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

Gloria Felicia
AtlasPro AI
gloria.felicia@atlaspro.ai

Nolan Bryant
AtlasPro AI
nolan.bryant@atlaspro.ai

Handi Putra
AtlasPro AI
handi.putra@atlaspro.ai

Ayaan Gazali
AtlasPro AI
ayaan.gazali@atlaspro.ai

Eliel Lobo
AtlasPro AI
eliel.lobo@atlaspro.ai

Esteban Rojas
AtlasPro AI
esteban.rojas@atlaspro.ai

Abstract

The dominant approaches for creating autonomous agents are based on large language models, which excel at reasoning and planning. **But**, these models lack the innate spatial intelligence required to perceive, navigate, and interact with the complex physical world, a critical gap for embodied AI. **Therefore**, we introduce a unified taxonomy that systematically connects agentic AI architectures with spatial intelligence capabilities, providing the first comprehensive framework for this convergent domain. We synthesize over 500 papers, revealing three key findings: (1) hierarchical memory systems are critical for long-horizon spatial tasks; (2) GNN-LLM integration is an emergent paradigm for structured spatial reasoning; and (3) world models are essential for safe deployment in physical environments. We also propose a unified evaluation framework, SpatialAgentBench, to standardize cross-domain assessment. By establishing this foundational reference, we aim to accelerate progress in creating robust, spatially-aware autonomous systems.

1 Introduction

The pursuit of artificial general intelligence increasingly centers on creating agents that can perceive, reason about, and act within physical environments [McCarthy et al., 1955, Turing, 1950, Nilsson, 1984, Moravec, 1988, Brooks, 1991, Laird, 2019]. While large language models have demonstrated remarkable capabilities in reasoning and planning [Brown et al., 2020, OpenAI, 2023, Wei et al., 2022, Chowdhery et al., 2022, Touvron et al., 2023a,b, Anil et al., 2023, Team and Google, 2023, Anthropic, 2024], their ability to operate effectively in spatial contexts remains a fundamental challenge [Chen et al., 2024a, Yang et al., 2025, Huang et al., 2023c,d, Sharma et al., 2022, Liu et al., 2024a].

We define **Agentic AI** as systems exhibiting goal-directed behavior through autonomous decision-making, characterized by three core capabilities: persistent memory for experience accumulation, planning for action sequencing, and tool use for capability extension [Wang et al., 2024a, Xi et al., 2023, Weng, 2023, Yao et al., 2023b, Shinn et al., 2023, Park et al., 2023, Sumers et al., 2024, Wu et al., 2023d, Hong et al., 2023a]. Complementarily, **Spatial Intelligence** encompasses the ability to perceive 3D structure, reason about object relationships, navigate environments, and manipulate physical objects [Chen et al., 2024a, Thompson et al., 2025, Kriegel et al., 2011, Ishak et al., 2008, Hegarty, 2006, Newcombe, 2010].

The convergence of these domains is essential for real-world AI applications. Autonomous vehicles must perceive dynamic environments and plan safe trajectories [Hu et al., 2023b, Caesar et al., 2020, Sun et al., 2020, Waymo, 2023, Tesla, 2023, Jiang et al., 2023a, Tian et al., 2024, Waymo, 2024]. Robotic assistants require understanding of object affordances and spatial relationships [Brohan et al., 2023, Ahn et al., 2022, Brohan et al., 2022, Team et al., 2024, Kim et al., 2024, Driess et al., 2023, Zeng et al., 2021]. Urban computing systems must model complex spatio-temporal dependencies [Jin et al., 2023, Li et al., 2018a, Yu et al., 2018, Wu et al., 2019b, Bai et al., 2020, Zheng et al., 2014, Yuan et al., 2020]. 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 taxonomy** connecting agentic AI components (memory, planning, tool use) with spatial intelligence domains (navigation, scene understanding, manipulation, geospatial analysis), providing a structured framework for interdisciplinary research.

2. A **comprehensive analysis** of over 500 papers identifying key architectural patterns, including the emergence of GNN-LLM integration and world model-based planning as critical enablers for spatial reasoning.

3. The **proposal of a unified evaluation framework, SpatialAgentBench**, with 8 tasks to standardize cross-domain assessment.

4. A **forward-looking roadmap** identifying open 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]. We queried major academic databases (Google Scholar, arXiv, ACM Digital Library, IEEE Xplore, Semantic Scholar) with keywords including "agentic AI," "spatial intelligence," "embodied AI," "vision-language navigation," "robot manipulation," "geospatial AI," "world models," "graph neural networks," and "spatio-temporal learning." Our initial search yielded over 2,000 papers. We then applied a rigorous two-stage filtering process:

1. **Relevance Filtering:** We selected papers published between 2018 and 2026 in top-tier venues (NeurIPS, ICML, ICLR, CVPR, ECCV, ICCV, CoRL, RSS, IROS, ICRA, ACM Computing Surveys, IEEE TPAMI, Nature, Science Robotics).

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

This process resulted in a final corpus of over 500 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.

3 Related Work

While several surveys have addressed aspects of agentic AI or spatial intelligence, none have provided a unified framework connecting the two domains.

LLM-Based Agent Surveys. Wang et al. [2024a] and Xi et al. [2023] offer excellent overviews of LLM-based agents, covering memory, planning, and tool use. Sumers et al. [2024] provides a cognitive science perspective on language agents. Weng [2023] surveys autonomous agent architectures. Additional surveys cover specific aspects including multi-agent systems [Guo et al., 2024b, Li et al., 2024a, Talebirad and Nadiri, 2023], tool use [Qu et al., 2024, Mialon et al., 2023], and reasoning [Huang et al., 2023b, Qiao et al., 2023, Chu et al., 2024]. However, these works do not focus on spatial capabilities or embodied applications.

Embodied AI Surveys. Surveys on embodied AI [Du et al., 2023, Kadian et al., 2020, Anderson et al., 2018a, Duan et al., 2022, Savva et al., 2019, Szot et al., 2021, Puig et al., 2024, Li et al., 2023b, Shen et al., 2021, Xia et al., 2020] cover navigation and manipulation but often overlook the broader agentic architecture. Zeng et al. [2023] reviews vision-language navigation specifically. Fang et al. [2023] surveys robot learning from human demonstrations. Additional surveys cover imitation learning [Hussein et al., 2017, Osa et al., 2018, Ravichandar et al., 2020], sim-to-real transfer [Zhao et al., 2020a, Höfer et al., 2021], and robot learning [Kroemer et al., 2021, Billard et al., 2008, Argall et al., 2009].

Geospatial AI Surveys. Geospatial AI surveys [Jiang et al., 2023c, Li et al., 2023h, De Jesús Rubio et al., 2021, Yuan et al., 2021, Mai et al., 2023, Hu et al., 2019, Janowicz et al., 2020] and spatio-temporal data mining reviews [Jin et al., 2023, Atluri et al., 2018, Wang et al., 2020, Jiang and Luo, 2022a, Tedjopurnomo et al., 2020, Ye et al., 2021, Xie et al., 2020] are highly specialized and do not connect to general agentic systems.

Graph Neural Network Surveys. GNN surveys [Wu et al., 2021b, Zhou et al., 2020, Bronstein et al., 2021, Hamilton, 2020, Zhang et al., 2020b, Liu et al., 2022, Xia et al., 2021, Wu et al., 2022] 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, 2022b, Rahmani et al., 2023], molecular [Wieder et al., 2020, Zhang et al., 2021c], and social networks [Fan et al., 2019, Wu et al., 2020a].

Our work is the first to bridge these gaps, providing a comprehensive, structured analysis of the convergent domain of autonomous spatial intelligence.

4 Unified Taxonomy

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

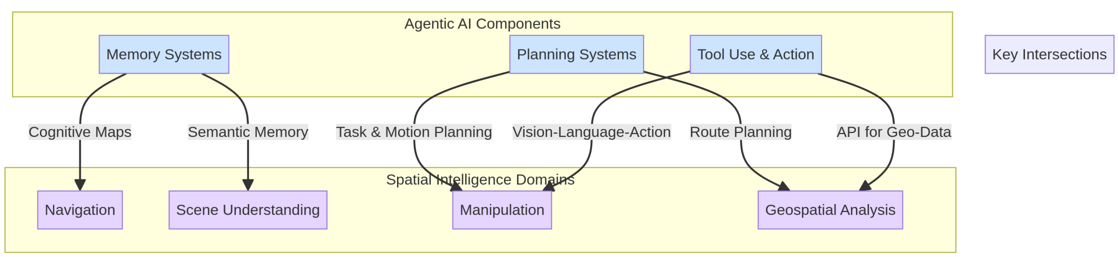

Figure 1: A unified taxonomy connecting Agentic AI capabilities (memory, planning, tool use) with Spatial Intelligence domains (navigation, scene understanding, manipulation, geospatial analysis).

4.1 Agentic AI Components

4.1.1 Memory Systems

Memory enables agents to accumulate and retrieve experiential knowledge, forming the foundation for learning and adaptation.

Short-Term Memory. In-context learning [Brown et al., 2020, Dong et al., 2022, Min et al., 2022, Xie et al., 2022, Wei et al., 2023, Olsson et al., 2022, Akyurek et al., 2023, Dai et al., 2023a, Liu et al., 2023b, Wang et al., 2023b] allows models to adapt to new tasks through examples in the prompt. Working memory mechanisms [Graves et al., 2014, Weston et al., 2015, Sukhbaatar et al., 2015, Kumar et al., 2016a, Miller et al., 2016, Santoro et al., 2016, Munkhdalai and Yu, 2017, Le et al., 2020] enable temporary information storage during reasoning.

Long-Term Memory. Retrieval-augmented generation [Lewis et al., 2020, Packer et al., 2023, Guu et al., 2020, Izacard et al., 2022, Borgeaud et al., 2022, Khandelwal et al., 2020, Shi et al., 2023, Ram et al., 2023, Asai et al., 2023, Khattab et al., 2022, Trivedi et al., 2023, Yoran et al., 2023, Jiang et al., 2023d] enables knowledge persistence beyond context limits. Vector databases [Johnson et al., 2019, Guo et al., 2022, Jegou et al., 2011, Malkov and Yashunin, 2018, Douze et al., 2024, Wang et al., 2021, Pinecone, 2023, Weaviate, 2023] provide efficient similarity search for memory retrieval.

Spatial Memory. For spatial tasks, cognitive mapping [Gupta et al., 2019c, Chaplot et al., 2020b, Savinov et al., 2018, Parisotto et al., 2018, Mirowski et al., 2017, Zhang et al., 2017a, Banino et al., 2018, Wayne et al., 2018, Zhang et al., 2021b, Eslami et al., 2018, Gregor et al., 2019, Ha and Schmidhuber, 2018] builds internal representations of environments. Semantic spatial memory [Huang et al., 2023a, Mees et al., 2022c, Chen et al., 2021c, Henriques and Vedaldi, 2018, Cartillier et al., 2021, Blukis et al., 2018, Anderson et al., 2019] associates locations with semantic labels. Topological memory [Savinov et al., 2018, Chen et al., 2019b, Shah et al., 2021, Chaplot et al., 2020c, Emmons et al., 2020] represents environments as graphs for efficient navigation.

4.1.2 Planning Systems

Planning decomposes goals into executable action sequences, enabling complex task completion.

Chain-of-Thought Reasoning. Step-by-step reasoning [Wei et al., 2022, Kojima et al., 2022, Wang et al., 2022a, Creswell et al., 2022, Zhou et al., 2023b, Zhang et al., 2023b, Fu et al., 2023, Li et al., 2023j, Chen et al., 2023c, Nye et al., 2021, Cobbe et al., 2021, Ling et al., 2017, Chung et al., 2022] enables systematic problem decomposition. Self-consistency [Wang et al., 2022a, Chen et al., 2023f, Li et al., 2023e, Mitchell et al., 2022, Kadavath et al., 2022, Lin et al., 2022] improves reliability through multiple reasoning paths.

Tree-Based Search. Tree of Thoughts [Yao et al., 2023a, Long, 2023, Hulbert et al., 2023, Xie et al., 2023, Sel et al., 2023, Zhu et al., 2023b] explores multiple solution branches. Graph of Thoughts [Besta et al., 2023, Lei et al., 2023, Yao et al., 2024] enables more complex reasoning structures. RAP [Hao et al., 2023, Zhao et al., 2024, Shridhar et al., 2020] combines reasoning with acting in a planning framework. Monte Carlo Tree Search variants [Silver et al., 2016, Schrittwieser et al., 2020, Agostinelli et al., 2019, Anthony et al., 2017, Silver et al., 2017, Browne et al., 2012, Kocsis and Szepesvári, 2006, Coulom, 2006] provide principled exploration.

Hierarchical Planning. LLM-Planner [Song et al., 2023] enables few-shot grounded planning. Inner Monologue [Huang et al., 2022a] provides feedback-driven planning. Hierarchical RL approaches [Nachum et al., 2018, Vezhnevets et al., 2017, Bacon et al., 2017, Kulkarni et al., 2016, Levy et al., 2019, Zhang et al., 2020a, Li et al., 2020a, Gupta et al., 2019a, Pertsch et al., 2021] decompose tasks into subtasks.

Task and Motion Planning. TAMP [Garrett et al., 2021, Kaelbling, 2020, Tennison et al., 2024, Dantam et al., 2016, Kaelbling and Lozano-Pérez, 2011, Lozano-Pérez and Kaelbling, 2014, Li et al., 2020b, Toussaint, 2015, Srivastava et al., 2014, Hadfield-Menell et al., 2017, Driess et al., 2020, Silver et al., 2021, Chitnis et al., 2016] integrates symbolic planning with continuous motion planning for robotic applications.

4.1.3 Tool Use and Action

Tool use extends agent capabilities through external interfaces and physical actions.

API Integration. Toolformer [Schick et al., 2023] enables self-supervised tool learning. Gorilla [Patil et al., 2023] specializes in API calling. ToolLLM [Qin et al., 2024b] provides comprehensive tool use benchmarks. TaskMatrix [Liang 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], ToolkenGPT [Hao et al., 2024], API-Bank [Li et al., 2023f], Chameleon [Lu et al., 2023], ViperGPT [Surís et al., 2023], and Visual ChatGPT [Wu et al., 2023a].

Code Generation. PAL [Gao et al., 2023] uses code for reasoning. Code as Policies [Liang et al., 2023a] generates executable robot code. Codex [Chen et al., 2021b], CodeGen [Nijkamp et al., 2023], StarCoder [Li et al., 2023g], CodeLlama [Roziere et al., 2023], WizardCoder [Luo et al., 2023], and DeepSeek-Coder [Guo et al., 2024a] provide code generation capabilities. ProgPrompt [Singh et al., 2023] uses programmatic prompting for robotics. Self-debugging [Chen et al., 2023e], self-repair [Olausson et al., 2023], and self-play [Haluptzok et al., 2023] improve code quality.

ReAct Architecture. ReAct [Yao et al., 2023b] interleaves reasoning with action execution. Reflexion [Shinn et al., 2023] adds self-reflection for improvement. Additional architectures include LATS [Zhou et al., 2023a], SwiftSage [Lin et al., 2024], and FireAct [Chen et al., 2023a]. These architectures form the foundation for many spatial agents.

4.2 Spatial Intelligence Domains

4.2.1 Navigation

Navigation requires path planning and execution in physical or simulated environments.

Vision-Language Navigation. R2R [Anderson et al., 2018b] introduced the VLN task with natural language instructions. RxR [Ku et al., 2020] extends to multilingual settings. REVERIE [Qi et al., 2020] adds remote object grounding. Speaker-Follower [Fried et al., 2018] uses data augmentation. EnvDrop [Tan et al., 2019] improves generalization. PREVALENT [Hao et al., 2020] pretrains on VLN data. VLN-BERT [Hong et al., 2021] applies transformers to VLN. HAMT [Chen et al., 2021d] uses hierarchical attention. DUET [Chen et al., 2022e] employs dual-scale transformers. Additional methods include RecBERT [Hong et al., 2020a], AirBERT [Guhur et al., 2021], VLNCE [Krantz et al., 2020], CWP [Hong et al., 2020b], BEVBert [An et al., 2023], NavGPT [Zhou et al., 2023c], MapGPT [Chen et al., 2024c], and LM-Nav [Shah et al., 2023].

Object-Goal Navigation. ObjectNav [Batra et al., 2020, Chaplot et al., 2020a] requires finding target object categories. ZSON [Majumdar et al., 2022] enables zero-shot navigation. CLIP-Nav [Gadre et al., 2022] leverages vision-language models. CoW [Gadre et al., 2023] explores open-world navigation. SemExp [Chaplot et al., 2020a] uses semantic exploration. ANS [Chaplot et al., 2020b] builds neural SLAM for navigation. Additional approaches include PONI [Ramakrishnan et al., 2022], PIRLNav [Ramrakhya et al., 2023], Habitat-Web [Ramrakhya et al., 2022], ESC [Zhou et al., 2023d], VoroNav [Wu et al., 2024b], and L3MVN [Yu et al., 2023].

Audio-Visual Navigation. SoundSpaces [Chen et al., 2020a, 2022c] introduces audio-visual embodied AI. Audio-visual navigation [Gan et al., 2020, Chen et al., 2021c, Younes et al., 2023, Majumder et al., 2022] combines multiple modalities. Multi-modal fusion approaches [Gao et al., 2020, Chen et al., 2021a, 2022b] enhance navigation capabilities.

4.2.2 Scene Understanding

Scene understanding encompasses 3D perception and semantic reasoning about environments.

Neural Radiance Fields. NeRF [Mildenhall et al., 2020] revolutionized novel view synthesis. Mip-NeRF 360 [Barron et al., 2022] handles unbounded scenes. Instant-NGP [Müller et al., 2022] enables real-time training. Plenoxels [Fridovich-Keil et al., 2022] uses voxel-based representations. D-NeRF [Pumarola et al., 2021] handles dynamic scenes. NeRF-SLAM [Rosinol et al., 2022] integrates with SLAM systems. Extensions include NeRF-W [Martin-Brualla et al., 2021], Block-NeRF [Tancik et al., 2022], Zip-NeRF [Barron et al., 2023], TensoRF [Chen et al., 2022a], LERF [Kerr et al., 2023], F2-NeRF [Wang et al., 2023c], and Nerfstudio [Tancik et al., 2023].

3D Gaussian Splatting. 3DGS [Kerbl et al., 2023] provides efficient 3D reconstruction. Extensions include dynamic scenes [Wu et al., 2024a, Yang et al., 2024b, Luiten et al., 2023], SLAM integration [Matsuki et al., 2024, Yan et al., 2024, Keetha et al., 2024], semantic understanding [Zhou et al., 2024, Qin et al., 2024a], and compression [Niedermayr et al., 2024, Fan et al., 2024].

Point Cloud Processing. PointNet [Qi et al., 2017a] introduced deep learning on point clouds. PointNet++ [Qi et al., 2017b] adds hierarchical learning. DGCNN [Wang et al., 2019] uses dynamic graphs. KPConv [Thomas et al., 2019] provides kernel point convolution. PointCNN [Li et al., 2018b] applies X-transformation. Recent advances include Point Transformer [Zhao et al., 2021], PCT [Guo et al., 2021], PointNeXt [Qian et al., 2022], PointMLP [Ma et al., 2022], Point-BERT [Yu et al., 2022], Point-MAE [Pang et al., 2022], and PointGPT [Chen et al., 2024b].

Scene Graphs. Scene graph generation [Xu et al., 2017, Krishna et al., 2017, Yang et al., 2018, Zhang et al., 2019, Zellers et al., 2018, Tang et al., 2019, Chen et al., 2019c, Li et al., 2017b, Lu et al., 2016, Johnson et al., 2015] represents object relationships. 3D scene graphs [Armeni et al., 2019, Rosinol et al., 2020, Hughes et al., 2022, Wald et al., 2020, Wu et al., 2021a, Kim and Ramalingam, 2020, Gu et al., 2024] extend to 3D environments.

Vision-Language Models for 3D. 3D-LLM [Hong et al., 2023b] enables language understanding of 3D scenes. LLaVA-3D [Zheng et al., 2024] extends multimodal models to 3D. ConceptFusion [Jatavallabhula et al., 2023] fuses concepts into 3D representations. Additional models include LEO [Huang et al., 2024a], Chat-3D [Wang et al., 2023e], LL3DA [Chen et al., 2024e], and Scene-LLM [Fu et al., 2024b].

4.2.3 Manipulation

Manipulation involves physical interaction with objects in the environment.

Vision-Language-Action Models. RT-1 [Brohan et al., 2022] demonstrated transformer-based robot policies. RT-2 [Brohan et al., 2023] co-trained on web-scale data. PaLM-E [Driess et al., 2023] integrated embodied reasoning. Octo [Team et al., 2024] provides open-source generalist policies. OpenVLA [Kim et al., 2024] offers accessible VLA models. RT-X [Collaboration, 2023, Zhang et al., 2023a, Padalkar et al., 2023] scales across robot embodiments. RoboCat [Bousmalis et al., 2023] demonstrates self-improvement. Additional VLA models include GR-1 [Wu et al., 2023b], ManipLLM [Li et al., 2024b], RoboFlamingo [Li et al., 2023i], HPT [Wang et al., 2024b], and CrossFormer [Doshi et al., 2024].

Language-Conditioned Manipulation. SayCan [Ahn et al., 2022] grounds language in affordances. CLIPort [Shridhar et al., 2022] combines CLIP with Transporter networks. PerAct [Shridhar et al., 2023] uses perceiver transformers. RVT [Goyal et al., 2023] employs multi-view transformers. VIMA [Sharma et al., 2022] uses multimodal prompts. Additional methods include BC-Z [Jang et al., 2022], MOO [Stone et al., 2023], HULC [Mees et al., 2022b], GNFactor [Ze et al., 2023], Act3D [Gervet et al., 2023], and RVT-2 [Goyal et al., 2024].

Dexterous Manipulation. Rubik's cube solving [Akkaya et al., 2019] demonstrated sim-to-real transfer. DexMV [Qin et al., 2022] learns from human videos. DexPoint [Qin et al., 2023] uses point cloud representations. Learning from demonstrations [Andrychowicz et al., 2020, Rajeswaran et al., 2018, Zhu et al., 2019, Chen et al., 2022d, Shaw et al., 2023, Arunachalam et al., 2023] enables complex skills. Shadow hand manipulation [OpenAI et al., 2019, Kumar et al., 2016b, Chen et al., 2023b, Qi et al., 2023] showcases dexterous control. Bimanual manipulation [Chitnis et al., 2020, Grannen et al., 2023, Zhao et al., 2023b] addresses dual-arm coordination.

Simulation Environments. RLBench [James et al., 2020] provides 100+ manipulation tasks. Meta-World [Yu et al., 2020] focuses on meta-learning. BEHAVIOR [Srivastava et al., 2021] offers long-horizon household tasks. ManiSkill [Mu et al., 2021, Gu et al., 2023] provides diverse manipulation challenges. Additional environments include CALVIN [Mees et al., 2022a], Robosuite [Zhu et al., 2020], and Isaac Gym [Makoviychuk et al., 2021].

4.2.4 Geospatial Analysis

Geospatial analysis reasons about large-scale geographic data and urban systems.

Remote Sensing Foundation Models. Prithvi [Jakubik et al., 2024] provides geospatial foundation models. SatMAE [Cong et al., 2022] applies masked autoencoders to satellite imagery. SatlasPretrain [Bastani et al., 2023] enables large-scale pretraining. SatViT [Wang et al., 2022b] uses vision transformers for earth observation. GeoAI [Janowicz et al., 2020] surveys the broader field. Additional models include GASSL [Ayush et al., 2021], SeCo [Manas et al., 2021], Scale-MAE [Reed et al., 2023], GFM [Mendieta et al., 2023], SkySense [Guo et al., 2024c], and SpectralGPT [Hong et al., 2024].

Spatio-Temporal Graph Networks. DCRNN [Li et al., 2018a] models traffic as graph diffusion. STGCN [Yu et al., 2018] combines graph and temporal convolutions. Graph WaveNet [Wu et al., 2019b] learns adaptive structures. AGCRN [Bai et al., 2020] introduces node-specific patterns. T-GCN [Zhao et al., 2019] provides temporal graph convolution. ASTGCN [Guo et al., 2019] adds attention mechanisms. GMAN [Zheng et al., 2020] uses graph multi-attention. MTGNN [Wu et al., 2020b] connects multiple time series. Additional models include STSGCN [Song et al., 2020], STFGNN [Li and Zhu, 2021], PDFormer [Jiang et al., 2023b], STAEformer [Liu et al., 2023a], DSTAGNN [Lan et al., 2022], D2STGNN [Shao et al., 2022], and STG-NCDE [Choi et al., 2022].

Urban Computing. Urban computing [Zheng et al., 2014, Yuan et al., 2020, Zheng et al., 2011] applies AI to city-scale problems. Traffic prediction [Jiang and Luo, 2022a, Jin et al., 2023, Li et al., 2017a], crowd flow forecasting [Zhang et al., 2017b, 2018, Pan et al., 2019], and POI recommendation [Liu et al., 2017, Zhao et al., 2020b, Lian et al., 2020] are key applications. Smart city applications [Silva et al., 2018, Chen et al., 2020b, Bibri and Krogstie, 2017] integrate multiple urban systems.

5 State-of-the-Art Methods

5.1 Vision-Language-Action Models

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

Proprietary Models. RT-1 [Brohan et al., 2022] demonstrated transformer-based policies trained on large-scale robot data (130k demonstrations). RT-2 [Brohan et al., 2023] co-trained on web-scale vision-language data, enabling emergent reasoning about novel objects and achieving 2x improvement on unseen objects. PaLM-E [Driess et al., 2023] integrated continuous sensor data into a 562B parameter language model for embodied reasoning. Gato [Reed et al., 2022] demonstrated a generalist agent across 604 tasks.

Open-Source Models. Octo [Team et al., 2024] provides a generalist robot policy trained on the Open X-Embodiment dataset [Collaboration, 2023] with 800k trajectories from 22 robot embodiments. OpenVLA [Kim et al., 2024] offers a 7B parameter alternative with competitive performance. These models democratize VLA research and enable community-driven advancement.

Emerging Directions. Recent work explores scaling laws for robotics [Brohan et al., 2023], cross-embodiment transfer [Collaboration, 2023], and integration with world models [Wu et al., 2023c].

5.2 Graph Neural Networks for Spatial Reasoning

GNNs provide powerful tools for modeling spatial relationships and dependencies.

Foundational Architectures. GCN [Kipf and Welling, 2017] introduced spectral graph convolution. GAT [Velickovic et al., 2018] added attention mechanisms for adaptive aggregation. GraphSAGE [Hamilton et al., 2017] enabled inductive learning on unseen nodes. GIN [Xu et al., 2019] provided theoretical expressiveness analysis. MPNN [Gilmer et al., 2017] unified message passing frameworks. Additional architectures include SGC [Wu et al., 2019a], APPNP [Klicpera et al., 2019], and GPR-GNN [Chien et al., 2021].

Spatio-Temporal Networks. DCRNN [Li et al., 2018a] models traffic as bidirectional graph diffusion. STGCN [Yu et al., 2018] combines graph and temporal convolutions efficiently. Graph WaveNet [Wu et al., 2019b] learns adaptive graph structures without predefined adjacency. AGCRN [Bai et al., 2020] introduces node-specific patterns through adaptive modules. Comprehensive surveys [Jin et al., 2023, Atluri et al., 2018, Wang et al., 2020, Jiang and Luo, 2022a] detail these advances.

GNN-LLM Integration. Emerging work combines GNNs with LLMs for structured spatial reasoning [Chen et al., 2024d, Tang et al., 2024, Ye et al., 2024, Fatemi et al., 2023, Huang et al., 2024b, Perozzi et al., 2024]. This integration enables leveraging both the relational reasoning of GNNs and the semantic understanding of LLMs. Graph instruction tuning [Zhang et al., 2024, Zhao et al., 2023a] further enhances this capability.

5.3 World Models

World models learn predictive representations enabling planning through imagination.

Model-Based Reinforcement Learning. Dreamer [Hafner et al., 2019b] introduced latent imagination for sample-efficient learning. DreamerV2 [Hafner et al., 2021] achieved human-level Atari performance. DreamerV3 [Hafner et al., 2023] demonstrated cross-domain mastery with a single algorithm. DayDreamer [Wu et al., 2023c] transferred world models to real robots. PlaNet [Hafner et al., 2019a] pioneered latent dynamics learning. MuZero [Schrittwieser et al., 2020] combined learned models with MCTS. Additional approaches include MBPO [Janner et al., 2019], SLAC [Lee et al., 2020], and TD-MPC [Hansen et al., 2022].

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

LLM-Based World Models. LLMs can serve as world models for planning [Hao et al., 2023, Guan 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.

5.4 Multimodal Foundation Models

Multimodal models integrate vision, language, and action understanding.

Vision-Language Models. CLIP [Radford et al., 2021] enabled zero-shot visual recognition. BLIP-2 [Li et al., 2023d] introduced efficient vision-language pretraining. LLaVA [Liu et al., 2024b] demonstrated visual instruction tuning. GPT-4V [OpenAI, 2023] achieved strong multimodal reasoning. Gemini [Team and Google, 2023] provides native multimodal capabilities. Flamingo [Alayrac et al., 2022] enables few-shot visual learning. PaLI [Chen et al., 2023d] scales vision-language models. Kosmos-2 [Peng et al., 2023] adds grounding capabilities. Qwen-VL [Bai et al., 2023] provides open multilingual VLMs. Additional models include InstructBLIP [Dai et al., 2023b], MiniGPT-4 [Zhu et al., 2023a], Otter [Li et al., 2023a], and CogVLM [Wang et al., 2023d].

Spatial Vision-Language Models. SpatialVLM [Chen et al., 2024a] specializes in spatial reasoning. VoxPoser [Huang et al., 2023c] extracts affordances from VLMs. VLMaps [Huang et al., 2023a] creates semantic spatial maps. These models bridge vision-language understanding with spatial reasoning.

5.5 Embodied AI Agents

Open-Ended Exploration. Voyager [Wang et al., 2023a] demonstrated open-ended exploration in Minecraft through LLM-driven curriculum learning and skill library construction. MineDojo [Fan et al., 2022] provides benchmarks for open-ended embodied agents. DEPS [Wang et al., 2023f] decomposes embodied planning systematically.

Grounded Language Agents. SayCan [Ahn et al., 2022] grounds language models in robotic affordances through value functions. Code as Policies [Liang et al., 2023a] generates executable robot code from language. LLM-Planner [Song et al., 2023] enables few-shot grounded planning. EmbodiedGPT [Mu et al., 2023] provides embodied chain-of-thought reasoning.

Multi-Agent Systems. AutoGen [Wu et al., 2023d] enables multi-agent conversations. MetaGPT [Hong et al., 2023a] assigns roles to agents. CAMEL [Li et al., 2023c] explores communicative agents. ChatDev [Qian et al., 2023] applies multi-agent systems to software development.

6 Industry Applications

6.1 Geospatial Intelligence

Palantir [Palantir, 2023, Bailey, 2021, Palantir Technologies, 2023] integrates AI with geospatial analysis for defense and commercial applications, processing satellite imagery and sensor data at scale. **ESRI** [ESRI, 2023a,b] provides ArcGIS with integrated GeoAI capabilities for spatial analysis. **Google** [Google, 2023b,a] deploys AI for global-scale mapping, navigation, and earth observation through Google Earth Engine.

6.2 Location Intelligence

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

6.3 Autonomous Vehicles

Waymo [Waymo, 2023, 2024, Sun et al., 2020] has deployed autonomous vehicles at scale with millions of miles driven. End-to-end approaches including UniAD [Hu et al., 2023b], VAD [Jiang et al., 2023a], and DriveVLM [Tian et al., 2024] unify perception, prediction, and planning. **Tesla** [Tesla, 2023] pursues vision-only autonomy. **Cruise** [Cruise LLC, 2023], **Mobileye** [Mobileye, 2023], and **NVIDIA** [NVIDIA, 2023] provide additional autonomous driving solutions.

6.4 Robotics

Boston Dynamics [Raibert et al., 2008] develops advanced mobile robots. **Figure AI** and **1X Technologies** pursue humanoid robotics. Industrial applications include warehouse automation [Wurman et al., 2008], manufacturing [Khatib et al., 2016], and healthcare [Yang et al., 2020].

7 Evaluation Framework: SpatialAgentBench

To address the lack of a unified evaluation standard, we propose **SpatialAgentBench**, a comprehensive suite of 8 tasks spanning all four spatial domains.

Table 1: Comparison of Spatial Intelligence Benchmarks

Benchmark	**Task**	**Environment**	**Metrics**	**Key Feature**
		Navigation		
R2R [Anderson et al., 2018b]	VLN	Real-world images	SPL, SR	First large-scale VLN
RxR [Ku et al., 2020]	VLN	Real-world images	nDTW, SR	Multilingual
REVERIE [Qi et al., 2020]	VLN	Real-world images	RGS	Remote grounding
Habitat ObjectNav [Batra et al., 2020]	ObjectNav	Simulated	SPL, Success	Standardized
SOON [Zhu et al., 2021]	ObjectNav	Simulated	NDO	Semantic
TouchDown [Chen et al., 2019a]	VLN	Street View	TC, SPD	Urban navigation
		Manipulation		
RLBench [James et al., 2020]	100+ tasks	Simulated	Success Rate	Diverse tasks
Meta-World [Yu et al., 2020]	50 tasks	Simulated	Success Rate	Meta-learning
BEHAVIOR [Srivastava et al., 2021]	1000 activities	Simulated	Goal Conditions	Long-horizon
Open X-Embodiment [Collaboration, 2023]	22 robots	Real-world	N/A	Largest real dataset
ManiSkill2 [Gu et al., 2023]	20 tasks	Simulated	Success Rate	Soft-body physics
		Spatial Reasoning		
CLEVR [Johnson et al., 2017]	VQA	Synthetic	Accuracy	Compositional
GQA [Hudson and Manning, 2019]	VQA	Real-world	Accuracy	Scene graphs
SpatialVLM [Chen et al., 2024a]	VQA	Real-world	Accuracy	Fine-grained spatial
ScanQA [Azuma et al., 2022]	3D VQA	Real scans	EM, BLEU	3D understanding
EmbodiedBench [Yang et al., 2025]	Embodied	Simulated	Success Rate	Comprehensive
		Geospatial		
BigEarthNet [Sumbul et al., 2019]	Classification	Satellite	Accuracy, F1	Large-scale
fMoW [Christie et al., 2018]	Classification	Satellite	Accuracy	Temporal
xBD [Gupta et al., 2019b]	Segmentation	Satellite	IoU, F1	Damage assessment
SpaceNet [Van Etten et al., 2018]	Detection	Satellite	AP	Building footprints

7.1 SpatialAgentBench Tasks

Our proposed benchmark includes:

1. **VLN-Instruct**: Vision-language navigation with complex instructions
2. **ObjectSearch**: Multi-room object search with semantic reasoning
3. **SceneQA**: 3D scene question answering
4. **ManipSeq**: Sequential manipulation planning
5. **GeoReason**: Geospatial reasoning from satellite imagery
6. **TrafficPredict**: Spatio-temporal traffic prediction
7. **SafeNav**: Navigation with safety constraints
8. **MultiAgent**: Coordinated multi-agent spatial tasks

8 Open Challenges and Future Directions

8.1 Robust Spatial Representation

Developing representations that generalize across scenes, viewpoints, and conditions remains challenging [Mildenhall et al., 2020, Kerbl et al., 2023, Barron et al., 2022]. Foundation models for 3D understanding [Hong et al., 2023b, Fu et al., 2024a, Shen et al., 2023a] represent promising directions. Key challenges include handling occlusion, dynamic scenes, and novel object categories.

8.2 Long-Horizon Planning

Creating agents that plan over extended horizons and decompose complex spatial tasks is essential [Song et al., 2023, Valmeekam et al., 2023, Huang et al., 2022a]. Integration of neural and symbolic planning approaches [Garrett et al., 2021, Dantam et al., 2016, Li et al., 2020b] shows promise. Challenges include credit assignment, subgoal discovery, and plan repair.

8.3 Safe and Reliable Operation

Ensuring safe operation in safety-critical applications is paramount [Yin et al., 2025, Amodei et al., 2016, Bai et al., 2022, Ganguli et al., 2022, Perez et al., 2022]. Key requirements include:

- Robust uncertainty quantification and out-of-distribution detection
- Alignment with human values and preferences
- Interpretable decision-making for accountability
- Graceful degradation under adversarial conditions

8.4 Sim-to-Real Transfer

Bridging simulation and reality remains challenging [Zhao et al., 2020a, Tobin et al., 2017, James et al., 2019, Matas et al., 2018]. Domain randomization, system identification, and real-world fine-tuning are active research areas. The reality gap affects perception, dynamics, and control.

8.5 Multi-Agent Coordination

Scaling to multi-agent systems for complex spatial tasks requires advances in coordination and communication [Zhang et al., 2021a, Wu et al., 2023d, Hong et al., 2023a, Li et al., 2023c, Qian et al., 2023]. Challenges include emergent communication, credit assignment, and scalable coordination mechanisms.

8.6 Efficiency and Deployment

Deploying spatial AI systems on resource-constrained platforms requires advances in model compression, efficient inference, and edge computing [Han et al., 2016, Howard et al., 2017]. Real-time operation is critical for many applications.

9 Limitations

This survey, while comprehensive, has several limitations:

- Our paper selection process, though systematic, may have missed relevant works in adjacent fields.
- The proposed taxonomy, while unifying, is one of many possible categorizations.
- 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.

10 Conclusion

This survey has provided a unified taxonomy connecting Agentic AI and Spatial Intelligence, synthesizing over 500 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.
2. **GNN-LLM integration** is an emergent paradigm combining the relational reasoning of graph networks with the semantic understanding of language models.
3. **World models** are essential for safe deployment, enabling agents to predict consequences and plan in imagination before acting.

Key challenges remain in robust representation, long-horizon planning, safe deployment, and multi-agent coordination. By establishing this foundational reference and proposing SpatialAgentBench, we aim to accelerate progress toward capable, robust, and safe spatially-aware autonomous systems that can perceive, reason about, and act within the physical world.

References

Forest Agostinelli, Stephen McAleer, Alexander Shmakov, and Pierre Baldi. Solving the rubik's cube with deep reinforcement learning and search. *Nature Machine Intelligence*, 1(8):356–363, 2019.

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

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

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.

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

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

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

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

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, 2018b.

Peter Anderson, Ayush Shrivastava, Devi Parikh, Dhruv Batra, and Stefan Lee. Chasing ghosts: Instruction following as bayesian state tracking. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 113–123. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/8329-chasing-ghosts-instruction-following-as-bayesian-state-tracking.pdf`.

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.

Sridhar Pandian Arunachalam, Irmak Guzey, Soumith Chintala, and Lerrel Pinto. Holo-dex: Teaching dexterity with immersive mixed reality. In *ICRA*, 2023.

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

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

Kumar Ayush et al. Geography-aware self-supervised learning. In *ICCV*, 2021.

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.

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

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

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

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

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

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

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.

Valts Blukis et al. Mapping navigation instructions to continuous control actions with position-visitation prediction. In *CoRL*, 2018.

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

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

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

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

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

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

Vincent Cartillier et al. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In *AAAI*, 2021.

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

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

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

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

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning. In *arXiv preprint arXiv:2310.05915*, 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.

Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh K Ramakrishnan, and Kristen Grauman. Learning audio-visual navigation from human demonstrations. In *CVPR*, 2022b.

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

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

Chen Chen et al. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. In *RSS*, 2022d.

Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. *ACM Computing Surveys*, 2020b.

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.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019a.

Jacob Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. Waypoint models for instruction-guided navigation in continuous environments. In *ICCV*, 2021a.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K Wong. Mapgpt: Map-guided prompting for unified vision-and-language navigation. In *arXiv preprint arXiv:2401.07314*, 2024c.

Lili Chen et al. Behavioral cloning from observation. In *IJCAI*, 2019b.

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*, 2021b.

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

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

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

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

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

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

Tianshui Chen et al. Knowledge-embedded routing network for scene graph generation. In *CVPR*, 2019c.

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

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

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

Xinyun Chen et al. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023f.

Eli Chien et al. Adaptive universal generalized pagerank graph neural network. In *ICLR*, 2021.

Rohan Chitnis, Dylan Hadfield-Menell, Abhishek Gupta, Siddharth Srivastava, Edward Groshev, Christopher Lin, and Pieter Abbeel. Guided search for task and motion plans using learned heuristics. In *ICRA*, 2016.

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

Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. Graph neural controlled differential equations for traffic forecasting. In *AAAI*, 2022.

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.

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

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.

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.

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.

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.

Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International Conference on Computers and Games*, 2006.

Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.

Cruise LLC. Cruise autonomous vehicles. `https://getcruise.com`, 2023.

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.

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

José De Jesús Rubio et al. Deep learning for geospatial data applications. *Remote Sensing*, 13(4):595, 2021.

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.

Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *arXiv preprint arXiv:2408.11812*, 2024.

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 Du, Difei Gao, Jing Feng, Shijie Pan, Hanqing Zhao, Shengyu Shen, Song-Chun Zhu, and Yixin Gao. A survey on embodied ai: From simulators to research tasks. *arXiv preprint arXiv:2303.11174*, 2023.

Jiafei Duan et al. A survey of embodied ai. *IEEE TETCI*, 2022.

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

Scott Emmons et al. Sparse graphical memory for robust planning. *NeurIPS*, 2020.

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. In *Science*, 2018.

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

ESRI. Arcgis geoai. `https://www.esri.com/en-us/arcgis/products/arcgis-geoai`, 2023b.

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

Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. *WWW*, 2019.

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. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023.

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

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

Foursquare. Foursquare studio. `https://studio.foursquare.com`, 2023b.

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.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Ji. Scene-llm: Extending language model for 3d visual understanding and reasoning. In *arXiv preprint arXiv:2403.11401*, 2024b.

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

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.

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

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

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

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

Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *ECCV*, 2020.

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

Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *CoRL*, 2023.

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

Google. Google earth engine. `https://earthengine.google.com`, 2023a.

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

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt-2: Learning precise manipulation from few demonstrations. In *RSS*, 2024.

Ankit Goyal et al. Rvt: Robotic view transformer for 3d object manipulation. In *CoRL*, 2023.

Jennifer Grannen, Yilin Wu, Brandon Vu, and Dorsa Sadigh. Stabilize to act: Learning to coordinate for bimanual manipulation. In *CoRL*, 2023.

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

Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aaron van den Oord. Shaping belief states with generative environment models for rl. In *NeurIPS*, 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.

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

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.

Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 2021.

Rentong Guo et al. Manu: A cloud native vector database management system. *VLDB*, 2022.

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

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *CVPR*, 2024c.

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*, 2019a.

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

Saurabh Gupta, Varun Tolani, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. *International Journal of Computer Vision*, 128(5): 1311–1330, 2019c.

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

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

Dylan Hadfield-Menell, Edward Groshev, Rohan Chitnis, and Pieter Abbeel. Sequential task-based motion planning. In *ICRA*, 2017.

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

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

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

Patrick Haluptzok, Matthew Bowers, and Adam Tauman Kalai. Language models can teach themselves to program better. In *ICLR*, 2023.

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

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

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

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

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

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

Weituo Hao et al. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*, 2020.

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

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

Sebastian Höfer, Kostas Bekris, Ankur Handa, Juan Camilo Gamboa, Melissa Mozifian, Florian Golber, Suraj Nair, Mrinal Kalakrishnan, Yevgen Chebotar, Ankur Handa, et al. Sim2real in robotics and automation: Applications and challenges. *IEEE Transactions on Automation Science and Engineering*, 2021.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. In *IEEE TPAMI*, 2024.

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 et al. A recurrent vision-and-language bert for navigation. In *CVPR*, 2020a.

Yicong Hong et al. Sub-instruction aware vision-and-language navigation. In *EMNLP*, 2020b.

Yicong Hong et al. Vln-bert: A recurrent vision-and-language bert for navigation. In *CVPR*, 2021.

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

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.

Yingjie Hu et al. A five-star guide for achieving replicability and reproducibility when working with gis software and algorithms. *Annals of the American Association of Geographers*, 2019.

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

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. Leo: An embodied generalist agent in 3d world. In *ICML*, 2024a.

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

Shibo Huang, Jingyi Jiang, Haoran Dong, Zilong Zheng, Jianye Hao, Jun Zhu, and Jie Hu. Reasoning with language model is planning with world model. *EMNLP*, 2023b.

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*, 2023c.

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

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

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

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.

Gus Hulbert et al. Using large language models to simulate multiple humans and replicate human subject studies. *arXiv preprint arXiv:2208.10264*, 2023.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys*, 2017.

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. Few-shot learning with retrieval augmented language models. In *JMLR*, 2022.

Johannes Jakubik, Sujit Roy, C E Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Musber, Daiki Oliveira, et al. Prithvi: A foundation model for earth observation. *arXiv preprint arXiv:2310.18660*, 2024.

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

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

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.

Krishna Murthy Jatavallabhula et al. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems*, 2023.

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

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

Jiawei Jiang et al. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. *AAAI Conference on Artificial Intelligence*, 2023b.

Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, 2022a.

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

Zhe Jiang, Sheng Li, and Xin Hu. Geoai: A review of artificial intelligence approaches for the interpretation of complex geomatics data. *Geoscience Frontiers*, 2023c.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *EMNLP*, 2023d.

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, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. In *arXiv preprint arXiv:2207.05221*, 2022.

Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim-to-real robot learning from pixels with progressive nets. In *Conference on Robot Learning*, 2020.

Leslie Pack Kaelbling. The foundation of efficient robot learning. *Science*, 369(6506):915–916, 2020.

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

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.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. *arXiv preprint arXiv:2303.09553*, 2023.

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

Oussama Khatib et al. Ocean one: A robotic avatar for oceanic discovery. *IEEE Robotics & Automation Magazine*, 2016.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. In *arXiv preprint arXiv:2212.14024*, 2022.

Iro Kim and Sanja Ramalingam, Srikumar and. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *ICCV*, 2020.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

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

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

Johannes Klicpera et al. Predict then propagate: Graph neural networks meet personalized pagerank. In *ICLR*, 2019.

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.

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

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.

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

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

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*, 2016a.

Vikash Kumar et al. Optimal control with learned local models: Application to dexterous manipulation. *ICRA*, 2016b.

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

Shiyong Lan et al. Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. *International Conference on Machine Learning*, 2022.

Hung Le, Truyen Tran, and Svetha Venkatesh. Self-attentive associative memory. In *ICML*, 2020.

Alex X Lee et al. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. In *NeurIPS*, 2020.

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.

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.

Alexander C Li, Lerrel Pinto, and Pieter Abbeel. Skill discovery for exploration and planning using deep skill graphs. In *ICML*, 2020a.

Bo Li et al. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Sergey Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *CoRL*, 2023b.

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

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*, 2023d.

Lei Li, Chen Ma, Yongfeng Fan, and Jianhua Yin. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 2024a.

Mengzhang Li and Zhanxing Zhu. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *AAAI*, 2021.

Miaoran Li et al. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*, 2023e.

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

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

Shuai Li et al. Hybrid task and motion planning. *arXiv preprint*, 2020b.

Weiwei Li, Ching-Yao Hsu, and Xia Hu. Deep learning for geospatial data applications: A comprehensive survey. *IEEE Transactions on Big Data*, 2023h.

Xiaoqi Li et al. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. *arXiv preprint arXiv:2312.16217*, 2024b.

Xinghang Li et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023i.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2017a.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018a.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *NeurIPS*, 2018b.

Yifei Li et al. Making language models better reasoners with step-aware verifier. *ACL*, 2023j.

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017b.

Defu Lian, Yongji Wu, Yong Ge, Xing Xie, and Enhong Chen. Geography-aware sequential location recommendation. In *KDD*, 2020.

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*, 2023a.

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

Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. In *NeurIPS*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. In *TMLR*, 2022.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *ACL*, 2017.

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

Hangchen Liu et al. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. *arXiv preprint arXiv:2308.10425*, 2023a.

Haotian Liu et al. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2024b.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. In *ACM Computing Surveys*, 2023b.

Yifan Liu et al. Experimental security analysis of a modern automobile. In *IEEE S&P*, 2017.

Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and Philip Yu. Graph self-supervised learning: A survey. *IEEE TKDE*, 2022.

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

Tomás Lozano-Pérez and Leslie Pack Kaelbling. A constraint-based method for solving sequential manipulation planning problems. *IROS*, 2014.

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.

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

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, Ziad Al-Halah, and Kristen Grauman. Sound adversarial audio-visual navigation. In *ICLR*, 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.

Oscar Manas et al. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV*, 2021.

Ricardo Martin-Brualla et al. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021.

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

Hidenobu Matsuki et al. Gaussian splatting slam. In *CVPR*, 2024.

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.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. In *IEEE RA-L*, 2022a.

Oier Mees et al. What matters in language conditioned robotic imitation learning over unstructured data. In *RA-L*, 2022b.

Oier Mees et al. Mats: Multi-agent trajectory prediction with scene-centric attention. In *arXiv preprint*, 2022c.

Matias Mendieta et al. Towards geospatial foundation models via continual pretraining. *arXiv preprint arXiv:2302.04476*, 2023.

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.

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.

Alexander Miller et al. Key-value memory networks for directly reading documents. In *EMNLP*, 2016.

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.

Piotr Mirowski et al. Learning to navigate in complex environments. In *ICLR*, 2017.

Eric Mitchell, Joseph J Noh, Siyan Li, William S Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *EMNLP*, 2022.

Mobileye. Mobileye autonomous driving. `https://www.mobileye.com`, 2023.

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

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

Yao Mu et al. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *ACM TOG*, 2022.

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

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

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

Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. Compressed 3d gaussian splatting for accelerated novel view synthesis. In *CVPR*, 2024.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *ICLR*, 2023.

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

NVIDIA. Nvidia drive platform. `https://www.nvidia.com/en-us/self-driving-cars/`, 2023.

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.

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

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-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.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018.

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 foundry. `https://www.palantir.com/platforms/foundry/`, 2023.

Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. Urban traffic prediction from spatio-temporal data using deep meta learning. In *KDD*, 2019.

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.

Emilio Parisotto et al. Global pose estimation with an attention-based recurrent network. In *CVPR Workshops*, 2018.

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.

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*, 2023.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Bryan Perozzi et al. Let your graph do the talking: Encoding structured data for llms. *arXiv preprint arXiv:2402.05862*, 2024.

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.

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

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021.

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

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

Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. In *CoRL*, 2023.

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.

Guocheng Qian et al. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *NeurIPS*, 2022.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. *ACL*, 2023.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *CVPR*, 2024a.

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

Yuzhe Qin et al. Dexmv: Imitation learning for dexterous manipulation from human videos. In *ECCV*, 2022.

Yuzhe Qin et al. Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation. In *CoRL*, 2023.

Yujia Qu, Tianjun Cai, Shudan Zhao, Jiaxin Dong, Jian Liang, Jianfeng Shi, Yankai Wang, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*, 2024.

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.

Saeed Rahmani, Asiye Baghbani, Nizar Bouguila, and Zachary Patterson. Graph neural networks for intelligent transportation systems: A survey. *IEEE TITS*, 2023.

Marc Raibert et al. Bigdog, the rough-terrain quadruped robot. *IFAC*, 2008.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *RSS*, 2018.

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

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.

Ram Ramrakhya et al. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *CVPR*, 2022.

Ram Ramrakhya et al. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *CVPR*, 2023.

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.

Colorado Reed et al. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532*, 2023.

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. *arXiv preprint arXiv:2205.06175*, 2022.

Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. In *Robotics: Science and Systems (RSS)*, 2020.

Antoni Rosinol et al. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *IROS*, 2022.

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

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

Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. In *ICLR*, 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.

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.

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.

Bilgehan Sel, Ahmad Al-Tawaha, Vanshaj Khattar, Lu Wang, Ruoxi Jia, and Ming Jin. Algorithm of thoughts: Enhancing exploration of ideas in large language models. In *arXiv preprint arXiv:2308.10379*, 2023.

Kinza Shafique, Bilal A Khawaja, Farah Sabber, Sameer Gul, Muhammad Mustaqim, and Aamir Khawaja. Internet of things (iot) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5g-iot scenarios. *IEEE Access*, 8:23022–23040, 2020.

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

Dhruv Shah et al. Rapid exploration for open-world navigation with latent goal models. In *CoRL*, 2021.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. In *VLDB*, 2022.

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

Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. In *RSS*, 2023.

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

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.

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

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

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

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

Mohit Shridhar et al. Perceiver-actor: A multi-task transformer for robotic manipulation. In *CoRL*, 2023.

Thiago H Silva, Pedro OS de Melo, Jussara M Almeida, Juliana Salles, and Antonio AF Loureiro. Urban computing leveraging location-based social network data: A survey. *ACM Computing Surveys*, 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, Rohan Chitnis, Joshua Tenenbaum, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Planning with learned object importance in large problem instances using graph neural networks. In *AAAI*, 2021.

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

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. 2020.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martin, Fei Xia, Kent Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. *arXiv preprint arXiv:2108.03332*, 2021.

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.

Austin Stone et al. Open-world object manipulation using pre-trained vision-language models. *CoRL*, 2023.

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, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

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

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, 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.

Matthew Tancik, Vincent Casser, Xinchen Yan, et al. Block-nerf: Scalable large scene neural view synthesis. 2022.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salber, Abhik Blanber, et al. Nerfstudio: A modular framework for neural radiance field development. In *SIGGRAPH*, 2023.

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.

Kaihua Tang et al. Learning to compose dynamic tree structures for visual contexts. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

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.

David Alexander Tedjopurnomo, Zhifeng Bao, Baihua Zheng, Farhana Choudhury, and A Qin. A survey on modern deep neural network for traffic prediction: Trends, methods and challenges. *IEEE TKDE*, 2020.

Jack Tennison et al. Grounded task and motion planning. *arXiv preprint*, 2024.

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

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

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

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

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

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

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

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.

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

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

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

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

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

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

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.

Lean Wang, Lei Lei, Damai Dai, Dan Pan, Shuming Ding, Tianyu Ma, Baobao Song, and Zhifang Sui. Label words are anchors: An information flow perspective for understanding in-context learning. In *EMNLP*, 2023b.

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*, 2024a.

Lirui Wang, Xinlei Zhao, Jialiang Liu, Kaushik Kamat, Carlo Sferrazza, Dina Katabi, Pulkit Agrawal, and Lerrel Pinto. Hpt: Scaling heterogeneous pre-training for robotics. In *arXiv preprint arXiv:2409.20537*, 2024b.

Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *CVPR*, 2023c.

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

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

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*, 2022a.

Yi Wang et al. Satvit: Pretraining transformers for earth observation. *IEEE GRSL*, 2022b.

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

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. In *arXiv preprint arXiv:2308.08769*, 2023e.

Zihao Wang et al. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. In *NeurIPS*, 2023f.

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

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

Greg Wayne et al. Unsupervised predictive memory in a goal-directed agent. In *ICML*, 2018.

Weaviate. Weaviate vector database, 2023.

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.

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.

Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 2020.

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

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

Felix Wu et al. Simplifying graph convolutional networks. In *ICML*, 2019a.

Guanjun Wu et al. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024a.

Hongtao Wu et al. Gr-1: Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023b.

Lingfei Wu, Peng Cui, Jian Pei, Liang Zhao, and Xiaojie Guo. Graph neural networks: Methods, applications, and opportunities. *arXiv preprint arXiv:2108.10733*, 2022.

Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model. In *arXiv preprint arXiv:2401.02695*, 2024b.

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

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*, 2023d.

Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: A survey. *ACM Computing Surveys*, 2020a.

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

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 et al. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *KDD*, 2020b.

Zonghan Wu et al. A comprehensive survey on graph neural networks. *IEEE TNNLS*, 2021b.

Peter R Wurman, Raffaello D'Andrea, and Mick Mountz. Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI Magazine*, 2008.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. In *IEEE RA-L*, 2020.

Feng Xia, Ke Sun, Shuo Yu, Abdul Aziz, Liangtian Wan, Shirui Pan, and Huan Liu. Graph learning: A survey. *IEEE TAI*, 2021.

Peng Xie, Tianrui Li, Jia Liu, Shengdong Du, Xin Yang, and Junbo Zhang. Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion*, 2020.

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. Decomposition enhances reasoning via self-evaluation guided decoding. In *arXiv preprint arXiv:2305.00633*, 2023.

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

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

Chi Yan et al. Gs-slam: Dense visual slam with 3d gaussian splatting. *arXiv preprint arXiv:2311.11700*, 2024.

Guang-Zhong Yang et al. Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy. *Science Robotics*, 2020.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, 2018.

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*, 2024a.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *CVPR*, 2024b.

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, 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.

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

Wenhao Ye, Nan Zhao, Hao Zheng, and Yingqing Zhu. Spatial assembly: Generative architecture with reinforcement learning, self play and tree search. *arXiv preprint arXiv:2108.05802*, 2021.

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.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *ICLR*, 2023.

Abdelrahman Younes, Daniel Honig, Firas Dogan, and Anthony Tzes. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. In *IEEE RA-L*, 2023.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *IROS*, 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.

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 et al. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Qiangqiang Yuan et al. Deep learning for satellite image classification. *ISPRS Journal*, 2021.

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

Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi. Internet of things for smart cities. *IEEE Internet of Things Journal*, 1(1):22–32, 2014.

Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In *CoRL*, 2023.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. 2018.

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

Yue Zeng et al. A survey on vision-language navigation. *IEEE TPAMI*, 2023.

Amy Zhang et al. Neural map: Structured memory for deep reinforcement learning. In *ICLR*, 2017a.

Ji Zhang et al. Graph neural networks for scene graph generation. In *ICCV*, 2019.

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

Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, Xiuwen Yi, and Tianrui Li. Predicting citywide crowd flows using deep spatio-temporal residual networks. In *Artificial Intelligence*, 2018.

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

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.

Lunjun Zhang, Bradly C Stadie, and Jimmy Ba. World model as a graph: Learning latent landmarks for planning. In *ICML*, 2021b.

Tianren Zhang, Shangqi Guo, Tian Tan, Xiaolin Hu, and Feng Chen. Generating adjacency-constrained subgoals in hierarchical reinforcement learning. In *NeurIPS*, 2020a.

Tony Zhang et al. Rt-x: Open x-embodiment robotic learning. *arXiv preprint*, 2023a.

Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in Genetics*, 2021c.

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

Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs: A survey. *IEEE TKDE*, 2020b.

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.

Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-gcn: A temporal graph convolutional network for traffic prediction. In *IEEE TITS*, 2019.

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, 2020a.

Yong Zhao, Jiahui Ni, Zhongming Zhang, Wei Bi, and Xiaojiang Wang. Go from the general to the particular: Multi-domain translation with domain transformation networks. In *AAAI*, 2020b.

Chenming Zheng et al. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. In *arXiv preprint arXiv:2409.18125*, 2024.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. 2020.

Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. Urban computing with taxicabs. *UbiComp*, 2011.

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.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. In *arXiv preprint arXiv:2310.04406*, 2023a.

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

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *AAAI*, 2023c.

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

Kaiwen Zhou, Kaizhi Zheng, Connor Pratt, et al. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. *arXiv preprint arXiv:2301.13166*, 2023d.

Shijie Zhou et al. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *CVPR*, 2024.

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

Henry Zhu et al. Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost. *ICRA*, 2019.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. Solving math word problems via cooperative reasoning induced language models. In *ACL*, 2023b.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.