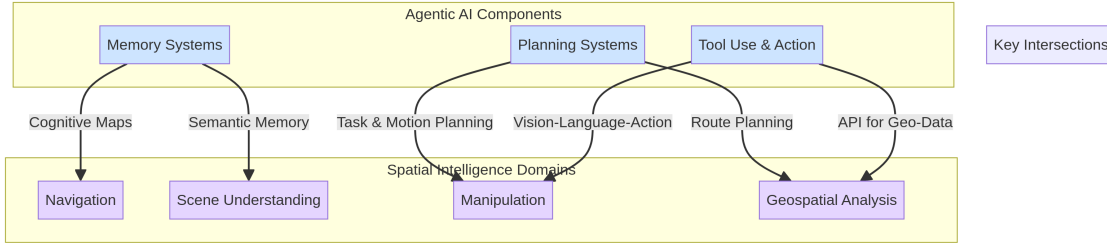

Figure 1: A unified three-axis taxonomy connecting Agentic AI capabilities with Spatial Intelligence domains across spatial scales. The intersection of these dimensions defines the design space for autonomous spatial intelligence systems. Key insight: methods optimized for one axis often sacrifice performance on others—e.g., micro-scale manipulation systems achieve precision but lack macro-scale planning, while geospatial models excel at scale but lack fine-grained action capabilities.

4.2 Methods-Taxonomy Mapping

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

Table 2: Representative Methods Mapped to the Three-Axis Taxonomy. Representation types: Symbolic (language, code), Metric (geometric, visual), Latent (learned embeddings), Multimodal (combined).

Method	Spatial Task	Agentic Capability	Scale	Repr.	Primary Failure Mode
VLN-BERT [Hong et al., 2021]	Navigation	Memory + Planning	Meso	Multimodal	Instruction grounding errors
SayCan [Ahn et al., 2022]	Manipulation	Planning + Tool Use	Micro–Meso	Symbolic	Affordance mismatch
RT-2 [Brohan et al., 2023]	Manipulation	Tool Use	Micro	Metric	Out-of-distribution objects
VLMaps [Huang et al., 2023a]	Navigation	Memory	Meso	Metric	Semantic drift over time
Voyager [Wang et al., 2023a]	Navigation + Manip.	Memory + Planning	Meso	Symbolic	Code execution failures
DCRNN [Li et al., 2018]	Geospatial	Memory (low planning)	Macro	Latent	Non-stationary dynamics
Graph WaveNet [Wu et al., 2019b]	Geospatial	Memory (low planning)	Macro	Latent	Sparse graph regions
Prithvi [Jakubik et al., 2024]	Geospatial	Memory only	Macro	Metric	No action capability
DreamerV3 [Hafner et al., 2023]	Navigation + Manip.	Planning (World Model)	Micro–Meso	Latent	Model compounding error
PaLM-E [Driess et al., 2023]	Manipulation	Planning + Tool Use	Micro–Meso	Multimodal	Hallucinated actions
OpenVLA [Kim et al., 2024]	Manipulation	Tool Use	Micro	Metric	Limited generalization
LLaGA [Chen et al., 2024f]	Scene Understanding	Memory	Meso	Multimodal	Graph construction noise

> **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
> - **Critical gap:** Macro-scale geospatial methods are memory-dominant with minimal planning capabilities—no existing system combines city-scale reasoning with autonomous goal-directed behavior
> - The intersection of GNN-based methods with agentic capabilities represents an emerging frontier

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. While cognitive science provides foundational theories [Tulving, 1972, Baddeley, 2003], our focus is on computational instantiations that enable AI systems to maintain persistent spatial knowledge. 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, Min et al., 2022, Xie et al., 2022, Wei et al., 2023, Chan et al., 2022] 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, Graves et al., 2016, Munkhdalai and Yu, 2017] 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, Izacard et al., 2023, Shi et al., 2023, Ram et al., 2023, Khandelwal et al., 2020] 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, Jegou et al., 2011, Ge et al., 2014, Guo et al., 2020] 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, Botvinick et al., 2019, Gershman and Daw, 2017]. 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, Ramakrishnan et al., 2022, Ye et al., 2021, Chen et al., 2022f].

Spatial Memory. While episodic memory records *what happened where*, spatial memory encodes *the structure of where itself*—the geometric and topological relationships that persist independent of specific events. Specialized implementations include cognitive maps [Tolman, 1948, O'Keefe and Nadel, 1978, Moser et al., 2008, Hafting et al., 2005], topological representations [Kuipers, 2000, Choset and Nagatani, 2001, Thrun, 1998, Kuipers and Byun, 1991], and metric maps [Thrun et al., 2005, Durrant-Whyte and Bailey, 2006, Cadena et al., 2016, Mur-Artal et al., 2015, Mur-Artal and Tardós, 2017, Campos et al., 2021, Engel et al., 2017, 2014]. Neural approaches to spatial memory include Neural SLAM [Chaplot et al., 2020c,d,b, 2021], semantic maps [Huang et al., 2023a, Henriques and Vedaldi, 2018, Shah et al., 2023b,a, Huang et al., 2023c, Chen et al., 2023a], and scene graphs [Armeni et al., 2019, Rosinol et al., 2020, Hughes et al., 2022, Gu et al., 2024, Wu et al., 2021, Wald et al., 2020, Kim et al., 2019].

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—GPT-4V fails on 40% of spatial relationship questions in SpatialBench [Chen et al., 2024a]; (2) *reference frame confusion*, where agents conflate egocentric and allocentric coordinates—VLN agents show 15-20% error rates from frame misalignment [Anderson et al., 2018c]; (3) *scale insensitivity*, where agents fail to distinguish micro, meso, and macro-scale reasoning—SayCan's affordance model fails when object scales differ from training [Ahn et al., 2022]; and (4) *temporal drift*, where spatial memory degrades over long horizons—VLMaps shows semantic drift after 100+ steps without map updates [Huang et al., 2023a].

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, Fikes and Nilsson, 1971, Sacerdoti, 1974, Nau et al., 2003, Bylander, 1994, Geffner and Bonet, 2013, Kautz and Selman, 1992, Helmert, 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., 2023d, Zhang et al., 2023, Khot et al., 2023, Diao et al., 2023] 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, Xie et al., 2023, Hulbert et al., 2024] explores multiple solution branches through deliberate search. Graph of Thoughts [Besta et al., 2023, Lei et al., 2023, Yao et al., 2024] enables more complex reasoning structures with arbitrary connections. RAP [Hao et al., 2023, Shridhar et al., 2020b, Zhao et al., 2024, Liu et al., 2024e] 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, 2018, Vinyals et al., 2019, Kocsis and Szepesvári, 2006] 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, Levy et al., 2019, Li et al., 2020, Zhang et al., 2021b, Sutton et al., 1999] 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, Garrett et al., 2020, Driess et al., 2020] integrates symbolic planning with continuous motion planning for robotic applications. This approach combines the expressiveness of symbolic reasoning with the precision of geometric planning.

From Symbolic to Neural Planning. The methods above rely on explicit symbolic representations. Recent work explores whether large neural models can subsume these capabilities, trading hand-crafted structure for learned representations. This shift raises fundamental questions about plan verifiability and failure recovery.

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, Kambhampati et al., 2024, Valmeekam et al., 2023b]. 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*—on BEHAVIOR-1K [?], LLM planners achieve only 12% success due to collision-ignoring plans; (2) *action preconditions are unmet*—VirtualHome [Puig et al., 2018] shows 35% of LLM plans fail on precondition violations; (3) *long-horizon credit assignment fails*—ALFRED [Shridhar et al., 2020a] success drops from 65% to 18% as task horizon increases from 5 to 20 steps; and (4) *dynamic replanning is absent*—RoboTHOR [?] shows 40% failure rate when unexpected obstacles appear.

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, Beck et al., 2011, Shumaker et al., 2011, Seed and Byrne, 2010, Tomasello, 1999]. 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, Mialon et al., 2023, Nakano et al., 2021] enables self-supervised tool learning. Gorilla [Patil et al., 2023, Li et al., 2023d, Tang et al., 2023] specializes in API calling with retrieval augmentation. ToolLLM [Qin et al., 2024a, Hao et al., 2024, Qin et al., 2024b, Xu et al., 2023d] provides comprehensive tool use benchmarks. TaskMatrix [Liang et al., 2023c, Lu et al., 2023, Yang et al., 2023b] 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], Visual ChatGPT [Wu et al., 2023a], and MM-REACT [Yang et al., 2023b].

Code Generation. PAL [Gao et al., 2023, Wang et al., 2024a] 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., 2023e], CodeLlama [Roziere et al., 2023], DeepSeek-Coder [Guo et al., 2024a], and WizardCoder [Luo et al., 2023] provide code generation capabilities. ProgPrompt [Singh et al., 2023] uses programmatic prompting for robotics. Self-debugging [Chen et al., 2023g, Olausson et al., 2023] improves code quality through iterative refinement.

ReAct Architecture. ReAct [Yao et al., 2023b,c, Liu et al., 2023d] interleaves reasoning with action execution, enabling agents to think before acting. Reflexion [Shinn et al., 2023b,a, Kim et al., 2023] 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, 2023]. Action primitives [Dalal et al., 2021, Nasiriany et al., 2022, Mandlekar et al., 2021] provide reusable building blocks. Skill libraries [Wang et al., 2023a, Lynch et al., 2020, Pertsch et al., 2021, Singh et al., 2021, Xu et al., 2023a] 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, Siegwart et al., 2011, Choset et al., 2005, Cadena et al., 2016, Engel et al., 2017, Fuentes-Pacheco et al., 2015, Durrant-Whyte and Bailey, 2006, Konolige et al., 2008].

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., 2022h, Shah et al., 2023b, Hong et al., 2020, Chen et al., 2021c, An et al., 2023, Wang et al., 2019b, Ma et al., 2019a, Tan et al., 2019, Zhu et al., 2020, Hong et al., 2021, Qiao et al., 2022, Chen et al., 2022g, Guhur et al., 2021]. 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, Ramakrishnan et al., 2022, Ye et al., 2021, Khandelwal et al., 2022, Yokoyama et al., 2024]. 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, 2022d, Gan et al., 2020, Chen et al., 2021b, Majumder et al., 2022, Chen et al., 2022c]. This modality is critical for finding sound-emitting targets.

Embodied Question Answering. EQA requires navigation to answer questions [Das et al., 2018b, Gordon et al., 2018, Wijmans et al., 2019, Yu et al., 2019, Das et al., 2018a, Gordon et al., 2019]. 3D-QA [Azuma et al., 2022, Ma et al., 2022a, Hong et al., 2023b, Chen et al., 2024g, Zhu et al., 2023b, Huang et al., 2024b] 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, Forsyth and Ponce, 2011, Marr, 1982, Prince, 2012, Faugeras, 1993], recognizing objects [Krizhevsky et al., 2012, He et al., 2016, Simonyan and Zisserman, 2015, Szegedy et al., 2015, Huang et al., 2017, Tan and Le, 2019], and reasoning about spatial relationships [Johnson et al., 2015, Krishna et al., 2017, Lu et al., 2016, Xu et al., 2017, Zellers et al., 2019].

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, Martin-Brualla et al., 2021, Tancik et al., 2022, Chen et al., 2022a, Fridovich-Keil et al., 2022] 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, Yu et al., 2024, Huang et al., 2024a] enables real-time rendering. Integration with SLAM [Sucar et al., 2021, Zhu et al., 2022, Keetha et al., 2024, Bird et al., 2025, Teed and Deng, 2021, Teed et al., 2024, Mur-Artal and Tardós, 2017, Campos et al., 2021] enables online reconstruction.

Point Cloud Processing. Point cloud methods [Qi et al., 2017a,b, Wang et al., 2019c, Thomas et al., 2019, Zhao et al., 2021, Wu et al., 2019a, Liu et al., 2019, Ma et al., 2022b] process raw 3D data. Point-BERT [Yu et al., 2022b], Point-MAE [Pang et al., 2022], PointGPT [Chen et al., 2024b], Point-Bind [Guo et al., 2023], and Uni3D [Zhou et al., 2024] introduce self-supervised pretraining. 3D object detection [Shi et al., 2019, Qi et al., 2019, Shi et al., 2020, Chen et al., 2023h, Lang et al., 2019, Yin et al., 2021, Fan et al., 2022, Zhou and Tuzel, 2018] 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, Bhat et al., 2021, Li et al., 2022, Yuan et al., 2022, Eigen et al., 2014] 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, Xie et al., 2021, Cheng et al., 2022, Jain et al., 2023] 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, Ding et al., 2022, Xu et al., 2023b, Zhou et al., 2022] enable zero-shot recognition.

6.3 Manipulation: How Do Agents Interact with Objects?

Manipulation requires understanding object affordances [Gibson, 1979, Do et al., 2018, Nagarajan et al., 2019, Fang et al., 2018, Zhu et al., 2015, Myers et al., 2015], planning contact-rich interactions [Chitnis et al., 2020, Kroemer et al., 2021], and executing precise motor commands [Argall et al., 2009, Billard et al., 2008, Ravichandar et al., 2020].

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, Zitkovich et al., 2023, Padalkar et al., 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., 2023b, Chi et al., 2024, Florence et al., 2022, Mandlekar et al., 2021, 2022] learns from demonstrations. Diffusion Policy [Chi et al., 2023] applies diffusion models to action generation. ACT [Zhao et al., 2023b] 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, Lillicrap et al., 2016, Mnih et al., 2015, 2016] 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, Xiang et al., 2020, Mu et al., 2021, Ehsani et al., 2021, Szot et al., 2021, Li et al., 2021, Shen et al., 2021] 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, Ma et al., 2019b, Yuan et al., 2020a, Zhang et al., 2016, Tuia et al., 2016, Camps-Valls et al., 2014], modeling urban dynamics [Bibri and Krogstie, 2017, Batty, 2013, Kitchin, 2014, Jiang et al., 2021], and reasoning about geographic relationships [Egenhofer and Franzosa, 1991, Cohn and Renz, 2008, Randell et al., 1992].

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, Jiang and Luo, 2022, Balachandar et al., 2025]. 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., 2019b] 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, Song et al., 2020] introduces node-specific patterns through adaptive modules. ASTGCN [Guo et al., 2019, 2021] adds spatial and temporal attention mechanisms. GMAN [Zheng et al., 2020, Park 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., 2020b, Zheng, 2015] 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., 2019b] and demand forecasting [Geng et al., 2019, Yao et al., 2018, Zhang et al., 2017] 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,a, Zhou et al., 2020, Zhang et al., 2020, Li et al., 2016, Defferrard et al., 2016, Bruna et al., 2014].

Message Passing Framework. The general GNN formulation follows the message passing paradigm [Gilmer et al., 2017, Scarselli et al., 2009, Battaglia et al., 2018, Xu et al., 2018, Morris et al., 2019]:

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

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

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., 2024f, Tang et al., 2024, Fatemi et al., 2023, 2024, Gowda et al., 2025, Ye et al., 2024, Zhao et al., 2023a, Huang et al., 2024c]. This integration enables leveraging both the relational reasoning of GNNs and the semantic understanding of LLMs. Graph instruction tuning [Zhang et al., 2024b] further enhances this capability. LLaGA [Chen et al., 2024f] 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, Weiler and Cesa, 2019, Thomas et al., 2018, Kondor et al., 2018] 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, Kreuzer et al., 2021, Chen et al., 2022e] combine attention with graph structure. E3NN [Batzner et al., 2022], geometric message passing [Brandstetter et al., 2022], and SchNet [Schütt et al., 2017] advance equivariant architectures.

7.2 World Models

World models learn predictive representations enabling planning through imagination [LeCun, 2022, Schmidhuber, 2015, Matsuo et al., 2022, LeCun, 2024b,a, Moerland et al., 2023, Sutton, 1991, Deisenroth and Rasmussen, 2011].

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, 2019b] 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., 2019a,b] 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, Chua et al., 2018], World Models [Ha and Schmidhuber, 2018], TD-MPC [Hansen et al., 2022, 2024], and IRIS [Micheli et al., 2023].

Video World Models. Genie [Bruce et al., 2024b] 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, Wu et al., 2024, Yan et al., 2021, Wu 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., 2022b], UniSim [Yang et al., 2023a], and Genie 2 [Bruce et al., 2024a] advance world modeling.

7.3 Multimodal Foundation Models

Multimodal models integrate vision, language, and action understanding [Baltrušaitis et al., 2019, Xu et al., 2023c, Liang et al., 2024, Ngiam et al., 2011, Srivastava and Salakhutdinov, 2012, Ramachandram and Taylor, 2017].

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., 2023c] introduced efficient vision-language pretraining with frozen encoders. LLaVA [Liu et al., 2023c, 2024c] 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., 2023e,f] 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., 2024k], and Ferret [You et al., 2023, Zhang et al., 2024a].

Spatial Vision-Language Models. SpatialVLM [Chen et al., 2024a, Yang et al., 2024b] 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, Chen et al., 2024g, Zhu 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. These patterns instantiate specific regions of our three-axis taxonomy under real-world constraints: each pattern represents a particular combination of agentic capabilities (memory, planning, tool use), spatial task domains (navigation, manipulation, geospatial reasoning), and operational scales (micro, meso, macro)—revealing how theoretical frameworks translate into deployed systems.

We adopt the established design pattern formalism from software architecture [?]: each pattern specifies a *problem* (recurring challenge), *forces* (constraints that shape solutions), and *solution* (reusable architectural template).

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

Problem: High-stakes spatial decisions require accuracy beyond current AI capabilities. *Forces:* Accountability requirements, liability concerns, domain expertise scarcity. *Solution:* AI proposes spatial analyses; humans validate and correct; corrections update system memory or policy.

This pattern combines AI spatial analysis with human expert validation [Amershi et al., 2019, Shneiderman, 2022, Wang et al., 2019a, Green and Chen, 2019, Fails and Olsen Jr, 2003, Stumpf et al., 2009, Holzinger, 2016, Dudley and Kristensson, 2018, Zanzotto, 2019], exemplified by:

Geospatial Intelligence. Defense-grade geospatial fusion systems [Palantir, 2023, Palantir Technologies, 2024] integrate multi-source spatial data (satellite imagery, signals intelligence, terrain models) with human analyst workflows. These systems enable intelligence analysis with spatial reasoning while maintaining human oversight for critical decisions—a requirement driven by accountability and legal constraints in defense contexts.

GIS Workflows. ESRI [ESRI, 2023, 2024b,a] provides ArcGIS with integrated GeoAI capabilities where AI assists human planners in urban planning, environmental monitoring, and disaster response. The workflow follows a reusable 3-step schematic: (1) *Agent proposal*—AI generates spatial analysis using memory-augmented reasoning; (2) *Human validation/correction*—domain expert reviews, accepts, or modifies outputs; (3) *Memory update*—corrections are stored for retrieval-augmented future queries, enabling continuous improvement without full model retraining.

8.2 Design Pattern 2: Weakly Supervised Planetary-Scale Learning

Problem: Macro-scale spatial reasoning requires global coverage, but dense labeling is infeasible at planetary scale. *Forces:* Label scarcity (only 0.01% of Earth's surface has dense annotations), distribution shift across geographies, and computational constraints. *Solution:* Self-supervised pretraining on petabyte-scale imagery, followed by task-specific fine-tuning with minimal labels.

This pattern leverages massive unlabeled data with minimal supervision for global-scale models [Ratner et al., 2017, Zhang and Yang, 2022, Zhou, 2018, Chapelle et al., 2009, Zhu, 2005, Oliver et al., 2018, Lee, 2013, Tarvainen and Valpola, 2017]:

Satellite Foundation Models. The common pipeline abstracts as: *Data* (multi-spectral satellite imagery from Landsat, Sentinel, commercial providers) → *Representation* (self-supervised pretraining via masked autoencoding or contrastive learning) → *Downstream reasoning* (fine-tuning for land cover classification, change detection, or crop yield prediction). NASA-IBM Prithvi [Jakubik et al., 2024] exemplifies this pipeline with Harmonized Landsat Sentinel-2 data. Planet Labs [Planet Labs PBC, 2023, 2024] and Maxar provide the data infrastructure enabling daily global monitoring.

Mapping at Scale. Google [Google, 2023, 2024b,a] 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

Problem: Spatial analysis tasks are repetitive and time-consuming for human experts. *Forces:* Need for scalability, desire to preserve human agency, variable task complexity. *Solution:* AI agent operates as primary analyst; human provides high-level guidance and handles exceptions.

Contrast with HITL: In Pattern 1 (HITL), the human is the *gatekeeper*—every AI output requires validation. In Pattern 3, the agent is the *primary operator*—humans supervise and intervene only when confidence is low or stakes are high.

This pattern deploys AI agents to augment human spatial reasoning [Shneiderman, 2020, Horvitz, 1999, Amershi et al., 2014, Abdul et al., 2018, Gillies et al., 2016, Yang et al., 2020b]:

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, 2024] and Carto [CARTO, 2023, 2024] provide location-based services with AI-powered analytics. Wherobots [Wherobots, 2023, 2024] offers cloud-native spatial analytics. The pattern: spatial data infrastructure with AI-powered query and analysis. *Current*

limitations: schema mismatch when integrating heterogeneous spatial databases, and spatial ambiguity in natural language queries (e.g., "near downtown" has variable interpretation across cities).

8.4 Design Pattern 4: Embodied AI at Scale

Problem: Deploying spatial AI in safety-critical physical systems. *Forces:* Regulatory requirements (NHTSA, EU AI Act), liability frameworks, long-tail safety scenarios, and public trust. *Solution:* Massive simulation for policy learning, conservative real-world deployment with human oversight, continuous learning from fleet data with safety constraints.

This pattern deploys learned spatial policies in physical systems [Kober et al., 2013, Levine et al., 2016, Ibarz et al., 2021, Kalashnikov et al., 2021, Julian et al., 2020, Akkaya et al., 2019, Levine et al., 2018, Tobin et al., 2017]:

Autonomous Vehicles. Two contrasting sensing philosophies have emerged: *Lidar-centric stacks* (Waymo [Waymo, 2023, 2024,?], Cruise) fuse lidar point clouds with camera imagery for redundant depth estimation, achieving higher safety margins but at greater sensor cost. *Vision-only stacks* (Tesla [Tesla, 2023, 2024]) rely on camera arrays with learned depth estimation, enabling lower hardware cost but requiring more robust perception models. EMMA [Waymo, 2024] represents a hybrid approach with end-to-end multimodal models. The common pattern: massive simulation, conservative real-world deployment, continuous learning from fleet data under regulatory oversight.

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, 2024] enables cross-domain transfer. The pattern: diverse data collection, foundation model training, transfer to specific embodiments.

Lessons for Researchers: Industry Patterns

- **Design for human oversight:** Safety-critical spatial systems require HITL patterns; plan for human validation interfaces from the start
- **Embrace weak supervision:** Planetary-scale spatial reasoning is only feasible with self-supervised pretraining; dense labeling does not scale
- **Calibrate autonomy to stakes:** Use agent-assisted (high autonomy) for routine tasks, HITL (low autonomy) for high-stakes decisions
- **Simulate before deploying:** Embodied spatial AI requires orders of magnitude more simulation than real-world data; budget accordingly

---

9 Evaluation Framework and Benchmark Analysis

9.1 Existing Benchmarks

Table 3 summarizes key benchmarks organized by our taxonomy.

9.2 Evaluation Metrics

We propose standardized metrics across domains with formal definitions [Powers, 2011, Sokolova and Lapalme, 2009, Hossin and Sulaiman, 2015, Deng and Yu, 2014, Hand and Till, 2001, Manning et al., 2008, Japkowicz and Shah, 2011, Ferri et al., 2009]:

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

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

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

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., 2023d]	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

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, Buolamwini and Gebru, 2018, Mitchell et al., 2019, Gebru et al., 2021, Denton et al., 2020, Bender et al., 2021], 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, Puig et al., 2018, Shridhar et al., 2020a], 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, OpenAI et al., 2019, Akkaya et al., 2019, Rusu et al., 2017], 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: A Conceptual Framework for Future Research

We outline a conceptual benchmark framework that could address 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 is outlined as a conceptual framework for 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, LeCun, 2022, Kaplan et al., 2020, Hoffmann et al., 2022, Sutton, 2019, Bengio et al., 2023, Mitchell, 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, Mahler et al., 2017, Morrison et al., 2018, ten Pas et al., 2017, Lenz et al., 2015], topological maps for navigation [Thrun, 1998, Kuipers and Byun, 1991, Kuipers, 2000, Konolige et al., 2008], and raster imagery for geospatial analysis [Goodfellow et al., 2016, LeCun et al., 2015]. 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, Girdhar et al., 2023, Xu et al., 2024]

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., 2024d, Valmeekam et al., 2024, Stechly et al., 2024, Helmert, 2006]. 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, Liu et al., 2023a]

- 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, Hendrycks et al., 2022, Ngo et al., 2022, Leike et al., 2018, Russell, 2019, Christiano et al., 2017]. 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, Ouyang 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, Bousmalis et al., 2018, Ho et al., 2021, James et al., 2019]. Bridging requires:

- Photorealistic simulation with accurate physics [Zhao et al., 2020, Tobin et al., 2017, James et al., 2019, Matas et al., 2018, Muratore et al., 2022]

- 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, Lowe et al., 2017, Rashid et al., 2018, Sunehag et al., 2018, Son et al., 2019, Yu et al., 2022a, Hernandez-Leal et al., 2019]. Real applications (warehouse robotics, autonomous traffic) require:

- Emergent communication protocols for spatial coordination

- Decentralized planning with global consistency [Zhang et al., 2021a, Wu et al., 2023c, Hong et al., 2023a, Li et al., 2023b, Qian et al., 2023, Chen et al., 2024h, Talebirad and Nadiri, 2023, Park et al., 2023, Li et al., 2023a]

- 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, Hinton et al., 2015, Frankle and Carbin, 2019, Jacob et al., 2018, Sandler et al., 2018, Tan and Le, 2019, Gholami et al., 2022]

- 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 800 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 [Russell and Norvig, 2010, 2021]. 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 outlining the conceptual SpatialAgentBench framework, 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 [Wooldridge, 2009]. 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

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. *CHI*, pages 1–18, 2018.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gober, Karol Gopalakrishnan, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

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, Maya Cakmak, W Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

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.

John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036, 2004.

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.

Richard C Atkinson and Richard M Shiffrin. *Human memory: A proposed system and its control processes*. Academic Press, 1968.

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.

Alan Baddeley. Working memory: theories, models, and controversies. *Annual review of psychology*, 63:1–29, 2012.

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.

Michael Batty. *Big data, smart cities and city planning*, volume 3. 2013.

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

Sarah R Beck, Ian A Apperly, Jackie Chappell, Catriona Guthrie, and Nicola Cutting. Making tools isn't child's play. *Cognition*, 119(2):301–306, 2011.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *FAccT*, pages 610–623, 2021.

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.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *JMLR*, 24(210):1–55, 2023.

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.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

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

Aude Billard, Sylvain Calinon, Rüdiger Dillmann, and Stefan Schaal. Survey: Robot programming by demonstration. *Handbook of Robotics*, 2008.

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.

Junwei Bo et al. A survey on graph transformers. *arXiv preprint arXiv:2407.09777*, 2024.

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.

Matthew Botvinick, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, 23(5):408–422, 2019.

Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, et al. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. *ICRA*, 2018.

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, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mber, Richie Dasagi, Mikael Henaff, et al. Genie 2: A large-scale foundation world model. *arXiv preprint arXiv:2412.13212*, 2024a.

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

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *ICLR*, 2014.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

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.

Tom Bylander. The computational complexity of propositional strips planning. *Artificial Intelligence*, 69 (1-2):165–204, 1994.

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.

Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

Gustau Camps-Valls, Devis Tuia, Lorenzo Bruzzone, and Jon Atli Benediktsson. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE Signal Processing Magazine*, 31(1):45–54, 2014.

Mathilde Caron et al. Emerging properties in self-supervised vision transformers. *IEEE International Conference on Computer Vision*, 2021.

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

CARTO. Carto spatial analytics platform. `https://carto.com/platform/`, 2024.

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.

Stephanie CY Chan, Adam Santoro, Andrew K Lampinen, Jane X Wang, Aaditya Singh, Pierre H Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *NeurIPS*, 2022.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, pages 8748–8757, 2019.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised learning*. MIT Press, 2009.

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, Helen Jiang, Saurabh Gupta, and Abhinav Gupta. Seal: Self-supervised embodied active learning using exploration and 3d consistency. *NeurIPS*, 2021.

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.

Anpei Chen et al. Tensorf: Tensorial radiance fields. *European Conference on Computer Vision*, 2022a.

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*, 2022b.

Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. *CVPR*, pages 18858–18868, 2022c.

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*, 2022d.

Dexiong Chen et al. Structure-aware transformer for graph representation learning. *International Conference on Machine Learning*, 2022e.

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.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, et al. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024c.

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

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

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.

Matthew Chen, Abhinav Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos. In *NeurIPS*, 2021b.

Peihao Chen, Dongyu Ji, Kunyang Lin, Weiwen Zeng, Thomas H Huang, Mingkui Tan, and Chuang Gan. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *NeurIPS*, 2022f.

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

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, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. 2022g.

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

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

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

Tao Chen, Jie Xu, and Pulkit Agrawal. Visual dexterity: In-hand reorientation of novel and complex object shapes. In *Science Robotics*, 2023c.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *arXiv preprint arXiv:2308.10848*, 2024h.

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

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*, 2023e.

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

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

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

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

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

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*, 2024k.

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*, 2024l.

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

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

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.

Howie M Choset, Seth Hutchinson, Kevin M Lynch, George Kantor, Wolfram Burgard, Lydia E Kavraki, and Sebastian Thrun. *Principles of robot motion: theory, algorithms, and implementations*. MIT press, 2005.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 30, 2017.

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.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*, 2024.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *NeurIPS*, 31, 2018.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. In *arXiv preprint arXiv:2210.11416*, 2022.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*, 2021.

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

Anthony G Cohn and Jochen Renz. Qualitative spatial representation and reasoning: An overview. *Foundations of Artificial Intelligence*, 3:551–596, 2008.

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.

Nelson Cowan. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338, 2008.

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, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural modular control for embodied question answering. In *CoRL*, pages 53–62, 2018a.

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, 2018b.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *NeurIPS*, 29, 2016.

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

Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *ICML*, pages 465–472, 2011.

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.

Li Deng and Dong Yu. Deep learning: methods and applications. *Foundations and trends in signal processing*, 7(3–4):197–387, 2014.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*, 2020.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

Shizhe Diao, Pengcheng Wang, Yong Lin, Hanxiao Tong, and Tong Zhang. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*, 2023.

Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, pages 11583–11592, 2022.

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

Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, pages 5882–5889, 2018.

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, Jung-Su Ha, and Marc Toussaint. Deep visual reasoning: Learning to predict action sequences for task and motion planning from an initial scene image. In *RSS*, 2020.

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.

John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems*, 8(2):1–37, 2018.

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.

Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Manipulathor: A framework for visual object manipulation. *CVPR*, 2021.

Howard Eichenbaum. A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience*, 1(1):41–50, 2000.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, 2014.

Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017.

Jakob Engel et al. Lsd-slam: Large-scale direct monocular slam. *European Conference on Computer Vision*, 2014.

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

ESRI. Arcgis pro: Professional gis software. `https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview`, 2024a.

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

Scott Ettinger et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. *IEEE International Conference on Computer Vision*, 2021.

Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. *IUI*, pages 39–45, 2003.

Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. *CVPR*, 2022.

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, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. *CVPR*, pages 2139–2147, 2018.

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.

Olivier Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, 1993.

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

César Ferri, José Hernández-Orallo, and R Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38, 2009.

Richard E Fikes and Nils J Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208, 1971.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *CoRL*, 2022.

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.

David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. Prentice Hall, 2011.

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.

Foursquare. Foursquare places: The world's most trusted location data. Technical report, Foursquare Labs Inc., 2024.

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

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *ICLR*, 2019.

Sara Fridovich-Keil et al. Plenoxels: Radiance fields without neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 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.

Jorge Fuentes-Pacheco, José Ruiz-Ascencio, and Juan Manuel Rendón-Mancha. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, 43(1):55–81, 2015.

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, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning. In *ICAPS*, pages 440–448, 2020.

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

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):744–755, 2014.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Hector Geffner and Blai Bonet. *A concise introduction to models and methods for automated planning*. Morgan and Claypool Publishers, 2013.

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.

Samuel J Gershman and Nathaniel D Daw. Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, 68:101–128, 2017.

Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 2023.

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.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *Low-Power Computer Vision*, pages 291–326, 2022.

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

Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. Human-centred machine learning. *CHI Extended Abstracts*, pages 3558–3565, 2016.

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

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023.

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. Google earth engine: Planetary-scale geospatial analysis. `https://earthengine.google.com/`, 2024a.

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

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.

Daniel Gordon, Abhishek Kadian, Devi Parikh, Judy Hoffman, and Dhruv Batra. Splitnet: Sim2sim and task2task transfer for embodied visual navigation. In *ICCV*, pages 1022–1031, 2019.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3.2 collection: Multimodal open foundation models. *arXiv preprint*, 2024.

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

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *CSCW*, 3: 1–24, 2019.

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.

Pierre-Louis Guhur et al. Airbert: In-domain pretraining for vision-and-language navigation. In *ICCV*, 2021.

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.

Kan Guo, Yongli Hu, Zhen Qian, Yanfeng Sun, Junbin Gao, and Baocai Yin. Hierarchical graph convolution network for traffic forecasting. *AAAI*, 35(1):151–159, 2021.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. *ICML*, pages 3887–3896, 2020.

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.

Xudong Guo et al. Embodied llm agents learn to cooperate in organized teams. *arXiv preprint arXiv:2403.12482*, 2024c.

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*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *ICML*, pages 2555–2565, 2019b.

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.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

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.

David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *ICLR*, 2024.

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.

Kaiming He et al. Masked autoencoders are scalable vision learners. *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Mary Hegarty. Spatial thinking in undergraduate science education. *Spatial Cognition and Computation*, 6 (3):209–223, 2006.

Malte Helmert. The fast downward planning system. *Journal of Artificial Intelligence Research*, 26:191–246, 2006.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*, 2022.

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

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. A survey of multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Daniel Ho, Kanishka Rao, Zhuo Xu, Eric Jang, Mohi Khansari, and Yunfei Bai. Retinagan: An object-aware approach to sim-to-real transfer. *ICRA*, pages 10920–10926, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *NeurIPS*, 35:30016–30030, 2022.

Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.

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.

John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *CoRL*, pages 409–418, 2021.

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.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. *SIGGRAPH*, 2024a.

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

Jiangyong Huang et al. An embodied generalist agent in 3d world. *International Conference on Machine Learning*, 2024b.

Jin Huang et al. Can llms effectively leverage graph structural information: When and why. *arXiv preprint arXiv:2309.16595*, 2024c.

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*, 2024d.

Yushi Huang et al. Visual instruction tuning. *arXiv preprint*, 2023c.

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.

Michael Hulbert, Matthias Feurer, and Frank Hutter. Using large language models for hyperparameter optimization. *arXiv preprint arXiv:2312.04528*, 2024.

Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

Ismail Ishak et al. The role of spatial intelligence in engineering education. *International Journal of Engineering Education*, 24(4):714, 2008.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *JMLR*, 24(251):1–43, 2023.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *CVPR*, pages 2704–2713, 2018.

Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orber, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *CVPR*, pages 2989–2998, 2023.

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.

Samireh Jalali and Claes Wohlin. Systematic literature studies: database searches vs. backward snowballing. *ESEM*, pages 29–38, 2012.

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.

Eric Jang et al. Bc-z: Zero-shot task generalization with robotic imitation learning. In *CoRL*, 2022.

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.

Nathalie Japkowicz and Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

Herve Jegou et al. Product quantization for nearest neighbor search. *IEEE TPAMI*, 2011.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

Yuhao Jiang et al. Deepurbandownscale: A physics informed deep learning framework for high-resolution urban surface temperature estimation from coarse-scale satellite observations. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:1–15, 2021.

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.

Ryan Julian, Benjamin Swanson, Gaurav S Sukhatme, Sergey Levine, Chelsea Finn, and Karol Hausman. Scaling data-driven robotics with reward sketching and batch reinforcement learning. *RSS*, 2020.

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.

Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *CoRL*, 2021.

Aishwarya Kamath, Jack Hessel, and Kai-Wei Chang. What's left? concept grounding with logic-enhanced foundation models. *Advances in Neural Information Processing Systems*, 2023.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Henry Kautz and Bart Selman. Planning as satisfiability. *ECAI*, 92:359–363, 1992.

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.

Remi Kesten, Mayank Usman, John Houston, et al. Lyft level 5 av dataset. *Lyft*, 2019.

Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *CVPR*, 2022.

Urvashi Khandelwal et al. Generalization through memorization: Nearest neighbor language models. In *ICLR*, 2020.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *ICLR*, 2023.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36, 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.

Ue-Hwan Kim, Jin-Man Park, Taek-Jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. In *IEEE transactions on cybernetics*, volume 50, pages 4921–4933, 2019.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pages 2688–2697, 2018.

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.

Rob Kitchin. *The real-time city? Big data and smart urbanism*, volume 79. 2014.

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

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *ECML*, 2006.

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.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. In *NeurIPS*, pages 10117–10126, 2018.

Kurt Konolige, Motilal Agrawal, Robert C Bolles, Cregg Cowan, Martin Fischler, and Brian Gerkey. Outdoor mapping and navigation using stereo vision. In *Experimental Robotics*, pages 179–190. Springer, 2008.

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.

Devin Kreuzer et al. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 2021.

Hans-Peter Kriegel, Peer Kroger, Jorg Sander, and Arthur Zimek. Spatial data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):1–13, 2011.

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.

Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *JMLR*, 2021.

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.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019.

Pat Langley, John E Laird, and Seth Rogers. Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2):141–160, 2009.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

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

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

Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 2024a.

Yann LeCun. Objective-driven ai. *Nature Machine Intelligence*, 2024b.

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

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *ICML Workshop on challenges in representation learning*, 3(2):896, 2013.

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.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *IJRR*, 34(4-5): 705–724, 2015.

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.

Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. Learning multi-level hierarchies with hindsight. In *ICLR*, 2019.

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.

Chengshu Li, Fei Xia, Roberto Martín-Martín, and Silvio Savarese. Hrl4in: Hierarchical reinforcement learning for interactive navigation with mobile manipulators. *CoRL*, pages 603–616, 2020.

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

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

Guohao Li et al. Camel: Communicative agents for mind exploration of large language model society. *arXiv preprint arXiv:2303.17760*, 2023b.

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*, 2023c.

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

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

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.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *ICLR*, 2016.

Yunzhu Li, Jiajun Wu, Jun-Yan Zhu, Joshua B Tenenbaum, Antonio Torralba, and Russ Tedrake. Propagation networks for model-based control under partial observation. In *IEEE International Conference on Robotics and Automation*, pages 1205–1211, 2019.

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.

Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022.

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.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations of multimodal learning. *arXiv preprint arXiv:2209.03430*, 2024.

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.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *ICLR*, 2016.

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.

Fangchen Liu et al. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *RSS*, 2024b.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023b.

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

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

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*, 2023d.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Luo, Yang Gao, and Zhiwei Xiong. Aligning cyber space with physical world: A comprehensive survey on embodied ai. In *arXiv preprint arXiv:2407.06886*, 2024d.

Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. *CVPR*, 2019.

Zhihan Liu, Hao Hu, Shenao Zhang, Hongyi Guo, Shuqi Ke, Boyi Liu, and Zhaoran Wang. Reason for future, act for now: A principled framework for autonomous llm agents with provable sample efficiency. *ICML*, 2024e.

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.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*, pages 6379–6390, 2017.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016.

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.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. In *ICLR*, 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.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. *ICLR*, 2019a.

Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019b.

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*, 2022a.

Xu Ma et al. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In *ICLR*, 2022b.

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.

Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *NeurIPS*, 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.

Ajay Mandlekar, Danfei Xu, Josiah Wong, et al. What matters in learning from offline human demonstrations for robot manipulation. 2021.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *CoRL*, pages 1678–1690, 2022.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

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.

Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, pages 7210–7219, 2021.

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.

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

James L McClelland, Bruce L McNaughton, and Randall C OReilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

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

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *TMLR*, 2023.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *ICLR*, 2023.

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.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Arber, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Marvin Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *FAT**, pages 220–229, 2019.

Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, pages 1928–1937, 2016.

Thomas M Moerland, Joost Broekens, Aske Plaat, and Catholijn M Jonker. Model-based reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 16(1):1–118, 2023.

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.

Hans P Moravec. Sensor fusion in certainty grids for mobile robots. *AI Magazine*, 9(2):61, 1988.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. *AAAI*, 33(01): 4602–4609, 2019.

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.

Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annual review of neuroscience*, 31:69–89, 2008.

Tongzhou Mu et al. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In *NeurIPS*, 2021.

Yao Mu, Qinglong Zhang, Mengkang Hu, et al. Embodiedgpt: Vision-language pre-training via embodied chain of thought. 2023.

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.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, 2017.

Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017.

Raul Mur-Artal et al. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 2015.

Fabio Muratore, Fabio Ramos, Greg Turk, Wenhao Yu, Michael Gienger, and Jan Peters. Robot learning from randomized simulations: A review. *Frontiers in Robotics and AI*, 9:799893, 2022.

Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *ICRA*, pages 1374–1381, 2015.

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

Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, pages 8688–8697, 2019.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

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.

Dana S Nau, Tsz-Chiu Au, Okhtay Ilghami, Ugur Kuter, J William Murdock, Dan Wu, and Fusun Yaman. Shop2: An htn planning system. *Journal of artificial intelligence research*, 20:379–404, 2003.

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

Allen Newell, J Cliff Shaw, and Herbert A Simon. The logic theory machine–a complex information processing system. *IRE Transactions on Information Theory*, 2(3):61–79, 1956.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. *ICML*, 2011.

Richard Ngo, Lawrence Chan, and Sören Geiger. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Nils J Nilsson. Shakey the robot. *Technical Note 323, AI Center, SRI International*, 1984.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Biber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. In *arXiv preprint arXiv:2112.00114*, 2021.

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

Chitu Okoli and Kira Schabram. A guide to conducting a systematic literature review of information systems research. *Sprouts: Working Papers on Information Systems*, 10(26), 2010.

Theo X Olausson et al. Is self-repair a silver bullet for code generation? *arXiv preprint arXiv:2306.09896*, 2023.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *NeurIPS*, 31, 2018.

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.

OpenAI et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022.

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.

Abhishek Padalkar et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

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

Palantir Technologies. Palantir aip: Artificial intelligence platform. `https://www.palantir.com/platforms/aip/`, 2024.

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.

Cheonbok Park, Chunggi Lee, Hyojin Bahng, Yunwon Tae, Seungmin Jin, Kihwan Kim, Sungahn Ko, and Jaegul Choo. St-grat: A novel spatio-temporal graph attention networks for accurately forecasting dynamically changing road speed. In *CIKM*, 2020.

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.

Planet Labs PBC. Planet daily monitoring: Global earth observation. `https://www.planet.com/products/monitoring/`, 2024.

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.

Simon JD Prince. *Computer vision: models, learning, and inference.* Cambridge University Press, 2012.

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.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. *arXiv preprint arXiv:1806.07011*, 2018.

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.

Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop: History-and-order aware pre-training for vision-and-language navigation. *CVPR*, 2022.

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.

Ori Ram et al. In-context retrieval-augmented language models. *TACL*, 2023.

Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. 2021.

Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18890–18900, 2022.

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

David A Randell, Zhan Cui, and Anthony G Cohn. A spatial logic based on regions and connection. *KR*, 92:165–176, 1992.

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.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*, pages 4295–4304, 2018.

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

H Ravichandar, A S Polydoros, S Chernova, and A Billard. Recent advances in robot learning from demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:297–330, 2020.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *TMLR*, 2022.

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. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

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

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 4th edition, 2021.

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.

Earl D Sacerdoti. Planning in a hierarchy of abstraction spaces. *Artificial intelligence*, 5(2):115–135, 1974.

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

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, pages 8459–8468, 2020.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *CVPR*, pages 4510–4520, 2018.

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.

Daniel L Schacter, Anthony D Wagner, and Randy L Buckner. *Cognitive neuroscience of memory*. MIT Press, 2007.

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.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *NeurIPS*, 30, 2017.

Amanda Seed and Richard Byrne. Cognition in tool use. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1545):1405–1412, 2010.

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

Dhruv Shah, Michael Equi, Blazej Osinski, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning. *CoRL*, 2023a.

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*, 2023b.

Mohit Sharma et al. Vima: General robot manipulation with multimodal prompts. In *arXiv preprint*, 2022.

Ahsan Shehzad, Feng Xia, Shagufta Abid, Chao Peng, Shuo Yu, Dongyu Zhang, and Karin Verspoor. Graph transformers: A survey. *arXiv preprint arXiv:2407.09777*, 2024.

Bokui Shen, Fei Xia, Chengshu Li, Roberto Martin, Linxi Fan, Guanzhi Wang, Shyamal Buch, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021.

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.

Weijia Shi et al. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

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, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. *arXiv preprint arXiv:1912.01734*, 2020a.

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

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799, 2023.

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

Robert W Shumaker, Kristina R Walkup, and Benjamin B Beck. *Animal tool behavior: the use and manufacture of tools by animals*. JHU Press, 2011.

Roland Siegwart, Illah Reza Nourbakhsh, and Davide Scaramuzza. *Introduction to autonomous mobile robots*. MIT press, 2011.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

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.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Avi Singh, Huihan Liu, Gaoyue Zhou, Albert Yu, Nicholas Rhinehart, and Sergey Levine. Parrot: Data-driven behavioral priors for reinforcement learning. *ICLR*, 2021.

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

Hannah Snyder. Literature review as a research methodology: An overview and guidelines. *Journal of business research*, 104:333–339, 2019.

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

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. *ICML*, pages 5887–5896, 2019.

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.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. *AAAI*, 34(01):914–921, 2020.

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

Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. *NeurIPS*, 25, 2012.

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.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv preprint arXiv:2402.08115*, 2024.

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

Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662, 2009.

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.

Gencer Sumbul, Marcela Charfuelan, Begum Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *arXiv preprint arXiv:1902.06148*, 2019.

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

Pei Sun et al. Scalability in perception for autonomous driving: Waymo open dataset. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. *AAMAS*, pages 2085–2087, 2018.

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

Rich Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1), 2019.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

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.

Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*, 2023.

Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *NAACL*, 2019.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. *CVPR*, 2022.

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.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised learning results. In *NeurIPS*, pages 1195–1204, 2017.

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.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems*, 2021.

Zachary Teed, Lahav Lipson, and Jia Deng. Deep visual odometry with adaptive memory. *TPAMI*, 2024.

Andreas ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp pose detection in point clouds. *The International Journal of Robotics Research*, 36(13-14):1455–1473, 2017.

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

Tesla. Tesla autopilot and full self-driving: Ai-powered driver assistance. Technical report, Tesla Inc., 2024.

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

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. In *arXiv preprint arXiv:1802.08219*, 2018.

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

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.

Michael Tomasello. *The cultural origins of human cognition*. Harvard University Press, 1999.

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.

David Tranfield, David Denyer, and Palminder Smart. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, 14(3): 207–222, 2003.

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.

Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2): 41–57, 2016.

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.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *NeurIPS*, 2024.

Petar Veličković. Everything is connected: Graph neural networks. *Current Opinion in Structural Biology*, 79:102538, 2023.

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.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

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

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

Homer Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. *CoRL*, 2024.

Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *CSCW*, 3:1–24, 2019a.

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.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *ICLR*, 2024a.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 2024b.

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

Weihan Wang, Qingsong Sun, Zhe Lv, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024c.

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

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, 2019b.

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*, 2019c.

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

Waymo. Waymo driver: Autonomous driving technology. https://waymo.com/waymo-driver/, 2024.

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

Jane Webster and Richard T Watson. Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, pages xiii–xxiii, 2002.

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.

Jerry Wei et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, 2019.

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

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

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

Wherobots. Apache sedona: Spatial data system. https://wherobots.com/apache-sedona/, 2024.

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.

Benjamin Wilson et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *Advances in Neural Information Processing Systems*, 2023.

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

Michael Wooldridge. *An Introduction to MultiAgent Systems*. Wiley, 2nd edition, 2009.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. *ECCV*, pages 720–736, 2022.

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.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *NeurIPS*, 2024.

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.

Shun-Cheng Wu et al. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *CVPR*, 2021.

Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. *CVPR*, 2019a.

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, 2019b.

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.

Gui-Song Xia et al. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *CVPR*, pages 11097–11107, 2020.

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.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

Sang Michael Xie et al. An explanation of in-context learning as implicit bayesian inference. In *ICLR*, 2022.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. Self-evaluation guided beam search for reasoning. *NeurIPS*, 2023.

Danfei Xu et al. Scene graph generation by iterative message passing. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. *ICML*, pages 5453–5462, 2018.

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

Mengda Xu, Zhenjia Huang, Yinghao Li, and Shuran Song. Xskill: Cross embodiment skill discovery. *CoRL*, 2023a.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023b.

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

Qiantong Xu, Fenglu Hong, Bo Li, Changran Hu, Zhengyu Chen, and Jian Zhang. On the tool manipulation capability of open-source large language models. *arXiv preprint arXiv:2305.16504*, 2023d.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *ECCV*, 2024.

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.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

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.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. *arXiv preprint arXiv:1908.02660*, 2020a.

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*, 2023a.

Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *CHI*, pages 1–13, 2020b.

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.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023b.

Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. Deep multi-view spatial-temporal network for taxi demand prediction. In *AAAI*, 2018.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2024.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yue Feng, Le Xue, Rithesh Murber, Zeyuan Chen, Jianguo Zhang, Devansh Arber, et al. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*, 2023d.

Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary tasks and exploration enable objectgoal navigation. In *ICCV*, 2021.

Qinghao Ye, Haiyang Xu, Jiabo Ye, et al. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

Ruosong Ye et al. Language is all a graph needs. *arXiv preprint arXiv:2308.07134*, 2024.

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.

Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021.

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.

Naoki Yokoyama, Dhruv Batra, et al. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. *arXiv preprint arXiv:2312.03275*, 2024.

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.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Baez, Claudio Fischetti, Maxim Egorov, and Cathy Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *NeurIPS*, 35:24611–24624, 2022a.

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*, 2022b.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. *CVPR*, 2024.

Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. *CVPR*, 2022.

Yifu Yuan, Zhizheng Xiong, and Qin Wang. Deep learning for multiple object tracking: a survey. *IET Computer Vision*, 14(8):601–612, 2020a.

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*, 2020b.

Fabio Massimo Zanzotto. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3301–3366, 2019.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *CVPR*, pages 6720–6731, 2019.

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

Jiawei Zha et al. How to enable llm with 3d capacity? a survey of spatial intelligence in large language models. 2025.

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.

Haotian Zhang, Haochen Zhang, Pengchuan Zhao, Haoxuan You, Yinfei Jiang, Jianwei Lee, Trevor Darrell, and Jianfeng Gao. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024a.

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

Junbo Zhang et al. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*, 2017.

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

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, 2021a.

Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.

Sixian Zhang, Xinhang Song, Yubing Bai, Weijie Li, Yakun Chu, and Shuqiang Jiang. Hierarchical object-to-zone graph for object navigation. *ICCV*, 2021b.

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.

Zhuosheng Zhang et al. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2023.

Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):249–270, 2020.

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.

Jianan Zhao et al. Graphtext: Graph reasoning in text space. *arXiv preprint arXiv:2310.01089*, 2023a.

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

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. Trajectory data mining: An overview. *ACM TIST*, 2015.

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.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. *ECCV*, pages 696–712, 2022.

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

Jie Zhou et al. Graph neural networks: A review of methods and applications. *AI Open*, 2020.

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.

Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 2018.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023a.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *CVPR*, pages 10012–10022, 2020.

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.

Xiaojin Jerry Zhu. Semi-supervised learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*, 2005.

Yixin Zhu, Yibiao Zhao, and Song-Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. *CVPR*, pages 2855–2864, 2015.

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

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. *ICCV*, 2023b.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *CoRL*, 2023.