

Large Language Models and Agentic AI Architectures for Aerial and Ground Robots: A Systematic Review, Taxonomy, and Design Guidelines

Abdelrahman Mahmoud

`abdelrahman.m.mohammed@gmail.com`

University of East London

Research Article

Keywords: large language models, agentic AI, embodied AI, autonomous robots, vision-language-action models, systematic review

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11246266/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Large Language Models and Agentic AI Architectures for Aerial and Ground Robots: A Systematic Review, Taxonomy, and Design Guidelines

Abdelrahman Mahmoud

School of Architecture, Computing and Engineering, University of East London, London, United Kingdom

Corresponding author: Abdelrahman Mahmoud. Email: abdelrahman.m.mohammed@gmail.com

Abstract. Large language models and agentic artificial intelligence provide the reasoning layer of autonomous aerial and ground robots, but the architectures remain fragmented across the two communities. This systematic review, reported according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses 2020 statement, searched ScienceDirect, IEEE Xplore, ProQuest, EBSCO, and Google Scholar from January 2020 through April 2026. Of 5298 records, 2228 remained after duplicates were removed and 124 studies were included: 67 aerial, 41 ground, and 16 on both platforms or a generic architecture. Six patterns appear, from a language model used only as a task-specific tool to an end-to-end foundation model, such as vision-language-action models. Open-loop pipelines, closed-loop pipelines, hierarchical executives, and multi-agent systems lie between those extremes. The model carries semantic reasoning, task decomposition, and tool orchestration, while classical planners and controllers retain trajectories, safety, and real-time control. Memory appears in 62 of 124 studies (50%) and function calling in 37 (30%), yet only 13 (10%) run the model onboard and 44 (35%) validate a physical robot. Aerial systems emphasize long-range vision and mission planning; ground systems emphasize local perception and interaction; both follow the same hybrid pattern. A suitability matrix shows that specialized vision models remain preferable for geometric perception. Designers should match the pattern to the task dynamics and choose the model for latency, safety, and where inference runs, rather than for benchmark score alone. Hallucination, spatial reasoning, inference latency, and the lack of standardized evaluation still limit deployment.

Keywords. large language models; agentic AI; embodied AI; autonomous robots; vision-language-action models; systematic review

Acknowledgments

The author thanks the researchers whose studies are synthesized in this review. This article is based on the author's MSc thesis at the University of East London, which has not been published in a journal or conference.

1. Introduction

Autonomous robotics has entered a transformative phase driven by advances in Artificial Intelligence (AI), particularly the emergence of Large Language Models (LLMs). Conventional robotic systems rely on carefully engineered perception, localization, planning, and control pipelines that perform well in structured environments but often struggle to generalize to dynamic or previously unseen scenarios. By providing natural language understanding, semantic reasoning, planning, and knowledge integration, LLMs enable robots to perform diverse tasks with substantially less task-specific engineering.

The rapid evolution of LLMs has further led to Agentic AI, where language models act as autonomous cognitive agents capable of reasoning, maintaining memory, using external tools, coordinating with other agents, and executing long-horizon plans. When integrated into robotic systems, these capabilities allow robots to translate high-level human goals into executable actions, adapt to environmental changes, recover from failures, and interact naturally with users rather than simply executing predefined commands.

These capabilities are particularly relevant to Unmanned Aerial Vehicles (UAVs) and Unmanned Ground Vehicles (UGVs), two rapidly growing domains of autonomous robotics. Although both pursue autonomous operation across applications such as search and rescue, inspection, logistics, agriculture, environmental monitoring, and defense, they operate under different constraints. UAVs are limited by payload, battery life, onboard computation, and stringent real-time flight safety requirements, whereas UGVs generally offer greater computational capacity and endurance but must navigate complex terrain and interact physically with their environment. These differences strongly influence architectural design decisions.

Since Google's PaLM-SayCan (Ichter et al. 2023) demonstrated how LLMs could ground natural language instructions into executable robotic actions through affordance-aware planning, research has rapidly expanded to include tool-augmented reasoning, retrieval-augmented generation (RAG), long-term memory, hierarchical planning, multi-agent systems, Vision-Language Models (VLMs), and more recently Vision-Language-Action (VLA) models. Collectively, these advances have shifted robotic intelligence from task-specific pipelines toward modular cognitive architectures capable of increasingly sophisticated autonomous behavior.

Despite this rapid progress, the literature remains fragmented. Most studies introduce new architectures or evaluate specific applications, while aerial and ground robotics have largely evolved as separate research communities. Existing surveys typically focus on individual topics such as embodied AI, task planning, manipulation, or human-robot interaction, rather than providing a unified architectural perspective spanning both UAVs and UGVs.

Furthermore, modern LLM-based robotic systems exhibit substantial variation in reasoning strategies, memory mechanisms, planning approaches, deployment models, tool integration, and multi-agent coordination. Foundation model selection has likewise become increasingly complex with the emergence of numerous proprietary and open-source LLMs, VLMs, and VLA models offering different strengths in reasoning, perception, planning, multimodal understanding, and code generation. Existing studies generally evaluate these models within isolated applications, offering limited guidance on their suitability for different robotic functional modules. At the same time, persistent challenges that include hallucinations, reasoning reliability, uncertainty handling, real-time performance, safety, explainability, computational efficiency, and deployment on resource-constrained platforms, continue to limit practical adoption.

These observations reveal a significant research gap. To the best of the author's knowledge, no previous study simultaneously (i) develops a unified taxonomy of LLM-based agentic AI architectures across both UAV and UGV platforms, (ii) systematically compares their architectural design patterns, (iii) analyzes the suitability of foundation models for different robotic functional modules, and (iv) synthesizes the major architectural limitations and open research challenges. Addressing this gap is essential for consolidating the rapidly expanding literature and providing practical design guidance for future intelligent robotic systems.

Accordingly, this study presents a structured evidence synthesis of LLM-based agentic AI architectures for autonomous UAVs and UGVs. Rather than proposing a new architecture, it synthesizes existing research to establish a comprehensive reference for researchers and practitioners. Through systematic literature analysis, the study develops a unified architectural taxonomy, compares existing design approaches, evaluates the suitability of foundation models for different robotic functional modules, and synthesizes the principal limitations and open research challenges reported across the literature.

The research is guided by the following primary research question: What LLM-based agentic AI architectures have been applied to autonomous aerial and ground robots, and how can they be systematically classified into a unified taxonomy?

This question is supported by the following research questions:

- RQ1: How are Large Language Models and other foundation models architecturally integrated into UAV and UGV robotic systems?
- RQ2: What are the dominant architectural patterns used in LLM-based agentic reasoning for autonomous robots, and how can they be systematically classified into a unified taxonomy?
- RQ3: How do these architectures differ in design and application between aerial (UAV) and ground (UGV) robotic platforms?
- RQ4: Which foundation models are most suitable for different robotic functional modules?
- RQ5: What are the key open challenges and research gaps in applying LLM-based agentic reasoning to autonomous robots across both domains?

To answer these questions, this study makes four primary contributions:

1. A unified taxonomy of LLM-based agentic AI architectures for autonomous UAV and UGV systems, providing a structured classification of contemporary architectural patterns.
2. A comparative analysis of existing architectures, examining reasoning strategies, planning approaches, memory mechanisms, deployment models, tool integration, multi-agent coordination, and platform-specific design choices.

3. An evaluation of foundation model suitability for key robotic functional modules, including perception, reasoning, planning, and human–robot interaction, together with the trade-offs influencing model selection.
4. A synthesis of architectural limitations, open research challenges, and design guidelines to support the development of more reliable, scalable, and effective agentic robotic systems.

2. Literature Review

The rapid growth of LLM-enabled robotics has led to numerous review papers covering topics such as robotic agents, embodied AI, UAVs, UGVs, multimodal foundation models, and human–robot interaction. While these surveys provide valuable insights, they differ substantially in scope and emphasis. This section positions the present study within the existing review literature by identifying its key contributions and remaining gaps. Unlike the systematic analysis of primary studies presented later, this review focuses on prior surveys to motivate the need for a comparative synthesis of LLM-enabled robotic architectures, particularly across unmanned aerial vehicles (UAVs) and unmanned ground vehicles (UGVs).

2.1. Existing Review Literature

2.1.1. General Reviews of LLM-Based Robotic Agents

Early review papers established the conceptual foundations of LLM-enabled robotic agents. Wang et al. (2024) introduced a unified framework comprising four core modules: profile, memory, planning, and action, with tool use part of the action module, which has become a widely adopted reference architecture for agentic AI systems. Extending these concepts to physical robots, Raptis et al. (2025) reviewed LLM integration across navigation, manipulation, multi-agent systems, and general-purpose robotics while proposing an "agenticness" taxonomy based on autonomy, adaptability, and decision-making capability. They also noted that much of the literature emphasizes architectural design over empirical evaluation in real-world deployments. Although these surveys define the fundamental building blocks of LLM-enabled robotics, they remain largely platform-agnostic and provide limited comparison between different robotic domains.

2.1.2. UAV-Specific Review Literature

Several surveys examine the integration of LLMs into UAVs, highlighting their role in semantic reasoning, mission planning, natural language interaction, and autonomous flight. Tian et al. (2025) proposed an Agentic UAV reference architecture consisting of perception, memory, reasoning, planning, and tool-use modules, together with a roadmap toward increasingly autonomous systems. Sapkota et al. (2025) reviewed agentic UAV applications across domains such as agriculture, inspection, search and rescue, logistics, and environmental monitoring, illustrating the transition from rule-based autonomy to language-guided decision-making. Likewise, Javaid et al. (2024) discussed how LLMs enhance navigation, mission execution, and human–drone collaboration. While these surveys provide comprehensive coverage of aerial robotics, their analyses are confined to UAV platforms.

2.1.3. UGV and Mobile Robot Review Literature

Reviews of ground robotics primarily investigate the application of generative AI and LLMs to navigation, perception, manipulation, planning, control, and human–robot interaction (Rashidieranjbar et al. 2025; Waga et al. 2025; Lisondra et al. 2026; Aremu et al. 2026). They show how foundation models increasingly complement conventional robotic algorithms through semantic understanding and flexible task planning while proposing taxonomies organized around robotic capabilities and applications. Commonly reported challenges include hallucinations, latency, safety, and computational cost. Reviews of foundation models for autonomous vehicles report similar trends in perception and scene understanding (Boroujeni et al. 2026; Ashqar et al. 2025). However, these surveys focus almost exclusively on terrestrial robots and provide little comparison with aerial systems or cross-platform architectural patterns. The present review differs by classifying aerial and ground systems within a single architectural taxonomy and comparing how the same patterns are adapted to each platform's constraints.

2.1.4. Reviews of Heterogeneous UAV–UGV Systems

Compared with platform-specific reviews, surveys covering heterogeneous UAV–UGV systems remain limited. Earlier reviews primarily addressed multi-robot coordination, communication, and

cooperative control, largely without considering LLMs (Ding et al. 2021; Munasinghe et al. 2024). More recent work has begun to apply LLMs and vision-language models across aerial and ground platforms, for example to schedule vehicle–drone deliveries (Geng and Chen 2025) or to ground language in 3D scenes captured by vehicles, drones, and quadrupeds (Li et al. 2025a). However, these studies generally emphasize applications rather than systematically analyzing shared architectural principles, leaving heterogeneous LLM-enabled robotics as an emerging research area.

2.2. Comparative Assessment of Existing Reviews

Existing surveys collectively establish the core functional modules of LLM-enabled robotic systems and consistently recognize LLMs as key enablers of semantic reasoning, natural language interaction, and adaptive decision-making.

Nevertheless, several gaps remain. Most reviews adopt a platform-specific perspective, focusing exclusively on either UAVs or UGVs, limiting cross-platform architectural comparisons. Many organize the literature by application domains rather than software architectures, leaving recurring architectural patterns and design strategies insufficiently explored. Existing reviews also rarely compare the suitability of different foundation models for specific robotic functions such as perception, planning, memory, or tool use. In addition, deployment considerations like latency, computational requirements, safety, hallucinations, and real-time performance are typically discussed qualitatively rather than synthesized across implementations. Finally, although future research directions are commonly proposed, the literature lacks a consolidated analysis of recurring challenges and open research gaps.

These limitations motivate the present review, which provides a unified taxonomy of LLM-enabled robotic architectures, systematically compares UAV and UGV systems, evaluates foundation model suitability across functional modules, and synthesizes the principal research challenges identified throughout the literature.

3. Materials and Methods

This study employed a Systematic Literature Review (SLR) to identify, analyze, and synthesize research on the integration of Large Language Models (LLMs), Vision-Language Models (VLMs), and other foundation models into autonomous robotic systems. The review focused on Unmanned Aerial Vehicles (UAVs) and Unmanned Ground Vehicles (UGVs) in which foundation models perform robotic functions such as reasoning, planning, task decomposition, decision-making, or agentic behavior. The review followed the PRISMA 2020 reporting guidelines. As the review was conducted by a single researcher, all screening decisions were performed by one reviewer. Generative artificial intelligence was not used for the literature search, study screening, data extraction, or analysis. ChatGPT (OpenAI) was used to generate the schematic diagrams in Figs. 1–9 from the author’s descriptions of those architectures and copy editing. The author reviewed and edited every diagram. Grok 4.7 was used for formatting. The author takes full responsibility for the content.

3.1. Review Protocol

The review protocol was not registered or published in advance; its elements are reported in Sections 3.1–3.6. The completed PRISMA 2020 checklist is Online Resource 2. The review addressed the following research questions (Table 1):

Table 1. Research questions.

ID	Type	Research Question
RQ1	Exploratory	How are Large Language Models and other foundation models architecturally integrated into UAV and UGV robotic systems?
RQ2	Descriptive	What are the dominant architectural patterns used in LLM-based agentic reasoning for autonomous robots, and how can they be systematically classified into a unified taxonomy?
RQ3	Comparative	How do these architectures differ in design and application between aerial (UAV) and ground (UGV) robotic platforms?

ID	Type	Research Question
RQ4	Comparative	Which foundation models are most suitable for different robotic functional modules?
RQ5	Evaluative	What are the key open challenges and research gaps in applying LLM-based agentic reasoning to autonomous robots across both domains?

Included studies describe a concrete robotic architecture, except for the peer-reviewed reviews and the three non-peer-reviewed documents reported in Sections 3.3 and 3.5.

3.2. Literature Search Strategy

The literature search covered ScienceDirect, IEEE Xplore, ProQuest, EBSCO, and Google Scholar. Searches covered publications from January 2020 to April 2026, and all databases were last searched on 1 May 2026, using Boolean combinations of terms related to foundation models, agentic reasoning, and autonomous mobile robots.

The search combined three groups of keywords corresponding to:

1. Foundation models (LLMs and VLMs)
2. Agentic reasoning and autonomous cognition
3. Mobile robotic platforms (UAVs and UGVs)

The search expression used throughout the databases was:

("large language model" OR "LLM" OR "foundation model" OR "vision language model" OR "VLM") AND ("agentic" OR "autonomous reasoning" OR "cognitive architecture" OR "agent") AND ("UAV" OR "drone" OR "unmanned aerial" OR "UGV" OR "unmanned ground" OR "autonomous robot" OR "mobile robot")

The search identified 5,298 publications. Duplicate records retrieved from multiple databases were removed, resulting in 2,228 unique publications for screening.

3.3. Study Selection

Studies were selected through a four-stage process consisting of database retrieval, duplicate removal, automated title and abstract screening, and full text manual verification for candidate papers. Screening employed a deterministic keyword-based approach in which each publication's title and abstract were evaluated against five predefined inclusion criteria. A study was retained only if all criteria were satisfied. Preprints and other non-peer-reviewed publications were subsequently excluded, with three exceptions: two patent documents (Niforatos 2025; Petersen et al. 2025) and a NASA technical report (Shetye et al. 2026) were retained because they describe concrete, implemented architectures. The complete selection process is illustrated in the PRISMA flow diagram (Fig. 1).

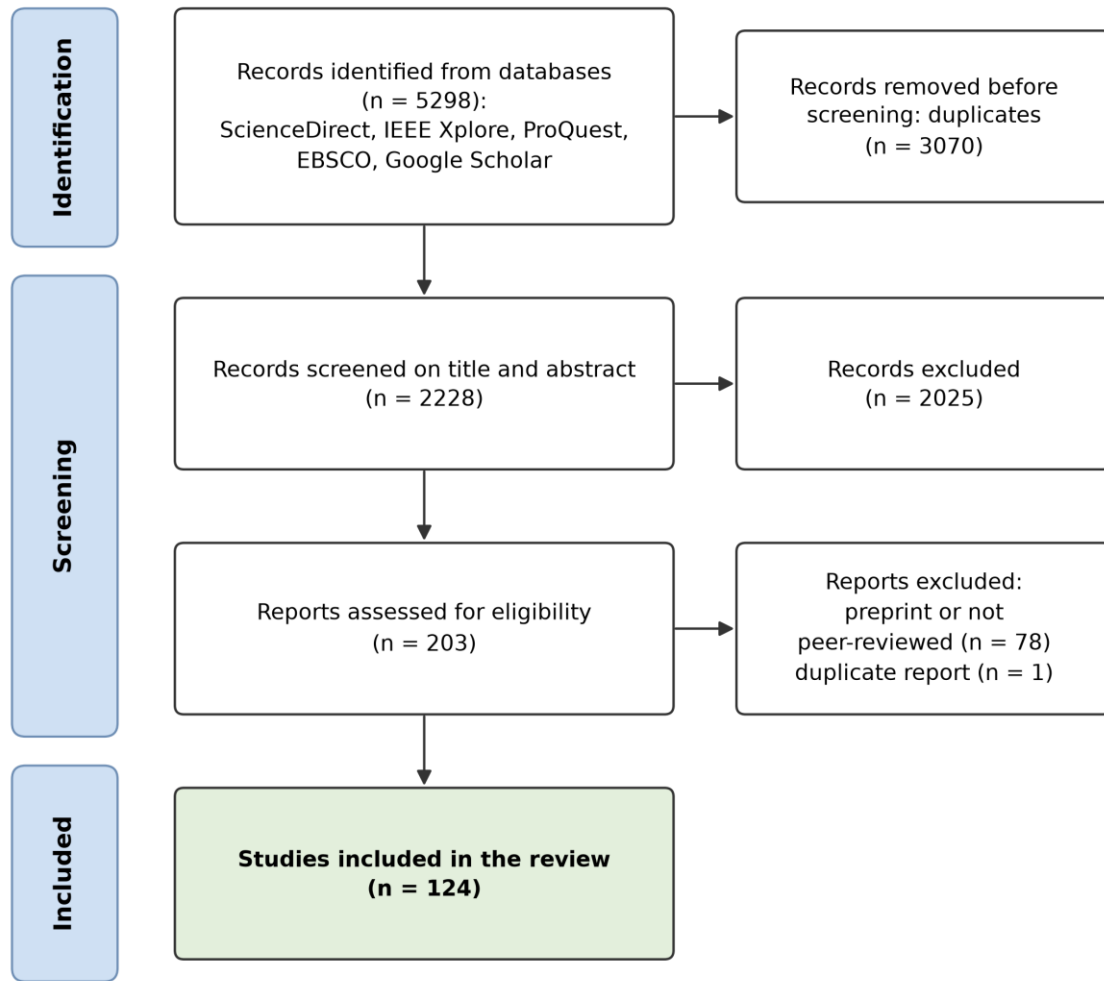


Fig. 1 PRISMA flow diagram for this study

3.4. Inclusion Criteria

Studies were included when they:

- identified a specific LLM, VLM, or foundation model;
- described an embodied UAV or UGV platform;
- used the foundation model for reasoning, planning, decision-making, or other cognitive functions;

- presented a concrete robotic architecture or framework, or analyzed such architectures in a peer-reviewed review (Section 3.5); and
- involved a physical robotic context rather than purely language-processing tasks.

Studies were excluded if they relied solely on conventional machine learning, lacked embodied robotic systems, used LLMs only for conversational interaction, presented purely conceptual work, or were non-peer-reviewed preprints (Table 2).

Table 2. Representative Keywords Used for Screening.

Criterion	Representative Keywords
C1	"large language model", "llm", "foundation model", "vision language model", "vision-language model", "vlm", "gpt-4", "gpt-3", "gpt4", "chatgpt", "gpt", "llama", "gemini", "palm", "claude", "mistral", "falcon", "qwen", "phi", "gemma", "rt-x", "rt-2", "deepseek", "llama-3", "llama-3.1", "llama-3.2", "language model", "vision-language", "multimodal llm", "pretrained language", "instruction-tuned", "generative model", "transformer-based", "generative"
C2	"unmanned aerial", "uav", "unmanned ground", "ugv", "drone", "quadrotor", "multirotor", "hexarotor", "fixed-wing", "aerial robot", "ground robot", "legged robot", "wheeled robot", "autonomous drone", "robotic drone", "quadruped", "legged locomotion", "mobile manipulator", "service robot", "warehouse robot", "ground vehicle", "tracked robot", "rover", "mobile robot platform", "skid-steer", "differential drive", "clearpath", "husky", "jackal", "mobile robot", "ground", "vehicle", "vehicles"

C3	"llm-based planning", "llm-based reasoning", "llm for robot", "language model for", "llm agent", "agentic", "autonomous reasoning", "llm-driven", "language-driven", "natural language instruction", "task planning", "high-level planning", "semantic navigation", "chain of thought", "tool use", "tool-augmented", "memory-augmented", "multi-agent", "llm controller", "agent", "reasoning", "motion planning", "path planning", "decision making", "decision-making", "cognitive", "goal-directed", "multi-step", "scene understanding"
C4	"physical robot", "real robot", "real world", "real-world", "real-world deployment", "hardware", "onboard", "outdoor", "indoor navigation", "flight", "fly", "takeoff", "landing", "hovering", "locomotion", "navigation task", "manipulation task", "sensor", "camera", "lidar", "gps", "imu", "embodied", "deployment", "field", "experiment", "test", "platform", "testbed", "prototype"
C5	"architecture", "model", "framework", "system", "proposed", "approach", "module", "hierarchical", "pipeline", "workflow", "end-to-end", "ablation", "integration", "system architecture", "consists of", "comprises", "framework consisting of", "modules include", "pipeline consisting of", "end-to-end system", "ablation study"

3.5. Final Study Selection

The screening process reduced the initial 5,298 records to 203 reports assessed for eligibility, of which 124 studies form the dataset analyzed throughout this study (Table 3). Seven of these are peer-reviewed reviews rather than original system studies; they analyze foundation-model or agent architectures for aerial, ground, autonomous-vehicle, or related robotic systems, and they are marked as “Survey” in Online Resource 1.

Table 3. PRISMA summary.

Stage	Records
Records identified	5,298
Duplicates removed	3,070
Records screened	2,228
Records excluded	2,025
Reports excluded at full-text assessment	79
Studies included	124

3.6. Data Extraction

Following study selection, a structured data extraction process was conducted to collect information relevant to the objectives of this review. A standardized extraction template was developed to ensure consistency across all included studies and to facilitate subsequent comparative analysis.

For each publication, the following information was extracted when available (Table 4):

Table 4. Data Extraction Fields.

Category	Extracted Information
Robotic Platform	Robot type (UAV or UGV)
Architecture	Architecture type and overall system organization
Foundation Model	LLM, VLM, or foundation model employed
Memory	Memory mechanism or memory architecture
Tool Integration	External tools, APIs, function calling, or software integration
Agent Organization	Single-agent or multi-agent architecture
Deployment	On-board, off-board, or hybrid computation
UAV Characteristics	UAV-specific architectural features or constraints
UGV Characteristics	UGV-specific architectural features or constraints
Physical Constraints	Hardware, sensing, computational, communication, or environmental limitations

Validation	Real-world deployment or simulation-based evaluation
Evaluation	Performance metrics and evaluation methodology
Limitations	Limitations reported by the authors
Research Gaps	Open challenges and future research directions identified in the publication

The extracted information was organized into a structured dataset to enable systematic comparison across studies. Particular emphasis was placed on architectural organization, the functional role of the foundation model, deployment strategy, and interactions between reasoning, perception, memory, planning, and execution modules. These characteristics formed the basis for the subsequent comparative analysis and taxonomy development.

3.7. Data Analysis and Taxonomy Development

The selected studies were analyzed using both quantitative and qualitative methods. Quantitative analysis summarized publication trends, platform distribution, model usage, deployment strategies, and evaluation approaches. Qualitative analysis decomposed each architecture into its principal functional modules, including perception, reasoning, planning, memory, tool integration, and execution, to identify recurring organizational patterns.

The architectural taxonomy was developed using an inductive, bottom-up approach. Rather than defining categories in advance, studies exhibiting similar structural characteristics and information flow were grouped into common architectural families through iterative comparison until stable categories emerged.

3.8. Validity and Limitations

Review rigor was supported through the use of multiple academic databases, PRISMA reporting guidelines, deterministic keyword-based screening, standardized data extraction, and the exclusion of preprints, with the three documented exceptions described in Section 3.3.

This review was conducted by a single reviewer, and therefore inter-reviewer agreement was not assessed. Although the automated screening procedure improves consistency by applying deterministic inclusion rules, keyword-based filtering may fail to identify relevant studies that employ uncommon terminology or alternative descriptions. Furthermore, the proposed taxonomy was developed through qualitative comparative analysis and therefore reflects the architectural patterns represented within the selected literature rather than an exhaustive classification of all possible robotic architectures. Studies were not individually assigned to the six categories, so Section 5.2 describes each pattern with representative examples but does not report how often each occurs. Finally, Google Scholar limits query length and returns at most 1,000 results per query, so its contribution to the search is less reproducible than that of the bibliographic databases.

4. Analysis of Existing Research

4.1. Perception

Perception forms the foundation of autonomous robotic systems by transforming raw sensor observations into structured representations used by reasoning, planning, memory, and control modules. Across the surveyed literature, the perception pipeline consistently comprises three stages: (1) sensing the environment, (2) processing sensor data using AI models, and (3) generating machine-readable representations for downstream modules. Sensor inputs originate either from egocentric sensors (e.g., RGB cameras, LiDAR, IMUs, depth cameras) mounted on the robot or allocentric sensors (e.g., fixed cameras or motion capture systems) observing it externally (Rashidieranjbar et al. 2025; Mower et al. 2026). However, nearly all surveyed UAV and UGV systems rely on egocentric sensing because it is suitable for real-world autonomous deployment.

4.1.1. Perception Architectures

Two dominant perception architectures emerge from the literature.

The first employs specialized perception models, with each model optimized for a specific task. Common examples include YOLO and GroundingDINO for object detection, SAM for segmentation, Depth Anything for depth estimation, CLIP, SigLIP, and DINOv2 for visual representation, and Recognize Anything or OpenSEED for object recognition (Zhao et al. 2026; Hu et al. 2025c; Wang et al. 2025c; Park et al. 2026; Cao et al. 2025; Xu et al. 2026). These models are typically organized as sequential or parallel pipelines whose outputs are fused before being passed to downstream modules.

The second architecture replaces multiple specialized models with a Vision-Language Model (VLM) capable of performing multiple perception tasks through prompting. Models such as GPT-4V, Qwen2.5-VL, LLaVA, Gemini Vision, and Vicuna-based multimodal models are used for scene understanding, object recognition, landmark extraction, visual question answering, and semantic interpretation (Wen et al. 2024; Wang et al. 2025d; Zhu et al. 2026; Zhang et al. 2025c; Ashqar et al. 2024).

Rather than competing approaches, these architectures offer complementary trade-offs. Specialized pipelines provide superior localization accuracy, lower latency, and predictable computational requirements, making them well suited to real-time robotic applications. In contrast, VLMs simplify system design by consolidating multiple perception capabilities into a single model while enabling zero-shot semantic understanding. Consequently, architecture selection is primarily determined by application requirements, including task definition, accuracy, inference latency, available onboard computing resources, and whether inference is performed onboard or offloaded to edge or cloud infrastructure.

Beyond using foundation models as perception components, several recent works seek to develop foundation perception models themselves. Li et al. (2025a) propose a universal perception model for 3D grounded object detection across multiple robotic platforms including UAVs, UGVs, and quadrupeds. The architecture, in addition to visual encoding, combines a PointNet++ LiDAR

encoder with a frozen RoBERTa language encoder before predicting three-dimensional bounding boxes through a transformer decoder. Joint training across heterogeneous robotic platforms substantially improves performance over single-platform baselines, although reported accuracy remains below 81%, indicating that generalized perception foundation models remain an active research challenge.

4.1.2. Environmental and Spatial Understanding

Beyond detecting objects, perception must identify environmental constraints that influence robot behavior, including workspace boundaries, restricted areas, safety margins, traffic rules, and mission-specific constraints. The surveyed literature adopts three complementary strategies: constraints may be encoded directly within global planners, enforced by local planning or control modules, or perceived from environmental cues such as signs and markers before being incorporated into planning (Wen et al. 2024; Zhou et al. 2024; Zhang et al. 2024). In all cases, perception converts raw observations into representations that downstream modules can reason over.

Spatial understanding is equally important, requiring the robot to infer relationships such as next to, inside, behind, or on top of. Most systems represent these relationships using semantic or scene graphs, where nodes correspond to detected entities and edges encode spatial or semantic relationships, enabling planners to reason about object interactions rather than isolated detections (Cao et al. 2025; Long et al. 2026). Despite recent advances in VLMs, precise geometric and metric spatial reasoning remains a known limitation. Consequently, many architectures supplement VLMs with specialized perception modules or explicitly constructed scene graphs for spatial reasoning (Cao et al. 2025; Park et al. 2026).

4.1.3. Sensor Modalities

A notable trend across the surveyed literature is the shift toward camera-centric perception. While traditional robotic systems often fuse cameras with LiDAR, RADAR, infrared, or depth sensors, many recent LLM-based architectures achieve competitive performance using only RGB or RGB-D cameras combined with powerful vision and agentic reasoning models. Recent benchmark

studies report state-of-the-art performance for camera-only systems on autonomous driving datasets such as nuScenes and Bench2Drive (Boroujeni et al. 2026), suggesting that advances in vision models increasingly compensate for reduced sensor diversity. Nevertheless, LiDAR and RADAR remain advantageous under adverse weather, poor illumination, and other safety-critical conditions.

4.1.4. Perception Optimizations

Several studies improve perception efficiency through architectural optimizations. Zhao et al. (2026) enhance small-object detection by combining GroundingDINO, SAM, and aspect-ratio-constrained image partitioning to generate regions of interest that are analyzed individually by a VLM, reducing unnecessary computation while improving detection accuracy. Zhang et al. (2026b) instead reduce perception frequency by executing perception only when actions are initiated, logging observations and robot states, and converting these logs into natural-language memories for future planning. These approaches illustrate two complementary optimization strategies: improving perception quality through targeted processing and reducing computational cost through event-driven perception.

4.1.5. Alternative Roles of Foundation Models

Although foundation models are commonly integrated into perception pipelines, several studies employ them in more specialized roles. Webb et al. (2026) use CLIP solely for fire and smoke detection, with navigation handled independently by a PPO reinforcement learning controller. Niforatos (2025) employs GPT4All as a semantic translation interface that converts sensor data into natural-language descriptions and translates operator instructions into MAVLink commands. Other systems expose perception models as callable tools, allowing an LLM to invoke YOLO, SAM, or Recognize Anything models dynamically during task execution rather than replacing them (Hu et al. 2025c).

4.1.6. Model Trends Across UAVs and UGVs

Across both UAV and UGV studies, YOLO remains the dominant real-time object detector, while GroundingDINO is the preferred open-vocabulary detector (Zhang et al. 2026b; Găitan et al. 2025; Azzouni and Pujolle 2026; Chang et al. 2025; Zhao et al. 2026; Wang et al. 2025c; Park et al. 2026). SAM is consistently paired with detection models for segmentation, Depth Anything is the most frequently adopted monocular depth estimator, and CLIP, SigLIP, and DINOv2 are widely used as visual encoders (Xu et al. 2026; Zhu et al. 2026). General-purpose VLMs such as including GPT-4V, Qwen2.5-VL, LLaVA, Gemini Vision, and Vicuna-based multimodal models are primarily used for semantic scene understanding, instruction grounding, landmark extraction, and visual question answering rather than precise geometric perception (Wang et al. 2025d; Cao et al. 2025; Zhu et al. 2026; Zhang et al. 2025c; Ashqar et al. 2024; Wen et al. 2024). The UAVBench benchmark (Ferrag et al. 2026) similarly identifies Qwen3-235B-A22B as the strongest of the evaluated models on its UAV scenario-reasoning tasks.

Overall, the surveyed literature shows clear architectural convergence toward hybrid perception systems. Specialized vision models continue to dominate tasks requiring accurate localization, segmentation, and geometric understanding, whereas foundation models increasingly provide semantic interpretation, contextual reasoning, and orchestration of perception tools. Rather than replacing conventional perception pipelines, state-of-the-art robotic architectures combine both approaches to balance accuracy, efficiency, and generalization.

4.2. Reasoning and Planning

Reasoning and planning form the cognitive layer of an autonomous robotic system, transforming a high-level user objective into executable behaviors that can be carried out by the robot. Unlike perception, which estimates the current state of the environment, the planning module reasons over the mission objective, perception outputs, memory, and the robot's available capabilities and constraints to determine the most appropriate next action or sequence of actions. Depending on the application and the underlying architecture, planning outputs may take several forms, including high-level action sequences (e.g., *pick the cup* → *move to the sink* → *pour water*), landmark

sequences (e.g., *blue door* $\rightarrow$ *stairs* $\rightarrow$ *white table*), executable robot API calls (e.g., *move_to(x, y)*, *rotate(45°)*), behavior trees, finite-state transitions, or waypoint trajectories consisting of positions and orientations. Although these representations differ in abstraction, they all aim to bridge semantic task descriptions with robot-executable behaviors.

4.2.1. Current Approaches

Most surveyed systems employ general-purpose LLMs (e.g., GPT-4, Claude, Gemini, Llama, Qwen, DeepSeek) as high-level planners that decompose user instructions into executable subtasks (Cao et al. 2025; Azzouni and Pujolle 2026; Găitan et al. 2025). These models are commonly used with zero-shot, few-shot, or retrieval-augmented prompting, while several recent studies fine-tune smaller open-source models for lower latency and onboard deployment.

A growing trend is integrating perception and planning through Vision-Language Models (VLMs), where models such as Qwen2.5-VL, InternVL2.5, and LLaVA jointly interpret images and generate context-aware plans (Wang et al. 2025d; Zhao et al. 2026; Yao et al. 2026). More recent Vision-Language-Action (VLA) models, including RT-2, Gemini Robotics, OpenVLA, and AeroVLA, further combine perception, planning, and control into end-to-end policies, simplifying system architecture but reducing interpretability (Xu et al. 2026).

4.2.2. Architectural Patterns Observed

Three planning architectures consistently emerge across the literature.

LLM-only planners generate explicit plans such as action sequences, JSON, executable code, or behavior trees, preserving interpretable intermediate representations (Mower et al. 2026; Zhang et al. 2026b).

Hybrid LLM-classical planners are the dominant approach. Here, LLMs perform semantic high-level reasoning while deterministic planners ensure geometric feasibility and safety. Representative examples combine Qwen2.5-VL with A* and Ego-Planner for navigation (Wang et al. 2025d), integrate LLM reasoning with PDDL planning for experience-guided optimization

(Chang et al. 2025), or embed the LLM within predefined finite-state workflows for robust recovery from failures (Cao et al. 2025).

Task-specific foundation models replace large language models as lightweight planners trained for robotic tasks. ViNT and similar navigation foundation models demonstrate competitive navigation performance while remaining suitable for real-time onboard deployment (Mukhametzianov and Nambo 2025; Shah et al. 2023).

4.2.3. Empirical Observations Across the Literature

Several consistent trends emerge across the surveyed literature. Chain-of-Thought prompting generally produces more reliable plans than direct prompting (Tabrizian et al. 2025), while planning performance remains sensitive to prompt wording (Hu et al. 2025c).

Few-shot prompting exhibits diminishing returns, with approximately three to nine demonstrations consistently outperforming both zero-shot prompting and large retrieval sets (Mahmud et al. 2025; Zhu et al. 2026; Song et al. 2023). Similarly, moderate decoding randomness often yields more reliable plans than fully deterministic decoding for ambiguous instructions (Al-Safi et al. 2025).

Another notable trend is the increasing use of lightweight fine-tuned models for latency-sensitive applications. For example, Sezgin (2025b) reports 90–96% decision accuracy with approximately 120 ms inference latency using a fine-tuned, quantized Llama 3.2 1B model in a scenario-driven evaluation of UAV mission reliability, while ViNT demonstrates that compact planners can achieve strong navigation performance on embedded hardware.

Safety is also increasingly enforced through deterministic safety monitors rather than relying solely on LLM reasoning. For example, Ewald et al. (2025) employ obstacle-based safety overrides, while replacing a text-only planner with LLaVA eliminates safety terminations caused by missing visual context.

Finally, long-horizon planning remains difficult. For example, Zhao et al. (2026) report that one-shot planning performs well for short missions but degrades significantly as the number of execution steps increases, making iterative replanning more reliable in dynamic environments.

4.3. Motion Planning

Motion planning converts high-level decisions into collision-free, dynamically feasible trajectories that satisfy the robot's kinematic and environmental constraints. While the reasoning module determines what the robot should do, the motion planner determines how it should physically reach the objective, typically by transforming waypoints, target poses, or navigation goals into executable trajectories.

4.3.1. Current Approaches

Across the surveyed literature, motion planning remains dominated by traditional robotics algorithms rather than foundation models. Classical graph-search methods (e.g., A*), sampling-based planners (e.g., RRT), optimization-based planners, and reinforcement learning (RL/DRL) policies continue to generate trajectories because continuous motion planning requires satisfying geometric and dynamic constraints that current LLMs are not designed to optimize.

Consequently, most systems adopt a hybrid architecture in which LLMs perform semantic reasoning while conventional planners handle trajectory generation. For example, Wang et al. (2025d) use Qwen2.5-VL to decompose missions and identify navigation landmarks, whereas Ego-Planner generates and refines safe flight paths using perception updates.

Some systems omit an explicit motion planner when the reasoning module directly outputs executable robot actions or code. This approach is mainly used for robots with discrete action spaces or predefined APIs, where commands can be executed without intermediate trajectory generation (Hassan et al. 2024; Mower et al. 2026).

Rather than replacing planners, several studies use LLMs to improve planning offline. Examples include guiding RL training using aviation knowledge (Han et al. 2024), partitioning tasks during multi-agent trajectory optimization (Zhu et al. 2024), and generating high-quality initial solutions for genetic algorithms (Geng and Chen 2025). In addition, some work replaces handcrafted local planners with lightweight neural trajectory predictors trained specifically for continuous motion generation (Wang et al. 2025c).

4.3.2. Architectural Patterns Observed

Three recurring patterns emerge across the literature:

- **Hybrid planning (dominant):** Foundation models perform task reasoning, while classical planners generate feasible trajectories (Wang et al. 2025d).
- **Direct action execution:** The reasoning module outputs executable actions or robot code, bypassing explicit motion planning (Hassan et al. 2024; Mower et al. 2026).
- **Offline LLM assistance:** LLMs improve reinforcement learning or optimization during training, while conventional planners are used during deployment (Han et al. 2024; Zhu et al. 2024; Geng and Chen 2025).

4.3.3. Empirical Observations Across the Literature

A consistent finding across the surveyed studies is that no architecture relies on an LLM to directly generate continuous motion trajectories. Instead, trajectory generation is consistently delegated to classical planners or specialized neural networks, reflecting the complementary strengths of foundation models in semantic reasoning and conventional planners in geometric optimization. This hybrid separation of responsibilities is observed across both UAV and UGV systems. Furthermore, lightweight neural trajectory predictors are emerging as a practical alternative for real-time local planning, offering lower latency than general-purpose foundation models while remaining specialized for continuous trajectory generation.

4.4. Control

The control module represents the final stage of the robotic autonomy pipeline, translating planned trajectories or robot actions into actuator commands. Given a reference trajectory, waypoint sequence, velocity profile, or discrete robot action, the controller continuously computes low-level motor commands that compensate for disturbances while ensuring stability, tracking accuracy, and dynamic feasibility. Unlike reasoning and motion planning, which operate at relatively low frequencies and reason over abstract task representations, controllers operate at high frequency and directly interact with the robot's physical dynamics.

4.4.1. Current Approaches

Across the surveyed literature, low-level control remains almost entirely based on conventional robotics and control theory. Controllers such as Proportional-Integral-Derivative (PID), Model Predictive Control (MPC), geometric controllers, and platform-specific flight controllers continue to execute robot motions without incorporating foundation models. Whether the robot receives waypoints from a motion planner or discrete actions from an LLM, execution is ultimately delegated to controllers responsible for maintaining stability and accurately following the desired trajectory.

4.5. Memory

Memory enables robots to retain information beyond a single perception–planning cycle, supporting long-horizon autonomy across multiple interactions. Unlike the limited context window of LLMs, external memory persistently stores observations, environmental knowledge, task progress, prior actions, and domain-specific knowledge. This improves reasoning consistency, reduces redundant computation, and enables robots to accumulate experience over time.

4.5.1. Current Approaches

Across the reviewed literature, perception outputs are typically transformed into structured memory rather than consumed directly by the planner foundation model. The most common representation is the semantic map, implemented as object grids, scene graphs, or key-value stores describing objects and their relationships (Cao et al. 2025; Zhang et al. 2026b; Long et al. 2026). Navigation systems additionally maintain occupancy maps for collision avoidance, while exploration systems employ uncertainty maps to guide exploration. Some architectures also maintain episodic memories that record actions and observations, enabling planning to leverage past experiences (Mukhametzianov and Nambo 2025; Zhang et al. 2026b).

Memory increasingly serves a dual purpose: storing both environmental knowledge (maps, landmarks, navigation history, sensor observations) and operational knowledge such as manuals, regulations, maintenance procedures, API specifications, and tool documentation. These

knowledge sources are retrieved on demand through Retrieval-Augmented Generation (RAG), allowing LLMs to reason using domain-specific information beyond their parametric knowledge. For example, Yu et al. (2025) report that retrieving maintenance documents reduced hallucinations by approximately 75% compared with a standalone LLM.

4.5.2. Architectural Patterns Observed

A recurring pattern is the separation of memory management from planning through dedicated memory agents responsible for updating, validating, organizing, and retrieving stored information. Retrieval is predominantly implemented using RAG with embedding models and vector databases such as FAISS, although alternative methods such as k-nearest neighbor retrieval over BERT embeddings have also been explored (Song et al. 2023).

Graph-based memory is another common design, particularly for navigation and long-horizon interaction. Instead of storing isolated observations, graph memories explicitly represent entities and their relationships, enabling robots to reuse previously explored routes and maintain consistent object identities across extended tasks (Zhang et al. 2024).

4.5.3. Empirical Observations Across Papers

The reviewed literature shows a clear shift from storing raw sensory data toward semantic memory composed of entities, landmarks, graphs, and symbolic descriptions, allowing reasoning modules to operate on compact models of the environment instead of low-level observations. Recent systems also increasingly use LLMs to filter perceived information before committing it to memory, retaining only task-relevant knowledge while preserving obstacle information needed for navigation (Park et al. 2026).

Another emerging trend is memory validation. Because similarity-based retrieval can return contextually incorrect information, several architectures verify retrieved memories before incorporating them into planning. For example, Liu et al. (2025b) validate retrieved historical incidents before reasoning, improving recommendation accuracy by approximately 26%. Other studies improve semantic consistency by encouraging planners to reuse previously established

object names, reducing inconsistencies during planning and tool invocation (Song et al. 2023). Together, these findings indicate that modern robotic memory is evolving from passive storage into an actively managed semantic knowledge base that supports reliable long-horizon reasoning.

4.6. Tool Calling

Tool calling extends LLM capabilities by allowing them to invoke external functions, APIs, and software modules instead of performing perception, computation, or robot control internally. In robotics, tools provide the interface between high-level reasoning and the physical robot, enabling LLMs to access perception models, sensors, memories, motion controllers, and external algorithms.

4.6.1. Current Approaches

Across the reviewed literature, tools are exposed as callable APIs through frameworks such as MCP, LangChain tools, or custom function-calling interfaces. Common tools include perception inference, localization, memory access, motion execution, and optimization algorithms. A notable trend is treating perception as an on-demand service rather than a continuously running pipeline. For example, Hu et al. (2025c) expose YOLO, SAM, and RAM as callable functions, allowing the LLM to invoke perception only when required while generating UAV control code.

Tool calling also enables LLMs to orchestrate specialized AI models rather than replace them. Raiyn (2026) uses GANs and VAEs to generate possible traffic scenarios, while GPT-4.1 interprets the results and recommends driving actions that avoid the predicted hazards. Similarly, Ma et al. (2025) fine-tune Llama3-8B to select the most suitable network topology planning algorithm from the Topo toolkit, based on mission requirements, instead of solving the optimization problem directly.

4.6.2. Architectural Patterns Observed

The reviewed architectures increasingly separate reasoning from tool execution. Rather than allowing planning agents to invoke APIs directly, many systems employ dedicated tool-use agents

responsible for function selection, parameter generation, and validation. Xue et al. (2026) exemplify this approach with Translator, Conformer, and Verifier agents that retrieve API documentation from memory, generate simulator calls, validate parameters, and detect errors before execution. Retrieval-Augmented Generation (RAG) is also commonly integrated to retrieve only the relevant API documentation, reducing prompt complexity and improving function selection.

Overall, tool calling shifts the LLM's role from computation to orchestration. Instead of performing perception or optimization itself, the model coordinates specialized tools and integrates their outputs into decision making. However, reliable language-to-API translation remains challenging. Xue et al. (2026) achieve approximately 71% language-to-API accuracy despite using retrieval and multi-stage validation, whereas Sezgin (2025a) reports 91.3% endpoint accuracy, 87.6% parameter matching, and 95.2% schema conformity by retrieving API documentation and validating generated calls against platform schemas. Collectively, these studies indicate that robust tool use depends not only on selecting the correct function but also on grounding generation with API documentation and validating generated calls before execution.

4.7. Communication Protocols

As LLMs become the reasoning core of robotic systems, communication has shifted from transmitting raw sensor data to exchanging semantic information that can be directly interpreted by language models. This evolution spans communication between functional modules, multiple robots, and human operators, with protocols increasingly designed to maximize semantic richness while minimizing token usage, latency, and bandwidth.

4.7.1. Current Approaches

A clear trend across the literature is semantic communication rather than conventional data transmission. Instead of sending images or sensor measurements, perception modules such as Vision-Language Models (VLMs) first convert observations into structured descriptions (e.g., *"Ball detected next to Door"*), typically encoded as compact JSON messages. This reduces token consumption while remaining directly usable by LLM-based reasoning modules. Zhang et al.

(2025a) demonstrated that transmitting semantic scene descriptions extracted by LLaVA instead of images reduced communication volume by over 90% without degrading task performance. Similarly, Du et al. (2026) employ semantic entity representations with edge-based caching in disaster evacuation scenarios, achieving over 70% lower communication latency and reducing evacuation time by approximately 11.3% compared with a nearest-rule baseline.

4.7.2. Architectural Patterns Observed

Three recurring communication patterns emerge across the reviewed studies. First, an intermediate semantic abstraction layer converts perception outputs into semantic representations before transmission, making language the shared knowledge representation among agents. Second, structured formats such as JSON are consistently preferred over free-form natural language because they reduce ambiguity, inference cost, and token usage while improving interoperability between conventional software modules and LLMs. Third, communication architectures increasingly adopt distributed and edge-assisted designs to reduce latency and improve scalability in multi-agent systems.

4.7.3. Empirical Observations Across Papers

Across the reviewed papers, semantic communication consistently reduces bandwidth requirements while preserving task-relevant information, suggesting that semantic compression may be more scalable than conventional multimedia compression for bandwidth-constrained robots (Zhang et al. 2025a). In multi-agent systems, richer semantic communication also improves coordination efficiency. Jia et al. (2024) report that LLM-enhanced swarms increased communication frequency while simultaneously reducing average decision time, with domain-specific LLMs providing the best performance, indicating that more informative semantic exchanges can accelerate collaborative planning.

5. Results

5.1. Literature Statistics

Publication activity accelerated substantially in 2025, which accounted for 72 papers (58.1%), followed by 30 papers (24.2%) in 2026, 18 papers (14.5%) in 2024, and only 4 papers (3.2%) in 2023. Years refer to the version of each study cited in this review (Table A1); the 2026 count therefore includes later peer-reviewed versions of studies retrieved earlier. This trend demonstrates the explosive increase in interest following the emergence of capable foundation models.

Aerial robotics dominated the literature, particularly during 2025, with 44 UAV-focused studies published that year. Across the entire dataset, 67 papers focused on aerial robots, 41 on ground robots, while 16 addressed both domains or proposed generic robotic architectures (Fig. 2).

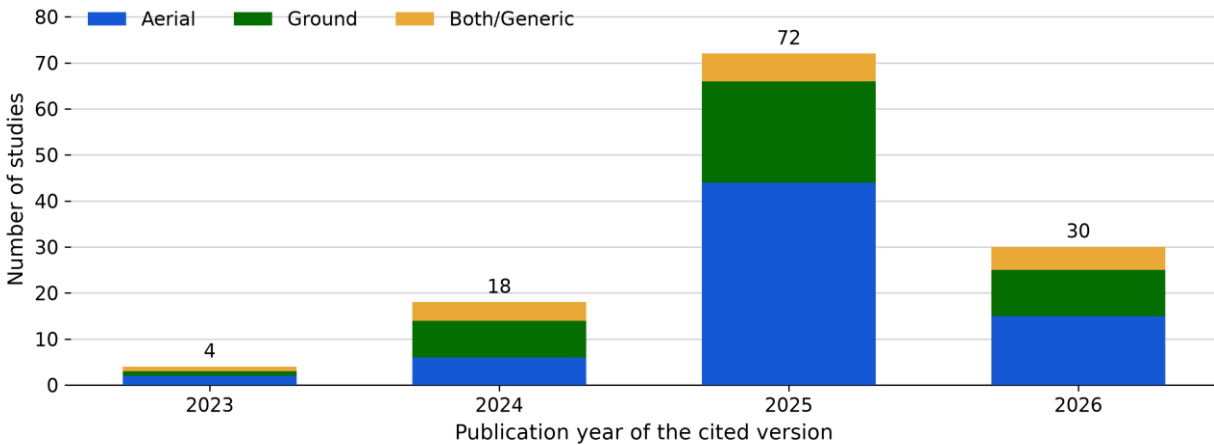


Fig. 2 Included studies by year and platform

Analysis of architectural characteristics shows several recurring implementation trends. Memory mechanisms were the most common architectural enhancement, appearing in 62 studies (50%), indicating that maintaining context and long-term task information is becoming a standard component of LLM-based robotic systems. Function/API calling was employed in 37 studies (30%), enabling language models to invoke perception modules, motion planners, or external services. Multi-agent architectures were reported in 21 studies (17%), reflecting growing interest

in collaborative planning and distributed reasoning. Only 13 studies (10%) deployed the LLM or VLM directly on board the robot, with most systems relying on remote or cloud-based inference due to computational constraints.

Regarding experimental evaluation, 44 papers (35%) validated their systems in real-world robotic environments, alone or alongside simulation, while 63 (51%) relied on simulation only, 1 used recorded real-world data without robot hardware, and 16 did not state a validation setting (Table A1). Success rate was the most frequently reported performance metric, used in 39 studies (31%), whereas 21 studies (17%) additionally reported latency or task completion time, highlighting increasing attention to the real-time constraints of robotic deployment (Fig. 3).

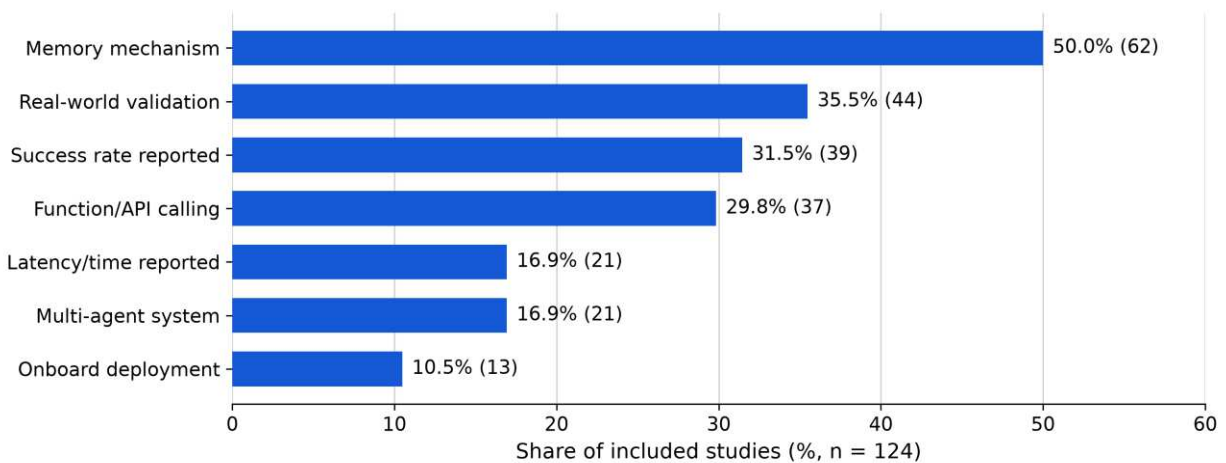


Fig. 3 Architectural and evaluation characteristics of the included studies

Model usage analysis reveals the continued dominance of proprietary and open-source foundation models. GPT-based models were used or evaluated in 30 studies, making them the most frequently tested family. Among open-source alternatives, LLaMA appeared in 17 studies, Qwen in 13, and LLaVA in 8, primarily for multimodal perception tasks. The diversity of evaluated models reflects the ongoing shift toward open-source, multimodal, and edge-deployable language models for robotics applications.

5.2. Architectural Taxonomy

The reviewed literature reveals that, despite considerable diversity in implementation details, most LLM-enabled robotic systems converge toward a small number of recurring architectural patterns. Rather than each work proposing a fundamentally new architecture, most systems represent variations of several common design paradigms that differ primarily in how cognitive responsibilities are distributed between the LLM and traditional robotic software.

Based on the analyzed studies, six major architectural categories were identified. These architectures form a spectrum ranging from minimal LLM integration, where the model performs only a narrow language-processing task, to fully end-to-end foundation models that replace nearly the entire classical robotics pipeline. Across all categories, several optional cognitive enhancement modules may be incorporated independently of the core architecture, which include memory, reflection, tool use, safety mechanisms, and human feedback

5.2.1. Open-Loop Sequential Pipeline

The simplest architecture identified in the literature is the open-loop sequential pipeline, in which information flows through a fixed sequence of processing stages (Fig. 4):

Perception → LLM Reasoning/Planning → Motion Planning → Robot Controller

Each stage executes once before passing its output to the next module, with no subsequent feedback to the LLM from execution. The LLM receives a representation of the environment together with the user instruction, generates a task plan or sequence of actions, and the remaining robotics stack executes the generated plan. For example, Sharma et al. (2025) translate a free-form instruction into a complete, schema-constrained UAV mission that is validated once and then executed without further feedback to the LLM.

This architecture is characterized by:

- single-pass execution;
- absence of execution feedback;

- no persistent internal state;
- no long-term memory;
- no reasoning revision after execution begins.

However, its limitations become apparent in dynamic environments. Since planning occurs only once, the system cannot recover from execution failures, respond to unexpected environmental changes, or correct reasoning errors. Consequently, open-loop pipelines are primarily suitable for relatively static tasks with predictable execution.

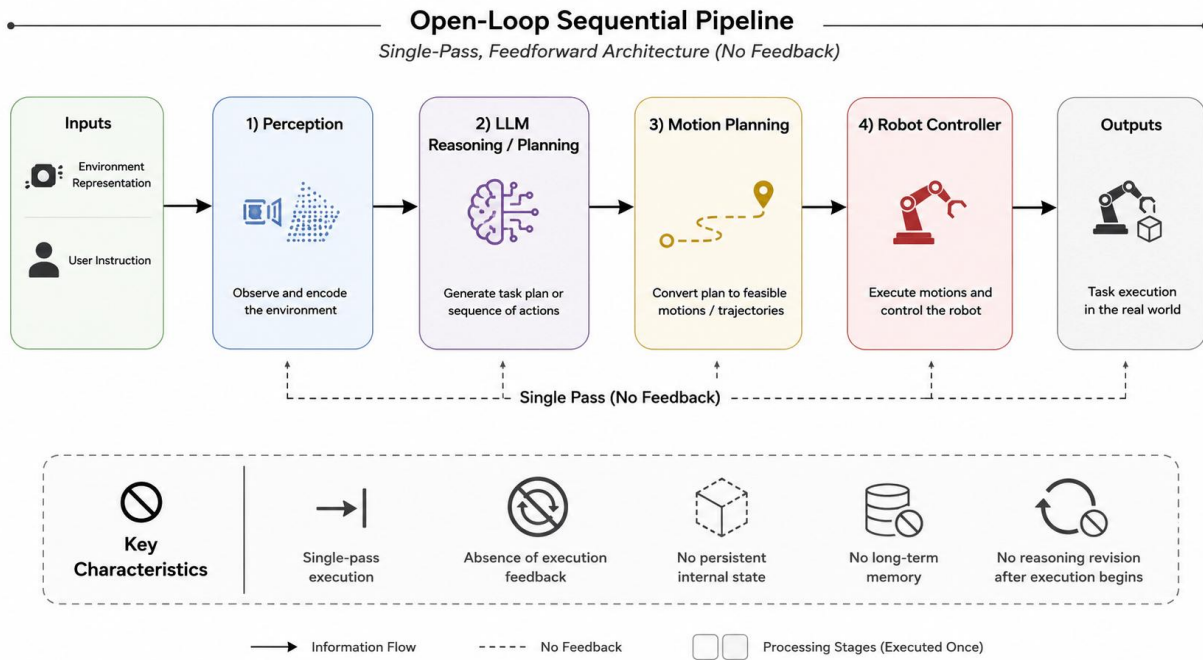


Fig. 4 Open-loop sequential pipeline

5.2.2. Closed-Loop Sequential Pipeline

Many systems extend the previous architecture by introducing feedback from execution, resulting in a closed-loop sequential pipeline. Instead of reasoning only once, the LLM repeatedly observes updated environmental information and revises its decisions throughout task execution. Examples include LLM-Planner, which re-plans when the agent fails or observes new objects (Song et al.

2023), and SRDrone, which refines drone task plans using execution feedback (Zhang et al. 2026b) (Fig. 5).

Closed-loop execution significantly improves robustness in partially observable and dynamic environments because failures, perception updates, and newly discovered obstacles can be incorporated into subsequent planning cycles. Closed-loop execution therefore suits the dynamic, partially observable tasks addressed by many of the reviewed studies.

Despite these enhancements, the architecture remains fundamentally sequential, with a single reasoning agent responsible for overall decision making.

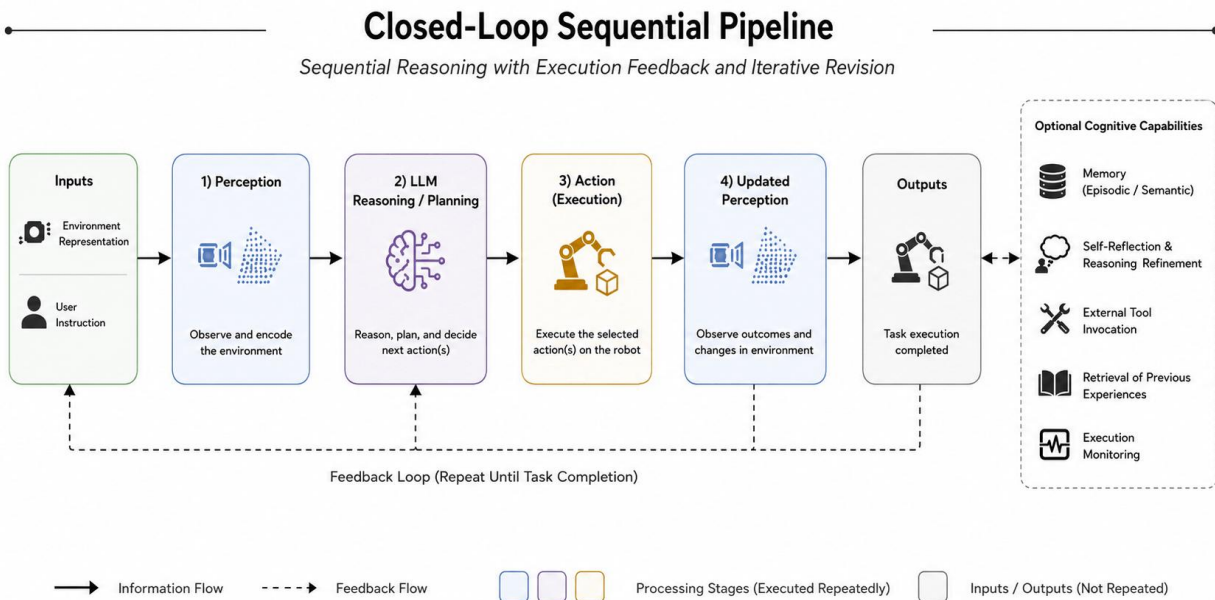


Fig. 5 Closed-loop sequential pipeline

5.2.3. Hierarchical Executive Architecture

As robotic tasks become more complex, many systems adopt a hierarchical executive architecture in which the LLM no longer performs every cognitive function directly. Instead, it acts as a high-level executive that coordinates specialized subsystems (Fig. 6).

The executive is responsible for:

- interpreting user intent;
- decomposing tasks;
- assigning subtasks;
- coordinating execution;
- monitoring overall progress.

Specialized modules then perform domain-specific computation, such as:

- perception,
- navigation,
- motion planning,
- manipulation,
- control.

These lower-level modules may themselves contain LLMs or may rely entirely on conventional robotics algorithms. Consequently, intelligence is distributed across multiple specialized components rather than concentrated in a single model. Representative examples include an executive LLM that invokes perception, navigation, and manipulation tools on a mobile manipulator (Park et al. 2026), multi-layer LLM planners that refine a coarse plan using VLM perception (Luan et al. 2024), and a hierarchical agentic framework for drone-based visual inspection (Herron et al. 2025).

This hierarchical separation reduces cognitive load on the executive LLM, allows reuse of mature robotics software, and enables independent optimization of perception, planning, and control. It also improves scalability, since specialized modules can evolve independently while maintaining a common supervisory layer.

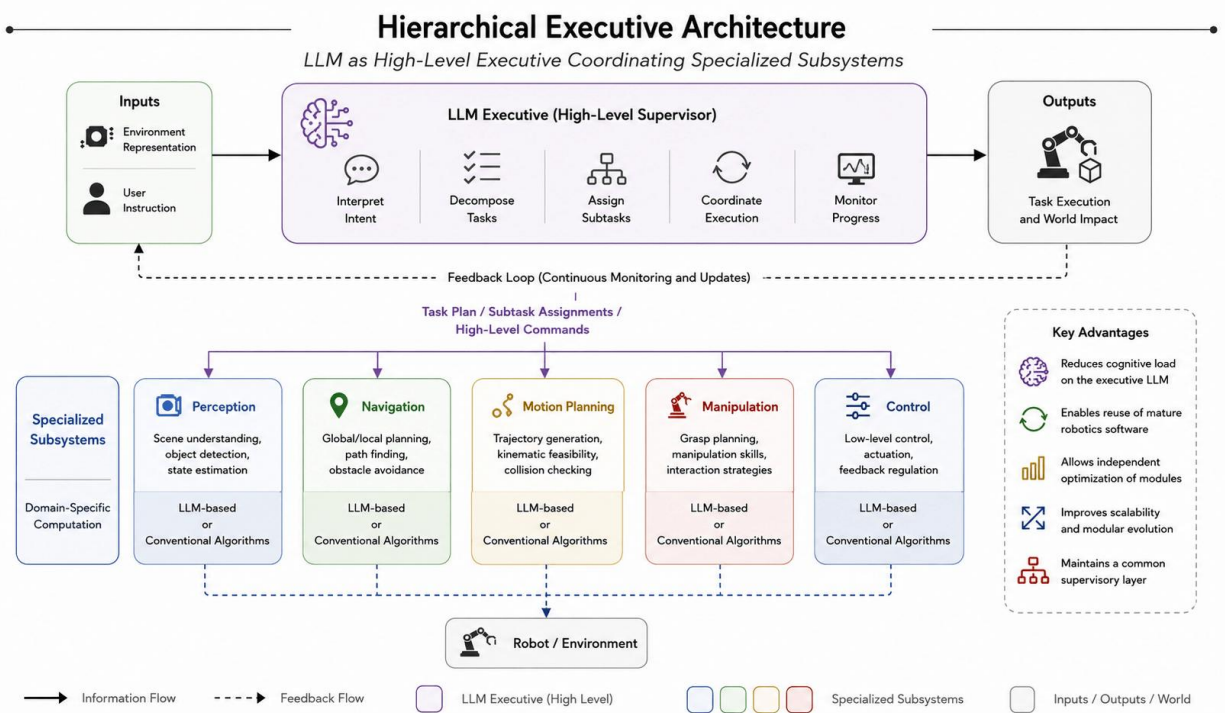


Fig. 6 Hierarchical Executive Architecture

5.2.4. Multi-Agent Architecture

The hierarchical architecture naturally extends into multi-agent systems, where multiple autonomous reasoning agents collaborate to accomplish a shared objective (Fig. 7).

Instead of a single reasoning module, several specialized agents possess their own reasoning and planning capabilities. Examples include navigation agents, perception agents, manipulation agents, communication agents, and task-specific expert agents. For example, Găitan et al. (2025) assign reasoning, correction, memory retrieval, and action to separate LLM agents on a rescue drone; Jarabo-Peñas et al. (2025) separate perception, task analysis, and command execution on a logistics UGV; Yu et al. (2025) pair a VLM inspection agent with a retrieval-augmented maintenance agent; and Azzouni and Pujolle (2026) compare peer-to-peer, centralized, and hybrid organizations for drone swarms.

Coordination typically occurs through one of two strategies:

- a dedicated coordinator responsible solely for communication and synchronization; or
- an executive planning agent that generates a global task decomposition while individual agents solve their assigned subtasks independently.

Compared with hierarchical systems, multi-agent architectures distribute not only computation but also reasoning itself. This enables greater specialization and parallelism, making the architecture particularly suitable for complex missions involving heterogeneous robots or multiple concurrent objectives.

However, these advantages introduce additional challenges related to communication protocols, synchronization, conflict resolution, shared memory, and coordination overhead.

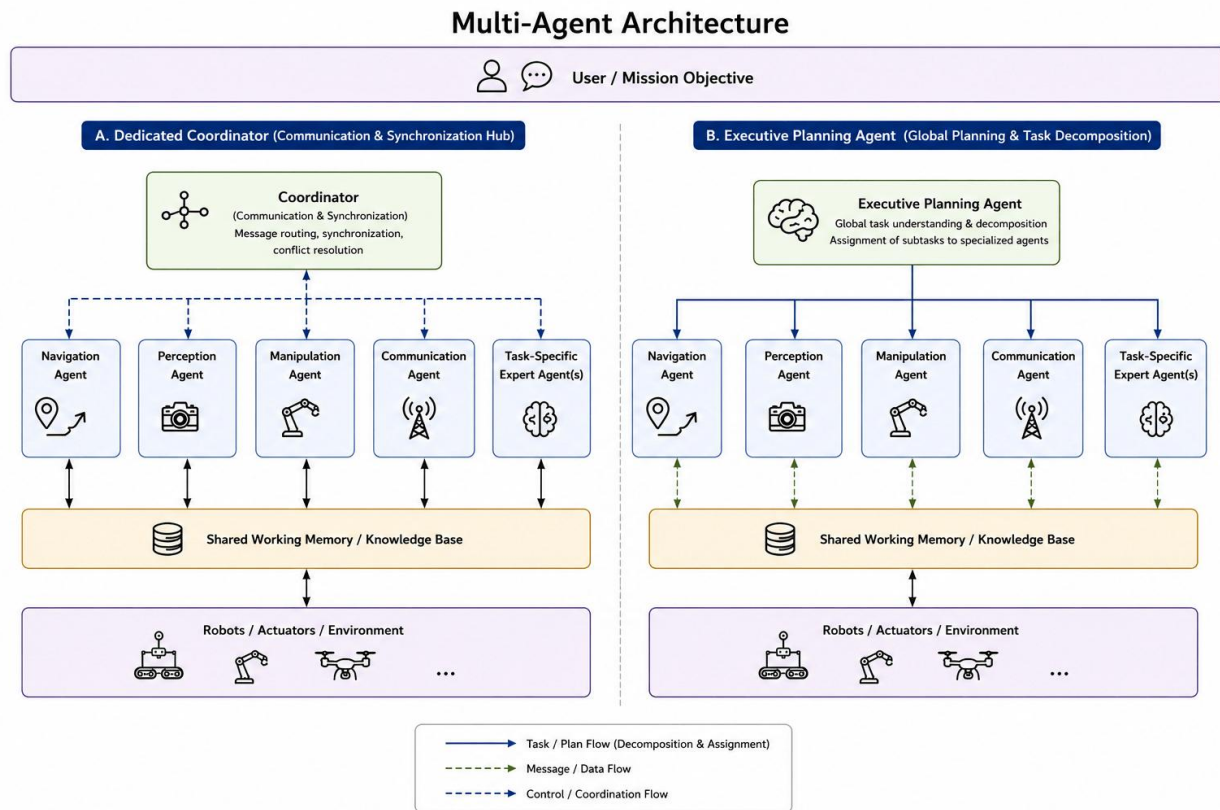


Fig. 7 Multi-agent Architecture

5.2.5. End-to-End Foundation Model

A fundamentally different design philosophy is represented by end-to-end foundation models, where a single large neural network directly maps sensory observations to robot actions (Fig. 8).

Although foundation models continue to improve rapidly, the reviewed literature indicates that purely end-to-end approaches remain uncommon, likely because of the high computational cost of training foundation models end-to-end on the large datasets needed for generalizable, state-of-the-art performance. Examples include ViNT, a navigation foundation model that maps camera observations to waypoints (Shah et al. 2023), OpenFly-Agent, which maps aerial observations and instructions to UAV actions (Gao et al. 2026), and AeroVLA, which outputs UAV control commands directly from images and instructions (Xu et al. 2026).

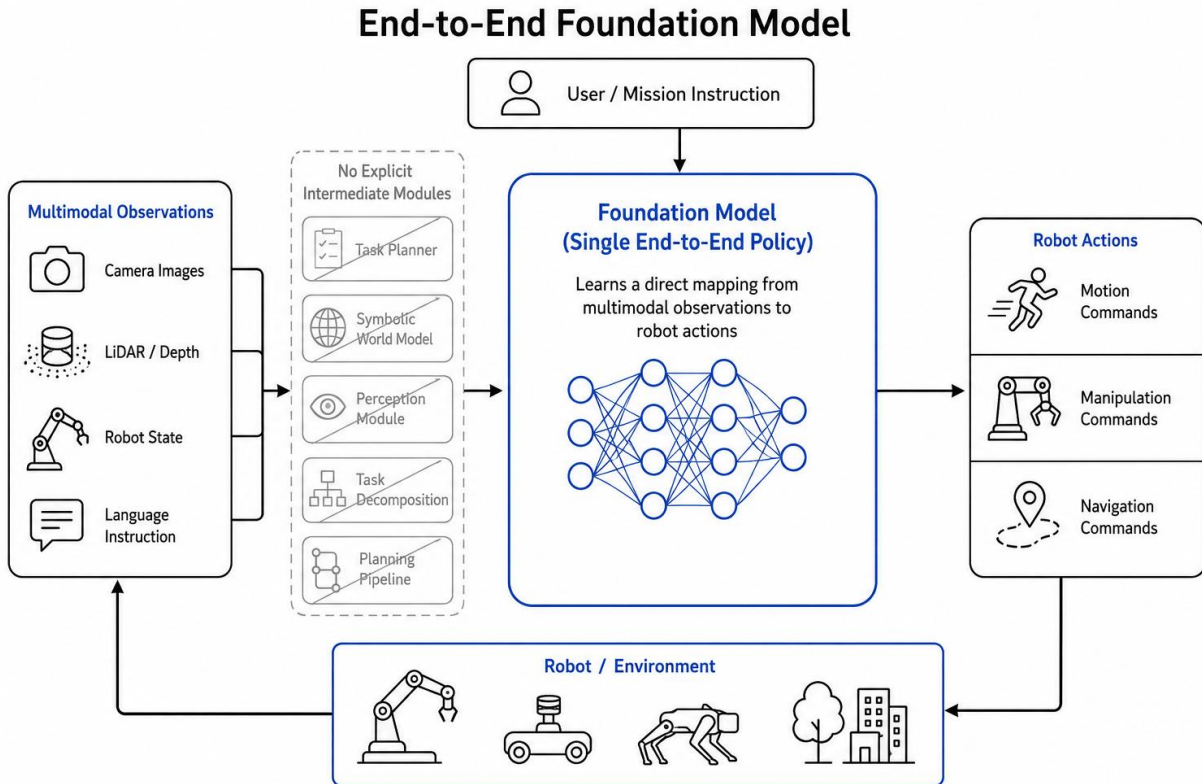


Fig. 8 End-to-End Model Architecture

5.2.6. LLM as a Task-Specific Tool

At the opposite end of the spectrum from end-to-end foundation models are systems in which the LLM serves only as a task-specific language-processing component rather than the central reasoning engine (Fig. 9).

In these architectures, traditional robotics algorithms continue to perform perception, planning, localization, navigation, and control. The LLM is used only for narrowly defined language-related functions, such as:

- translating natural-language commands into structured task or mission specifications;
- extracting entities, constraints, or task parameters from human instructions;
- summarizing sensor logs in natural language for interpretability.

Once this translation step is complete, the remainder of the robotic pipeline proceeds using conventional methods without further LLM involvement. Examples include an offline LLM that translates operator instructions into MAVLink commands and sensor data into natural-language descriptions (Niforatos 2025), an LLM that translates natural-language task descriptions into formal specifications for a task-and-motion planner (Chen et al. 2024), and an LLM that compresses inter-UAV messages into compact semantic representations (Lin et al. 2025b).

LLM as a Task-Specific Tool

LLM used only for language-related tasks; traditional robotics pipeline handles the rest.

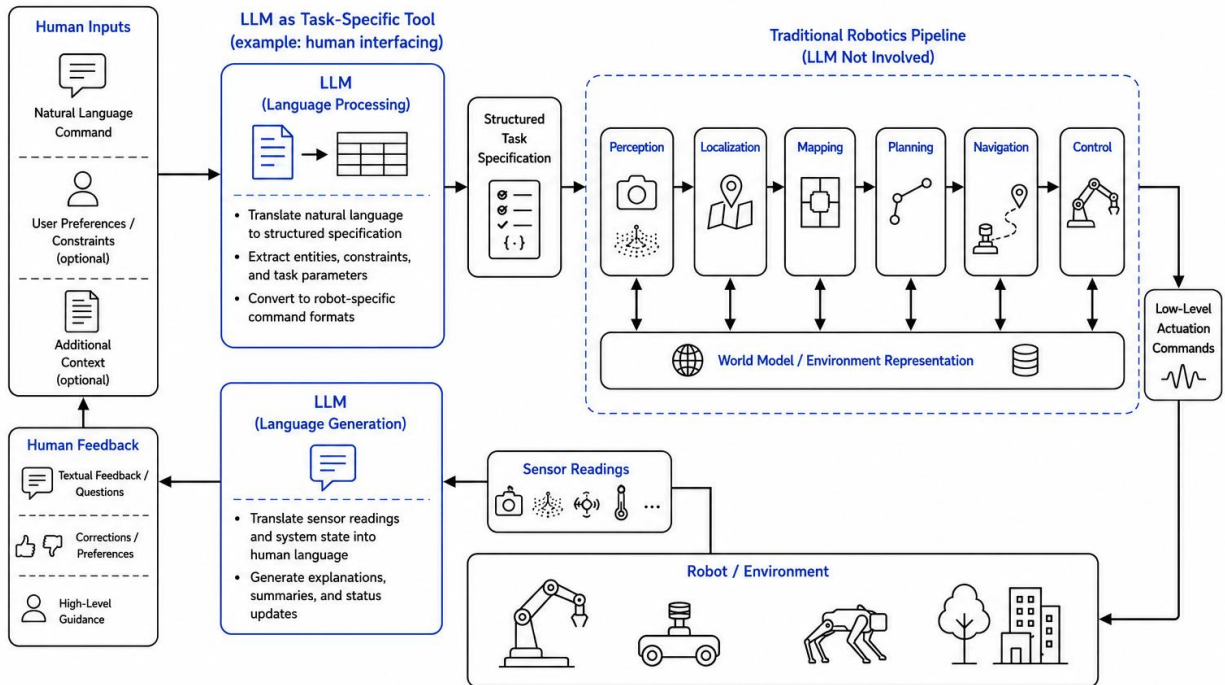


Fig. 9 LLM as task-specific tool

5.2.7. Cross-Cutting Cognitive Enhancement Modules

Beyond the core architectural patterns, several cognitive capabilities appear consistently across multiple architectures. Rather than defining new architectural categories, these components function as optional modules that augment the reasoning capabilities of the system.

The most frequently observed enhancements include:

- **Memory:** enabling storage and retrieval of previous experiences, environmental knowledge, or user preferences.
- **Reflection:** allowing the system to evaluate and revise its own reasoning after execution.
- **Tool use:** where the LLM invokes external software, APIs, perception models, databases, or robotic functions to obtain information or perform specialized computation.

- **Safety mechanisms:** including rule checking, action validation, constraint verification, or secondary model evaluation before execution.
- **Human feedback:** enabling users to supervise planning, approve actions, provide corrections, or intervene during execution.

5.3. LLM Suitability Matrix

The reviewed literature indicates that no single foundation model consistently outperforms others across all robotic functions. Instead, different model families have emerged for different functional modules, with general-purpose LLMs dominating reasoning and planning, multimodal LLMs (VLMs) enabling scene understanding, and specialized vision models continuing to outperform general VLMs on specific perception tasks. Likewise, conventional robotics algorithms remain the preferred solution for motion planning and closed-loop control. Table 5 summarizes the observed suitability of model categories for the major robotic modules identified throughout this review. Suitability is inferred from how often each model family is used for a module and the results reported for it, not from controlled head-to-head comparisons; the maturity column is a qualitative rating of how widely, and how often on physical robots, a family is used for that module.

Table 5. Model Suitability Matrix.

Robotic Module	Most Suitable Models	Findings from the Literature	Maturity
Human interaction	GPT, Llama, Qwen	Strong language understanding, dialogue generation, and instruction following	High

High-level reasoning and planning	GPT, Claude, Qwen, InternVL, LLaVA, Llama, Gemma	Most frequently adopted for task decomposition, decision making, and mission planning	High
Visual understanding	GPT-4V, Qwen2.5-VL, LLaVA, InternVL	Effective for scene understanding, VQA/EQA, landmark recognition, and semantic reasoning	High
Object detection	YOLO, GroundingDINO	Specialized detectors consistently outperform general VLMs for real-time perception	Very High
Image segmentation	SAM/SAM2	Dominant segmentation model across aerial and ground robotic systems	Very High
Image embeddings	DINOv2, SigLIP, CLIP	Common visual backbones for navigation and multimodal reasoning	High
Depth estimation	Depth Anything	Preferred dedicated depth estimation model	High

End-to-end action generation	OpenVLA, NavFoM	Emerging capability with promising results but limited deployment	Medium
------------------------------	-----------------	---	--------

One of the strongest findings across both UAV and UGV systems is the continued dominance of specialized vision models for perception. Object detection is overwhelmingly performed using YOLO or GroundingDINO, while segmentation is almost universally delegated to SAM or SAM2. Similarly, Depth Anything has become the preferred depth estimation model, whereas DINOv2, SigLIP, and CLIP are commonly adopted as visual encoders for downstream navigation and reasoning tasks. This specialization reflects a consistent architectural trend in which perception and reasoning are separated, allowing each component to exploit models optimized for its specific function rather than relying on a single general-purpose VLM.

Several studies further demonstrate that task-specific fine-tuning can substantially improve performance within narrow domains. Fine-tuned variants of Llama, GPT4All, OpenVLA, and Qwen successfully perform specialized tasks including command translation, task offloading, scene understanding, navigation policy learning, and action prediction. In contrast, general-purpose foundation models remain preferred for open-ended reasoning and planning because of their broader knowledge and stronger language capabilities.

5.4. Limitation Analysis

Despite rapid progress, the reviewed literature reveals several recurring limitations that remain largely unresolved, indicating that current LLM-enabled robotic architectures are not yet capable of reliable and fully autonomous operations in complex, real-world environments. These challenges span reasoning, perception, planning, reliability, computational efficiency, and evaluation.

5.4.1. Reliability and Hallucination

The most frequently reported limitation is the imperfect reliability of LLM reasoning. Models may generate plausible but incorrect plans, misinterpret environmental constraints, or hallucinate objects, actions, or world knowledge. Since robotic outputs directly influence physical actions, such errors can result in unsafe behavior or task failure. Consequently, most architectures do not execute LLM outputs directly, instead incorporating rule-based verification, secondary validation models, traditional motion planners, or human approval before execution (Sharma et al. 2025; Ewald et al. 2025; Xue et al. 2026). This widespread use of safety layers reflects the current role of LLMs as advisory reasoning components rather than fully trusted autonomous decision-makers.

5.4.2. Long-Horizon and Spatial Reasoning

Although LLMs exhibit strong commonsense reasoning, they continue to struggle with long-horizon planning and precise spatial reasoning. Planning performance degrades as task complexity, environmental uncertainty, and the number of constraints increase, with early errors often propagating through subsequent actions. Recent systems increasingly address this limitation through hierarchical task decomposition, memory, reflection, and closed-loop replanning.

Spatial reasoning remains one of the weakest capabilities of current foundation models. Multiple studies report difficulties in estimating metric distances, understanding three-dimensional relationships, reasoning about object locations, and navigating cluttered environments (Wen et al. 2024; Majumdar et al. 2024). Similarly, multimodal models frequently struggle with small-object detection, crowded scenes, object counting, fine-grained visual details, thermal imagery, and precise localization. These limitations contribute to the persistent performance gap between current embodied foundation models and human operators on navigation and planning benchmarks.

5.4.3. Perception and Error Propagation

Current LLMs cannot process raw robotic sensor data directly and therefore depend on external perception modules, including VLMs, object detectors, SLAM, and conventional computer vision

algorithms. This modular design introduces error propagation, where inaccurate perception directly degrades downstream reasoning and planning.

The literature also highlights the limitations of general-purpose VLMs for precise perception. For example, Liu et al. (2025a) showed that Qwen2-VL-7B failed to accurately count multiple drones, while LLaVA correctly counted them but produced inaccurate object localization. The authors improved performance by combining an LLM with YOLO-World, allowing language understanding and object localization to be handled by specialized models, illustrating the continued need for hybrid perception architectures.

5.4.4. Prompt Sensitivity and Model Scaling

The reviewed studies consistently identify prompt sensitivity as a practical limitation. Small changes in prompt wording can significantly affect planning quality, action generation, and code synthesis (Hu et al. 2025c). Likewise, excessive in-context examples or retrieved samples may reduce performance by encouraging overfitting to demonstrations rather than the current environment. Several studies report that a medium number of carefully selected demonstrations, typically three to nine examples, provides the best balance between contextual guidance and reasoning performance (Song et al. 2023; Mahmud et al. 2025; Zhu et al. 2026).

Increasing model size also raises inference latency and computational cost, and the reviewed studies provide no consistent evidence that larger models alone improve robotic performance. Consequently, many recent architectures employ smaller task-specific models for perception or control while reserving larger foundation models for high-level reasoning and mission planning.

5.4.5. Computational Efficiency and Evaluation

Large foundation models introduce considerable computational and communication overhead, particularly when cloud-hosted. Latency increases with model size, prompt length, and iterative reasoning, limiting real-time performance in dynamic environments. The challenge is amplified in closed-loop and multi-agent systems, motivating research into model compression, edge deployment, selective reasoning, and lightweight specialized models.

Finally, evaluation remains fragmented across the literature. Most studies compare proposed methods on fragmented datasets and against classical robotic planners or learning-based baselines, while relatively few perform standardized comparisons across LLM-, VLM-, and VLA-based agentic architectures. This lack of consistent benchmarking limits reproducibility and makes it difficult to quantify the relative strengths of competing foundation models.

6. Discussion

6.1. Interpretation of Findings

The reviewed literature demonstrates that LLM-enabled robotic systems increasingly favor modular, hybrid architectures rather than fully end-to-end foundation model solutions. This appears to be driven by the complementary strengths and weaknesses of current foundation models. LLMs excel at semantic reasoning, task decomposition, natural language understanding, and decision making, while specialized perception models remain substantially more accurate and computationally efficient for object detection, segmentation, depth estimation, and feature extraction. Likewise, traditional motion planners and feedback controllers continue to provide deterministic behavior, real-time execution, and safety guarantees that current LLMs cannot reliably achieve. Consequently, contemporary robotic systems increasingly employ foundation models as cognitive reasoning modules embedded within conventional robotics pipelines rather than replacing the entire autonomy stack.

The review also indicates a gradual evolution from sequential pipelines toward closed-loop and agentic architectures. The observed designs range from one-pass perception–planning–control pipelines with little or no feedback to multi-agent architectures with extensive world modeling and feedback.

Model specialization emerged as another dominant trend. Rather than searching for a universal model capable of solving every robotic task, recent systems increasingly assign different foundation models to different functional modules. General-purpose LLMs perform planning and reasoning, multimodal LLMs interpret visual information, specialized vision models execute

perception tasks, and fine-tuned vision-language-action models generate low-level actions in constrained settings. This division of responsibilities reflects both practical engineering considerations and the current maturity of available foundation models.

Finally, despite rapid progress, several limitations consistently constrain current architectures. Spatial reasoning, precise localization, continuous control, and real-time operation remain challenging for foundation models. As a result, most architectures continue to rely on conventional robotics algorithms for safety-critical components, suggesting that hybrid cognitive architectures are likely to remain the dominant design paradigm in the near future.

6.2. Architectural Differences Between UAVs and UGVs

Although UAVs and UGVs share a common high-level architecture consisting of perception, reasoning, planning, execution, and feedback, the reviewed literature reveals notable differences in how these components are implemented. These differences arise primarily from the operating environments, sensing requirements, vehicle dynamics, and safety constraints associated with each platform. The extracted data reflect some of these differences (Online Resource 1): of the 67 UAV and 41 UGV studies, 27% and 44%, respectively, validated their systems on physical robots; 12% and 12% ran the foundation model onboard; 55% and 46% used a memory mechanism; 27% and 34% used function or tool calling; and 18% and 17% used multiple agents.

6.2.1. UAV Architectures

UAV systems place greater emphasis on visual perception and mission-level planning. Because aerial robots typically operate in open three-dimensional environments with limited physical interaction, onboard cameras become the primary sensing modality, often complemented by GPS and inertial measurements. Consequently, UAV architectures extensively employ vision-language models for scene understanding, landmark recognition, visual navigation, inspection, environmental monitoring, and aerial question answering.

Navigation in UAV systems is generally hierarchical. Foundation models interpret mission objectives and generate semantic navigation plans, while classical planners and flight controllers

translate these plans into executable trajectories. Most systems restrict foundation model outputs to high-level navigation decisions or action spaces in four degrees of freedom, such as waypoint generation or position and yaw commands, leaving stabilization and continuous flight control entirely to conventional flight controllers.

UAV applications also exhibit greater adoption of multimodal reasoning because aerial observations naturally encompass large-scale environments requiring scene interpretation, landmark extraction, and semantic mapping. This explains the widespread use of VLMs in inspection, surveillance, search-and-rescue, infrastructure monitoring, and autonomous exploration.

6.2.2. UGV Architectures

UGV architectures generally exhibit stronger integration between perception and interaction with complex environments. Ground robots frequently operate in structured or semi-structured spaces populated by pedestrians, vehicles, buildings, and manipulable objects, requiring richer semantic reasoning about local surroundings.

Compared with UAVs, UGV systems place greater emphasis on object interaction, obstacle avoidance, traffic understanding, human interaction, and socially aware navigation. Consequently, perception pipelines often integrate multiple specialized vision models, including object detectors, segmentation networks, pedestrian detectors, and localization algorithms before semantic reasoning occurs.

6.2.3. Shared Architectural Trends

Despite these platform-specific differences, several architectural characteristics are remarkably consistent across both UAVs and UGVs.

First, both platforms increasingly adopt modular hybrid architectures that combine foundation models with specialized AI models and conventional robotics algorithms. Second, semantic communication has largely replaced direct transmission of raw sensor data between cognitive modules. Third, memory, tool invocation, and iterative feedback are becoming standard

components of modern robotic systems. Finally, perception, reasoning, and control remain clearly separated, reflecting the current limitations of end-to-end foundation models in real-world robotic deployment.

Overall, the similarities between UAV and UGV architectures substantially outweigh their differences. The principal distinction lies not in the overall architectural organization but in the relative emphasis placed on individual modules. UAV systems prioritize long-range visual perception, semantic navigation, and mission planning, whereas UGV systems require richer local perception, human interaction, and rich contextual reasoning. Both platforms, however, lean more towards the same fundamental design philosophy: foundation models serve as high-level cognitive engines integrated with specialized perception modules and conventional robotic control systems rather than replacing them entirely.

6.3. Design Guidelines

Drawing on the taxonomy (Section 5.2), the functional analysis (Section 4), and the platform comparison (Section 6.2), the following guidelines summarize recurring design choices of the reviewed systems. They are derived from the reviewed evidence rather than from controlled comparisons and should therefore be read as starting points for architecture selection rather than as validated rules.

G1. Match the architecture to the task dynamics. Open-loop pipelines suit short tasks in static environments with predictable execution. Tasks in partially observable or changing environments need closed-loop feedback, missions with several heterogeneous subtasks benefit from a hierarchical executive, and multi-agent organizations are justified mainly for multi-robot or multi-objective missions, where their coordination overhead pays off.

G2. Keep geometry and low-level control in classical modules. Restrict the foundation model to semantic, high-level decisions, such as mission decomposition, landmark selection, or waypoint and yaw commands, and delegate trajectory generation, stabilization, and low-level control to deterministic planners and controllers, as most reviewed UAV and UGV systems do.

G3. Use specialized models for geometric perception and VLMs for semantics. Detectors, segmenters, and depth estimators remain more accurate and efficient for precise geometric perception, whereas VLMs are best used for scene understanding, instruction grounding, and visual question answering. Exposing perception models as callable tools lets the LLM invoke them only when needed (Hu et al. 2025c).

G4. Communicate semantically between modules and robots. Passing compact, structured descriptions, for example in JSON, instead of raw sensor data reduces bandwidth and token use, which matters most for swarms with constrained links (Lin et al. 2025b; Du et al. 2026).

G5. Validate every model output before execution. Constrain outputs with schemas or finite-state workflows and check them with rule-based validators or verifier agents before they reach the actuators. Retrieving API documentation and validating calls against platform schemas substantially improves language-to-API reliability (Xue et al. 2026; Sezgin 2025a; Sharma et al. 2025).

G6. Keep safety layers independent of the LLM. Enforce geofences, obstacle-based overrides, and emergency behaviors in deterministic monitors that the model cannot bypass (Ewald et al. 2025), and keep a human in the loop for safety-critical missions.

G7. Budget latency and choose the deployment location explicitly. Most reviewed systems run the foundation model offboard. Latency-critical loops call for small fine-tuned or quantized models onboard or at the edge (Sezgin 2025b; Picano et al. 2025), with large cloud-hosted models reserved for infrequent, high-level reasoning.

G8. Add memory and reflection where the task requires them. Persistent memory supports long-horizon and multi-session tasks, and self-reflection supports recovery from failures, but both add latency and context length. Retrieving memories selectively, for example from semantic maps or key-value stores, limits this cost (Zhang et al. 2026b).

G9. Report system-level metrics under realistic conditions. Report success rate together with latency, onboard computation, communication load, and failure cases, and validate on physical

robots where possible: only 44 of the 124 reviewed studies (35%) included real-world experiments (Section 5.1).

7. Conclusions

Large Language Models (LLMs) have rapidly evolved from natural language interfaces into general-purpose reasoning engines that increasingly serve as the cognitive layer of autonomous robotic systems. This systematic literature review examined recent advances in LLM-enabled ground and aerial robots, synthesizing 124 studies to analyze architectural evolution, functional decomposition, model selection, and emerging engineering practices. The findings indicate that LLMs are not replacing traditional robotics algorithms but are instead enabling a new generation of hybrid architectures in which language models coordinate specialized AI models and conventional robotics modules responsible for perception, localization, planning, and control.

The review presents a taxonomy of six LLM-based robotic architectures: open-loop and closed-loop sequential pipelines, hierarchical executive architectures, multi-agent architectures, end-to-end foundation models, and LLMs used as task-specific tools. These architectures are complemented by optional cognitive enhancement modules such as memory, reflection, and tool use. Across the literature, architectural designs consistently evolve toward increased modularity, feedback, persistent memory, tool use, and self-reflection, enabling robots to perform long-horizon reasoning, recover from execution failures, adapt to dynamic environments, and interact more naturally with humans. Functional analysis further demonstrates a clear division of responsibilities: specialized perception and control algorithms remain essential for real-time, safety-critical operation, while LLMs contribute semantic reasoning, task decomposition, knowledge integration, decision support, and orchestration of external tools. This complementary relationship suggests that hybrid, agentic architectures will remain the dominant paradigm for intelligent robotic systems in the foreseeable future.

The review also provides a comparative analysis of foundation model families and their suitability for different robotic functions. Large proprietary models generally deliver superior reasoning and planning performance, whereas open-weight models offer advantages in privacy, customization,

and edge deployment. Vision-language models have become fundamental for grounding language in visual observations, while smaller language models are emerging as practical alternatives for resource-constrained platforms. These findings demonstrate that model selection should be driven by application requirements that include reasoning complexity, latency, computational resources, deployment constraints, and safety rather than benchmark performance alone.

Despite rapid progress, several challenges continue to limit practical deployment. Current systems remain constrained by inference latency, fragmented memory mechanisms, limited robustness, hallucinations, and the absence of standardized evaluation methodologies for complete robotic systems. Existing research also reveals the need for stronger integration between symbolic reasoning and continuous robotic control, more efficient reasoning methods for edge deployment, standardized architectural frameworks and communication protocols, and safety mechanisms that incorporate uncertainty estimation, runtime monitoring, verification, and human oversight. Addressing these challenges will be essential for deploying trustworthy autonomous robots in complex real-world environments.

This review contributes to the field in four ways. First, it establishes a comprehensive taxonomy of LLM-enabled robotic architectures and their evolution toward agentic systems. Second, it provides a comparative analysis of architectural components, functional modules, communication mechanisms, and memory systems across autonomous robotic applications. Third, it evaluates the suitability of foundation model families for key robotic functional modules and the trade-offs that govern model selection. Fourth, it synthesizes current design principles into practical design guidelines (Section 6.3) and identifies open research challenges that can guide future research and system development.

Abbreviations: The following abbreviations are used in this manuscript.

Abbreviation	Meaning
A*	A-Star Search Algorithm
AI	Artificial Intelligence
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CLIP	Contrastive Language–Image Pre-training
DRL	Deep Reinforcement Learning
EQA	Embodied Question Answering
FAISS	Facebook AI Similarity Search
GAN	Generative Adversarial Network

GPS	Global Positioning System
GPT	Generative Pre-trained Transformer
IMU	Inertial Measurement Unit
JSON	JavaScript Object Notation
LiDAR	Light Detection and Ranging
LLM	Large Language Model
MAVLink	Micro Air Vehicle Link
MCP	Model Context Protocol
MPC	Model Predictive Control
PDDL	Planning Domain Definition Language
PID	Proportional–Integral–Derivative

PPO	Proximal Policy Optimization
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RADAR	Radio Detection and Ranging
RAG	Retrieval-Augmented Generation
RAM	Recognize Anything Model
RGB	Red–Green–Blue
RGB-D	Red–Green–Blue plus Depth
RL	Reinforcement Learning
RQ	Research Question
RRT	Rapidly Exploring Random Tree
SAM	Segment Anything Model

SigLIP	Sigmoid Loss for Language–Image Pre-training
SLAM	Simultaneous Localization and Mapping
SLR	Systematic Literature Review
UAV	Unmanned Aerial Vehicle
UGV	Unmanned Ground Vehicle
VAE	Variational Autoencoder
VLA	Vision-Language-Action Model
VLM	Vision-Language Model
VQA	Visual Question Answering
YOLO	You Only Look Once

Online Resources

Online Resource 1: Data extracted from the 124 studies included in the review.

Online Resource 2: Completed PRISMA 2020 checklist for this review.

Appendix A

Table A1 lists the 124 studies included in the review. Citation labels match the reference list. The full extraction data are in Online Resource 1.

Table A1 Studies included in the review (n = 124)

Citation	Title	Platform	Validation
Afzal et al. 2025	Intelligent UAV Base Station Placement with Generative AI and Retrieval-Augmented LLMs	Aerial	Simulation
Al-Safi et al. 2025	Vega: LLM-Driven Intelligent Chatbot Platform for Internet of Things Control and Development	Both/Generic	Real-world
Aremu et al. 2026	Autonomous Mobile Robot Path Planning Techniques—A Review: Metaheuristic and Cognitive Techniques	Ground	Not stated
Ashqar et al. 2024	Leveraging Multimodal Large Language Models (MLLMs) for Enhanced Object Detection and Scene Understanding in Thermal Images for Autonomous Driving Systems	Ground	Simulation
Ashqar et al. 2025	Advancing Object Detection in Transportation with Multimodal Large Language Models (MLLMs): A Comprehensive Review and Empirical Testing	Ground	Not stated
Azzouni and Pujolle 2026	Autonomous Drone Swarms Using Lightweight LLMs	Aerial	Both
Bai et al. 2025	UAV Dual-Agent: A Framework Integrating LLM and Knowledge Graphs for Low-Altitude UAV Operation and Maintenance	Aerial	Simulation
Basit et al. 2025	ROVER: Autonomous Open-Vocabulary Object Searching in Unexplored Environments Using VLM-Driven Scene Understanding	Ground	Real-world
Blavette et al. 2025	Acceptability and Usability of a Socially Assistive Robot Integrated With a Large Language Model for	Ground	Real-world

	Enhanced Human-Robot Interaction in a Geriatric Care Institution: Mixed Methods Evaluation		
Boroujeni et al. 2026	All you need for object detection: From pixels, points, and prompts to Next-Gen fusion and multimodal LLMs/VLMs in autonomous vehicles	Ground	Not stated
Cao et al. 2025	CogNav: Cognitive Process Modeling for Object Goal Navigation with LLMs	Ground	Both
Cao et al. 2026	Toward Cognitive Augmentation: Personalization and Standardization of Humanoid Social Robot Instruction	Ground	Real-world
Chai et al. 2025	InGen: An LLM-Agent Based Approach for Automatically Generating Evaluation Indicator System for Avionics	Aerial	Simulation
Chang et al. 2025	Experience-Driven NeuroSymbolic System for Efficient Robotic Bolt Disassembly	Ground	Real-world
Chang et al. 2026	Multimodal Intelligent Perception at an Intersection: Pedestrian and Vehicle Flow Dynamics Using a Pipeline-Based Traffic Analysis System	Ground	Real-world
Chen et al. 2024	AutoTAMP: Autoregressive Task and Motion Planning with LLMs as Translators and Checkers	Both/Generic	Both
Chen et al. 2025	GRaD-Nav++: Vision-Language Model Enabled Visual Drone Navigation with Gaussian Radiance Fields and Differentiable Dynamics	Aerial	Both
Chen et al. 2026	CoDrone: Autonomous Drone Navigation Assisted by Edge and Cloud Foundation Models	Aerial	Simulation
Chib and Singh 2025	LG-Traj: LLM Guided Pedestrian Trajectory Prediction	Both/Generic	Simulation
Costa et al. 2024	Experimenting with Planning and Reasoning in Ad Hoc Teamwork Environments with Large Language Models	Ground	Simulation
Dong et al. 2025	Large Language Model-Enabled Multimodal UAV for Rescue	Aerial	Real-world

Du et al. 2026	TaskON: Task-Oriented Networking for Agentic AI	Ground	Simulation
Duvvuru et al. 2025	LLM-Agents Driven Automated Simulation Testing and Analysis of small Uncrewed Aerial Systems	Aerial	Simulation
Ewald et al. 2025	Exploring the cognitive capabilities of large language models in autonomous and swarm navigation systems	Ground	Real-world
Fan et al. 2023	Aerial Vision-and-Dialog Navigation	Aerial	Simulation
Ferrag et al. 2026	UAVBench: An Open Benchmark Dataset for Autonomous and Agentic AI UAV Systems via LLM-Generated Flight Scenarios	Aerial	Simulation
Gao et al. 2026	OpenFly: A Comprehensive Platform for Aerial Vision-Language Navigation	Aerial	Real-world
Geng and Chen 2025	Optimized Scheduling for Multi-Drop Vehicle-Drone Collaboration with Delivery Constraints Using Large Language Models and Genetic Algorithms with Symmetry Principles	Both/Generic	Simulation
Guo et al. 2026	BEDI: a comprehensive benchmark for evaluating embodied agents on UAVs	Aerial	Not stated
Găitan et al. 2025	AI-Enhanced Rescue Drone with Multi-Modal Vision and Cognitive Agentic Architecture	Aerial	Simulation
Han et al. 2024	Large Language Model Guided Reinforcement Learning Based Six-Degree-of-Freedom Flight Control	Aerial	Simulation
Hassan et al. 2024	Integrating Vision and Olfaction via Multi-Modal LLM for Robotic Odor Source Localization	Ground	Real-world
Herron et al. 2025	Hierarchical Agentic Framework for Autonomous Drone-Based Visual Inspection	Aerial	Simulation
Hu et al. 2025a	Agent technologies in smart manufacturing: A comprehensive review of evolution, architectures, and edge-cloud deployment	Ground	Not stated
Hu et al. 2025b	AI Agent for Task Assignment in End-Edge-Cloud Networks	Both/Generic	Not stated
Hu et al. 2025c	LLVM-drone: A synergistic framework integrating large language models and vision	Aerial	Both

	models for visual tasks in unmanned aerial vehicles		
Ieva et al. 2025	Enhancing Last-Mile Logistics: AI-Driven Fleet Optimization, Mixed Reality, and Large Language Model Assistants for Warehouse Operations	Ground	Both
Jarabo-Peñas et al. 2025	Real-World Deployment of an LLM-Enabled Voice-Commanded UGV for Logistics in SAR Missions	Ground	Real-world
Ji et al. 2025	Human-AI Collaborative UAV Visual Object Search via Web Platform	Aerial	Simulation
Ji et al. 2026	Towards Autonomous UAV Visual Object Search in City Space: Benchmark and Agentic Methodology	Aerial	Simulation
Jia et al. 2024	Adversarial and Cooperation Tasks in Multi-Agent System with Large Language Models	Aerial	Both
Jin et al. 2026	HoverAI: An Embodied Aerial Agent for Natural Human-Drone Interaction	Aerial	Real-world
Joshi et al. 2025	Neuro-LIFT: A Neuromorphic, LLM-based Interactive Framework for Autonomous Drone Flight at the Edge	Aerial	Both
Kachaev et al. 2025	Mind and Motion Aligned: A Joint Evaluation IsaacSim Benchmark for Task Planning and Low-Level Policies in Mobile Manipulation	Ground	Simulation
Kallakuri et al. 2024	ATLAS: Adaptive landmark acquisition using LLM-guided navigation	Ground	Simulation
Kong et al. 2025	Embodied AI in Mobile Robot Simulation with EyeSim: Coverage Path Planning with Large Language Models	Ground	Simulation
Kumar and Kant 2025	TU-DAT: A Computer Vision Dataset on Road Traffic Anomalies	Ground	Not stated
Li et al. 2025a	3EED: Ground Everything Everywhere in 3D	Both/Generic	Not stated
Li et al. 2025b	SkyVLN: Vision-and-Language Navigation and NMPC Control for UAVs in Urban Environments	Aerial	Simulation
Lin et al. 2025a	AirVista-II: An Agentic System for Embodied UAVs Toward Dynamic Scene Semantic Understanding	Aerial	Simulation

Lin et al. 2025b	Talk Less, Fly Lighter: Autonomous Semantic Compression for UAV Swarm Communication via LLMs	Aerial	Simulation
Lisondra et al. 2026	Embodied AI with Foundation Models for Mobile Service Robots: A Systematic Review	Ground	Not stated
Liu et al. 2023	AerialVLN: Vision-and-Language Navigation for UAVs	Aerial	Simulation
Liu et al. 2025a	EdLLM: A Novel Entity Detection Method for Test Data Integrating YOLO-World and LLM	Aerial	Simulation
Liu et al. 2025b	RADAR: Agent-guided retrieval with physics-based consistency filtering for aviation fault diagnosis	Aerial	Simulation
Liu et al. 2026a	A Visual Target Navigation Method for Quadrotor Based on Large Language Model in Unknown Environment	Aerial	Both
Liu et al. 2026b	IndoorUAV: Benchmarking Vision-Language UAV Navigation in Continuous Indoor Environments	Both/Generic	Not stated
Long et al. 2026	Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory	Both/Generic	Simulation
Lovaco 2025	The Large Language Model Pilot: Generative Artificial Intelligence for Cyber-Physical Flight Simulations	Aerial	Simulation
Luan et al. 2024	Enhancing Robot Task Planning and Execution through Multi-Layer Large Language Models	Ground	Real-world
Luo 2025	Sensing User Intent: An LLM-Powered Agent for On-the-Fly Personalized Virtual Space Construction from UAV Sensor Data	Aerial	Real-world
Luo et al. 2024	Integration of LLMs and the Physical World: Research and Application	Both/Generic	Not stated
Ma et al. 2025	TopoLLM: LLM-driven adaptive tool learning for real-time emergency network topology planning	Aerial	Simulation
Macdonald et al. 2024	Language, Camera, Autonomy! Prompt-engineered Robot Control for Rapidly Evolving Deployment	Both/Generic	Both

Mahmud et al. 2025	A knowledge-driven framework for Robotic Odor Source Localization using large language models	Ground	Both
Majumdar et al. 2024	OpenEQA: Embodied Question Answering in the Era of Foundation Models	Both/Generic	Not stated
Maletić et al. 2025	Spatial-Semantic Reasoning using Large Language Models for Efficient UAV Search Operations	Aerial	Both
Masri et al. 2025a	Large Language Models (LLMs) as Traffic Control Systems at Urban Intersections: A New Paradigm	Ground	Simulation
Masri et al. 2025b	Leveraging Bird Eye View Video and Multimodal Large Language Models for Real-Time Intersection Control and Reasoning	Aerial	Real-world data only
Mekrache and Ksentini 2024	LLM-enabled Intent-driven Service Configuration for Next Generation Networks	Aerial	Real-world
Mower et al. 2026	A robot operating system framework for using large language models in embodied AI	Both/Generic	Real-world
Mukhametzianov and Nambo 2025	History-Aware Multimodal Instruction-Oriented Policies for Navigation Tasks	Ground	Both
Narchetty et al. 2025	Towards AI Drone Assistants in Extended Reality	Aerial	Simulation
Niforatos 2025	Offline large language model for drone control and monitoring (U.S	Aerial	Not stated
Nunes et al. 2025	A Framework Leveraging Large Language Models for Autonomous UAV Control in Flying Networks	Aerial	Simulation
Park et al. 2026	EXAONE-VLA: a Unified Vision-Language Framework for mobile manipulation via semantic Topology and Hierarchical LLM reasoning	Ground	Real-world
Petersen et al. 2025	Method and system for specifying an active inference based agent using natural language (U.S	Both/Generic	Both
Picano et al. 2025	A Matching Game for LLM Layer Deployment in Heterogeneous Edge Networks	Ground	Both
Raiyn 2026	A generative AI-driven intelligent rules framework for traffic accident detection and analysis	Ground	Simulation

Rashid et al. 2025	MAVREN: A multilayered adaptive framework for deploying vision-language models on resource-constrained unmanned aerial vehicles for autonomous search and rescue	Aerial	Not stated
Rashidieranjbar et al. 2025	A systematic literature review: Generative artificial intelligence applications for ground mobile robot navigation	Ground	Not stated
Sezgin 2025a	Natural Language Interfaces for Structured Query Generation in IoD Platforms	Aerial	Simulation
Sezgin 2025b	Scenario-Driven Evaluation of Autonomous Agents: Integrating Large Language Model for UAV Mission Reliability	Aerial	Simulation
Sezgin and Boyacı 2025	LLM-Powered UAVs: A RAG-Based Approach for Safety-Critical Operations	Aerial	Both
Shah et al. 2023	ViNT: A foundation model for visual navigation	Both/Generic	Both
Sharma et al. 2025	Prompted to Fly: Translating Free-Form Instructions into Schema-Constrained Mission Generation for UAVs Using LLMs	Aerial	Simulation
Shetye et al. 2026	AMPLIFY: A Modular LLM Platform for Mission-Critical Retrieval and Reasoning (Technical report)	Both/Generic	Real-world
Skaug and Nojournian 2025	A Multimodal Artificial Intelligence Framework for Intelligent Geospatial Data Validation and Correction	Ground	Real-world
Song et al. 2023	LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models	Ground	Simulation
Su et al. 2025	Thinking Like an Expert: Aligning LLM Thought Processes for Automated Safety Modeling of High-Altitude Solar Drones	Aerial	Simulation
Sun et al. 2024	Generative AI for Advanced UAV Networking	Aerial	Simulation
Tabrizian et al. 2025	Chain-of-thought flight planner: End-to-end LLM routing under wind hazards	Aerial	Simulation
Trabelsi et al. 2026	Dynamic task offloading in vehicular networks using large language models for adaptive low latency decision making	Ground	Simulation

Waga et al. 2025	A survey on autonomous navigation for mobile robots: From traditional techniques to deep learning and large language models	Ground	Not stated
Wang and Deng 2025	Outdoor Multi-UAV Collaborative Task Planning System Based on Multi-Agent and Improved PPO Algorithm	Aerial	Simulation
Wang et al. 2025a	GSCE: A Prompt Framework with Enhanced Reasoning for Reliable LLM-Driven Drone Control	Aerial	Simulation
Wang et al. 2025b	PuppetLine: An Interactive System for Embodied Storytelling with LLM-driven Swarm Robots	Ground	Real-world
Wang et al. 2025c	Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology	Aerial	Simulation
Wang et al. 2025d	“Hi AirStar, Guide Me to the Badminton Court”	Aerial	Real-world
Webb et al. 2026	FIRE-VLM: A vision-language-driven reinforcement learning framework for UAV wildfire tracking in a physics-grounded fire digital twin	Aerial	Simulation
Wen et al. 2024	On the road with GPT-4V(ision): Explorations of utilizing visual-language model as autonomous driving agent	Ground	Simulation
Xiao et al. 2025	UAV-ON: A Benchmark for Open-World Object Goal Navigation with Aerial Agents	Aerial	Simulation
Xu et al. 2026	AeroVLA: A Vision-Language-Action Model for UAV Navigation via Minimalist End-to-End Control	Aerial	Simulation
Xue et al. 2026	Towards Naturalistic Human-Machine Teaming with LLM Agents: A Case Study in Air Traffic Management	Aerial	Simulation
Yao et al. 2026	AeroVerse: UAV-Agent Benchmark Suite for Simulating, Pre-training, Finetuning, and Evaluating Aerospace Embodied Foundation Models	Aerial	Simulation
Yu et al. 2025	Lightweight multimodal LLM-empowered dual-agent	Ground	Simulation

	collaboration for reliable defect detection and maintenance recommendation in tunnel infrastructure		
Yuan et al. 2025	Seeing With Words: Interpretable Language-Guided Drone Geo-Localization via LLM-Enriched Semantic Attribute Alignment	Aerial	Simulation
Zha et al. 2026	AirCopBench: A Benchmark for Multi-drone Collaborative Embodied Perception and Reasoning	Aerial	Simulation
Zhang et al. 2024	Demo abstract: Embodied aerial agent for city-level visual language navigation using large language model	Aerial	Simulation
Zhang et al. 2025a	Embodied AI-Enhanced Vehicular Networks: An Integrated Vision Language Models and Reinforcement Learning Method	Ground	Simulation
Zhang et al. 2025b	Is your VLM sky-ready? A comprehensive spatial intelligence benchmark for UAV navigation	Aerial	Simulation
Zhang et al. 2025c	Open3D-VQA: A Benchmark for Embodied Spatial Concept Reasoning with Multimodal Large Language Model in Open Space	Aerial	Simulation
Zhang et al. 2026a	Embodied navigation foundation model	Both/Generic	Both
Zhang et al. 2026b	SRDrone: LLM-Driven Self-Refinement for Embodied Drone Task Planning	Aerial	Both
Zhao et al. 2025a	AirScape: An Aerial Generative World Model with Motion Controllability	Aerial	Simulation
Zhao et al. 2025b	CityEQA: A Hierarchical LLM Agent on Embodied Question Answering Benchmark in City Space	Aerial	Simulation
Zhao et al. 2025c	FlexiFly: Interfacing the Physical World with Foundation Models Empowered by Reconfigurable Drone Systems	Aerial	Real-world
Zhao et al. 2025d	UrbanVideo-Bench: Benchmarking Vision-Language Models on Embodied Intelligence with Video Data in Urban Spaces	Aerial	Real-world

Zhao et al. 2026	EmbodiedFly: Embodied LLM Agent with an Autonomous Reconfigurable Drone	Aerial	Real-world
Zhou et al. 2024	HAZARD challenge: Embodied decision making in dynamically changing environments	Ground	Simulation
Zhou et al. 2025	AirSpatialBot: A Spatially Aware Aerial Agent for Fine-Grained Vehicle Attribute Recognition and Retrieval	Aerial	Simulation
Zhu et al. 2024	Task Offloading with LLM-Enhanced Multi-Agent Reinforcement Learning in UAV-Assisted Edge Computing	Aerial	Simulation
Zhu et al. 2026	CA-VLN: Collaborative Agents in MLLM-Powered Visual-Language Navigation	Ground	Simulation
Zu et al. 2024	Language and Sketching: An LLM-driven Interactive Multimodal Multitask Robot Navigation Framework	Ground	Both

Declarations

Funding: No funds, grants, or other support was received.

Competing interests: The author has no competing interests to declare that are relevant to the content of this article.

Ethics approval: Not applicable. This article is a systematic review of published literature and did not involve human participants or animals.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Data availability: The data extracted from the included studies are available in Online Resource 1. The included studies are listed in Appendix A and in the reference list. The completed PRISMA 2020 checklist is Online Resource 2.

Authors' contributions: Abdelrahman Mahmoud: conceptualization, methodology, investigation, formal analysis, visualization, and writing of the original draft. The author read and approved the final manuscript.

Artificial intelligence: ChatGPT (OpenAI) was used to generate the schematic diagrams in Figs. 1–9 from the author's descriptions of those architectures. The author reviewed and edited every diagram. Grok 4.7 was used for formatting. These tools were not used for the search, screening, data extraction, or analysis. The author takes full responsibility for the content.

References

- Afzal M, Guo S, Saeed N (2025) Intelligent UAV Base Station Placement with Generative AI and Retrieval-Augmented LLMs. In 2025 IEEE International Workshop on Radio Frequency and Antenna Technologies (iWRF&AT) (pp. 392–397).
<https://doi.org/10.1109/iwrfat65352.2025.11103391>
- Al-Safi H, Ibrahim H, Steenson P (2025) Vega: LLM-Driven Intelligent Chatbot Platform for Internet of Things Control and Development. *Sensors*, 25(12), 3809.
<https://doi.org/10.3390/s25123809>
- Aremu MB, Ahmed G, Elferik S, Saif A-WA (2026) Autonomous Mobile Robot Path Planning Techniques—A Review: Metaheuristic and Cognitive Techniques. *Robotics*, 15(1), 23.
<https://doi.org/10.3390/robotics15010023>
- Ashqar HI, Alhadidi TI, Elhenawy M, Khanfar NO (2024) Leveraging Multimodal Large Language Models (MLLMs) for Enhanced Object Detection and Scene Understanding in Thermal Images for Autonomous Driving Systems. *Automation*, 5(4), 508–526.
<https://doi.org/10.3390/automation5040029>
- Ashqar HI, Jaber A, Alhadidi TI, Elhenawy M (2025) Advancing Object Detection in Transportation with Multimodal Large Language Models (MLLMs): A Comprehensive

- Review and Empirical Testing. *Computation*, 13(6), 133.
<https://doi.org/10.3390/computation13060133>
- Azzouni A, Pujolle G (2026) Autonomous Drone Swarms Using Lightweight LLMs. In M. Andreoni & S. Thakkar (Eds.), *Proceedings of the 1st GENZERO Workshop* (pp. 21–31). Springer. https://doi.org/10.1007/978-981-95-1050-4_3
- Bai G, Luo X, Wang Y, Zhou J, Shen J, Sun Z (2025) UAV Dual-Agent: A Framework Integrating LLM and Knowledge Graphs for Low-Altitude UAV Operation and Maintenance. In *Proceedings of the International Symposium on Intelligent Computing and Networking 2025 (Lecture Notes in Networks and Systems, pp. 590–602)*. Springer. https://doi.org/10.1007/978-3-032-09694-4_45
- Basit A, Pudasaini N, Shafique M (2025) ROVER: Autonomous Open-Vocabulary Object Searching in Unexplored Environments Using VLM-Driven Scene Understanding. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). <https://doi.org/10.1109/ijcnn64981.2025.11228732>
- Blavette L, Dacunha S, Alameda-Pineda X, García DH, Gannot S, Gras F, Gunson N, Lemaignan S, Polic M, Tandaitnik P, Tonini F, Rigaud A, Pino M (2025) Acceptability and Usability of a Socially Assistive Robot Integrated With a Large Language Model for Enhanced Human-Robot Interaction in a Geriatric Care Institution: Mixed Methods Evaluation. *JMIR Human Factors*, 12, e76496. <https://doi.org/10.2196/76496>
- Boroujeni SPH, Mehrabi N, Alzorgan H, Fazeli M, Razi A (2026) All you need for object detection: From pixels, points, and prompts to Next-Gen fusion and multimodal LLMs/VLMs in autonomous vehicles. *Image and Vision Computing*, 169, 105944. <https://doi.org/10.1016/j.imavis.2026.105944>
- Cao Y, Zhang J, Yu Z, Liu S, Qin Z, Zou Q, Du B, Xu K (2025) CogNav: Cognitive Process Modeling for Object Goal Navigation with LLMs. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9550-9560. <https://doi.org/10.1109/iccv51701.2025.00891>

- Cao H, Chen S, Braun G, Stahre J (2026) Toward Cognitive Augmentation: Personalization and Standardization of Humanoid Social Robot Instruction. *IOP Conference Series Materials Science and Engineering*, 1342(1), 012045. <https://doi.org/10.1088/1757-899x/1342/1/012045>
- Chai R, Wang L, Fan R, Liu S (2025) InGen: An LLM-Agent Based Approach for Automatically Generating Evaluation Indicator System for Avionics. In *2025 4th International Symposium on Aerospace Engineering and Systems (ISAES)* (pp. 138–143). <https://doi.org/10.1109/isaes66870.2025.11274390>
- Chang P, Wang Z, Peng Y, He Z, Chen M (2025) Experience-Driven NeuroSymbolic System for Efficient Robotic Bolt Disassembly. *Batteries*, 11(9), 332. <https://doi.org/10.3390/batteries11090332>
- Chang BR, Tsai H, Chen CH (2026) Multimodal Intelligent Perception at an Intersection: Pedestrian and Vehicle Flow Dynamics Using a Pipeline-Based Traffic Analysis System. *Electronics*, 15(2), 353. <https://doi.org/10.3390/electronics15020353>
- Chen Y, Arkin J, Dawson C, Zhang Y, Roy N, Fan C (2024) AutoTAMP: Autoregressive Task and Motion Planning with LLMs as Translators and Checkers. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6695–6702). <https://doi.org/10.1109/icra57147.2024.10611163>
- Chen Q, Gao N, Huang S, Low J, Chen T, Sun J, Schwager M (2025) GRaD-Nav++: Vision-Language Model Enabled Visual Drone Navigation with Gaussian Radiance Fields and Differentiable Dynamics. *IEEE Robotics and Automation Letters*, 11(2), 1418–1425. <https://doi.org/10.1109/lra.2025.3643290>
- Chen P, Ouyang T, Luo K, Hong W, Chen X (2026) CoDrone: Autonomous Drone Navigation Assisted by Edge and Cloud Foundation Models. *IEEE Internet of Things Journal*, 13(4), 5593–5609. <https://doi.org/10.1109/jiot.2025.3648721>

- Chib PS, Singh P (2025) LG-Traj: LLM Guided Pedestrian Trajectory Prediction. In 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (pp. 6861–6871). <https://doi.org/10.1109/iccvw69036.2025.00709>
- Costa PBD, Santos PTC, Boaro JMC, Moraes D, Duarte J, Colcher S (2024) Experimenting with Planning and Reasoning in Ad Hoc Teamwork Environments with Large Language Models. In Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) (Vol. 1, pp. 438–445). <https://doi.org/10.5220/0012472600003636>
- Ding Y, Xin B, Chen J (2021) A Review of Recent Advances in Coordination Between Unmanned Aerial and Ground Vehicles. *Unmanned Systems*, 9(2), 97–117. <https://doi.org/10.1142/s2301385021500084>
- Dong A, Choot X, Seow B, Trinh R, Wang Z (2025) Large Language Model-Enabled Multimodal UAV for Rescue. In 2025 Regional Student Conferences. American Institute of Aeronautics and Astronautics. <https://doi.org/10.2514/6.2025-106062>
- Du B, Li R, Niyato D, Su Z (2026) TaskON: Task-Oriented Networking for Agentic AI. *IEEE Network*, 40(3), 75–82. <https://doi.org/10.1109/mnet.2026.3657581>
- Duvvuru VSA, Zhang B, Vierhauser M, Agrawal A (2025) LLM-Agents Driven Automated Simulation Testing and Analysis of small Uncrewed Aerial Systems. In 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE) (pp. 385–397). <https://doi.org/10.1109/icse55347.2025.00223>
- Ewald D, Rogowski F, Suśniak M, Bartkowiak P, Blumensztajn P (2025) Exploring the cognitive capabilities of large language models in autonomous and swarm navigation systems. *Electronics*, 15(1), 35. <https://doi.org/10.3390/electronics15010035>
- Fan Y, Chen W, Jiang T, Zhou C, Zhang Y, Wang X (2023) Aerial Vision-and-Dialog Navigation. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 3043–3061). <https://doi.org/10.18653/v1/2023.findings-acl.190>

- Ferrag MA, Lakas A, Debbah M (2026) UAVBench: An Open Benchmark Dataset for Autonomous and Agentic AI UAV Systems via LLM-Generated Flight Scenarios. *IEEE Open Journal of Vehicular Technology*, 7, 2542–2558.
<https://doi.org/10.1109/ojvt.2026.3721777>
- Găitan NC, Batinas BI, Ursu C (2025) AI-Enhanced Rescue Drone with Multi-Modal Vision and Cognitive Agentic Architecture. *AI*, 6(10), 272. <https://doi.org/10.3390/ai6100272>
- Gao Y, Li C, You Z, Liu J, Li Z, Chen P, Chen Q, Tang Z, Wang L, Yang P, Tang Y, Tang Y, Liang S, Zhu S, Xiong Z, Su Y, Ye X, Li J, Ding Y, Wang D, Li X, Wang Z, Zhao B (2026) OpenFly: A Comprehensive Platform for Aerial Vision-Language Navigation. In *Proceedings of the International Conference on Learning Representations (ICLR 2026)*.
https://proceedings.iclr.cc/paper_files/paper/2026/hash/a6fcfd15cd01e4a550808c3e01f5583d-Abstract-Conference.html
- Geng M, Chen A (2025) Optimized Scheduling for Multi-Drop Vehicle-Drone Collaboration with Delivery Constraints Using Large Language Models and Genetic Algorithms with Symmetry Principles. *Symmetry*, 17(6), 934. <https://doi.org/10.3390/sym17060934>
- Guo M, Wu M, He J, Li S, Li H, Tao C (2026) BEDI: a comprehensive benchmark for evaluating embodied agents on UAVs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232, 910–936. <https://doi.org/10.1016/j.isprsjprs.2026.01.013>
- Han Y, Yang M, Ren Y, Li W (2024) Large Language Model Guided Reinforcement Learning Based Six-Degree-of-Freedom Flight Control. *IEEE Access*, 12, 89479-89492.
<https://doi.org/10.1109/access.2024.3411015>
- Hassan S, Wang L, Mahmud KR (2024) Integrating Vision and Olfaction via Multi-Modal LLM for Robotic Odor Source Localization. *Sensors*, 24(24), 7875.
<https://doi.org/10.3390/s24247875>
- Herron E, Lee XY, Sin G, Gonzalez Diaz T, Farahat A, Gupta C (2025) Hierarchical Agentic Framework for Autonomous Drone-Based Visual Inspection. In *Proceedings of the Annual*

Conference of the PHM Society (Vol. 17, No. 1).
<https://doi.org/10.36001/phmconf.2025.v17i1.4328>

Hu K, Yao X, Qiao F (2025a) Agent technologies in smart manufacturing: A comprehensive review of evolution, architectures, and edge-cloud deployment. *Advanced Engineering Informatics*, 70, 104235. <https://doi.org/10.1016/j.aei.2025.104235>

Hu H, Zhang Y, Huang X, Zhang K, Wu F, Ning Z (2025b) AI Agent for Task Assignment in End-Edge-Cloud Networks. *IEEE Network*, 40(3), 67–74.
<https://doi.org/10.1109/mnet.2025.3643628>

Hu Y, Zhou Y, Zhu Z, Yang X, Zhang H, Bian K, Han H (2025c) LLVM-drone: A synergistic framework integrating large language models and vision models for visual tasks in unmanned aerial vehicles. *Knowledge-Based Systems*, 327, 114190.
<https://doi.org/10.1016/j.knosys.2025.114190>

Ichter B, Brohan A, Chebotar Y, Finn C, Hausman K, Herzog A, Ho D, Ibarz J, Irpan A, Jang E, Julian R, Kalashnikov D, Levine S, Lu Y, Parada C, Rao K, Sermanet P, Toshev AT, Vanhoucke V et al (2023) Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Proceedings of the 6th Conference on Robot Learning* (Vol. 205, pp. 287–318). PMLR. <https://proceedings.mlr.press/v205/ichter23a.html>

Ieva S, Bilenchi I, Gramegna F, Pinto A, Scioscia F, Ruta M, Loseto G (2025) Enhancing Last-Mile Logistics: AI-Driven Fleet Optimization, Mixed Reality, and Large Language Model Assistants for Warehouse Operations. *Sensors*, 25(9), 2696.
<https://doi.org/10.3390/s25092696>

Jarabo-Peñas A, Bravo-Arrabal J, Lin-Yang D, Pastor F, Ladig R, Fernández-Lozano JJ, Christensen AL, García-Cerezo A (2025) Real-World Deployment of an LLM-Enabled Voice-Commanded UGV for Logistics in SAR Missions. In *2025 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)* (pp. 57–63).
<https://doi.org/10.1109/ssrr68451.2025.11391314>

- Javaid S, Fahim H, He B, Saeed N (2024) Large language models for UAVs: Current state and pathways to the future. *IEEE Open Journal of Vehicular Technology*, 5, 1166–1192.
<https://doi.org/10.1109/ojvt.2024.3446799>
- Ji Y, Qiu S, Zhu Z, Ju R (2025) Human-AI Collaborative UAV Visual Object Search via Web Platform. In *Web Engineering (Lecture Notes in Computer Science*, pp. 385–390). Springer.
https://doi.org/10.1007/978-3-031-97207-2_32
- Ji Y, Zhu Z, Zhao Y, Liu B, Gao C, Zhao Y, Qiu S, Hu Y, Yin Q (2026) Towards Autonomous UAV Visual Object Search in City Space: Benchmark and Agentic Methodology. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(22), 18342–18350.
<https://doi.org/10.1609/aaai.v40i22.38898>
- Jia S, Cui Z, Wang K, Ding Y, Han D, Ma T (2024) Adversarial and Cooperation Tasks in Multi-Agent System with Large Language Models. In *2024 IEEE International Conference on Unmanned Systems (ICUS)* (pp. 1840–1847).
<https://doi.org/10.1109/icus61736.2024.10839948>
- Jin Y, Kuzmin N, Demianchuk G, Lezina M, Mehboob F, Tokmurziyev I, Altamirano Cabrera M, Mustafa MA, Tsetserukou D (2026) HoverAI: An Embodied Aerial Agent for Natural Human-Drone Interaction. *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, 1130-1134.
<https://doi.org/10.1145/3776734.3794571>
- Joshi A, Sanyal S, Roy K (2025) Neuro-LIFT: A Neuromorphic, LLM-based Interactive Framework for Autonomous Drone Flight at the Edge. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–9).
<https://doi.org/10.1109/ijcnn64981.2025.11228467>
- Kachaev NE, Spiridonov AN, Gorodetsk AS, Muravyev KF, Oskolkov NS, Narendra A, Shakhuro VI, Makarov DA, Panov AI, Fedotova PD, Kovalev AK (2025) Mind and Motion Aligned: A Joint Evaluation IsaacSim Benchmark for Task Planning and Low-Level

Policies in Mobile Manipulation. *Doklady Mathematics*, 112(2), 400–408.
<https://doi.org/10.1134/s1064562425700450>

Kallakuri U, Prakash B, Mazumder AN, Al-Rashid H, Waytowich NR, Mohsenin T (2024) ATLAS: Adaptive landmark acquisition using LLM-guided navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (VLADR 2024)*. <https://openreview.net/forum?id=VhpxzSWTWj>

Kong X, Zhang W, Hong JB, Bräunl T (2025) Embodied AI in Mobile Robot Simulation with EyeSim: Coverage Path Planning with Large Language Models. In *Proceedings of the 15th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH 2025)* (pp. 185–192). SCITEPRESS.
<https://doi.org/10.5220/0013455100003970>

Kumar PP, Kant K (2025) TU-DAT: A Computer Vision Dataset on Road Traffic Anomalies. *Sensors*, 25(11), 3259. <https://doi.org/10.3390/s25113259>

Li R, Dong Y, Hu T, Liang A, Liu Y, Lu DR, Pan L, Kong L, Liang J, Liu Z (2025a) 3EED: Ground Everything Everywhere in 3D. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)* (pp. 172663–172704). <https://doi.org/10.52202/085713-5198>

Li T, Huai T, Li Z, Gao Y, Li H, Zheng X (2025b) SkyVLN: Vision-and-Language Navigation and NMPC Control for UAVs in Urban Environments. *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 17199-17206.
<https://doi.org/10.1109/iros60139.2025.11247609>

Lin F, Tian Y, Zhang T, Huang J, Guan S, Wang F-Y (2025a) AirVista-II: An Agentic System for Embodied UAVs Toward Dynamic Scene Semantic Understanding. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 6319–6324). <https://doi.org/10.1109/smc58881.2025.11342598>

Lin F, Zhang T, Ni Q, Huang J, Ma S, Tian Y, Lv Y, Wu N (2025b) Talk Less, Fly Lighter: Autonomous Semantic Compression for UAV Swarm Communication via LLMs. *2025 21st*

- IEEE International Conference on Mechatronic and Embedded Systems and Applications (MESA), 29-34. <https://doi.org/10.1109/mesa68091.2025.11278876>
- Lisondra M, Benhabib B, Nejat G (2026) Embodied AI with Foundation Models for Mobile Service Robots: A Systematic Review. *Robotics*, 15(3), 55. <https://doi.org/10.3390/robotics15030055>
- Liu S, Zhang H, Qi Y, Wang P, Zhang Y, Wu Q (2023) AerialVLN: Vision-and-Language Navigation for UAVs. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 15338–15348). <https://doi.org/10.1109/iccv51070.2023.01411>
- Liu J, Feng C, Jin G, Fan J (2025a) EdLLM: A Novel Entity Detection Method for Test Data Integrating YOLO-World and LLM. In *Advanced Intelligent Computing Technology and Applications (Lecture Notes in Computer Science*, pp. 195–206). Springer. https://doi.org/10.1007/978-981-96-9815-8_17
- Liu K, Zhang S, Wang Y, Zhang H, Zhang W (2025b) RADAR: Agent-guided retrieval with physics-based consistency filtering for aviation fault diagnosis. In 2025 7th International Conference on Frontier Technologies of Information and Computer (ICFTIC) (pp. 1152-1159). IEEE. <https://doi.org/10.1109/ICFTIC68075.2025.11324487>
- Liu Y, Ma Z, Guo J, Sun H, Niu Y, Zhang H, Wang M (2026a) A Visual Target Navigation Method for Quadrotor Based on Large Language Model in Unknown Environment. *Drones*, 10(2), 134. <https://doi.org/10.3390/drones10020134>
- Liu X, Liu Y, Qiu H, Qirong Y, Lian Z (2026b) IndoorUAV: Benchmarking Vision-Language UAV Navigation in Continuous Indoor Environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28), 23864–23872. <https://doi.org/10.1609/aaai.v40i28.39562>
- Long L, He Y, Ye W, Pan Y, Lin Y, Li H, Zhao J, Li W (2026) Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *Proceedings of the International Conference on Learning Representations (ICLR 2026)*. <https://mlanthology.org/iclr/2026/long2026iclr-seeing/>

- Lovaco J (2025) The Large Language Model Pilot: Generative Artificial Intelligence for Cyber-Physical Flight Simulations. In FT2025: Proceedings of the 12th Swedish Aerospace Technology Congress (Linköping Electronic Conference Proceedings, No. 215). <https://doi.org/10.3384/ecp215.1196>
- Luan Z, Lai Y, Huang R, Bai S, Zhang Y, Zhang H, Wang Q (2024) Enhancing Robot Task Planning and Execution through Multi-Layer Large Language Models. *Sensors*, 24(5), 1687. <https://doi.org/10.3390/s24051687>
- Luo S (2025) Sensing User Intent: An LLM-Powered Agent for On-the-Fly Personalized Virtual Space Construction from UAV Sensor Data. *Sensors*, 25(24), 7610. <https://doi.org/10.3390/s25247610>
- Luo X, Liu D, Dang F, Luo H (2024) Integration of LLMs and the Physical World: Research and Application. In ACM Turing Award Celebration Conference 2024 (pp. 1-5). ACM. <https://doi.org/10.1145/3674399.3674402>
- Ma Y, Wang R, Liang S, Luo G, Qin K (2025) TopoLLM: LLM-driven adaptive tool learning for real-time emergency network topology planning. *Digital Communications and Networks*, 12(2), 273–282. <https://doi.org/10.1016/j.dcan.2025.10.002>
- Macdonald JP, Mallick R, Wollaber AB, Peña JD, McNeese N, Siu HOC (2024) Language, Camera, Autonomy! Prompt-engineered Robot Control for Rapidly Evolving Deployment. In Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (pp. 717-721). ACM. <https://doi.org/10.1145/3610978.3640671>
- Mahmud KR, Wang L, Hassan S, Zhang Z (2025) A knowledge-driven framework for Robotic Odor Source Localization using large language models. *Robotics and Autonomous Systems*, 186, 104915. <https://doi.org/10.1016/j.robot.2025.104915>
- Majumdar A, Ajay A, Zhang X, Putta P, Yenamandra S, Henaff M, Silwal S, Mcvay P, Maksymets O, Arnaud S, Yadav K, Li Q, Newman B, Sharma M, Berges V, Zhang S, Agrawal P, Bisk Y, Batra D et al (2024) OpenEQA: Embodied Question Answering in the

- Era of Foundation Models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16488-16498. <https://doi.org/10.1109/cvpr52733.2024.01560>
- Maletić M, Peti M, Petrović T, Bogdan S (2025) Spatial-Semantic Reasoning using Large Language Models for Efficient UAV Search Operations. 2025 European Conference on Mobile Robots (ECMR), 1-8. <https://doi.org/10.1109/ecmr65884.2025.11163229>
- Masri S, Ashqar HI, Elhenawy M (2025a) Large Language Models (LLMs) as Traffic Control Systems at Urban Intersections: A New Paradigm. *Vehicles*, 7(1), 11. <https://doi.org/10.3390/vehicles7010011>
- Masri S, Ashqar HI, Elhenawy M (2025b) Leveraging Bird Eye View Video and Multimodal Large Language Models for Real-Time Intersection Control and Reasoning. *Safety*, 11(2), 40. <https://doi.org/10.3390/safety11020040>
- Mekrache A, Ksentini A (2024) LLM-enabled Intent-driven Service Configuration for Next Generation Networks. 2024 IEEE 10th International Conference on Network Softwarization (NetSoft), 253-257. <https://doi.org/10.1109/netsoft60951.2024.10588881>
- Mower CE, Wan Y, Yu H, Grosnit A, Gonzalez-Billandon J, Zimmer M, Liu P, Palenicek D, Tateo D, Peters J, Qu K, Zhang M, Lan G, Cramariuc A, Cadena C, Hutter M, Tian G, Zhuang Y, Shao K et al (2026) A robot operating system framework for using large language models in embodied AI. *Nature Machine Intelligence*, 8(3), 313-325. <https://doi.org/10.1038/s42256-026-01186-z>
- Mukhametzianov R, Nambo H (2025) History-Aware Multimodal Instruction-Oriented Policies for Navigation Tasks. *AI*, 6(4), 75. <https://doi.org/10.3390/ai6040075>
- Munasinghe I, Perera A, Deo RC (2024) A Comprehensive Review of UAV-UGV Collaboration: Advancements and Challenges. *Journal of Sensor and Actuator Networks*, 13(6), 81. <https://doi.org/10.3390/jsan13060081>

- Narchetty S, Caetano A, Sra M (2025) Towards AI Drone Assistants in Extended Reality. 2025 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), 885-886. <https://doi.org/10.1109/ismar-adjunct68609.2025.00239>
- Niforatos AT (2025) Offline large language model for drone control and monitoring (U.S. Patent Application No. US20250378284A1). <https://patents.google.com/patent/US20250378284A1/en>
- Nunes D, Amorim R, Ribeiro P, Coelho A, Campos R (2025) A Framework Leveraging Large Language Models for Autonomous UAV Control in Flying Networks. 2025 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), 12-17. <https://doi.org/10.1109/meditcom64437.2025.11104434>
- Park J, Lee Y, Park J, Park S, Woo J, Lim M (2026) EXAONE-VLA: a Unified Vision-Language Framework for mobile manipulation via semantic Topology and Hierarchical LLM reasoning. *Applied Sciences*, 16(5), 2600. <https://doi.org/10.3390/app16052600>
- Petersen C, Heins C, Albarracin M, Pitlya R, Verbelen T, Tschantz A, Constant A, Buckley CL, Kiefer AB, Friston K, Swanson S, Rene G, Salvatori T (2025) Method and system for specifying an active inference based agent using natural language (U.S. Patent No. 12,393,581 B2). <https://patents.google.com/patent/US12393581B2/en>
- Picano B, Hoang DT, Nguyen DN (2025) A Matching Game for LLM Layer Deployment in Heterogeneous Edge Networks. *IEEE Open Journal of the Communications Society*, 6, 3795–3805. <https://doi.org/10.1109/ojcoms.2025.3561605>
- Raiyn J (2026) A generative AI-driven intelligent rules framework for traffic accident detection and analysis. *Transportation Research Interdisciplinary Perspectives*, 36, 101916. <https://doi.org/10.1016/j.trip.2026.101916>
- Raptis EK, Kapoutsis AC, Kosmatopoulos EB (2025) Agentic LLM-based robotic systems for real-world applications: A review on their agenticness and ethics. *Frontiers in Robotics and AI*, 12, Article 1605405. <https://doi.org/10.3389/frobt.2025.1605405>

- Rashid MT, Siddique MJ, Shaqur A (2025) MAVREN: A multilayered adaptive framework for deploying vision-language models on resource-constrained unmanned aerial vehicles for autonomous search and rescue. *Engineering Applications of Artificial Intelligence*, 162, 112498. <https://doi.org/10.1016/j.engappai.2025.112498>
- Rashidieranjbar F, Farhadi A, Zamanifar A, Sorouri M (2025) A systematic literature review: Generative artificial intelligence applications for ground mobile robot navigation. *Computers & Electrical Engineering*, 130, 110906. <https://doi.org/10.1016/j.compeleceng.2025.110906>
- Sapkota R, Roumeliotis KI, Karkee M (2025) UAVs meet agentic AI: A multidomain survey of autonomous aerial intelligence and agentic UAVs. *arXiv*. <https://arxiv.org/abs/2506.08045>
- Sezgin A (2025a) Natural Language Interfaces for Structured Query Generation in IoD Platforms. *Drones*, 9(6), 444. <https://doi.org/10.3390/drones9060444>
- Sezgin A (2025b) Scenario-Driven Evaluation of Autonomous Agents: Integrating Large Language Model for UAV Mission Reliability. *Drones*, 9(3), 213. <https://doi.org/10.3390/drones9030213>
- Sezgin A, Boyacı A (2025) LLM-Powered UAVs: A RAG-Based Approach for Safety-Critical Operations. In *Intelligent and Fuzzy Systems (Lecture Notes in Networks and Systems*, pp. 577–584). Springer. https://doi.org/10.1007/978-3-031-97992-7_64
- Shah D, Sridhar A, Dashora N, Stachowicz K, Black K, Hirose N, Levine S (2023) ViNT: A foundation model for visual navigation. In J. Tan, M. Toussaint, & K. Darvish (Eds.), *Proceedings of the 7th Conference on Robot Learning (Vol. 229, pp. 711-733)*. *Proceedings of Machine Learning Research*. <https://proceedings.mlr.press/v229/shah23a.html>
- Sharma AK, Ravikiran M, Chakrabarty S, Saluja R, Limaye M, Kolhe HN, Samiron (2025) Prompted to Fly: Translating Free-Form Instructions into Schema-Constrained Mission Generation for UAVs Using LLMs. In *2025 IEEE International Conference on Big Data (BigData)* (pp. 7278–7286). <https://doi.org/10.1109/bigdata66926.2025.11401432>

- Shetye S, AbdelBaky M, Nolasco SG, Prakash KS, Liddell D, Kotchavong P, Schuet S, Jones A, Mujeci B, Alexander O, Ishihara J (2026) AMPLIFY: A Modular LLM Platform for Mission-Critical Retrieval and Reasoning (Technical report). NASA Ames Research Center. <https://doi.org/10.64631/syeg6685>
- Skaug L, Nojournian M (2025) A Multimodal Artificial Intelligence Framework for Intelligent Geospatial Data Validation and Correction. *Inventions*, 10(4), 59. <https://doi.org/10.3390/inventions10040059>
- Song CH, Sadler BM, Wu J, Chao W-L, Washington C, Su Y (2023) LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2986-2997. <https://doi.org/10.1109/iccv51070.2023.00280>
- Su Q, Li X, Ren Y, Fu B, Hu C, Yin Y (2025) Thinking Like an Expert: Aligning LLM Thought Processes for Automated Safety Modeling of High-Altitude Solar Drones. *Drones*, 9(11), 780. <https://doi.org/10.3390/drones9110780>
- Sun G, Xie W, Niyato D, Du H, Kang J, Wu J, Sun S, Zhang P (2024) Generative AI for Advanced UAV Networking. *IEEE Network*, 39(4), 244–253. <https://doi.org/10.1109/mnet.2024.3494862>
- Tabrizian A, Ghazanfari M, Wei P (2025) Chain-of-thought flight planner: End-to-end LLM routing under wind hazards. In AIAA AVIATION Forum and ASCEND 2025 (AIAA Paper No. 2025-3711). American Institute of Aeronautics and Astronautics. <https://doi.org/10.2514/6.2025-3711>
- Tian Y, Lin F, Li Y, Zhang T, Zhang Q, Fu X, Wang FY (2025) UAVs meet LLMs: Overviews and perspectives towards agentic low-altitude mobility. *Information Fusion*, 122, Article 103158. <https://doi.org/10.1016/j.inffus.2025.103158>

- Trabelsi Z, Ali M, Qayyum T, Tariq A (2026) Dynamic task offloading in vehicular networks using large language models for adaptive low latency decision making. *Scientific Reports*, 16(1), 9144. <https://doi.org/10.1038/s41598-026-39791-y>
- Waga A, Benhlila S, Bekri A, Abdouni J, Saber FZ (2025) A survey on autonomous navigation for mobile robots: From traditional techniques to deep learning and large language models. *Journal of King Saud University – Computer and Information Sciences*, 37(7), 198. <https://doi.org/10.1007/s44443-025-00216-x>
- Wang R, Deng Z (2025) Outdoor Multi-UAV Collaborative Task Planning System Based on Multi-Agent and Improved PPO Algorithm. In *Proceedings of the 2025 5th International Conference on Automation Control, Algorithm and Intelligent Bionics* (pp. 193-200). ACM. <https://doi.org/10.1145/3760269.3760299>
- Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao WX, Wei Z, Wen J-R (2024) A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), Article 186345. <https://doi.org/10.1007/s11704-024-40231-1>
- Wang W, Li Y, Jiao L, Yuan J (2025a) GSCE: A Prompt Framework with Enhanced Reasoning for Reliable LLM-Driven Drone Control. In *2025 International Conference on Unmanned Aircraft Systems (ICUAS)* (pp. 441–448). <https://doi.org/10.1109/icuas65942.2025.11007864>
- Wang R, Li S, Zhang P, Huang D, Guo Y, Mi H (2025b) PuppetLine: An Interactive System for Embodied Storytelling with LLM-driven Swarm Robots. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (pp. 1-3). ACM. <https://doi.org/10.1145/3746058.3759016>
- Wang X, Yang D, Wang Z, Kwan H, Chen J, Wu W, Li H, Liao Y, Liu S (2025c) Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In *Proceedings of the International Conference on Learning Representations (ICLR 2025)*. <https://openreview.net/forum?id=rUvCIvI4eB>

- Wang Z, Chen J, Zheng X, Liao Q, Huang L, Liu S (2025d) “Hi AirStar, Guide Me to the Badminton Court”. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 13477-13479). ACM. <https://doi.org/10.1145/3746027.3754491>
- Webb C, Habibpour M, Raha MH, Tavakkoli AR, Coen J, Afghah F (2026) FIRE-VLM: A vision-language-driven reinforcement learning framework for UAV wildfire tracking in a physics-grounded fire digital twin. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW 2026) (pp. 1493-1502). https://openaccess.thecvf.com/content/WACV2026W/CV4EO/html/Webb_FIRE-VLM_A_Vision-Language-Driven_Reinforcement_Learning_Framework_for_UAV_Wildfire_Tracking_WACVW_2026_paper.html
- Wen L, Yang X, Fu D, Wang X, Cai P, Li X, Ma T, Li Y, Xu L, Shang D, Zhu Z, Sun S, Bai Y, Cai X, Dou M, Hu S, Shi B, Qiao Y (2024) On the road with GPT-4V(ision): Explorations of utilizing visual-language model as autonomous driving agent. In Proceedings of the ICLR 2024 Workshop on Large Language Model (LLM) Agents. <https://openreview.net/forum?id=2UBexKm8TE>
- Xiao J, Sun Y, Shao Y, Gan B, Liu R, Wu Y, Guan W, Deng X (2025) UAV-ON: A Benchmark for Open-World Object Goal Navigation with Aerial Agents. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 13023–13029). ACM. <https://doi.org/10.1145/3746027.3758251>
- Xu P, Deng Z, Deng J, Gu Z, Wan S (2026) AeroVLA: A Vision-Language-Action Model for UAV Navigation via Minimalist End-to-End Control. In P. Favaro, Z. Kukelova, A. Maki, A. Rohrbach, K. Schindler, & F. Tombari (Eds.), Computer Vision – ECCV 2026 (Lecture Notes in Computer Science, Vol. 17057, pp. 289–305). Springer. https://doi.org/10.1007/978-3-032-37335-9_16

- Xue N, Verma D, Srihari V, Piotrowski W, Chour K, Kalyanam K (2026) Towards Naturalistic Human-Machine Teaming with LLM Agents: A Case Study in Air Traffic Management. In AIAA SCITECH 2026 Forum. <https://doi.org/10.2514/6.2026-2559>
- Yao F, Yue Y, Liu Y, Wang Z, Jin L, Zhao B, Zhao J, Sun X, Fu K (2026) AeroVerse: UAV-Agent Benchmark Suite for Simulating, Pre-training, Finetuning, and Evaluating Aerospace Embodied Foundation Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10), 12301–12318. <https://doi.org/10.1109/tpami.2026.3697634>
- Yu G, Ju R, Sugumaran V, Liu H (2025) Lightweight multimodal LLM-empowered dual-agent collaboration for reliable defect detection and maintenance recommendation in tunnel infrastructure. *Reliability Engineering & System Safety*, 268, 111973. <https://doi.org/10.1016/j.ress.2025.111973>
- Yuan C, Zhou Y, Guo C, Han D, Shi G, Wang W (2025) Seeing With Words: Interpretable Language-Guided Drone Geo-Localization via LLM-Enriched Semantic Attribute Alignment. *IEEE Transactions on Multimedia*, 28, 2132–2144. <https://doi.org/10.1109/tmm.2025.3642913>
- Zha J, Fan Y, Zhang T, Chen G, Chen Y, Gao C, Chen X (2026) AirCopBench: A Benchmark for Multi-drone Collaborative Embodied Perception and Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(2), 1507-1515. <https://doi.org/10.1609/aaai.v40i2.37126>
- Zhang W, Liu Y, Wang X, Chen X, Gao C, Chen X (2024) Demo abstract: Embodied aerial agent for city-level visual language navigation using large language model. In 2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN) (pp. 265-266). IEEE. <https://doi.org/10.1109/IPSN61024.2024.00033>
- Zhang R, Zhao C, Du H, Niyato D, Wang J, Sawadsitang S, Shen X, Kim DI (2025a) Embodied AI-Enhanced Vehicular Networks: An Integrated Vision Language Models and Reinforcement Learning Method. *IEEE Transactions on Mobile Computing*, 24(11), 11494–11510. <https://doi.org/10.1109/tmc.2025.3582864>

- Zhang L, Zhang Y, Li H, Fu H, Tang Y, Ye H, Chen L, Liang X, Hao X, Ding W (2025b) Is your VLM sky-ready? A comprehensive spatial intelligence benchmark for UAV navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025).
https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Is_Your_VLM_Sky-Ready_A_Comprehensive_Spatial_Intelligence_Benchmark_for_UAV_Navigation_CVPR_2025_paper.html
- Zhang W, Zhou Z, Zeng X, Liu X, Fang J, Gao C, Cui J, Li Y, Chen X, Zhang X-P (2025c) Open3D-VQA: A Benchmark for Embodied Spatial Concept Reasoning with Multimodal Large Language Model in Open Space. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 12784-12791). ACM.
<https://doi.org/10.1145/3746027.3758219>
- Zhang J, Li A, Qi Y, Li M, Liu J, Wang S, Liu H, Zhou G, Wu Y, Li X, Fan Y, Li W, Chen Z, Gao F, Wu Q, Zhang Z, Wang H (2026a) Embodied navigation foundation model. In Proceedings of the International Conference on Learning Representations (ICLR 2026).
<https://mlanthology.org/iclr/2026/zhang2026iclr-embodied/>
- Zhang D, Zhang X, Zhang J, Li J, Long T, Dai X, Fu Y, Ren J, Zhang Y (2026b) SRDrone: LLM-Driven Self-Refinement for Embodied Drone Task Planning. IEEE Transactions on Mobile Computing, 25(7), 10757–10772. <https://doi.org/10.1109/tmc.2026.3664806>
- Zhao B, Tang R, Jia M, Wang Z, Man F, Zhang X, Shang Y, Zhang W, Wu W, Gao C, Chen X, Li Y (2025a) AirScape: An Aerial Generative World Model with Motion Controllability. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 12519-12528). ACM. <https://doi.org/10.1145/3746027.3758180>
- Zhao Y, Xu K, Zhu Z, Hu Y, Zheng Z, Chen Y, Ji Y, Gao C, Li Y, Huang J (2025b) CityEQA: A Hierarchical LLM Agent on Embodied Question Answering Benchmark in City Space. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 12476-12491. <https://doi.org/10.18653/v1/2025.emnlp-main.630>

- Zhao M, Xia J, Hou K, Liu Y, Xia S, Jiang X (2025c) FlexiFly: Interfacing the Physical World with Foundation Models Empowered by Reconfigurable Drone Systems. In Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems (pp. 463-476). ACM. <https://doi.org/10.1145/3715014.3722081>
- Zhao B, Fang J, Dai Z, Wang Z, Zha J, Zhang W, Gao C, Wang Y, Cui J, Chen X, Li Y (2025d) UrbanVideo-Bench: Benchmarking Vision-Language Models on Embodied Intelligence with Video Data in Urban Spaces. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 32400–32423). <https://doi.org/10.18653/v1/2025.acl-long.1558>
- Zhao M, Hou K, Xia J, Liu Y, Xia S, Jiang X (2026) EmbodiedFly: Embodied LLM Agent with an Autonomous Reconfigurable Drone. ACM Transactions on Internet of Things, 7(2), 1-32. <https://doi.org/10.1145/3772079>
- Zhou Q, Chen S, Wang Y, Xu H, Du W, Zhang H, Du Y, Tenenbaum JB, Gan C (2024) HAZARD challenge: Embodied decision making in dynamically changing environments. In Proceedings of the International Conference on Learning Representations (ICLR 2024). <https://openreview.net/forum?id=n6mLhaBahJ>
- Zhou Y, Ding R, Yang X, Jiang X, Liu X (2025) AirSpatialBot: A Spatially Aware Aerial Agent for Fine-Grained Vehicle Attribute Recognition and Retrieval. IEEE Transactions on Geoscience and Remote Sensing, 63, 1-12. <https://doi.org/10.1109/tgrs.2025.3570895>
- Zhu F, Huang F, Yu Y, Liu G, Huang T (2024) Task Offloading with LLM-Enhanced Multi-Agent Reinforcement Learning in UAV-Assisted Edge Computing. Sensors, 25(1), 175. <https://doi.org/10.3390/s25010175>
- Zhu R, Li S, Zhu Z, Jia J, Yang M (2026) CA-VLN: Collaborative Agents in MLLM-Powered Visual-Language Navigation. Sensors, 26(4), 1254. <https://doi.org/10.3390/s26041254>
- Zu W, Song W, Chen R, Guo Z, Sun F, Tian Z, Pan W, Wang J (2024) Language and Sketching: An LLM-driven Interactive Multimodal Multitask Robot Navigation Framework. 2024 IEEE

International Conference on Robotics and Automation (ICRA), 1019-1025.

<https://doi.org/10.1109/icra57147.2024.10611462>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [OnlineResource1.xlsx](#)
- [OnlineResource2PRISMA2020.docx](#)