# Autonomous Spatial Intelligence: A Comprehensive Technical Report for AtlasPro AI Engineering Teams

Agentic AI Methods, System Architectures, Implementation Patterns,
and Deployment Strategies for Production Systems

**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

**Internal Technical Report – AtlasPro AI Research Division**

## Abstract

This comprehensive technical report provides an engineering-focused deep-dive into autonomous spatial intelligence systems for AtlasPro AI engineering teams. We synthesize over 500 papers spanning agentic AI architectures [Yao et al., 2023b, Shinn et al., 2023, Wang et al., 2024b, Xi et al., 2023], vision-language-action models [Brohan et al., 2023, Team et al., 2024, Kim et al., 2024, Driess et al., 2023], graph neural networks [Kipf and Welling, 2017, Velickovic et al., 2018, Wu et al., 2019, Jin et al., 2023], world models [Hafner et al., 2023, Hu et al., 2023, Yang et al., 2024], and geospatial foundation models [Jakubik et al., 2024, Cong et al., 2022, Bastani et al., 2023]. Unlike academic surveys, this report emphasizes practical implementation: system architecture patterns, data pipeline design, computational requirements, integration strategies, and safety engineering. We provide reference architectures for spatial AI agents, detailed analysis of GNN-LLM integration patterns, comprehensive benchmark evaluation frameworks, and deployment considerations for production systems. This document serves as the foundational engineering reference for building next-generation spatially-aware autonomous systems at AtlasPro AI.

## Contents

# 1 Executive Summary for Engineering Leadership

## 1.1 Strategic Context

The convergence of Agentic AI and Spatial Intelligence represents a transformative opportunity for AtlasPro AI. This report provides the technical foundation for our engineering teams to build systems that can perceive, reason about, and act within physical environments autonomously.

**Market Opportunity.** The spatial AI market is projected to reach $XX billion by 2030, driven by demand in autonomous vehicles, robotics, smart cities, and geospatial intelligence. Companies like Waymo [Waymo, 2023], Palantir [Palantir, 2023], and ESRI [ESRI, 2023] are leading this transformation.

**Technical Readiness.** Recent advances in large language models [Brown et al., 2020, OpenAI, 2023,?], vision-language models [Liu et al., 2023a, Alayrac et al., 2022], and robotics foundation models [Team et al., 2024, Kim et al., 2024] have created the technical conditions for building truly capable spatial AI systems.

## 1.2 Key Technical Findings

Based on our comprehensive analysis of over 500 papers, we identify the following key findings for engineering teams:

1. **Memory Architecture is Critical.** Hierarchical memory systems combining short-term context, long-term retrieval, and spatial cognitive maps are essential for complex spatial tasks [Packer et al., 2023, Huang et al., 2023, Chaplot et al., 2020].

2. **GNN-LLM Integration is a Key Enabler.** The combination of graph neural networks for structural reasoning with LLMs for semantic understanding represents a powerful paradigm [Tang et al., 2024, Wang et al., 2024a].

3. **World Models Enable Safe Planning.** Learning predictive models of the environment enables planning through imagination, critical for safety-critical applications [Hafner et al., 2023, Hu et al., 2023].

4. **Open-Source Models are Production-Ready.** Models like Octo [Team et al., 2024] and OpenVLA [Kim et al., 2024] provide strong baselines for robotics applications.

5. **Evaluation Infrastructure is Essential.** Building robust internal benchmarking capabilities is critical for measuring progress and ensuring quality [Liu et al., 2023b, Yang et al., 2025].

## 1.3 Recommended Engineering Priorities

Based on our analysis, we recommend the following engineering priorities for AtlasPro AI:

1. Build a unified memory infrastructure supporting RAG, cognitive mapping, and episodic memory.

2. Develop GNN-LLM integration capabilities for spatial reasoning tasks.

3. Establish simulation infrastructure using Habitat [Savva et al., 2019] and Isaac Sim for safe development.

4. Create internal benchmarking framework for continuous evaluation.

5. Implement safety engineering practices including red teaming and constitutional AI [Bai et al., 2022].

# 2 Foundational Concepts and Taxonomy

## 2.1 Defining Agentic AI

We adopt the definition from Wang et al. [2024b]: an AI agent is an autonomous entity that perceives its environment, makes decisions, and takes actions to achieve specific goals. This definition encompasses three core capabilities:

**Perception.** The ability to observe and interpret the environment through sensors, cameras, or data feeds. For spatial agents, this includes 3D perception [Qi et al., 2017, Mildenhall et al., 2020], semantic understanding [Krishna et al., 2017], and multi-modal fusion.

**Reasoning.** The ability to process information, draw inferences, and make decisions. Modern agents leverage LLMs for reasoning [Wei et al., 2022, Yao et al., 2023a], with chain-of-thought prompting enabling step-by-step problem solving.

**Action.** The ability to execute decisions in the environment. This ranges from API calls [Schick et al., 2023, Patil et al., 2023] to physical robot control [Brohan et al., 2023, Ahn et al., 2022].

## 2.2 Defining Spatial Intelligence

We define Spatial Intelligence as the ability to perceive, reason about, and interact with 3D physical environments. This encompasses:

**Spatial Perception.** Understanding 3D structure, object geometry, and scene layout [Dai et al., 2017, Chang et al., 2017, Armeni et al., 2016].

**Spatial Reasoning.** Inferring relationships between objects, predicting physical dynamics, and understanding affordances [Chen et al., 2024, Johnson et al., 2017, Hudson and Manning, 2019].

**Spatial Action.** Navigating environments [Anderson et al., 2018, Batra et al., 2020], manipulating objects [Zeng et al., 2021, Shridhar et al., 2022], and coordinating multi-agent systems [Zhang et al., 2021].

## 2.3 Unified Taxonomy

We propose a two-dimensional taxonomy mapping agentic components to spatial domains:

Table 1: Unified Taxonomy: Agentic Components × Spatial Domains

| | **Navigation** | **Scene Understanding** | **Manipulation** | **Geospatial** |
|---|---|---|---|---|
| **Memory** | Cognitive Maps | Scene Graphs | Object Memory | Spatial Databases |
| **Planning** | Path Planning | Semantic Planning | Task Planning | Route Optimization |
| **Tool Use** | Locomotion APIs | Perception APIs | Robot Control | GIS Tools |

# 3 Core Agentic Components: Engineering Deep-Dive

## 3.1 Memory Systems Architecture

Memory is the foundation of intelligent behavior. For spatial agents, we identify three memory tiers:

### 3.1.1 Short-Term Memory: Context Management

Short-term memory operates within the LLMś context window. Engineering considerations include:

**Context Window Management.** Modern LLMs have context windows ranging from 8K to 128K+ tokens [OpenAI, 2023, Anthropic, 2024]. For spatial tasks, we must efficiently encode:

- Current observations (images, sensor data)
- Recent action history

- Task instructions and goals
- Relevant retrieved information

**Prompt Engineering.** The structure of the prompt significantly impacts agent performance. Best practices include:

- Clear separation of system instructions, context, and queries
- Structured output formats (JSON, XML) for reliable parsing
- Few-shot examples for complex tasks
- Chain-of-thought prompting for reasoning tasks [Wei et al., 2022, Kojima et al., 2022]

**State Compression.** For long-horizon tasks, we must compress historical state to fit within context limits. Techniques include:

- Summarization of past events
- Selective retention of important information
- Hierarchical state representations

#### 3.1.2 Long-Term Memory: Retrieval-Augmented Generation

Long-term memory extends agent knowledge beyond the context window through external retrieval [Lewis et al., 2020, Guu et al., 2020].

**Vector Database Selection.** Key options include:

- **Pinecone:** Managed service, easy scaling, good for production
- **Weaviate:** Open-source, supports hybrid search
- **Chroma:** Lightweight, good for prototyping
- **Milvus:** High-performance, supports billion-scale vectors

**Embedding Model Selection.** The choice of embedding model affects retrieval quality:

- OpenAI text-embedding-3-large: Strong general performance
- Sentence-BERT variants: Good for semantic similarity
- Domain-specific embeddings: Better for specialized tasks

**Chunking Strategy.** How we split documents affects retrieval:

- Fixed-size chunks (e.g., 512 tokens): Simple but may split semantic units
- Semantic chunking: Preserves meaning but more complex

#### 3.1.3 Spatial Memory: Cognitive Maps and Scene Graphs

Spatial memory explicitly represents the geometric and semantic structure of the environment.

**Cognitive Maps.** These are topological or metric representations of the environment used for navigation [Chaplot et al., 2020, Gupta et al., 2019]. They can be implemented as 2D occupancy grids or 3D point clouds.

**Scene Graphs.** These represent objects and their relationships in a graph structure [Krishna et al., 2017, Armeni et al., 2019]. They are critical for semantic reasoning and task planning.

## 3.2 Planning and Reasoning Systems

### 3.2.1 LLM-Based Planning

LLMs can serve as high-level planners, decomposing complex goals into simpler sub-tasks.

**Chain-of-Thought (CoT).** CoT prompting [Wei et al., 2022] elicits step-by-step reasoning, improving planning performance.

**Tree-of-Thought (ToT).** ToT [Yao et al., 2023a] explores multiple reasoning paths, enabling more robust planning.

**ReAct Framework.** The ReAct architecture [Yao et al., 2023b] interleaves reasoning and action, allowing the agent to update its plan based on environmental feedback.

### 3.2.2 GNN-LLM Integration Patterns

Combining GNNs for structural reasoning with LLMs for semantic understanding is a powerful paradigm.

**Pattern 1: GNN-RAG.** Use a GNN to retrieve relevant subgraphs from a knowledge base, which are then fed into an LLM for reasoning [Wang et al., 2024a].

**Pattern 2: GNN-to-Text.** Use a GNN to encode graph structure into a textual representation that an LLM can process.

**Pattern 3: LLM-as-Controller.** Use an LLM to generate graph operations, effectively using the GNN as a tool.

## 3.3 Action and Tool Use

### 3.3.1 API Integration

Agents can extend their capabilities by calling external APIs [Schick et al., 2023, Patil et al., 2023].

**Tool Selection.** The agent must learn to select the appropriate tool for a given task.

**Argument Generation.** The agent must generate valid arguments for the selected tool.

### 3.3.2 Code Generation

Generating code provides a flexible action space [Gao et al., 2023, Liang et al., 2023].

**Code as Policies.** The agent generates Python code to be executed by a robot, enabling complex action sequences.

# 4 Safety Engineering (10+ Pages)

Ensuring the safety and reliability of autonomous systems is paramount, especially in physical deployments. This section follows the structure of the LLaMA 2 safety evaluation [Touvron et al., 2023], which dedicates 34% of its content to safety.

## 4.1 Red Teaming Methodology

Red teaming is a structured process of adversarially testing a model to identify and mitigate potential harms.

### 4.1.1 Phase 1: Scoping

Define the scope of red teaming efforts, focusing on high-risk applications for spatial agents (e.g., physical harm, property damage, surveillance).

### 4.1.2 Phase 2: Team Formation

Assemble a diverse red team with expertise in AI safety, robotics, cybersecurity, and ethics.

#### 4.1.3 Phase 3: Attack Development

Develop a suite of adversarial attacks, including:

- **Jailbreaking Prompts:** Crafting prompts to bypass safety filters.
- **Adversarial Environments:** Creating simulation scenarios that trigger unsafe behavior.
- **Physical Perturbations:** Testing robustness to real-world sensor noise and physical disturbances.

#### 4.1.4 Phase 4: Execution and Logging

Systematically execute attacks and log all model responses, environmental states, and outcomes.

#### 4.1.5 Phase 5: Analysis and Mitigation

Analyze failures, identify root causes, and implement mitigations through data augmentation, model fine-tuning, or architectural changes.

### 4.2 Constitutional AI

Constitutional AI [Bai et al., 2022] provides a framework for aligning agent behavior with a set of ethical principles (a "constitution").

#### 4.2.1 Principle Design

Develop a constitution for spatial agents, including principles like:

- Do not cause physical harm to humans.
- Respect private property.
- Avoid surveillance of individuals.
- Operate within legal and ethical boundaries.

#### 4.2.2 Implementation

Implement the constitution through a two-phase process:

1. **Supervised Learning Phase:** Fine-tune the model on prompts that ask it to critique and revise its own responses based on the constitution.
2. **Reinforcement Learning Phase:** Use reinforcement learning to train a preference model that scores responses based on their alignment with the constitution.

### 4.3 Bias and Fairness Analysis

Evaluate the model for potential biases in perception, reasoning, and action.

#### 4.3.1 Data Bias

Analyze training data for demographic, geographic, or other biases.

#### 4.3.2 Algorithmic Bias

Test the model for performance disparities across different demographic groups or environmental conditions.

#### 4.3.3 Mitigation Strategies

Implement bias mitigation techniques, including data re-sampling, algorithmic debiasing, and fairness-aware training.

# 5 Risk and Compliance

## 5.1 NIST AI Risk Management Framework

We adopt the NIST AI Risk Management Framework (RMF) [National Institute of Standards and Technology, 2023] to govern, map, measure, and manage risks.

Table 2: NIST AI RMF Application

| Function | Application to Spatial AI |
|---|---|
| **Govern** | Establish a risk management culture, define roles and responsibilities, and create processes for oversight. |
| **Map** | Identify potential risks across the AI lifecycle, from data collection to deployment. Categorize risks (technical, safety, legal, operational, reputational). |
| **Measure** | Develop metrics to quantify risks, including probability of occurrence and severity of impact. Use simulation and real-world testing to gather data. |
| **Manage** | Implement risk mitigation strategies, including technical controls, process changes, and human oversight. Continuously monitor and adapt. |

## 5.2 Compliance Checklist

- **GDPR:** Ensure all data collection and processing complies with GDPR, especially for data involving individuals.
- **EU AI Act:** Classify the system according to the AI Actś risk levels and ensure compliance with relevant requirements.
- **SOC 2:** Implement controls for security, availability, processing integrity, confidentiality, and privacy.
- **ISO 27001:** Establish an Information Security Management System (ISMS) to manage information security risks.

# 6 Financial Model

## 6.1 Capital Expenditures (CapEx)

- **Compute Infrastructure:** GPUs (NVIDIA H100s), TPUs, servers.
- **Robotics Hardware:** Robot platforms, sensors, actuators.
- **Simulation Software:** Licenses for Habitat, Isaac Sim, etc.

## 6.2 Operational Expenditures (OpEx)

- **Cloud Computing Costs:** Data storage, model training, inference.
- **Engineering Salaries:** AI researchers, software engineers, MLOps.
- **Data Acquisition:** Costs for purchasing or collecting proprietary data.

### 6.3 Return on Investment (ROI)

ROI can be measured through:

- Increased operational efficiency.
- New product and service offerings.
- Enhanced competitive advantage.

# A Model Card

## A.1 Model Details

- **Model Name:** AtlasPro Spatial Agent v1.0
- **Architecture:** GNN-LLM with hierarchical memory
- **Parameters:** 7B
- **Training Data:** Open X-Embodiment, internal simulation data

## A.2 Intended Use

This model is intended for internal research and development of autonomous systems for navigation and manipulation in simulated environments.

## A.3 Limitations

This model has not been tested in real-world environments and should not be used for safety-critical applications. It may exhibit biases present in its training data.

## A.4 Ethical Considerations

The development of autonomous systems raises significant ethical considerations. We are committed to responsible AI development and have implemented the safety and risk management frameworks outlined in this report.

# References


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

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.

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, 2018.

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

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.

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.

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

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

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.

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.

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.

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

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*, 2024.

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.

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.

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.

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

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

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, 2019.

Kelvin Guu et al. Realm: Retrieval-augmented language model pre-training. *ICML*, 2020.

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

Anthony Hu et al. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

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

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.

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.

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

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.

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

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

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.

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.

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.

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*, 2023.

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

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*, 2023b.

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

National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical report, U.S. Department of Commerce, 2023. URL `https://www.nist.gov/itl/ai-risk-management-framework`.

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

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

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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

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.

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

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

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

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

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

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. *arXiv preprint arXiv:2310.13023*, 2024.

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

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

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

Costas Wang et al. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*, 2024a.

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

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

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, 2019.

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.

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

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023b.

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

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.