Spatial Intelligence at Scale

AtlasPro AI's Approach to Building Agentic Geospatial Systems

Technical Report

Version 1.0 — January 2026

Gloria Felicia
Research Lead

Nolan Bryant
Systems Architect

Handi Putra
ML Engineer

Ayaan Gazali
Research Engineer

Eliel Lobo
Research Engineer

Esteban Rojas
Research Engineer

AtlasPro AI

Research Division

Correspondence: research@atlaspro.ai

Abstract

This technical report presents AtlasPro AI's research approach to building autonomous spatial intelligence systems. We introduce a unified three-axis taxonomy that organizes the intersection of agentic AI capabilities with spatial task domains across multiple scales. Our preliminary research, synthesizing over 800 papers from top venues, reveals critical gaps in existing approaches: current systems excel within narrow operational envelopes but fail systematically when tasks require cross-scale reasoning or long-horizon planning under geometric constraints.

We identify four key findings that inform our architectural approach: (1) memory systems must be explicitly spatial, encoding not just what happened but where; (2) planning under geometric constraints requires hybrid symbolic-neural representations; (3) graph neural networks provide the structural inductive bias necessary for spatial reasoning that transformers alone cannot capture; and (4) world models offer a path to safe deployment by enabling planning through imagination rather than trial-and-error in the physical world.

This report documents our research methodology, presents the three-axis taxonomy as a framework for system design, analyzes failure modes across 12 representative methods, and outlines AtlasPro AI's architectural principles for building spatially-aware autonomous agents. We release this report to establish priority on our methodological contributions and to invite collaboration from the research community.

Keywords: Spatial Intelligence, Agentic AI, World Models, Graph Neural Networks, Embodied AI, Geospatial AI, Vision-Language-Action Models

Contents

1 Introduction

Large language models have achieved remarkable success in symbolic reasoning, code generation, and natural language understanding. Yet these same models fail systematically when confronted with the physical world. Navigation agents hallucinate paths through walls. Manipulation planners propose grasps that violate basic physics. Embodied systems misjudge distances by orders of magnitude. The gap between linguistic competence and spatial competence represents one of the most significant barriers to deploying AI systems in real-world applications.

AtlasPro AI was founded to bridge this gap. Our research program investigates how to build autonomous systems that can perceive three-dimensional structure, reason about object relationships under physical constraints, and execute actions that respect the geometry of the world. This is not merely an incremental improvement over language understanding; it requires fundamentally different representations, architectures, and training paradigms.

This technical report documents our research approach. We present findings from a systematic analysis of over 800 papers spanning agentic AI, embodied AI, graph neural networks, world models, and geospatial foundation models. From this analysis, we derive a unified taxonomy and identify the architectural principles that will guide our system development.

1.1 Scope and Limitations of This Report

This report presents AtlasPro AI's research methodology and preliminary findings. It does not describe a deployed system or report experimental results from a novel architecture. Our contributions are:

1. A **three-axis taxonomy** (Task × Capability × Scale) that organizes the design space for spatial AI agents
2. A **systematic analysis** of failure modes across representative methods
3. **Architectural principles** derived from our literature synthesis
4. A **research roadmap** for AtlasPro AI's development program

We are transparent about what this report is not: it is not a peer-reviewed publication, it does not contain novel experimental results, and it does not describe production-ready systems. We release it to establish priority on our methodological contributions and to invite feedback from the research community.

1.2 Why Spatial Intelligence Matters Now

Three converging trends make spatial intelligence tractable and urgent:

Foundation Model Capabilities. Large language models now exhibit emergent reasoning capabilities [Wei et al., 2022, Brown et al., 2020]. Vision-language models can understand complex scenes [Liu et al., 2023a, Alayrac et al., 2022]. The question is no longer whether AI can reason, but whether it can reason about the physical world.

Robotics at Scale. Open-source robotics datasets [Collaboration, 2023, Walke et al., 2023] and foundation models [Team et al., 2024, Kim et al., 2024] have democratized embodied AI research. The barrier to entry has dropped dramatically.

Industry Demand. Autonomous vehicles, warehouse robotics, drone delivery, and smart city infrastructure all require spatial intelligence. The market opportunity exceeds $100 billion by 2030 [McKinsey and Company, 2023].

AtlasPro AI is positioned at this intersection. Our research program aims to develop the foundational capabilities for spatially-aware autonomous systems.

2 Research Methodology

Our findings derive from a systematic literature review conducted between September 2025 and January 2026.

2.1 Paper Selection

We applied a multi-stage filtering process:

Stage 1: Temporal Filter. We focused on papers published between 2020 and 2025, with selective inclusion of foundational works from earlier years.

Stage 2: Venue Filter. We prioritized top-tier venues: NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, CoRL, RSS, ICRA, and high-impact arXiv preprints with significant citations.

Stage 3: Relevance Filter. Papers were included if they addressed at least one of: agentic AI architectures, spatial reasoning, embodied AI, graph neural networks for spatial data, world models, or geospatial foundation models.

Stage 4: Quality Filter. Two researchers independently assessed each paper. Inter-annotator agreement was 94%. Disagreements were resolved through discussion.

2.2 Analysis Framework

For each included paper, we extracted:

- Primary spatial task (navigation, scene understanding, manipulation, geospatial)
- Agentic capabilities employed (memory, planning, tool use)
- Spatial scale of operation (micro, meso, macro)
- Representation type (symbolic, metric, latent, multimodal)
- Reported failure modes and limitations

This structured extraction enabled the taxonomy development and failure mode analysis presented in subsequent sections.

3 The Three-Axis Taxonomy

We propose a unified taxonomy that organizes the intersection of agentic AI and spatial intelligence. The taxonomy comprises three orthogonal axes: Spatial Task, Agentic Capability, and Spatial Scale.

3.1 Axis 1: Spatial Task

We identify four primary spatial task domains:

Navigation. Moving through environments toward goals. This includes vision-language navigation [Anderson et al., 2018], object-goal navigation [Batra et al., 2020], and point-goal navigation. The core challenge is grounding linguistic instructions in traversable paths.

Scene Understanding. Perceiving and representing 3D structure. This encompasses depth estimation, semantic segmentation, 3D reconstruction [Mildenhall et al., 2020, Kerbl et al., 2023], and scene graph construction [Armeni et al., 2019]. The core challenge is building representations that support downstream reasoning.

Manipulation. Interacting with objects through physical contact. This includes grasping [Mahler et al., 2017], pick-and-place [Zeng et al., 2021], and long-horizon task completion [Shridhar et al., 2022]. The core challenge is planning contact-rich interactions under uncertainty.

Geospatial Analysis. Reasoning about large-scale spatial phenomena. This includes remote sensing [Jakubik et al., 2024], traffic prediction [Li et al., 2018], and urban computing [Zheng et al., 2014]. The core challenge is handling heterogeneous data sources at city-to-global scales.

3.2 Axis 2: Agentic Capability

We identify three core agentic capabilities:

Memory Systems. How agents accumulate and retrieve spatial knowledge. This includes cognitive maps [Tolman, 1948, Chaplot et al., 2020], semantic maps [Huang et al., 2023], scene graphs [Rosinol et al., 2020], and retrieval-augmented generation [Lewis et al., 2020]. The central question is: *How can agents maintain persistent spatial knowledge across varying time horizons?*

Planning Systems. How agents decompose goals into executable action sequences. This includes task planning [Garrett et al., 2021], motion planning [LaValle, 2006], hierarchical planning [Nachum et al., 2018], and LLM-based planning [Huang et al., 2022]. The central question is: *How can agents plan under geometric constraints while accounting for uncertainty?*

Tool Use and Action. How agents translate decisions into physical effects. This includes API integration [Schick et al., 2023], code generation [Liang et al., 2023], and vision-language-action models [Brohan et al., 2023, Kim et al., 2024]. The central question is: *How can language-based reasoning be grounded in precise geometric actions?*

3.3 Axis 3: Spatial Scale

We distinguish three spatial scales with distinct computational and representational requirements:

Micro-spatial (<1m). Fine manipulation, grasping, and precise positioning. Requires millimeter-level accuracy. Representations must capture detailed geometry and contact dynamics.

Meso-spatial (1m–100m). Room-scale navigation, indoor exploration, and local scene understanding. Requires meter-level accuracy. Representations must balance detail with coverage.

Macro-spatial (>100m). City-scale planning, satellite imagery analysis, and infrastructure networks. Requires kilometer-level reasoning. Representations must handle sparse observations over large areas.

Key Insight: Scale Determines Architecture

Methods optimized for one scale often fail at others. Micro-scale manipulation systems achieve precision but lack macro-scale planning. Geospatial models handle city-scale reasoning but cannot guide fine manipulation. A unified spatial AI system must bridge these scales, which remains an open challenge.

3.4 Taxonomy Mapping

Table 1 maps representative methods to our taxonomy, demonstrating how the framework organizes the field.

Table 1: Representative Methods Mapped to the Three-Axis Taxonomy

Method	**Spatial Task**	**Agentic Capability**	**Scale**	**Primary Failure Mode**
VLN-BERT	Navigation	Memory + Planning	Meso	Instruction grounding
SayCan	Manipulation	Planning + Tool Use	Micro-Meso	Affordance mismatch
RT-2	Manipulation	Tool Use	Micro	Out-of-distribution
VLMaps	Navigation	Memory	Meso	Semantic drift
Voyager	Navigation + Manip.	Memory + Planning	Meso	Code execution
DCRNN	Geospatial	Memory (low planning)	Macro	Non-stationarity
Graph WaveNet	Geospatial	Memory (low planning)	Macro	Sparse regions
Prithvi	Geospatial	Memory only	Macro	No action capability
DreamerV3	Navigation + Manip.	Planning (World Model)	Micro-Meso	Model compounding
PaLM-E	Manipulation	Planning + Tool Use	Micro-Meso	Hallucination
OpenVLA	Manipulation	Tool Use	Micro	Limited generalization
LLaGA	Scene Understanding	Memory	Meso	Graph construction

4 Preliminary Findings: Failure Mode Analysis

Our literature synthesis reveals systematic failure patterns that inform AtlasPro AI's architectural decisions.

4.1 Spatial Hallucination

Language models describe spatial configurations that do not exist. GPT-4V fails on 40% of spatial relationship questions in SpatialBench [Chen et al., 2024]. The model confidently describes objects as "to the left of" when they are actually "behind," or claims paths exist through solid walls.

Root Cause. Language models learn spatial language from text, not from grounded experience. The word "left" appears in training data without consistent geometric grounding.

Implication for AtlasPro. Spatial representations must be grounded in metric observations, not derived solely from language.

4.2 Reference Frame Confusion

Agents conflate egocentric (body-relative) and allocentric (world-relative) coordinate systems. Vision-language navigation agents show 15–20% error rates from frame misalignment [Anderson et al., 2018].

Root Cause. Natural language instructions often leave reference frames implicit. "Turn left" is egocentric; "go to the kitchen" is allocentric. Agents must infer the intended frame from context.

Implication for AtlasPro. Systems must explicitly represent and transform between reference frames.

4.3 Scale Insensitivity

Models trained at one scale fail at others. SayCan's affordance model, trained on tabletop manipulation, fails when applied to room-scale tasks [Ahn et al., 2022]. The model cannot distinguish between "reachable" at arm's length versus "reachable" after navigation.

Root Cause. Scale is often implicit in training data. Models learn correlations that hold at one scale but not others.

Implication for AtlasPro. Architectures must explicitly encode scale and support cross-scale reasoning.

4.4 Temporal Drift

Spatial memory degrades over extended operation. VLMaps shows semantic drift after 100+ steps without map updates [Huang et al., 2023]. Accumulated localization error corrupts the spatial representation.

Root Cause. Spatial representations are updated incrementally. Small errors compound over time without correction mechanisms.

Implication for AtlasPro. Memory systems must include mechanisms for uncertainty estimation and correction.

4.5 Planning Constraint Violations

LLM-based planners propose actions that violate geometric constraints. On BEHAVIOR-1K, LLM planners achieve only 12% success due to collision-ignoring plans [Li et al., 2023]. The planner proposes "move the chair" without checking if the path is clear.

Root Cause. LLMs plan in language space, which does not enforce geometric consistency. The word "move" does not carry collision information.

Implication for AtlasPro. Planning must integrate geometric reasoning, not just linguistic sequencing.

4.6 Long-Horizon Credit Assignment

Performance degrades as task horizons extend. On ALFRED, success drops from 65% to 18% as task steps increase from 5 to 20 [Shridhar et al., 2020]. Agents cannot determine which early decisions caused late failures.

Root Cause. Reward signals are sparse and delayed. The agent receives feedback only at task completion, making it difficult to learn from intermediate mistakes.

Implication for AtlasPro. Systems need hierarchical decomposition and intermediate feedback mechanisms.

Key Insight: Failure Modes Are Architectural

These failures are not random errors but systematic consequences of architectural choices. Addressing them requires fundamental changes to how spatial AI systems are designed, not just better training data or larger models.

5 AtlasPro AI's Architectural Principles

Based on our analysis, we derive six architectural principles that will guide AtlasPro AI's system development.

5.1 Principle 1: Explicit Spatial Memory

Memory systems must encode *where*, not just *what*. This means:

- Cognitive maps that maintain metric relationships between locations
- Scene graphs that encode spatial relationships between objects
- Episodic memory that indexes experiences by location
- Uncertainty estimates that degrade gracefully over time

We draw inspiration from VLMaps [Huang et al., 2023], Neural SLAM [Chaplot et al., 2020], and 3D scene graphs [Rosinol et al., 2020], while addressing their limitations around temporal drift and scale.

5.2 Principle 2: Hybrid Planning Under Constraints

Planning must integrate symbolic task decomposition with geometric feasibility checking. This means:

- Task-and-motion planning (TAMP) integration [Garrett et al., 2021]
- Geometric constraint propagation during plan generation
- Continuous replanning as the world state changes
- Hierarchical abstraction to manage complexity

Pure LLM-based planning fails because language does not enforce geometric consistency. Pure motion planning fails because it cannot handle abstract goals. The hybrid approach combines their strengths.

5.3 Principle 3: Graph-Structured Representations

Graphs provide the inductive bias necessary for spatial reasoning. This means:

- Scene graphs for object relationships
- Road networks for navigation
- Spatial knowledge graphs for semantic reasoning
- GNN-LLM integration for combining structural and semantic understanding

Transformers treat all tokens uniformly; they do not inherently respect spatial adjacency. Graph neural networks encode relational structure directly, making them better suited for spatial reasoning [Kipf and Welling, 2017, Velickovic et al., 2018].

5.4 Principle 4: World Models for Safe Planning

Learning predictive models enables planning through imagination rather than trial-and-error. This means:

- Latent dynamics models that predict future states [Hafner et al., 2023]
- Video prediction for visual planning [Hu et al., 2023]
- Uncertainty-aware predictions for safe exploration
- Sim-to-real transfer for deployment

World models allow agents to "imagine" the consequences of actions before executing them. This is critical for safety-critical applications where trial-and-error is unacceptable.

5.5 Principle 5: Multi-Scale Architecture

Systems must explicitly handle multiple spatial scales. This means:

- Scale-aware representations that encode resolution explicitly
- Hierarchical processing from fine to coarse
- Cross-scale attention mechanisms
- Scale-conditioned action generation

A manipulation action at micro-scale may require navigation planning at meso-scale and route optimization at macro-scale. The architecture must support this seamlessly.

5.6 Principle 6: Grounded Action Generation

Language must be grounded in precise geometric actions. This means:

- Vision-language-action models that output continuous controls [Brohan et al., 2023]
- Affordance prediction to filter infeasible actions [Ahn et al., 2022]
- Code generation for programmatic control [Liang et al., 2023]
- Closed-loop execution with visual feedback

The gap between "pick up the cup" and the motor commands to actually grasp it is vast. Bridging this gap requires models that understand both language and geometry.

6 Technical Deep-Dive: Core Components

This section provides technical detail on the components central to AtlasPro AI's approach.

6.1 Memory Architecture

We propose a three-tier memory architecture:

Tier 1: Working Memory. Operates within the LLM context window. Stores current observations, recent actions, and active goals. Capacity: 8K–128K tokens depending on model.

Tier 2: Spatial Memory. Persistent storage of spatial knowledge. Implemented as a combination of:

- Vector database for semantic retrieval (e.g., embeddings of location descriptions)
- Metric map for geometric relationships (e.g., occupancy grid, point cloud)

- Scene graph for object relationships (e.g., “cup is on table”)

Tier 3: Episodic Memory. Long-term storage of experiences indexed by location and time. Enables learning from past successes and failures.

The key innovation is the integration layer that queries all three tiers based on the current context and retrieves relevant information for the LLM.

6.2 GNN-LLM Integration

We identify three integration patterns:

Pattern A: GNN as Encoder. The GNN encodes graph structure into embeddings that are projected into the LLM's token space. The LLM then reasons over the combined text and graph information.

Pattern B: LLM as Graph Enhancer. The LLM generates node features, edge labels, or graph augmentations that improve GNN performance.

Pattern C: Iterative Refinement. The GNN and LLM alternate, with each refining the other's outputs.

Our preliminary analysis suggests Pattern A is most suitable for spatial reasoning tasks, where the graph structure (e.g., scene graph, road network) provides critical inductive bias that the LLM alone cannot capture.

6.3 World Model Architecture

We adopt the Dreamer family architecture [Hafner et al., 2019, 2021, 2023] as our starting point:

Encoder. Maps observations to latent states: $z_t = \mathrm{enc}(o_t)$

Dynamics Model. Predicts future latent states: $\hat{z}_{t+1} = \mathrm{dyn}(z_t, a_t)$

Reward Predictor. Estimates rewards from latent states: $\hat{r}_t = \mathrm{rew}(z_t)$

Decoder. Reconstructs observations for training: $\hat{o}_t = \mathrm{dec}(z_t)$

The key advantage is that planning occurs entirely in latent space, enabling fast rollouts without expensive simulation.

6.4 Vision-Language-Action Models

For grounded action generation, we build on the RT-2 [Brohan et al., 2023] and OpenVLA [Kim et al., 2024] architectures:

Input. Image observation I and language instruction L

Processing. Vision encoder extracts features; language model processes instruction; cross-attention fuses modalities

Output. Action tokens decoded to continuous robot commands

The critical design choice is action representation. Discretizing the action space enables language model training but sacrifices precision. Continuous outputs preserve precision but require architectural modifications.

7 Benchmark Analysis

We analyze key benchmarks to identify evaluation gaps and inform AtlasPro AI's internal benchmarking strategy.

7.1 Navigation Benchmarks

Gap Identified. Existing benchmarks focus on single-episode navigation. They do not evaluate persistent spatial memory across multiple episodes or long-horizon exploration.

Table 2: Navigation Benchmark Comparison

Benchmark	Environment	Instruction Type	Key Metric
R2R [Anderson et al., 2018]	Matterport3D	Step-by-step	SPL
RxR [Ku et al., 2020]	Matterport3D	Detailed, multilingual	nDTW
REVERIE [Qi et al., 2020]	Matterport3D	High-level + object	SR
SOON [Zhu et al., 2021]	Matterport3D	Object-centric	SR
Habitat ObjectNav	Habitat	Object category	SPL

Table 3: Manipulation Benchmark Comparison

Benchmark	Tasks	Horizon	Key Challenge
RLBench [James et al., 2020]	100	Short	Diversity
Meta-World [Yu et al., 2020]	50	Short	Multi-task
CALVIN [Mees et al., 2022]	34	Long	Language grounding
ALFRED [Shridhar et al., 2020]	25K	Long	Embodied instruction
BEHAVIOR-1K [Li et al., 2023]	1000	Long	Realistic activities

7.2 Manipulation Benchmarks

Gap Identified. Benchmarks evaluate task success but not failure mode diagnosis. Understanding *why* agents fail is as important as measuring *how often*.

7.3 Agent Benchmarks

Table 4: Agent Benchmark Comparison

Benchmark	Environments	Focus	Spatial?
AgentBench [Liu et al., 2023b]	8	General LLM agents	Limited
WebArena [Zhou et al., 2024]	Web	Web navigation	No
OSWorld [Xie et al., 2024]	Desktop	OS interaction	No
EmbodiedBench [Yang et al., 2025]	Simulation	Embodied MLLM	Yes

Gap Identified. No benchmark comprehensively evaluates spatial reasoning across scales. EmbodiedBench is closest but focuses on meso-scale embodied tasks.

Key Insight: Benchmark Gaps

The field lacks benchmarks that evaluate: (1) cross-scale spatial reasoning, (2) persistent spatial memory, (3) failure mode diagnosis, and (4) safe exploration. AtlasPro AI's internal benchmarking will address these gaps.

8 Industry Landscape

We analyze industry leaders to identify design patterns and market positioning.

8.1 Autonomous Vehicles

Waymo has deployed the most mature autonomous driving system, with millions of miles of real-world operation. Their recent EMMA model [Waymo, 2024] demonstrates end-to-end multimodal driving. Key lesson: safety requires extensive real-world validation, not just simulation.

Tesla pursues a vision-only approach, betting that camera-based perception can match or exceed lidar. Key lesson: data scale can compensate for sensor limitations.

Cruise and Zoox focus on urban robotaxi deployment. Key lesson: constrained operational domains enable faster deployment.

8.2 Robotics

Boston Dynamics demonstrates state-of-the-art locomotion and manipulation. Key lesson: hardware and control co-design matters as much as perception.

Figure AI and 1X are developing humanoid robots for general-purpose tasks. Key lesson: the market is betting on generalist robots over specialized solutions.

Covariant applies foundation models to warehouse robotics. Key lesson: industrial applications provide data and revenue to fund research.

8.3 Geospatial Intelligence

Palantir provides data integration and analysis for defense and enterprise. Key lesson: the value is in integration, not just algorithms.

ESRI dominates enterprise GIS with ArcGIS. Key lesson: ecosystem lock-in creates durable competitive advantage.

Planet Labs operates the largest constellation of Earth-imaging satellites. Key lesson: data access is a moat.

8.4 Positioning for AtlasPro AI

Based on this analysis, AtlasPro AI will focus on:

1. **Cross-scale reasoning** that existing players do not address
2. **GNN-LLM integration** as a differentiating technical capability
3. **Safety-first development** using world models for planning
4. **Open research** to build community and attract talent

9 Research Roadmap

AtlasPro AI's research program proceeds in three phases.

9.1 Phase 1: Foundation (Q1–Q2 2026)

Objective. Establish core infrastructure and validate architectural principles.

Deliverables.

- Internal benchmarking framework covering navigation, manipulation, and geospatial tasks
- Prototype memory architecture with three-tier integration
- GNN-LLM integration baseline on scene graph reasoning
- Simulation environment setup (Habitat, Isaac Sim)

9.2 Phase 2: Integration (Q3–Q4 2026)

Objective. Integrate components into unified system and evaluate on benchmarks.

Deliverables.

- Unified spatial agent architecture
- World model integration for safe planning
- Cross-scale reasoning demonstration
- Internal benchmark results and analysis

9.3 Phase 3: Deployment (2027)

Objective. Transfer to real-world applications and establish production capabilities.

Deliverables.

- Sim-to-real transfer validation
- Pilot deployments in target applications
- Safety certification process
- Production infrastructure

10 Limitations and Risks

We acknowledge the following limitations of our current work and risks to our research program.

10.1 Limitations of This Report

- **No experimental validation.** The architectural principles are derived from literature analysis, not from experiments on our own systems.
- **Selection bias.** Our paper selection may have missed relevant work outside our search scope.
- **Rapidly evolving field.** Findings may be outdated as the field advances.

10.2 Research Risks

- **Integration complexity.** Combining memory, planning, GNNs, and world models is technically challenging. Integration may prove harder than anticipated.
- **Sim-to-real gap.** Performance in simulation may not transfer to real-world deployment.
- **Compute requirements.** Training world models and large VLAs requires significant computational resources.
- **Safety challenges.** Deploying autonomous systems in the physical world carries inherent risks.

10.3 Mitigation Strategies

- Incremental development with frequent evaluation checkpoints
- Collaboration with academic partners for validation
- Conservative deployment strategy with extensive testing
- Safety engineering practices including red teaming and formal verification where applicable

11 Conclusion

This technical report has presented AtlasPro AI's research approach to building autonomous spatial intelligence systems. Our key contributions are:

1. A **three-axis taxonomy** (Task × Capability × Scale) that organizes the design space for spatial AI
2. A **systematic failure mode analysis** revealing architectural limitations of current approaches
3. **Six architectural principles** that will guide our system development
4. A **research roadmap** for AtlasPro AI's development program

We release this report to establish priority on our methodological contributions, to invite feedback from the research community, and to signal our commitment to open research. Spatial intelligence represents one of the most important frontiers in AI, and we believe progress requires collaboration across academia and industry.

Acknowledgments. We thank the broader research community whose work made this analysis possible. This report synthesizes insights from hundreds of researchers; any errors in interpretation are our own.

Contact. Questions and feedback: research@atlaspro.ai

Version History.

- v1.0 (January 2026): Initial release

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.

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.

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.

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.

Open X-Embodiment Collaboration. Open x-embodiment, 2023.

Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrated task and motion planning. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:265–293, 2021.

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*, 2019.

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.

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

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

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

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuhler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

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

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 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*, 2020.

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

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.

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

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

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.

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. *arXiv preprint arXiv:1703.09312*, 2017.

McKinsey and Company. Industrial robotics: Insights into the sector's future growth dynamics. *McKinsey Global Institute Report*, 2023.

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.

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.

Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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.

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

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.

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

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

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

Edward C Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948.

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

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

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

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

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.

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.

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

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.

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*, 2024.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenarios for object navigation with natural language instructions. In *CVPR*, 2021.