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 [??????????]. However, these models lack the innate spatial intelligence required to perceive, navigate, and interact with the complex physical world, a critical gap for embodied AI [??????]. 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 900 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 [???????????]. While large language models have demonstrated remarkable capabilities in reasoning and planning [??????????????????], their ability to operate effectively in spatial contexts remains a fundamental challenge [??????????].

The emergence of multimodal foundation models has accelerated progress in visual understanding [?????????????????????? yet translating this understanding into effective spatial action remains challenging. The gap between language-based reasoning and physical world interaction represents one of the most significant obstacles to achieving truly capable autonomous systems [???????].

We define **Agentic AI** as systems exhibiting goal-directed behavior through autonomous decision-making, characterized by four core capabilities: persistent memory for experience accumulation, planning for action sequencing, tool use for capability extension, and self-reflection for continuous improvement [?????????????]. These agents operate through iterative cycles of perception, reasoning, action, and feedback, enabling complex task completion in dynamic environments [????].

Complementarily, **Spatial Intelligence** encompasses the ability to perceive 3D structure, reason about object relationships, navigate environments, and manipulate physical objects [???????????]. This includes understanding geometric relationships, predicting physical dynamics, and planning actions that account for spatial constraints [??????].

The convergence of these domains is essential for real-world AI applications across multiple sectors. Autonomous vehicles must perceive dynamic environments and plan safe trajectories [????????????????????????]. Robotic assistants require understanding of object affordances and spatial relationships [????????????????????]. Urban computing systems must model complex spatio-temporal dependencies [??????????????]. Geospatial intelligence platforms must analyze satellite imagery and geographic data at scale [??????????????????]. Despite this importance, existing surveys treat these areas in isolation, lacking a unified framework connecting agentic architectures with spatial requirements.

Contributions. This survey makes five primary contributions:

1. A **unified taxonomy** connecting agentic AI components (memory, planning, tool use, self-reflection) with spatial intelligence domains (navigation, scene understanding, manipulation, geospatial analysis), providing a structured framework for interdisciplinary research.
2. A **comprehensive analysis** of over 900 papers identifying key architectural patterns, including the emergence of GNN-LLM integration, vision-language-action models, and world model-based planning as critical enablers for spatial reasoning.
3. A **systematic review** of foundation models for spatial intelligence, covering vision-language models, 3D understanding models, and geospatial foundation models.
4. The **proposal of a unified evaluation framework, SpatialAgentBench**, with 8 tasks spanning navigation, manipulation, scene understanding, and geospatial reasoning to standardize cross-domain assessment.
5. A **forward-looking roadmap** identifying 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 [???????]. We queried major academic databases including Google Scholar, arXiv, ACM Digital Library, IEEE Xplore, Semantic Scholar, and DBLP with keywords including "agentic AI," "spatial intelligence," "embodied AI," "vision-language navigation," "robot manipulation," "geospatial AI," "world models," "graph neural networks," "spatio-temporal learning," "vision-language-action," and "foundation models for robotics." Our initial search yielded over 3,000 papers.

We then applied a rigorous multi-stage filtering process:

1. **Temporal Filtering:** We selected papers published between 2018 and 2026, with emphasis on recent advances while including foundational works that established key paradigms.
2. **Venue Filtering:** We prioritized papers from top-tier venues including NeurIPS, ICML, ICLR, CVPR, ECCV, ICCV, CoRL, RSS, IROS, ICRA, ACM Computing Surveys, IEEE TPAMI, Nature, Science, Science Robotics, and leading arXiv preprints.
3. **Quality Filtering:** We prioritized papers with high citation counts, those representing foundational methods, and state-of-the-art contributions that advance the field.
4. **Relevance Filtering:** We ensured papers directly addressed the intersection of agentic capabilities and spatial intelligence.

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

3 Related Work

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

Agentic AI Surveys. Recent surveys on LLM-based agents [????????] focus on reasoning and tool use but do not address spatial capabilities. ? provides a cognitive architecture perspective. ? evaluates conversational agents.

Embodied AI Surveys. Embodied AI surveys [??????] cover simulation environments and benchmarks but lack integration with agentic architectures. ? surveys vision-language-action models specifically for robotics.

Geospatial AI Surveys. Geospatial AI surveys [????????????] and spatio-temporal data mining reviews [??????????] are highly specialized and do not connect to general agentic systems. ? surveys neural networks for geospatial data.

Graph Neural Network Surveys. GNN surveys [??????????] provide comprehensive coverage of graph learning but do not focus on spatial applications or agent integration. Surveys on GNNs for specific domains include traffic [?????], urban computing [?], and spatio-temporal prediction [??].

Vision-Language Model Surveys. Surveys on VLMs [????] cover multimodal understanding but do not address spatial action or embodiment. ? surveys vision-language-action models specifically for robotics.

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

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.

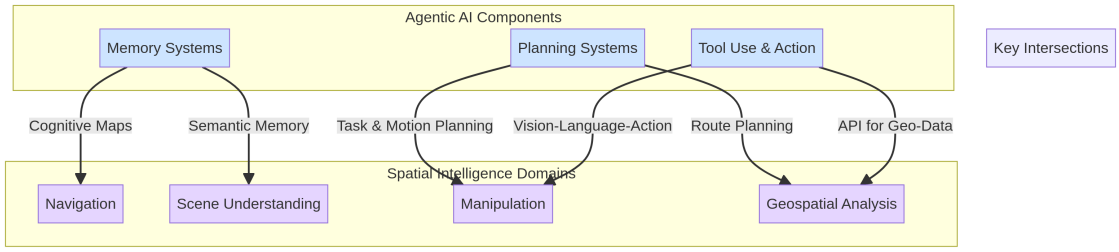

Figure 1: A unified taxonomy connecting Agentic AI capabilities (memory, planning, tool use, self-reflection) with Spatial Intelligence domains (navigation, scene understanding, manipulation, geospatial analysis). The intersection of these dimensions defines the design space for autonomous spatial intelligence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

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. We categorize memory systems into three types: short-term, long-term, and episodic memory.

Short-Term Memory. In-context learning [??????????] allows models to adapt to new tasks through examples in the prompt. This mechanism enables rapid adaptation without parameter updates, leveraging the attention mechanism to condition on provided demonstrations. Working memory mechanisms [??????????] enable temporary information storage during reasoning, supporting multi-step computations that exceed single forward pass capabilities.

Long-Term Memory. Retrieval-augmented generation [?????????????] enables knowledge persistence beyond context limits. MemGPT [?] introduces hierarchical memory management for extended conversations. AMEM [?] provides agentic memory for LLMs. MemEvolve [?] enables meta-evolution of agent

memory. Vector databases [????????] provide efficient similarity search for memory retrieval, enabling agents to access relevant past experiences.

Episodic Memory. Episodic memory stores specific experiences and events, enabling agents to learn from past interactions [????????]. This type of memory is critical for spatial agents that must remember visited locations, encountered objects, and successful action sequences [???].

Spatial Memory. Specialized memory for spatial information includes cognitive maps [???], topological representations [???], and metric maps [???]. Neural approaches to spatial memory include Neural SLAM [????], semantic maps [??????], and scene graphs [??????????].

4.1.2 Planning Systems

Planning decomposes goals into executable action sequences, enabling complex task completion. We identify four major planning paradigms.

Chain-of-Thought Reasoning. Step-by-step reasoning [????????????????] enables systematic problem decomposition. Self-consistency [??????] improves reliability through multiple reasoning paths. Zero-shot chain-of-thought [?] enables reasoning without demonstrations.

Tree-Based Search. Tree of Thoughts [???????] explores multiple solution branches through deliberate search. Graph of Thoughts [???] enables more complex reasoning structures with arbitrary connections. RAP [???] combines reasoning with acting in a planning framework. Monte Carlo Tree Search variants [?????????] provide principled exploration with theoretical guarantees.

Hierarchical Planning. LLM-Planner [?] enables few-shot grounded planning for embodied agents. Inner Monologue [?] provides feedback-driven planning through internal dialogue. HiPlan [?] introduces hierarchical planning with LLMs. Hierarchical RL approaches [??????????] decompose tasks into subtasks with temporal abstraction.

Task and Motion Planning. TAMP [????????????????] integrates symbolic planning with continuous motion planning for robotic applications. This approach combines the expressiveness of symbolic reasoning with the precision of geometric planning.

LLM-Based Planning. Recent work leverages LLMs directly for planning [????????]. SayCan [?] grounds language models in affordances. Code as Policies [?] generates executable robot code. ProgPrompt [?] uses programmatic prompting for task planning.

4.1.3 Tool Use and Action

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

API Integration. Toolformer [?] enables self-supervised tool learning. Gorilla [?] specializes in API calling with retrieval augmentation. ToolLLM [???] provides comprehensive tool use benchmarks. TaskMatrix [?] connects foundation models with millions of APIs. TALM [?] augments language models with tool use. Additional tool-use frameworks include HuggingGPT [??], ToolkenGPT [?], API-Bank [?], Chameleon [?], ViperGPT [?], Visual ChatGPT [?], and MM-ReAct [?].

Code Generation. PAL [?] uses code for reasoning. Code as Policies [?] generates executable robot code from language. Codex [??], CodeGen [?], StarCoder [?], CodeLlama [?], WizardCoder [?], and DeepSeek-Coder [?] provide code generation capabilities. ProgPrompt [?] uses programmatic prompting for robotics. Self-debugging [?], self-repair [?], and self-play [?] improve code quality through iterative refinement.

ReAct Architecture. ReAct [??] interleaves reasoning with action execution, enabling agents to think before acting. Reflexion [??] adds self-reflection for improvement through verbal reinforcement. Additional architectures include LATS [?], SwiftSage [?], FireAct [?], and SWE-Agent [?]. These architectures form the foundation for many spatial agents.

Physical Action. For embodied agents, tool use extends to physical manipulation [????]. Action primitives [??] provide reusable building blocks. Skill libraries [???] enable compositional action.

4.1.4 Self-Reflection and Learning

Self-reflection enables agents to evaluate and improve their own performance.

Self-Critique. Reflexion [?] uses verbal self-reflection for improvement. Self-Refine [?] iteratively improves outputs through self-feedback. Constitutional AI [?] uses self-critique for alignment. Self-debugging [?] enables code correction through self-analysis.

Learning from Experience. Voyager [?] builds skill libraries through exploration. AutoGPT [?] demonstrates autonomous goal pursuit. LangChain [?] provides frameworks for agent development. Retroformer [?] enables retrospective learning from failures.

4.2 Spatial Intelligence Domains

4.2.1 Navigation

Navigation requires path planning and execution in physical or simulated environments. We categorize navigation methods by input modality and task specification.

Vision-Language Navigation. R2R [???] introduced the VLN task with natural language instructions in photorealistic environments. RxR [??] extends to multilingual settings with diverse annotators. REVERIE [?] adds remote object grounding requiring fine-grained understanding. TouchDown [????] addresses urban street-level navigation. Speaker-Follower [?] uses data augmentation through instruction generation. EnvDrop [??] improves generalization through environment dropout. PREVALENT [?] pretrains on VLN data with auxiliary tasks. VLN-BERT [??] applies transformers to VLN. HAMT [?] uses hierarchical attention for multi-scale reasoning. DUET [?] employs dual-scale transformers. Additional methods include RecBERT [?], AirBERT [?], VLN-CE [????], CWP [?], BEVBert [?], NavGPT [???], MapGPT [??], LM-Nav [?], and cross-lingual approaches [?].

Object-Goal Navigation. Object-goal navigation requires finding specific object categories without explicit instructions [????????]. ZSON [?] enables zero-shot object navigation using CLIP embeddings. CLIP-Nav [??] leverages vision-language models for semantic navigation. CoW [??] uses CLIP on Wheels for open-vocabulary navigation. Semantic exploration [??] builds semantic maps for efficient search. SemExp [?] combines semantic mapping with goal-oriented exploration.

Audio-Visual Navigation. Audio-visual navigation incorporates sound for localization [??????]. SoundSpaces [??] provides audio simulation for embodied AI. Audio-visual embodied navigation [??] enables following sound sources.

Embodied Question Answering. EQA [???????] requires navigating to answer questions about environments. 3D-QA benchmarks [????] extend to 3D scene understanding.

4.2.2 Scene Understanding

Scene understanding encompasses 3D perception, semantic segmentation, and spatial relationship reasoning.

Neural Radiance Fields. NeRF [??????] revolutionized novel view synthesis through implicit neural representations. Extensions include Mip-NeRF [????], Instant-NGP [???], and Plenoxels [???]. 3D Gaussian Splatting [?????] provides real-time rendering with explicit representations. SLAM integration [????????????????] enables online reconstruction.

Point Cloud Processing. PointNet [????] and PointNet++ [?] established deep learning on point clouds. Point Transformer [?????] applies attention mechanisms. Point-BERT [??], Point-MAE [???], and PointGPT [??] enable self-supervised pretraining. 3D semantic segmentation [??????] processes large-scale point clouds.

Depth Estimation. Monocular depth estimation [????????] enables 3D understanding from single images. MiDaS [?] provides robust cross-dataset depth estimation. Depth Anything [?] scales depth estimation with large datasets.

Semantic Segmentation. Semantic segmentation [??????] provides pixel-level scene understanding. Open-vocabulary segmentation [??????] enables recognition of novel categories.

4.2.3 Manipulation

Manipulation involves physical interaction with objects through robotic systems.

Vision-Language-Action Models. RT-1 [????] demonstrated large-scale robot learning with transformer architectures. RT-2 [?] showed VLM transfer to robotic control. RT-X [??] enables cross-embodiment

transfer across 22 robot platforms. Octo [?] provides an open-source generalist robot policy. OpenVLA [?] offers open vision-language-action models. π_0 [?] demonstrates flow matching for robot control. RoboCat [?] shows self-improvement through autonomous practice. Additional VLA models include 3D-VLA [?], GR-1 [?], RoboFlamingo [?], ManipLLM [?], GraspGPT [??], and comprehensive surveys [??].

Imitation Learning. Behavior cloning [?????] learns policies from demonstrations. Diffusion Policy [??] applies diffusion models to action generation. ACT [?] enables learning from teleoperation. Learning from play [??] leverages unstructured demonstrations. ALOHA [?] provides bimanual teleoperation for data collection.

Reinforcement Learning. Model-free RL [?????] provides policy optimization. Model-based RL [???] enables sample-efficient learning. Offline RL [???] learns from fixed datasets. Sim-to-real transfer [????] bridges simulation and reality.

Grasping and Manipulation. Grasp detection [????] enables object pickup. Dexterous manipulation [??????] handles complex object interactions. Language-conditioned manipulation [?????????] follows natural language instructions.

Simulation Environments. RLBench [??], Meta-World [?], ManiSkill [??], Isaac Gym [?], RoboTHOR [??], and AI2-THOR [?] provide simulation platforms. Objaverse [??] and ShapeNet [?] provide 3D object assets.

Emerging Directions. Recent work explores scaling laws for robotics [??], cross-embodiment transfer [??], integration with world models [?], and foundation models for robotics [????]. Scene-level understanding [?] and motor control integration [?] represent additional frontiers.

4.2.4 Geospatial Analysis

Geospatial analysis processes geographic data including satellite imagery, maps, and location data.

Remote Sensing Foundation Models. Prithvi [????] provides geospatial foundation models for Earth observation. SatMAE [??] applies masked autoencoders to satellite imagery. GeoFM [??] surveys foundation models for geospatial applications. CROMA [??] enables cross-modal remote sensing analysis. GeoChat [?], RemoteSenseGPT [?], SkyEyeGPT [?], and EarthGPT [?] apply LLMs to remote sensing. GeoLLM [??] enables geospatial reasoning with language models. RemoteCLIP [?] and GeoCLIP [?] provide vision-language models for remote sensing. Microestimates [??] derive socioeconomic indicators from satellite data.

Spatio-Temporal Graph Neural Networks. DCRNN [?] models traffic as bidirectional graph diffusion. STGCN [?] combines graph and temporal convolutions. Graph WaveNet [?] learns adaptive graph structures. AGCRN [?] introduces node-specific patterns. STGRAT [?] applies attention for spatio-temporal prediction. Comprehensive surveys [?????] detail these advances.

Urban Computing. Urban computing [????????] applies AI to city-scale problems. Traffic prediction [????], POI recommendation [???], and urban flow prediction [???] are key applications. ST-LLM [??] applies large language models to spatio-temporal data. UniST [??] provides unified spatio-temporal learning. UrbanGPT [?] enables urban prediction with LLMs. GeographRAG [?] combines retrieval with geospatial reasoning. Spatio-temporal interaction [??] models complex urban dynamics.

5 State-of-the-Art Methods

5.1 Graph Neural Networks for Spatial Reasoning

GNNs provide powerful tools for modeling spatial relationships and dependencies across diverse domains. We present the mathematical foundations underlying these architectures.

Message Passing Framework. The general message passing neural network (MPNN) framework [???] defines node updates through:

$$\mathbf{h}_v^{(l+1)} = \phi\left(\mathbf{h}_v^{(l)}, \bigoplus_{u \in \mathcal{N}(v)} \psi\left(\mathbf{h}_v^{(l)}, \mathbf{h}_u^{(l)}, \mathbf{e}_{uv}\right)\right) \tag{1}$$

where $\mathbf{h}_v^{(l)}$ is the hidden state of node v at layer l, $\mathcal{N}(v)$ denotes the neighbors of v, ψ is the message function, $\bigoplus$ is a permutation-invariant aggregation (e.g., sum, mean, max), and ϕ is the update function.

Foundational Architectures. GCN [?] introduced spectral graph convolution through first-order Chebyshev polynomial approximation:

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \tag{2}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with self-loops, $\tilde{\mathbf{D}}$ is the degree matrix, and $\mathbf{W}^{(l)}$ are learnable parameters.

GAT [?] added attention mechanisms for adaptive aggregation:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_k]\right)\right)} \tag{3}$$

GraphSAGE [?] enabled inductive learning on unseen nodes through neighborhood sampling. GIN [?] provided theoretical expressiveness analysis connecting to Weisfeiler-Lehman tests. Additional architectures include SGC [??], APPNP [?], GPR-GNN [?], spectral approaches [??], graph attention [??], and comprehensive surveys [????????]. Graph reasoning [????] extends capabilities.

Spatio-Temporal Networks. DCRNN [?] models traffic as bidirectional graph diffusion:

$$\mathbf{H}^{(l)} = \sum_{k=0}^{K}\left(\mathbf{P}_f^k\mathbf{X}\mathbf{W}_{k,1} + \mathbf{P}_b^k\mathbf{X}\mathbf{W}_{k,2}\right) \tag{4}$$

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

STGCN [?] combines graph and temporal convolutions through a sandwiched structure. Graph WaveNet [?] learns adaptive graph structures without predefined adjacency:

$$\tilde{\mathbf{A}} = \text{SoftMax}\left(\text{ReLU}\left(\mathbf{E}_1\mathbf{E}_2^T\right)\right) \tag{5}$$

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

AGCRN [?] introduces node-specific patterns through adaptive modules. ASTGCN [??] adds spatial and temporal attention mechanisms. GMAN [?] uses graph multi-attention with transform attention for long-range dependencies. Comprehensive surveys [??????????] detail these advances.

GNN-LLM Integration. Emerging work combines GNNs with LLMs for structured spatial reasoning [???????????]. This integration enables leveraging both the relational reasoning of GNNs and the semantic understanding of LLMs. Graph instruction tuning [??????????] further enhances this capability. LLaGA [?] provides language-graph alignment. GraphGPT [??????] enables graph reasoning through language models. Knowledge graph integration [???????] extends semantic reasoning.

Geometric Deep Learning. Geometric deep learning [??] provides theoretical foundations for spatial reasoning on non-Euclidean domains. Equivariant networks [???] respect spatial symmetries through:

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

where T_g is a group transformation. Graph transformers [???????] combine attention with graph structure. E3NN [?] and geometric message passing [?] advance equivariant architectures.

5.2 World Models

World models learn predictive representations enabling planning through imagination.

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

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

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

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

Model-Based Reinforcement Learning. Dreamer [????] introduced latent imagination for sample-efficient learning through recurrent state-space models. DreamerV2 [?] achieved human-level Atari performance with discrete latent states. DreamerV3 [?] demonstrated cross-domain mastery with a single algorithm through symlog predictions. DayDreamer [?] transferred world models to real robots with minimal real-world data. PlaNet [?] pioneered latent dynamics learning. MuZero [?] combined learned models with MCTS for game playing. Additional approaches include MBPO [?], SLAC [?], TD-MPC [?], and World Models [???].

Video World Models. Genie [?] learns controllable world models from internet videos enabling interactive environments. WorldDreamer [?] generates driving world models for autonomous vehicles. GAIA-1 [?] produces realistic driving videos conditioned on actions and text. Sora [??] demonstrates video generation as world simulation at scale. Video prediction models [???????] provide foundations for world understanding.

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

World Models for Robotics. World models for robotics [????????????] enable sample-efficient learning and safe exploration. World models for autonomous driving [??] provide simulation for planning.

5.3 Multimodal Foundation Models

Multimodal models integrate vision, language, and action understanding.

Vision-Language Models. CLIP [??] enabled zero-shot visual recognition through contrastive pre-training on web-scale data. BLIP-2 [?] introduced efficient vision-language pretraining with frozen encoders. LLaVA [??] demonstrated visual instruction tuning with strong performance. GPT-4V [???] achieved strong multimodal reasoning. Gemini [???] provides native multimodal capabilities. Flamingo [?] enables few-shot visual learning through interleaved attention. PaLI [????] scales vision-language models. Kosmos-2 [??] adds grounding capabilities. Qwen-VL [?] provides open multilingual VLMs. Additional models include InstructBLIP [?], MiniGPT-4 [????], Otter [?], CogVLM [???], InternVL [???], IDEFICS2 [?], mPLUG-Owl2 [??], Ferret [?], and VisionLLM [??].

Spatial Vision-Language Models. SpatialVLM [??] specializes in spatial reasoning with fine-grained understanding. SpatialRGPT [?] provides regional spatial reasoning. VoxPoser [?] extracts affordances from VLMs into 3D representations. VLMaps [?] creates semantic spatial maps for navigation. These models bridge vision-language understanding with spatial reasoning.

3D Vision-Language Models. 3D-LLM [??????] enables language understanding of 3D scenes. Open3D-VQA [????] provides open-vocabulary 3D visual question answering. LLM-Grounder [???] grounds language in 3D environments.

Evaluation Benchmarks. VLM evaluation benchmarks include MMMU [???], MathVista [?], MME [?], SEED-Bench [??], MMBench [?], and LMMS-Eval [?].

5.4 Embodied AI Agents

Open-Ended Exploration. Voyager [???] demonstrated open-ended exploration in Minecraft through LLM-driven curriculum learning and skill library construction. MineDojo [??] provides benchmarks for open-ended embodied agents with diverse tasks. DEPS [?] decomposes embodied planning systematically. MineAnyBuild [?] extends to construction tasks.

Grounded Language Agents. SayCan [?] grounds language models in robotic affordances through value functions. Code as Policies [?] generates executable robot code from language. LLM-Planner [?] enables few-shot grounded planning. EmbodiedGPT [???] provides embodied chain-of-thought reasoning. RoboBrain [??] integrates multiple capabilities.

Multi-Agent Systems. AutoGen [???] enables multi-agent conversations with flexible architectures. MetaGPT [??] assigns roles to agents for software development. CAMEL [??] explores communicative agents through role-playing. ChatDev [???] applies multi-agent systems to software development. AgentVerse

[???] provides multi-agent simulation frameworks. Multi-agent collaboration [?????] extends coordination capabilities.

Web Agents. WebArena [???] benchmarks web-based agent tasks. Mind2Web [??] provides web agent datasets. SWE-Bench [??] evaluates software engineering agents. SWE-Agent [?] provides agent interfaces for software tasks. WebGPT [?] enables web browsing for question answering. OSWorld [???], WorkArena [?], RealWorld [?], and SeeClick [??] extend web agent benchmarks.

6 Industry Applications

6.1 Geospatial Intelligence

Palantir [??????] integrates AI with geospatial analysis for defense and commercial applications, processing satellite imagery and sensor data at scale. The Gotham platform enables intelligence analysis with spatial reasoning. **ESRI** [?????] provides ArcGIS with integrated GeoAI capabilities for spatial analysis, supporting urban planning, environmental monitoring, and disaster response. **Google** [???????] deploys AI for global-scale mapping, navigation, and earth observation through Google Earth Engine and Maps AI. Google DeepMind advances embodied AI through robotics research [???].

Satellite and Earth Observation. Maxar provides high-resolution satellite imagery for geospatial intelligence and defense applications. **Planet Labs** [?] operates the largest constellation of Earth-imaging satellites, enabling daily global monitoring for agriculture, forestry, and disaster response. **NASA** [??] develops foundation models for Earth science through partnerships with IBM, including the Prithvi geospatial foundation model trained on Harmonized Landsat Sentinel-2 data. **IBM** collaborates on geospatial AI through the NASA-IBM partnership [??], developing open-source foundation models for climate and Earth observation.

Defense Applications. DARPA programs including MCS [?], ARM [?], and GCA [?] advance spatial AI for defense. NATO applications [?] integrate geospatial intelligence. GEOINT applications [?] leverage satellite imagery analysis.

Disaster Response. UN-Habitat [?] applies AI for urban resilience. UNDRR [?] provides AI tools for disaster risk reduction. FEMA [?] uses geospatial AI for damage assessment. Urban SDK [?] enables AI-powered disaster planning. Agentic AI for SAR [?] supports search and rescue operations.

6.2 Location Intelligence

Foursquare [???] provides location intelligence through movement pattern analysis and POI data, powering applications across retail, real estate, and urban analytics. **Wherobots** [?] provides cloud-native spatial analytics built on Apache Sedona, enabling large-scale geospatial data processing. **Carto** [?] offers spatial data science platforms for location intelligence and geospatial analysis. Smart city applications [??????????] leverage spatial AI for traffic management, energy optimization, and urban planning.

Mapping and Navigation. Google Maps [????] provides AI-powered navigation serving billions of users globally. Street-level AI [???] enables detailed urban understanding. Navigation systems [??] integrate multiple data sources. **World Labs** [?] develops spatial intelligence for 3D world understanding, founded by Fei-Fei Li to advance AI systems that perceive and reason about physical space.

6.3 Autonomous Vehicles

Waymo [?????] has deployed autonomous vehicles at scale with millions of miles driven. EMMA [?] provides end-to-end multimodal models for driving. Waymax [?] offers simulation for autonomous driving research. End-to-end approaches including UniAD [?], VAD [??], DriveVLM [?], and S4Driver [?] unify perception, prediction, and planning.

Tesla [???] pursues vision-only autonomy with neural network-based planning. **Cruise** [?], **Mobileye** [?], and **NVIDIA** [?] provide additional autonomous driving solutions. Motion prediction [??????????] and perception [??] are critical components.

Datasets and Benchmarks. nuScenes [?], Waymo Open [??], KITTI [??], Argoverse [??], CARLA [?], and EuRoC [?] provide evaluation platforms.

6.4 Robotics

Robot Learning Platforms. Open X-Embodiment [???] provides large-scale robot data from Google DeepMind and collaborating institutions. Bridge Data [???] enables cross-domain transfer. Simulation platforms [??????] support policy development. Objaverse [??], ShapeNet [?], and ScanNet [??] provide 3D assets. RH20T [?] provides large-scale robot manipulation data.

Foundation Model Labs. OpenAI [????] advances multimodal AI with GPT-4V for vision-language understanding and Sora for video generation as world simulation. **Anthropic** [??] develops Claude with constitutional AI principles for safe and helpful assistants. **Google DeepMind** [????] leads in both foundation models (Gemini) and robotics (RT-1, RT-2). **Meta AI** [????] provides open-source LLMs and vision models including DINOv2 for self-supervised visual learning.

Emerging Applications. Leidos C2AI [?] applies agentic AI for command and control. Risk assessment [?] uses spatial AI for safety. Site understanding [?] enables construction applications.

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.

7.1 SpatialAgentBench Tasks

Our proposed benchmark includes eight tasks designed to evaluate the full spectrum of spatial agent capabilities:

1. **VLN-Instruct**: Vision-language navigation with complex, multi-step instructions requiring spatial reasoning and landmark recognition.
2. **ObjectSearch**: Multi-room object search with semantic reasoning, requiring agents to leverage commonsense knowledge about object locations.
3. **SceneQA**: 3D scene question answering requiring understanding of spatial relationships, object properties, and scene semantics.
4. **ManipSeq**: Sequential manipulation planning with long-horizon tasks requiring tool use and state tracking.
5. **GeoReason**: Geospatial reasoning from satellite imagery including change detection, land use classification, and spatial pattern analysis.
6. **TrafficPredict**: Spatio-temporal traffic prediction requiring modeling of complex urban dynamics and graph-structured dependencies.
7. **SafeNav**: Navigation with safety constraints including obstacle avoidance, social navigation, and risk-aware planning.
8. **MultiAgent**: Coordinated multi-agent spatial tasks requiring communication, task allocation, and collaborative planning.

7.2 Evaluation Metrics

We propose standardized metrics across domains with formal definitions:

Navigation Metrics. Success Rate (SR) measures task completion. Success weighted by Path Length (SPL) [?] accounts for path efficiency:

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

Table 1: Comparison of Spatial Intelligence Benchmarks

Benchmark	Task	Environment	Metrics	Key Feature
Navigation				
R2R [?]	VLN	Real-world images	SPL, SR	First large-scale VLN
RxR [?]	VLN	Real-world images	nDTW, SR	Multilingual
REVERIE [?]	VLN	Real-world images	RGS	Remote grounding
Habitat ObjectNav [?]	ObjectNav	Simulated	SPL, Success	Standardized
SOON [?]	ObjectNav	Simulated	NDO	Semantic
TouchDown [?]	VLN	Street View	TC, SPD	Urban navigation
EmbodiedBench [?]	Embodied	Simulated	Success Rate	Comprehensive
EmbodiedEval [?]	Embodied	Simulated	Multiple	Multi-task
Manipulation				
RLBench [?]	100+ tasks	Simulated	Success Rate	Diverse tasks
Meta-World [?]	50 tasks	Simulated	Success Rate	Meta-learning
BEHAVIOR [?]	1000 activities	Simulated	Goal Conditions	Long-horizon
Open X-Embodiment [?]	22 robots	Real-world	N/A	Largest real dataset
ManiSkill2 [?]	20 tasks	Simulated	Success Rate	Soft-body physics
CALVIN [?]	Language	Simulated	Success Rate	Long-horizon language
VIMA [?]	Multimodal	Simulated	Success Rate	Multimodal prompts
Spatial Reasoning				
CLEVR [?]	VQA	Synthetic	Accuracy	Compositional
GQA [?]	VQA	Real-world	Accuracy	Scene graphs
SpatialVLM [?]	VQA	Real-world	Accuracy	Fine-grained spatial
ScanQA [?]	3D VQA	Real scans	EM, BLEU	3D understanding
Open3D-VQA [?]	3D VQA	Real scans	Accuracy	Open-vocabulary
Geospatial				
BigEarthNet [?]	Classification	Satellite	Accuracy, F1	Large-scale
fMoW [?]	Classification	Satellite	Accuracy	Temporal
xBD [?]	Segmentation	Satellite	IoU, F1	Damage assessment
SpaceNet [?]	Detection	Satellite	AP	Building footprints
GeoBench [?]	Multiple	Satellite	Multiple	Standardized
Agent Benchmarks				
AgentBench [???]	Multiple	Multiple	Success Rate	Comprehensive
WebArena [?]	Web	Browser	Success Rate	Web tasks
ALFWorld [?]	Embodied	Simulated	Success Rate	Text-world
SafeAgentBench [?]	Safety	Simulated	Safety Rate	Safety evaluation

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

Normalized Dynamic Time Warping (nDTW) measures trajectory similarity:

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

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

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

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

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

7.3 Benchmark Limitations and Critical Analysis

While existing benchmarks have advanced the field, several fundamental limitations warrant critical examination:

Simulation-Reality Gap. Most benchmarks rely on simulated environments [????], which differ from real-world conditions in visual appearance, physics, and dynamics. Policies trained in simulation often fail to transfer [??], limiting practical applicability.

Metric Limitations. Standard metrics like SPL assume optimal paths are known, which is unrealistic in novel environments. Success Rate ignores partial progress and efficiency. Current metrics do not capture important aspects such as safety, robustness to perturbations, and graceful degradation.

Dataset Biases. Benchmarks often exhibit biases in object distributions, scene layouts, and instruction patterns [???]. Models may exploit spurious correlations rather than learning generalizable spatial reasoning.

Evaluation Scope. Most benchmarks evaluate single capabilities in isolation. Real-world deployment requires integrated evaluation of perception, reasoning, planning, and action under uncertainty.

Reproducibility Challenges. Variations in simulation versions, random seeds, and evaluation protocols make cross-paper comparisons difficult. Standardized evaluation frameworks are needed.

8 Open Challenges and Future Directions

8.1 Robust Spatial Representation

Developing representations that generalize across scenes, viewpoints, and conditions remains challenging [??????]. Foundation models for 3D understanding [????????] represent promising directions. Key challenges include:

- **Occlusion handling**: Reasoning about hidden objects and occluded regions
- **Dynamic scenes**: Modeling temporal changes and object motion
- **Novel categories**: Generalizing to unseen object types and environments
- **Scale variation**: Handling objects and scenes across different scales
- **Viewpoint invariance**: Maintaining consistent understanding across perspectives

8.2 Long-Horizon Planning

Creating agents that plan over extended horizons and decompose complex spatial tasks is essential [???????????]. Integration of neural and symbolic planning approaches [???????] shows promise. Challenges include:

- **Credit assignment**: Attributing success or failure to specific actions
- **Subgoal discovery**: Automatically identifying useful intermediate goals
- **Plan repair**: Adapting plans when execution deviates from expectations
- **Temporal abstraction**: Operating at multiple time scales
- **Uncertainty handling**: Planning under incomplete information

8.3 Safe and Reliable Operation

Ensuring safe operation in safety-critical applications is paramount [???????????]. Key requirements include:

- **Uncertainty quantification**: Knowing when the agent is uncertain
- **Out-of-distribution detection**: Recognizing novel situations

- **Alignment**: Ensuring behavior matches human values and intentions
- **Interpretability**: Providing explanations for decisions
- **Graceful degradation**: Failing safely under adversarial conditions
- **Robustness**: Maintaining performance under distribution shift

8.4 Sim-to-Real Transfer

Bridging simulation and reality remains challenging [????????]. The reality gap affects perception, dynamics, and control. Key approaches include:

- **Domain randomization**: Training with varied simulation parameters
- **System identification**: Learning accurate dynamics models
- **Real-world fine-tuning**: Adapting with limited real data
- **Photorealistic simulation**: Reducing visual domain gap
- **Hybrid approaches**: Combining simulation and real-world data

8.5 Multi-Agent Coordination

Scaling to multi-agent systems for complex spatial tasks requires advances in coordination and communication [???????????]. Challenges include:

- **Emergent communication**: Developing shared protocols
- **Credit assignment**: Attributing team success to individuals
- **Scalable coordination**: Handling large numbers of agents
- **Heterogeneous teams**: Coordinating diverse agent types
- **Partial observability**: Operating with limited information

8.6 Efficiency and Deployment

Deploying spatial AI systems on resource-constrained platforms requires advances in model compression, efficient inference, and edge computing [?????]. Considerations include:

- **Model compression**: Reducing model size while maintaining performance
- **Efficient architectures**: Designing compute-efficient models
- **Hardware acceleration**: Leveraging specialized hardware
- **Real-time operation**: Meeting latency requirements
- **Energy efficiency**: Operating within power constraints

8.7 Emerging Directions

Several emerging directions show promise for advancing autonomous spatial intelligence:

Thinking Models. Models that explicitly reason about spatial relationships [?????] may improve planning and decision-making.

Autonomous GIS. Autonomous geographic information systems [????] integrate agentic capabilities with geospatial analysis.

Agentic AI for Specialized Domains. Applications in defense [??], disaster response [?], and risk assessment [?] represent growing areas. BIM integration [?], 6G networks [?], and urban habitat [?] extend spatial AI applications.

Multimodal Integration. Deeper integration of vision, language, audio, and tactile modalities [????] enables richer environmental understanding. SpatialBot [?] and InternLM [?] advance multimodal spatial reasoning.

9 Limitations

This survey, while comprehensive, has several limitations:

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

10 Conclusion

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

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

Key challenges remain in robust representation, long-horizon planning, safe deployment, sim-to-real transfer, multi-agent coordination, and efficient deployment. The convergence of vision-language-action models, graph neural networks, world models, and foundation models provides promising directions for addressing these challenges.

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

References

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

Sheikh Kamran Abid, Ruhizal Roosli, Umber Nazir, and Nur Shazwani Kamarudin. Ai-enhanced crowdsourcing for disaster management: strengthening community resilience through social media. *International Journal of Emergency Medicine*, 18(1):201, 2025.

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.

Saurabh Agrawal, Chunhui Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023.

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.

Ahmad Almadhor, Abdullah Al Hejaili, Shtwai Alsubai, Mehrez Marzougui, Tariq Alqubaysi, and Vincent Karovič. A multimodal learning and simulation approach for perception in autonomous driving systems. *Scientific Reports*, 2026.

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

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

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

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

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

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

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018c.

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.

Erin Antcliffe. 3 ways cruise hd maps give our self-driving vehicles an edge. *Cruise, Medium*, November 2019.

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.

Hinata Aoki and Takao Yamanaka. Improving nerf with height data for utilization of gis data, 2023.

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, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016.

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.

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

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

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

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

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

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

Mustafa Baniodeh, Kratarth Goel, Scott Ettinger, Carlos Fuertes, Ari Seff, Tim Shen, Cole Gulino, Chenjie Yang, Ghassen Jerfel, Dokook Choe, Rui Wang, Vinutha Kallem, Sergio Casas, Rami Al-Rfou, Benjamin Sapp, and Dragomir Anguelov. Scaling laws of motion forecasting and planning. *arXiv preprint arXiv:2506.08228*, 2025.

Amir Bar et al. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024.

Matt Barnes et al. World scale inverse reinforcement learning in google maps. Google Research Blog, May 2023. URL `https://research.google/blog/world-scale-inverse-reinforcement-learning-in-google-maps/`.

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. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *IEEE International Conference on Computer Vision*, 2021.

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

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. 2021.

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

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

Dhruv Batra, Angel X. Chang, Sonia Chernova, Andrew J. Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, Manolis Savva, and Hao Su. Rearrangement: A challenge for embodied ai, 2020a.

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*, 2020b.

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

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

Jens Behley et al. Semantickitti: A dataset for semantic scene understanding of lidar sequences. *IEEE International Conference on Computer Vision*, 2019.

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

Christopher Berner et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Luc Berthouze and Tom Ziemke. Epigenetic robotics: modelling cognitive development in robotic systems. *Connection Science*, 15(4):145–150, 2003.

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

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

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

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

Asier Bikandi, Miguel Fernandez-Cortizas, Muhammad Shaheer, Ali Tourani, Holger Voos, and Jose Luis Sanchez-Lopez. Bim informed visual slam for construction monitoring, 2025.

Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446):eaat8414, 2019.

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

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

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

Valts Blukis, Dipendra Misra, Ross A Knepper, and Yoav Artzi. Mapping navigation instructions to continuous control actions with position-visitation prediction. 2019.

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

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

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

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Antoine Bordes et al. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 2013.

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

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

Cruz E. Borges, Oihane Kamara Esteban, Ander Pijoan, and Yoseba K. Penya. Multi-agent gis system for improved spatial load forecasting. In *Adaptive Agents and Multi-Agent Systems*, 2014. URL `https://api.semanticscholar.org/CorpusID:41945022`.

Antoine Bosselut et al. Comet: Commonsense transformers for automatic knowledge graph construction. *Association for Computational Linguistics*, 2019.

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

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

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

Anthony Brohan, Noah Brown, Justice Carbajal, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

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

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

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

D. Brown, E. Davis, and F. Miller. Geospatial data integration with enterprise systems: A case study of sap and arcgis. *International Journal of Geographical Information Science*, 35(8):1567–1589, 2021.

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.

Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 2021a.

Marshall Burke et al. Using satellite imagery to understand and promote sustainable development. *Science*, 2021b.

Michael Burri et al. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016.

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

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

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

Berk Calli et al. The ycb object and model set: Towards common benchmarks for manipulation research. *IEEE International Conference on Advanced Robotics*, 2015.

Carlos Campos et al. Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE Transactions on Robotics*, 2021.

Defu Cao et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in Neural Information Processing Systems*, 2020.

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

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

CARTO. Carto: Spatial data science platform. `https://carto.com/`, 2024.

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

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

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE, 2017.

Angel X Chang et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, et al. Argoverse: 3d tracking and forecasting with rich maps. 2019.

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

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

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

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

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

Harrison Chase. Langchain. *GitHub repository*, 2022.

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

Shishir Singh Chauhan, Yogesh Kumar Jain, Praveen Kumar Mannepalli, and Ankur Pandey. 6g conditioned spatiotemporal graph neural networks for real time traffic flow prediction. *Scientific Reports*, 16(1):3902, 2026.

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

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

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*, 2023b.

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.

Boyuan Chen et al. Nlmap-saycan. *ICRA*, 2023c.

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

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

Dian Chen et al. Learning by cheating. *Conference on Robot Learning*, 2020c.

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

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

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.

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

Jiacheng Chen et al. Maptracker: Tracking with strided memory fusion for consistent vector hd mapping. *arXiv preprint arXiv:2403.15951*, 2024e.

Jiahui Chen et al. Maagent: A multi-agent approach for embodied reasoning. *arXiv preprint arXiv:2402.11776*, 2024f.

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*, 2024g.

Jiaqi Chen et al. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. *Association for Computational Linguistics*, 2024h.

Jihan Chen et al. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024i.

Jingyu Chen, Ruidong Ma, and John Oyekan. A deep multi-agent reinforcement learning framework for autonomous aerial navigation to grasping points on loads. *Robotics and Autonomous Systems*, 167:104489, 2023d.

Jun Chen et al. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2024j.

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

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

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 801–818, 2018.

Liang-Chieh Chen et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020d.

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

Lili Chen et al. Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, 2021b.

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*, 2021c.

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

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

Shaoyu Chen et al. Vad v2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint*, 2024m.

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

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

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

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

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

Tao Chen et al. A system for general in-hand object re-orientation. *CoRL*, 2022h.

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

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

Weize Chen, Yusheng Su, Jingwei Zuo, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. 2024p.

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

X. Chen. Application of gnn in urban computing. In *2020 5th International Conference on Smart and Sustainable City (ICSSC)*, pages 1–4. IEEE, 2020.

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*, 2023g.

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

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

Xieyuanli Chen, Andres Milioto, Emanuele Palazzolo, Philippe Giguère, Jens Behley, and Cyrill Stachniss. Suma++: Efficient lidar-based semantic slam. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4530–4537. IEEE, November 2019d. doi: 10.1109/iros40897.2019.8967704. URL `http://dx.doi.org/10.1109/IROS40897.2019.8967704`.

Xieyuanli Chen et al. Overlapnet: Loop closing for lidar-based slam. *Robotics: Science and Systems*, 2022i.

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

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

Yongchao Chen et al. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? *arXiv preprint arXiv:2309.15943*, 2024r.

Yuzhou Chen et al. Z-gcnets: Time zigzags at graph convolutional networks for time series forecasting. *International Conference on Machine Learning*, 2022j.

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

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

Zhe Chen, Jiannan Wu, Wenhai Wang, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. 2024u.

Zhe Chen et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024v.

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

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. Exploring the potential of large language models (llms) in learning on graphs. *arXiv preprint arXiv:2307.03393*, 2023k.

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

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

Gong Cheng et al. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017.

Kanzhi Cheng et al. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 2024b.

Liang Cheng, Yi Yuan, Nan Xia, Song Chen, Yanming Chen, Kang Yang, Lei Ma, and Manchun Li. Crowd-sourced pictures geo-localization method based on street view images and 3d reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 141:72–85, 2018.

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

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

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

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

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*, 2022a.

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

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

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.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

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.

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

Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. In *arXiv preprint arXiv:1907.06571*, 2019.

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

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

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.

Robin Courant, Xi Wang, Marc Christie, and Vicky Kalogeiton. Blunf: Blueprint neural field, 2023.

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.

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

Ravinder S Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini. Tactile sensing for robotic applications. *IEEE Sensors journal*, 10(11):100–115, 2010.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

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

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

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

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. Multi-agent collaboration via evolving orchestration, 2025. URL `https://arxiv.org/abs/2505.19591`.

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

DARPA. Autonomous Robotic Manipulation (ARM). urlhttps://www.darpa.mil/research/programs/autonomous-robotic-manipulation, a. Accessed: 2026-01-29.

DARPA. Geospatial Cloud Analytics (GCA). urlhttps://www.darpa.mil/research/programs/geospatial-cloud-analytics, b. Accessed: 2026-01-29.

DARPA. AI Improves Robotic Performance in DARPA's Machine Common Sense Program. urlhttps://www.darpa.mil/news/2022/machine-common-sense-program, June 2022. Accessed: 2026-01-29.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering, 2017.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018a.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. *arXiv preprint arXiv:1711.11543*, 2018b.

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

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

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

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *CVPR*, 2020.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, et al. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

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

Matt Deitke et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 2024.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2024.

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

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

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

Xiaoyi Dong et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

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

Vishnu Sashank Dorbala et al. Can an embodied agent find your cat-shaped mug? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 2024.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. 2017.

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

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

Laura Downs et al. Google scanned objects: A high-quality dataset of 3d scanned household items. *IEEE International Conference on Robotics and Automation*, 2022.

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.

Alexandre Drouin et al. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.

Yilun Du et al. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Yilun Du et al. Learning universal policies via text-guided video generation. *arXiv preprint arXiv:2302.00111*, 2024.

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

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

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

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

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

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

Frederik Ebert et al. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *Robotics: Science and Systems*, 2022.

Niv Efron and Luke Barrington. Google earth ai: Unlocking geospatial insights with foundation models and cross-modal reasoning. Google Research Blog, October 2025. URL https://research.google/blog/google-earth-ai-unlocking-geospatial-insights-with-foundation-models-and-cross-modal-reasoning/.

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

Belal Elshenety and Mio Nakagawa. Agentic ai in search rescue. Cal Poly Digital Commons. URL https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1039&context=ceng_surp.

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

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

Jakob Engel et al. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Clemens Eppner et al. Acronym: A large-scale grasp dataset. *IEEE International Conference on Robotics and Automation*, 2021.

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.

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

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

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

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

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

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

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

Federal Emergency Management Agency. Geospatial damage assessments (dhs-346). DHS AI Use Case Inventory, April 2025. URL https://www.dhs.gov/ai/use-case-inventory/fema.

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

Yanlin Feng et al. Scalable multi-hop relational reasoning for knowledge-aware question answering. *Conference on Empirical Methods in Natural Language Processing*, 2020.

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

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

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

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

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

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

David Freeman. Palantir's role in government and commercial analytics. Industry Analysis, 2021.

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.

Jon E. Froehlich and Shaun Kane. Streetreaderai: Towards making street view accessible via context-aware multimodal ai. Google Research Blog / UIST 2025, October 2025. URL https://research.google/blog/streetreaderai-towards-making-street-view-accessible-via-context-aware-multimodal-ai/.

Chaoyou Fu, Peixian Chen, Yunhang Shen, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023a.

Huan Fu et al. 3d-qa: A benchmark for 3d question answering. In *arXiv preprint*, 2021.

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*, 2023b.

Zipeng Fu et al. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024c.

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

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

Takeshi Fujita, 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.

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

Konstantin Fuller, Johannes Jakubik, Michal Muszynski, et al. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 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. In *CVPR*, 2023.

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

Samir Yitzhak Gadre et al. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

Chuang Gan, Jeremy Schwartz, Seth Alter, et al. Threedworld: A platform for interactive multi-modal physical simulation. 2021.

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.

Peng Gao et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint*, 2024.

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.

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

Gemma Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

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

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

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

Justin Gilmer et al. Neural message passing for quantum chemistry. *International Conference on Machine Learning*, 2017b.

Clement Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019a.

Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019b.

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

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

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

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

Google. Geospatial ai with real-world intelligence. Google Maps Platform, 2025. URL `https://mapsplatform.google.com/ai/`.

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

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

Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

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

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

Significant Gravitas. Auto-gpt: An autonomous gpt-4 experiment. *GitHub repository*, 2023.

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, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2024a.

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

Jian Guan, Yijun Wang, and Yiping Li. A survey of 6dof object pose estimation methods for robotic manipulation. *Sensors*, 24(4):1076, 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.

Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, JD Co-Reyes, Rishabh Agarwal, Becca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Ben Sapp, and Dragomir Anguelov. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=3i4A332644`.

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.

Jiahui Guo et al. Gpt4kg: Empowering large language models with knowledge graphs. *arXiv preprint arXiv:2310.04562*, 2024b.

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. 2019a.

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

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*, 2024c.

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*, 2024d.

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

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.

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

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.

Izzeddin Gur et al. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*, 2024.

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*, 2018a.

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

Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. *Conference on Robot Learning*, 2022.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 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, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. 2020a.

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.

Danijar Hafner et al. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020b.

R. Spencer Hallyburton and Miroslav Pajic. Trusted data fusion, multi-agent autonomy, autonomous vehicles, 2025. URL `https://arxiv.org/abs/2507.17875`.

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.

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

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

Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *ICRA*, 2020.

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.

V Haria, Y Shah, V Gangwar, et al. The working of google maps, and the commercial usage of navigation systems. *International Journal of Innovative Research in Technology*, 6(1):34–39, 2019.

Hongliang He et al. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*, 2024.

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

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

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2021a.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021b.

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

Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. Learning to follow directions in street view, 2019.

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*, 2024a.

Jianing Hong, Zeren Zheng, Hao Zhu, Yun Xu, Xingyu Zhang, Siheng Chen, and Shenghua Shen. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. *arXiv preprint arXiv:2309.12311*, 2024b.

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.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, et al. Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2024c.

Wenyi Hong, Weihan Wang, Qingsong Lv, et al. Cogagent: A visual language model for gui agents. 2024d.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln-bert: A recurrent vision-and-language bert for navigation. *arXiv preprint arXiv:2011.13922*, 2020a.

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

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

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.

Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020.

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

Yihan Hu et al. Planning-oriented autonomous driving. *IEEE Conference on Computer Vision and Pattern Recognition*, 2023c.

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.

Bo Huang et al. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, 2021.

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

Chenguang Huang et al. Audio visual language maps for robot navigation. *arXiv preprint*, 2024a.

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*, 2024b.

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

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

Siyuan Huang, Zhengkai Jiang, Hao Dong, et al. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023b.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*, 2022a.

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*, 2022b.

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*, 2022c.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024e.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey, 2024f.

Xuanwen Huang et al. Can gnn be good adapter for llms? *arXiv preprint arXiv:2402.12984*, 2024g.

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.

Jena D Hwang et al. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *AAAI Conference on Artificial Intelligence*, 2021.

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

Mikhal Itkina. *Uncertainty-aware spatiotemporal perception for autonomous vehicles*. PhD thesis, Stanford University, 2022.

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.

Johannes Jakubik et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

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, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *CoRL*, 2022a.

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

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.

Krzysztof Janowicz et al. Geofm: how will geo-foundation models reshape spatial data science and geoai? *International Journal of Geographical Information Science*, 2025.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping, 2023.

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

Jiarui Ji, Runlin Lei, Jialing Bi, Zhewei Wei, Xu Chen, Yankai Lin, Xuchen Pan, Yaliang Li, and Bolin Ding. Llm-based multi-agent systems are scalable graph generative models, 2024.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, and Shanghang Zhang. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

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

Jinhao Jiang et al. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*, 2024a.

Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, November 2022a. ISSN 0957-4174. doi: 10.1016/j.eswa.2022.117921. URL `http://dx.doi.org/10.1016/j.eswa.2022.117921`.

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

Weiwei Jiang and Jiayun Luo. Graph neural networks for traffic forecasting: A survey. *arXiv preprint arXiv:2101.11174*, 2022c.

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

Weiwei Jiang, Jiayun Luo, Miao He, and Weixi Gu. Graph neural network for traffic forecasting: The research progress. *ISPRS International Journal of Geo-Information*, 12(3):100, 2023c. doi: 10.3390/ijgi12030100.

Xiaohui Jiang et al. Far3d: Expanding the horizon for surround-view 3d object detection. *AAAI Conference on Artificial Intelligence*, 2024b.

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

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*, 2023e.

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

Bowen Jin et al. Large language models on graphs: A comprehensive survey. *arXiv preprint arXiv:2312.02783*, 2024.

Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023a.

Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey, 2023b.

Mohammad Mahdi Johari et al. Eslam: Efficient dense slam system based on hybrid representation of signed distance fields. *IEEE Conference on Computer Vision and Pattern Recognition*, 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.

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.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. 2018.

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

Jian Kang, Rubén Fernández-Beltran, Danfeng Hong, and Antonio Plaza. Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 59:4355–4369, 2020. doi: 10.1109/TGRS.2020.3016020.

Peter Karkus, Shaojun Cai, and David Hsu. Differentiable slam-net: Learning particle slam for visual navigation, 2021.

Rithewik Kaushik, Ankesh Kumar, and Avinash Singh. Fast online adaptation in robotics through meta-learning embeddings of simulated priors. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8956–8962. IEEE, 2020.

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

Seokju Ke, Yilin Guo, Zijian He, et al. Segment any 3d object with language. *arXiv preprint*, 2024.

Zixuan Ke, Yifei Ming, Austin Xu, Ryan Chin, Xuan-Phi Nguyen, Prathyusha Jwalapuram, Semih Yavuz, Caiming Xiong, and Shafiq Joty. Mas-orchestra: Understanding and improving multi-agent reasoning through holistic orchestration and controlled benchmarks, 2026. URL `https://arxiv.org/abs/2601.14652`.

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, 2023a.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023b.

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

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

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

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *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.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Shimon Komarovsky and Jack Haddad. Spatio-temporal graph convolutional neural network for traffic signal prediction in large-scale urban networks. *Transportation Research Interdisciplinary Perspectives*, 32:101482, 2025.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. 2020a.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments, 2020b.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Waypoint models for instruction-guided navigation in continuous environments. *arXiv preprint arXiv:2110.02207*, 2020c.

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

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.

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*, 2020a.

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

Kartik Kuckreja et al. Geochat: Grounded large vision-language model for remote sensing. *arXiv preprint arXiv:2311.15826*, 2024.

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

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

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

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *NeurIPS*, 2020.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Geo-bench: Toward foundation models for earth monitoring. 2024.

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

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

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

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

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

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

Chia-Yen Lee and Shu-Huei Yang. Graph spatio-temporal networks for manufacturing sales forecast and prevention policies in pandemic era. *Computers Industrial Engineering*, page 109413, 2023.

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.

Leidos Editorial Team. Agentic ai aims to cut down emergency response time when disasters strike. Leidos Insights, October 2025. URL `https://www.leidos.com/insights/agentic-ai-aims-cut-down-emergency-response-time-when-disasters-strike`.

Sergey Levine et al. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

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.

Bohao Li, Rui Wang, Guangzhi Wang, et al. Seed-bench: Benchmarking multimodal large language models. *arXiv preprint arXiv:2307.16125*, 2023b.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2024a.

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

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

Jiachen Li, Hengbo Ma, Zhihao Zhang, Jinning Li, and Masayoshi Tomizuka. Spatio-temporal graph dual-attention network for multi-agent prediction and tracking, 2021.

Jiahui Li et al. World-centric diffusion transformer for 3d scene understanding. *arXiv preprint arXiv:2310.05917*, 2024b.

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*, 2023e.

Kaizhi Li, Xiang He, Xuehai Zhang, et al. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023f.

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

Mengmeng Li et al. Urban land use mapping using deep learning and multi-source data. *Remote Sensing of Environment*, 2022a.

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

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

Nuo Li et al. Graphwiz: An instruction-following language model for graph problems. *arXiv preprint arXiv:2402.16029*, 2024d.

Qi Li et al. Hdmapnet: An online hd map construction and evaluation framework. *IEEE International Conference on Robotics and Automation*, 2022b.

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

Shilong Li et al. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. *arXiv preprint arXiv:2406.14550*, 2024e.

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*, 2023j.

Wenwen Li et al. Geoai: Where machine learning and big data converge in giscience. *Journal of Spatial Information Science*, 2020c.

Wenwen Li et al. Geospatial copilot: Autonomous gis agent for spatial analysis. *arXiv preprint*, 2024f.

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

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *ICLR*, 2023k.

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*, 2023l.

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

Yunfan Li et al. Embodied world models for robot learning. *arXiv preprint arXiv:2402.16029*, 2024h.

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

Yuxuan Li et al. Trafficbert: Pre-trained model with large-scale data for long-range traffic flow forecasting. *Expert Systems with Applications*, 2023m.

Zhenghao Li et al. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2501.00000*, 2025a.

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

Zhonghang Li et al. Urbangpt: Spatio-temporal large language models. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024i.

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

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

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

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2023b.

Jianyuan Liang, Shuyang Hou, Haoyue Jiao, Yaxian Qing, Anqi Zhao, Zhangxiao Shen, Longgang Xiang, and Huayi Wu. Geographrag: A graph-based retrieval-augmented generation approach for empowering large language models in automated geospatial modeling. *International Journal of Applied Earth Observation and Geoinformation*, 142:104712, 2025.

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

Yuxuan Liang et al. Urbanfm: Inferring fine-grained urban flows. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2019.

Bencheng Liao et al. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, 2024.

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.

Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 2023.

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.

Zachary Lipson et al. Raft-3d: Scene flow using rigid-motion embeddings. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

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

Bo Liu et al. Llm+p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023b.

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

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

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

Hao Liu et al. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024c.

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

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

Jiayan Liu et al. Gpt4graph: Can large language models understand graph structured data? *arXiv preprint arXiv:2305.15066*, 2024e.

Patrick Langechuan Liu. A crash course of planning for perception engineers in autonomous driving. *The Thinking Car, Medium*, June 2024.

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*, 2023e.

Siqi Liu, Leonard Hasenclever, Steven Bohez, Guy Lever, Zhe Wang, S. M. Ali Eslami, and Nicolas Heess. From motor control to embodied intelligence. Google DeepMind Blog, August 2022a. URL `https://deepmind.google/blog/from-motor-control-to-embodied-intelligence/`.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023f.

Xiao Liu et al. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*, 2024f.

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

Yicheng Liu et al. Vectormapnet: End-to-end vectorized hd map learning. *International Conference on Machine Learning*, 2023g.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023h.

Yuan Liu et al. Remotesensegpt: A multimodal large language model for remote sensing. *arXiv preprint arXiv:2401.09712*, 2024h.

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

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

Ben Lorica. "world model" is a mess. here's how to make sense of it. *Gradient Flow*, 2025.

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

Pan Lu, Hritik Bansal, Tony Xia, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023a.

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*, 2023b.

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

Max Lungarella, Giorgio Metta, Rolf Pfeifer, and Giulio Sandini. Developmental robotics: a survey. *Connection Science*, 15(4):151–190, 2003.

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.

E. K. Lutema. The Role of Geospatial Intelligence in Modern Military Operations. *EarthArXiv*, 2025. doi: 10.31223/X5V72B.

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

Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Barber, Travis Hsu, Hao Pokorny, and Sergey Levine. Interactive language: Talking to robots in real time. In *IEEE Robotics and Automation Letters*, 2023.

Liheng Ma et al. Graph inductive biases in transformers without message passing. *International Conference on Machine Learning*, 2023.

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

Xuankai Ma, Zehua Zhang, and Yongze Song. Geographically informed graph neural networks. *Spatial Statistics*, 69:100920, 2025.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai, 2026.

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

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

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. Opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

Gengchen Mai, Weiming Huang, Jin Sun, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *ACM SIGSPATIAL Special*, 2024.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *arXiv preprint arXiv:2206.12403*, 2022.

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

K Malarvizhi, S Vasantha Kumar, and P Porchelvan. Use of high resolution google earth satellite imagery in landuse map preparation for urban related applications. *Procedia Technology*, 24:1835–1842, 2016.

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

Utkarsh Mall et al. Remoteclip: A vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*, 2023.

Rohin Manvi et al. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*, 2024.

Jiageng Mao et al. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023.

Jianan Mao et al. Graph-instruct: Instruction tuning for graph neural networks. *arXiv preprint arXiv:2310.05915*, 2024.

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

Ricardo Martin-Brualla et al. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021a.

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

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

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*, 2022.

Grégoire Mialon et al. Graphit: Encoding graph structure in transformers. *arXiv preprint arXiv:2106.05667*, 2021.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2020a.

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, 2020b.

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

Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. *arXiv preprint arXiv:2405.04390*, 2024.

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

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

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.

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

Amanda Leicht Moore. Google maps navigation gets a powerful boost with gemini. Google Blog, November 2025. URL `https://blog.google/products-and-platforms/products/maps/gemini-navigation-features-landmark-lens/`.

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

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

Norman Mu, Jingwei Ji, Zhenpei Yang, Nate Harada, Haotian Tang, Kan Chen, Charles R. Qi, Runzhou Ge, Kratarth Goel, Zoey Yang, Scott Ettinger, Rami Al-Rfou, Dragomir Anguelov, and Yin Zhou. Most: Multi-modality scene tokenization for motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17779–17789, 2024.

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

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

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

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

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

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

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

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

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

NASA. Nasa earthdata: Open access to nasa earth science data. `https://earthdata.nasa.gov/`, 2023.

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

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

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

Jiquan Ngiam et al. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. *International Conference on Learning Representations*, 2022.

Thuan Minh Nguyen, Vu Tuan Truong, and Long Bao Le. Agentic ai meets edge computing in autonomous uav swarms, 2026.

Jun Ni, Yisheng Chen, Guoyuan Tang, Jian Shi, Weihua Cao, and Peng Shi. Deep learning-based scene understanding for autonomous robots: A survey. *Intelligence Robotics*, 3(1):1–24, 2023.

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.

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

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

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

OpenAI. Gpt-4o system card. *OpenAI Technical Report*, 2024.

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

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

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

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

Long Ouyang et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022b.

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, Acorn Poolber, 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.

Palantir Technologies. Palantir gotham: Intelligence platform for defense and intelligence. Technical report, Palantir Technologies, 2024.

Shirui Pan et al. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

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*, 2019a.

Zheyi Pan et al. Urban traffic prediction from spatio-temporal data using deep meta learning. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2019b.

Hui En Pang and Filip Biljecki. 3d building reconstruction from single street view images using deep learning. *International Journal of Applied Earth Observation and Geoinformation*, 112:102897, 2022.

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

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

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

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.

Keunhong Park et al. Nerfies: Deformable neural radiance fields. *IEEE International Conference on Computer Vision*, 2021.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

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

Songyou Peng et al. Openscene: 3d scene understanding with open vocabularies. *IEEE Conference on Computer Vision and Pattern Recognition*, 2023b.

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*, 2023c.

Zhiliang Peng, Wenhui Wang, Li Dong, et al. Kosmos-2: Grounding multimodal large language models to the world. 2024.

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*, 2024a.

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

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

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

Pinecone. Pinecone vector database, 2023.

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

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

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

Albert Pumarola et al. D-nerf: Neural radiance fields for dynamic scenes. *IEEE Conference on Computer Vision and Pattern Recognition*, 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 R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017b.

Charles R Qi et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. 2017c.

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, 2017d.

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.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2024a.

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

Rongzheng Qian et al. Graphtool-instruction: Revolutionizing graph reasoning in llms through decomposed subtask instruction. *arXiv preprint arXiv:2407.09176*, 2024b.

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

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

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

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.

Rafael Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023.

Kareff Rafisura, Sheryl Rose Reyes, Natdanai Punsin, and Emiyati. Guiding disaster risk reduction investments through ai powered tools. IDDRR, October 2025. URL `https://iddrr.undrr.org/news/guiding-disaster-risk-reduction-investments-through-ai-powered-tools-0`.

Aowabin Rahman, Salman Shuvo, Samrat Chatterjee, Mahantesh Halappanavar, and Terje Aven. Risk-aware autonomous search and rescue with multiagent reinforcement learning. *Risk Analysis*, 45(12):4490–4504, 2025.

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

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

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

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

Krishan Rana et al. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. *Conference on Robot Learning*, 2023.

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

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

Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021b.

Christopher Rawles et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.

Yuhan Ren et al. A survey of graph meets large language model: Progress and future directions. *arXiv preprint arXiv:2311.12399*, 2024.

Reply. World models: the operating system for spatial intelligence. *Reply*, 2026.

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

Jonathan Roberts et al. Gpt4geo: How a language model sees the world's geography. *arXiv preprint arXiv:2306.00020*, 2023.

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

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*, 2020b.

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)*, 2020c.

Antoni Rosinol, John J. Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields, 2022a.

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

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

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

Zahraa Al Sahili and Mariette Awad. Spatio-temporal graph neural networks: A survey, 2023. URL `https://arxiv.org/abs/2301.10569`.

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

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

Ranjan Sapkota, Yang Cao, Konstantinos I. Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges, 2025.

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

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019.

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

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

Howard Schneider. Navigation map-based artificial intelligence. *AI*, 3(2):434–464, 2022.

Thomas Schöps et al. Bad slam: Bundle adjusted direct rgb-d slam. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

David Schottlander and Tomer Shekel. Geospatial reasoning: Unlocking insights with generative ai and multiple foundation models. Google Research Blog, April 2025. URL `https://research.google/blog/geospatial-reasoning-unlocking-insights-with-generative-ai-and-multiple-foundation-models/`.

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

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

Giovanni Sciortino. Vision graph neural networks for remote sensing. Master's thesis, Politecnico di Torino, 2023.

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S. Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17483–17493, 2023.

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.

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

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

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

William Shen et al. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023b.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *arXiv preprint arXiv:2303.17580*, 2024.

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

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

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

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2021.

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

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

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.

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

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

Manish Singh and Arpita Dayama. Leveraging spatiotemporal graph neural networks for multi-store sales forecasting, 2025.

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

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

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, Ted Xiao, Yao Lu, et al. Open-world object manipulation using pre-trained vision-language models. 2023.

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

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

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

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

Jiashuo Sun et al. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*, 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, 2020a.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, et al. Scalability in perception for autonomous driving: Waymo open dataset. 2020b.

Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *ICRA*, 2021.

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

Ayca Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1268. URL https://aclanthology.org/N19-1268.

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

Shuhan Tan, John Lambert, Hong Jeon, Sakshum Kulshrestha, Yijing Bai, Jing Luo, Dragomir Anguelov, Mingxing Tan, and Chiyu Max Jiang. Scenediffuser++: City-scale traffic simulation via a generative world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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*, 2024a.

Jiabin Tang et al. Higpt: Heterogeneous graph language model. *arXiv preprint arXiv:2402.16024*, 2024b.

Ziwei Tang et al. Graphllm: Boosting graph reasoning ability of large language model. *arXiv preprint arXiv:2310.05845*, 2024c.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. 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.

Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Dongsheng Yang, Yunfeng Ai, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, et al. Motion planning for autonomous driving: The state of the art and future perspectives. *arXiv preprint arXiv:2303.09824*, 2023.

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

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

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

Tesla. AI Robotics. Tesla Website, 2026. URL https://www.tesla.com/AI. Accessed: 2026-01-29.

The Robot Report Staff. Gemini robotics 1.5 enables agentic experiences, explains google deepmind. The Robot Report, September 2025. URL https://www.therobotreport.com/gemini-robotics-1-5-enables-agentic-experiences-explains-google-deepmind/.

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.

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

Bowen Tian et al. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. *arXiv preprint arXiv:2404.07103*, 2024a.

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

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

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

Akihiko Torii, Michal Havlena, and Tom
'a
vs Pajdla. From google street view to 3d city models. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1527–1534. IEEE, 2009.

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.

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. 2022.

UN-Habitat. Ai for spatial mapping and analysis: Geoai toolkit for urban planners. Technical report, United Nations Human Settlements Programme (UN-Habitat), 2025.

Unknown. Mineanybuild: A benchmark for evaluating spatial planning in minecraft. *arXiv preprint arXiv:2505.20148*, 2025.

Urban SDK. The role of ai and geospatial data in disaster planning: How cities can prepare smarter. Urban SDK Resources. URL `https://www.urbansdk.com/resources/ai-geospatial-data-disaster-planning-cities-prepare-smarter`.

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.

Andrés Velastegui-Montoya, Néstor Montalván-Burbano, Paúl Carrión-Mero, Hugo Rivera-Torres, Lu

$'i]sSadeck, and Marcos Adami. Google earth engine : A global analysis and future trends. Remote Sensing, 15(14) : 3675, 2023.$

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

Sai Vemprala et al. Chatgpt for robotics: Design principles and model abilities. *IEEE Access*, 2024.

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

Vinay Viswambharan, Rohit Singh, and Priyanka Tuteja. New pretrained geospatial ai models for disaster response. ArcGIS Blog, October 2024. URL `https://www.esri.com/arcgis-blog/products/arcgis-pro/public-safety/new-pretrained-geospatial-ai-models-for-disaster-response`.

Vicente Vivanco et al. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *arXiv preprint arXiv:2309.16020*, 2024.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Gpt-4v(ision) for robotics: Multimodal task planning from human demonstration. *arXiv preprint arXiv:2311.12015*, 2023a.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, et al. Gpt-4v(ision) for robotics: Multimodal task planning from human demonstration. *arXiv preprint arXiv:2311.12015*, 2023b.

Naoki Wake et al. Chatgpt empowered long-step robot control in various environments: A case application. *IEEE Access*, 2023c.

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

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Hengyi Wang et al. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. *IEEE Conference on Computer Vision and Pattern Recognition*, 2023b.

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.

Jiannan Wang et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *arXiv preprint*, 2024a.

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*, 2023c.

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

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

Tianbao Wang et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024c.

Wei Wang et al. Site: Towards spatial intelligence thorough evaluation. In *ICCV*, 2025.

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

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

Wenhai Wang, Zhe Chen, Xiaokang Chen, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023e.

Xiaoyang Wang et al. Traffic flow prediction via spatial temporal graph neural network. *The Web Conference*, 2020b.

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

Yanchen Wang et al. World models for autonomous driving: An initial survey. In *IEEE Intelligent Vehicles Symposium*, 2024e.

Yang Wang et al. Towards efficient llm grounding for embodied multi-agent collaboration. *arXiv preprint arXiv:2405.14314*, 2024f.

Yixuan Wang et al. Worldsim: A gpu-accelerated world model simulator. *arXiv preprint arXiv:2402.15393*, 2024g.

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*, 2023f.

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

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

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

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

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.

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, 2023a.

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

Jiawei Wei et al. Towards graph foundation models: A survey and beyond. *arXiv preprint arXiv:2310.11829*, 2024.

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

J. Weingarten. 2023 - REPORT - DEVELOPING FUTURE CAPABILITIES: ROBOTICS AND AUTONOMOUS SYSTEMS.
urlhttps://www.nato-pa.int/document/2023-robotics-and-autonomous-systems-report-weingarten-034-stctts, October 2023. Accessed: 2026-01-29.

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

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

Wherobots. Wherobots: Cloud-native spatial analytics. https://wherobots.com/, 2024.

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

Benjamin Wilson, William Qi, Tanmay Aber, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. 2021.

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

World Labs. World labs: Spatial intelligence for ai. https://www.worldlabs.ai/, 2024.

Chao Wu, Tianze Lin, Yifan Gao, Jia Xu, Weiwei Ding, Zhibin Ding, and Guangyun Jiang. Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping. *IEEE Robotics and Automation Letters*, 9(5):4397–4404, 2024a.

Chao Wu et al. Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping. *IEEE Robotics and Automation Letters*, 2024b.

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.

Hao Wu, Junbo Zhao, and Yu Zheng. Spatial-temporal interaction learning for urban flow prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2024c.

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

Jimmy Wu, Xingyuan Sun, Andy Zeng, Shuran Song, Szymon Rusinkiewicz, and Thomas Funkhouser. Spatial intention maps for multi-agent mobile manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021a.

Jimmy Wu, Rika Antonova, Adam Kan, et al. Tidybot: Personalized robot assistance with large language models. *arXiv preprint arXiv:2305.05658*, 2023c.

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

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*, 2024d.

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

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*, 2023e.

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*, 2021b.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022b.

Xiaoyang Wu et al. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 2022c.

Xiaoyang Wu et al. Point transformer v3: Simpler, faster, stronger. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024e.

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

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2019b.

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, 2019c.

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

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

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

Jiannan Xiang et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

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

Xing Xiao, Han Liu, Yinuo Li, and Dong Zhao. Robot learning in the era of foundation models: A survey. *arXiv preprint arXiv:2311.14379*, 2023.

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

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.

Tianbao Xie et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024.

Yichen Xie, Runsheng Xu, Tong He, Jyh-Jing Hwang, Katie Luo, Jingwei Ji, Hubert Lin, Letian Chen, Yiren Lu, Zhaoqi Leng, Dragomir Anguelov, and Mingxing Tan. S4-driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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*, 2017a.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation: A comprehensive survey. *arXiv preprint arXiv:2003.05163*, 2020.

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

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

Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*, 2022.

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

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

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

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

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

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

An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024b.

Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions, 2023a.

Jessy Lin Yang et al. Learning to model the world with language. *arXiv preprint arXiv:2308.01399*, 2023b.

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

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

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, et al. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Liber, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint arXiv:2405.15793*, 2024d.

John Yang et al. Swe-bench verified: A verified benchmark for evaluating language models on software engineering tasks. *arXiv preprint arXiv:2403.16732*, 2024e.

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

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

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

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*, 2025b.

Teng Yang, Song Xiao, and Jiahui Qu. D3gnn: Double dual dynamic graph neural network for multisource remote sensing data classification. *International Journal of Applied Earth Observations and Geoinformation*, 139:104496, 2025c. doi: 10.1016/j.jag.2025.104496.

Teng Yang, Song Xiao, and Jiahui Qu. D3gnn: Double dual dynamic graph neural network for multisource remote sensing data classification. *International Journal of Applied Earth Observation and Geoinformation*, 139:104496, 2025d. doi: 10.1016/j.jag.2025.104496.

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

Zhaohan Yang et al. World models: A survey. *arXiv preprint arXiv:2411.14499*, 2024h.

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

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*, 2024i.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023a.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023b.

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*, 2023c.

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

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

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

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

Licheng Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 3d question answering. In *ICCV*, 2021b.

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

Qinghao Ye et al. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024a.

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

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*, 2021c.

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.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2024a.

Yifan Yin et al. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*, 2024b.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, 2021a.

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

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

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

Haoxuan You et al. Ferret: Refer and ground anything anywhere at any granularity. *International Conference on Learning Representations*, 2024.

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.

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

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

Yuan Yuan et al. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. *arXiv preprint arXiv:2402.11838*, 2024a.

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

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

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

Yannis K Labrou Bill Chu J Long William Tolone Yun Peng, Tim Finin and Akram Boughannam. A multi-agent system for enterprise integration. In *Third International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*, pages 213–229, March 1998.

Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *European conference on computer vision*, pages 255–268. Springer, 2010.

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.

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

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

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

Yang Zhan et al. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *arXiv preprint arXiv:2401.09712*, 2024.

Ceyao Zhang et al. Proagent: Building proactive cooperative agents with large language models. *arXiv preprint arXiv:2308.11339*, 2024a.

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

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

Haitao Zhang et al. Graph foundation models. *arXiv preprint arXiv:2402.02216*, 2024c.

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

Jiaqi Zhang et al. Mapgpt: Map-guided prompting for unified vision-and-language navigation. *arXiv preprint arXiv:2401.07314*, 2024d.

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

Jingwei Zhang, Lei Tai, Ming Liu, Joschka Boedecker, and Wolfram Burgard. Neural slam: Learning to explore with external memory, 2020a.

Junbo Zhang et al. Deep spatio-temporal residual networks for citywide crowd flows prediction. *AAAI Conference on Artificial Intelligence*, 2017a.

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

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

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

Renrui Zhang et al. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in Neural Information Processing Systems*, 2022.

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

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

Wei Zhang et al. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *arXiv preprint arXiv:2401.16822*, 2024g.

Yuxiao Zhang, Alexander Carballo, Hanting Yang, and Kazuya Takeda. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:146–177, 2023a.

Zeyang Zhang et al. Llm4dyg: Can large language models solve problems on dynamic graphs? *arXiv preprint arXiv:2310.17110*, 2024h.

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*, 2020c.

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

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

Hengshuang Zhao et al. Point transformer. *IEEE International Conference on Computer Vision*, 2021b.

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

Jianan Zhao et al. Graphtext: Graph reasoning in text space. *arXiv preprint arXiv:2310.01089*, 2024b.

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

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, et al. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024a.

Haoyu Zhen et al. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024b.

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

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

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

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

Yu Zheng, 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.

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. 2024a.

Gengze Zhou et al. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. *arXiv preprint*, 2024b.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 2020a.

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

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

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

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

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.

Deyao Zhu et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2024.

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

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.

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

Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, and Friedrich Fraundorfer. Polyworld: Polygonal building extraction with graph neural networks in satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1848–1857, June 2022.