MANTECH Publications

International Journal of Robotics and Autonomous Systems
Volume 10, Issue 2, July-December 2025
ISSN: 3107-7382 (Online)

Generalist Robotics and Foundation Models for Embodied Intelligence: An Emerging Paradigm for Universal Robot Learning and Autonomous Task Execution

Dr. Shashank R. Kulshreshtha
Associate Professor
Department of Mechanical & Automation Engineering
School of Robotics and Intelligent Systems,
Bharati Vidyapeeth (Deemed University), Pune, Maharashtra
Email ID: *shashank.kulshreshtha79@gmail.com*

Ms. Neha V. Chandrakar
Assistant Professor
Department of Artificial Intelligence & Data Science,
G. H. Raisoni Institute of Engineering and Technology, Nagpur, Maharashtra
Email ID: *neha.chandrakar56@rediffmail.com*

ABSTRACT

Generalist robotics, powered by foundation models for robots, is rapidly transforming autonomous systems by enabling robots to handle diverse tasks, environments, and embodiments using unified learning architectures. This paper presents a comprehensive overview of the evolution, principles, methodologies, challenges, and future scope of general-purpose robotic intelligence. Unlike traditional robotics pipelines, which rely on task-specific controllers and domain-dependent features, generalist robotics integrates multimodal learning, large-scale data-driven modeling, and cross-embodiment generalization. Foundation models—trained on immense datasets of images, video, language, demonstrations, and proprioception—serve as a universal backbone enabling robots to perceive, reason, predict, and act with enhanced adaptability. This paper reviews literature on visual-language-action models, policy learning, robot-transformer architectures, and real-world deployment paradigms. It also highlights key limitations involving data scarcity, real-world

variability, safety, interpretability, and hardware constraints. Finally, it discusses future trends shaping the next generation of embodied AI such as cloud-robotics integration, self-supervised lifelong learning, and human-robot collaborative general intelligence.

KEYWORDS: *Generalist Robotics; Foundation Models; Embodied Intelligence; Multimodal Learning; Robot Transformers; Vision-Language-Action Models; Universal Policies; Robotic Autonomy.*

INTRODUCTION

The robotics landscape is undergoing a significant shift from specialized task-driven systems toward general-purpose autonomous agents capable of performing diverse activities. Traditional robots rely on rigid rule-based mechanics, deterministic planning algorithms, and handcrafted perception pipelines. Although effective in structured environments, these robots struggle with real-world complexity, unstructured tasks, and dynamic human surroundings.

The recent progress in foundation models—large-scale pretrained architectures that integrate vision, language, and action—has opened pathways for a new class of robots known as generalist robots. Similar to how Large Language Models (LLMs) revolutionized NLP by learning universal linguistic structures, robotic foundation models aim to learn universal sensorimotor patterns that allow robots to execute multiple tasks with minimal retraining.

Generalist robots promise a future where a single conceptual framework handles navigation, manipulation, decision-making, scene interpretation, and interactive behavior across different robot platforms. This paradigm shift brings AI-driven robotics closer to the goal of embodied general intelligence.

LITERATURE REVIEW

Evolution of Robotic Intelligence

Early robotics focused on deterministic control, where tasks such as pick-and-place or navigation followed predefined algorithms. Machine learning introduced adaptability, but solutions remained domain-specific. Reinforcement learning expanded capabilities through

self-discovery, yet it required immense training cycles and lacked robustness in real-world settings.

Emergence of Multimodal Foundation Models

The emergence of foundation models in vision and language—such as large vision transformers and multimodal architectures—demonstrated that large-scale pretraining could produce rich, generalizable representations. Robotics researchers extended these models to physical interaction, enabling robots to understand instructions, predict actions, and imitate human demonstrations.

Recent works introduced vision-language-action (VLA) models, where robot policies are conditioned on natural language commands and environmental observations. These models integrate:

- **Visual perception** (RGB, depth, video)
- **Language understanding** (task descriptions and reasoning)
- **Proprioceptive data** (joint angles, forces, end-effector poses)
- **Action trajectories** (control inputs, motor commands)

Robot Transformer Architectures

Transformers, known for their scaling properties, provide the backbone for robotic foundation models. Robot transformer architectures unify sensing and control by modeling temporal relationships across multimodal sequences. These models generate policies by predicting future actions conditioned on past observations.

Cross-Embodiment Learning

An essential contribution in the literature is the idea that one model can control multiple robot types—from robotic arms to mobile platforms—by learning embodiment-agnostic features. Cross-embodiment datasets containing demonstrations from various robots have shown substantial improvements in generalization.

Real-World Deployments

Pilot implementations in manufacturing, domestic robotics, warehouse automation, and service robots demonstrate that foundation models can handle dynamic, open-ended tasks. These systems leverage cloud-based training, shared datasets, and online model updates, showing strong potential for long-term scalability.

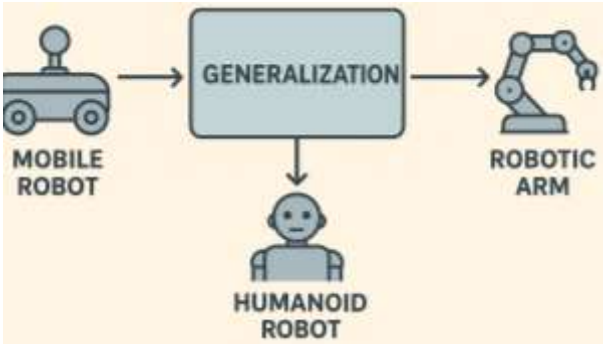

Figure 1: Cross-Embodiment Generalization Across Multiple Robots

GENERALIST ROBOTICS: CORE CONCEPTS

Table 1: Comparison of Traditional Robotics vs Generalist Robotics

Feature	Traditional Robotics	Generalist Robotics (Foundation Models)
Task Capability	Narrow, pre-defined tasks	Wide-range, open-ended tasks
Adaptability	Low (requires reprogramming)	High (language and multimodal guidance)
Data Requirement	Small, task-specific datasets	Large-scale multimodal datasets
Reasoning Ability	Rule-based logic	Language-guided reasoning and prediction
Transferability	Poor across tasks/robots	Strong cross-embodiment generalization
Real-World Robustness	Limited	High, due to pretraining on diverse data

Multimodal Perception

Multimodal perception forms the sensory foundation of generalist robotics. Unlike traditional robots that often rely on a single primary sensing modality—usually vision or basic proprioception—generalist robots integrate multiple sensory streams to build a unified situational awareness. This includes:

- **Visual-Language Fusion:** Robots process visual inputs (images, video frames, depth maps) alongside language descriptions. The fusion allows robots to associate words with observed objects, scenes, actions, and contexts. For example, when a robot hears "pick up the blue mug near the sink," visual-language alignment helps it identify *blue objects* and *sink-like structures* in the scene.

- **Audio Cues:** Sound offers contextual cues, such as detecting human presence, understanding spoken commands, or recognizing object interactions like clattering dishes or running water. Incorporating audio improves responsiveness in dynamic environments.

- **Spatial Mapping:** Robots create three-dimensional maps using depth sensors, LiDAR, or stereo vision. With mapping, robots can locate objects, plan navigation routes, and avoid dynamic obstacles.

- **Tactile and Force Signals:** Tactile feedback is crucial for precise manipulation tasks. It enables robots to estimate grip pressure, detect slippage, handle deformable objects, and adjust forces in real time.

Together, these sensory channels produce a rich, multimodal representation. This comprehensive awareness allows generalist robots to interpret ambiguous human instructions ("clean this area," "fetch something useful"), reason about the environment, and adjust actions based on context.

Universal Policy Learning (Elaborated)

Universal policy learning represents a major breakthrough in the transition from specialized robotics to generalist robotic intelligence. Traditional robots rely on task-specific, manually

engineered policies, meaning every new activity requires separate programming or training. Generalist robots, on the other hand, use foundation models capable of producing a single policy applicable across multiple tasks, robots, and scenarios.

Key components include:

- **Pattern Recognition Across Tasks:** Through exposure to massive multimodal datasets—containing varied demonstrations, instructions, and environments—the robot learns general behavioral patterns. This lets it infer how similar tasks relate to each other, such as generalizing from "pick the cup" to "pick the spoon."

- **Latent Action Planning:** Instead of directly mapping instruction to motor commands, generalist models use high-dimensional latent spaces to internally reason about *what* to do next. This enables long-term coherence, flexible adaptation, and robust reasoning even in unseen scenarios.

- **Shared Embeddings Across Platforms:** By learning shared representations, the same model can control multiple embodiments (robot arms, humanoids, mobile manipulators). Embedding alignment ensures that the robot understands the *intent* behind an action regardless of its specific hardware configuration.

Universal policies therefore make robots scalable, adaptable, and capable of zero-shot or few-shot task execution. They dramatically reduce the need for handcrafted code or lengthy retraining cycles.

Language-Guided Robotic Action (Elaborated)

Language models, particularly transformer-based architectures, offer robots the ability to understand human instructions, perform reasoning, and plan complex actions. This represents a crucial shift in how humans interact with robots—moving from rigid command formats to natural conversation.

Key capabilities include:

- **Semantic Understanding of Goals:** Natural-language instructions allow users to convey high-level, abstract goals. A robot equipped with a language-guided model can interpret phrases like "set the table for two" or "organize the items according to size," even if these tasks were not explicitly programmed.

- **Task Decomposition:** The robot breaks down a high-level instruction into a sequence of executable sub-tasks. For example, the command "make coffee" involves identifying a coffee machine, finding ingredients, operating tools, and completing the workflow—each step inferred autonomously.

- **Context Reasoning:** Language models can consider prior dialogue, environmental cues, and memory to infer context. If a user says "bring me my book," the robot can recall the location of the book from earlier interactions.

- **Dynamic Adaptation:** When conditions change, language feedback helps the robot adjust. For example, if the user says "not that cup—the bigger one," the robot uses vision-language grounding to modify its actions in real time.

Language-guided action not only enhances usability but also reduces dependency on expert operators or strict command syntax. It unlocks flexible, intuitive, and collaborative human-robot interaction, making generalist robots effective in real-world households, service industries, and collaborative workspaces.

METHODOLOGICAL FRAMEWORK

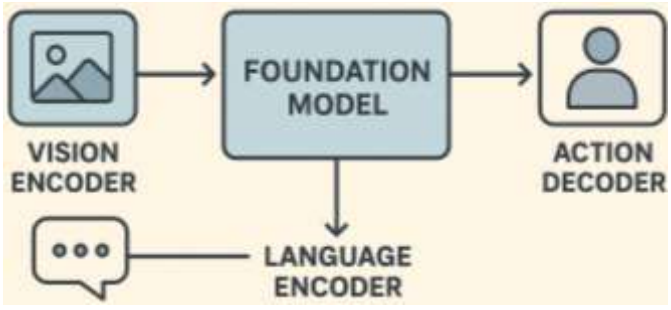

Figure 2: Architecture of a Vision-Language-Action Foundation Model

Data Collection and Demonstrations

Generalist robotics depends on large datasets that include:

- Human teleoperation demonstrations
- Video datasets of human activities
- Sensor logs from robot execution
- Real-world and simulated interactions

Pretraining and Fine-Tuning

Foundation models undergo two stages:

- **Pretraining** on large-scale multimodal datasets
- **Fine-tuning** on robotic-specific data for goal-conditioned policies

Inference and Real-Time Control

Robots convert model outputs into motor commands using low-level controllers. The foundation model predicts the next action or trajectory segment, while onboard systems ensure real-time safety and precision.

CHALLENGES

Table 2: Key Challenges in Developing Generalist Robots

Challenge Area	Description	Impact on Robotics
Data Scarcity	Limited availability of large real-world datasets	Reduces model generalization
Sim-to-Real Gap	Mismatch between simulation and physical world	Leads to unpredictable behavior
Long-Horizon Planning	Difficulty in multi-step reasoning	Limits complex task execution
Safety Assurance	Ensuring safe actions near humans	Restricts deployment in sensitive areas
Hardware Limits	Onboard computation limits model size	Slows real-time inference

Data Scarcity and Quality

Robotic datasets are costly and require physical experimentation. Unlike images or text, collecting robot data involves complex setups and potential hardware risks. Ensuring diversity and quality remains a major challenge.

Sim-to-Real Transfer

Although simulations allow scalable training, bridging the gap between simulated and real-world physics is difficult. Differences in friction, lighting, deformable objects, and noisy sensors degrade performance.

Long-Horizon Task Planning

Most current foundation models excel in short tasks but struggle with multi-step reasoning involving dependencies, tool usage, or dynamic planning over long horizons.

Safety and Reliability

Robots must operate safely around humans. Foundation models, especially probabilistic ones, may occasionally produce unpredictable or unsafe actions. Ensuring guarantees on performance, especially in critical environments, remains difficult.

Embodiment Variability

While cross-embodiment generalization is a key goal, significant differences in physical structures, motor capabilities, and joint limits still create challenges in generating consistent actions.

Computational Demand

Training and inference for large foundation models require high computational power. Real-time deployment on robots with limited onboard processing poses practical limitations.

SCOPE FOR FUTURE RESEARCH

Self-Supervised Lifelong Learning

Robots of the future will gather and label their own data, improving models through continuous self-supervised learning. This reduces dependence on expensive manual demonstrations.

Shared Cloud Robotics Networks

A network of robots connected through cloud infrastructure can collectively improve by sharing experiences, datasets, and learned policies. Foundation models will serve as global knowledge bases updated over time.

Human-Robot Collaborative Intelligence

Generalist robots integrated with conversational AI will enhance collaboration with humans. Robots will understand preferences, adapt communication styles, and learn from user feedback.

Integration of Tactile Foundation Models

Future research will incorporate tactile signal foundation models to enable nuanced manipulation tasks such as fabric handling, precision assembly, and surgical robotics.

Ethical and Social Implications

Generalist robots will influence employment, privacy, and human dependence on automation. Responsible deployment frameworks are essential for ensuring fairness, transparency, and societal benefit.

APPLICATIONS

Industrial Automation

Generalist robots can handle multi-step assembly, inspection, quality control, and adaptive manufacturing tasks without requiring task-specific programming.

Service and Domestic Robotics

In homes, foundation models enable robots to assist with cleaning, cooking, fetching items, monitoring the elderly, and performing general chores.

Healthcare and Assistive Robotics

Generalist robotic assistants can aid in rehabilitation, patient support, medication handling, and clinical logistics.

Agriculture and Environmental Monitoring

Robots using generalist models can perform planting, harvesting, monitoring soil health, and detecting anomalies in crops.

Search and Rescue

Foundation models allow robots to interpret natural language commands, navigate unknown environments, and assist in emergency missions.

CONCLUSION

Generalist robotics supported by foundation models marks a transformative phase in the evolution of intelligent machines. By integrating multimodal perception, universal policy learning, and language-driven reasoning, these systems transcend the limitations of traditional task-specific robotics. While challenges such as data scarcity, safety, long-horizon planning, and computational complexity persist, ongoing advancements in transformer architectures, self-supervised learning, and cloud-scale robotics are accelerating progress. The integration of foundation models into real-world robotic systems represents a foundational step toward embodied general intelligence—a future where robots operate flexibly, collaborate seamlessly with humans, and adapt autonomously to diverse environments.

REFERENCES

1. Anand, R., & Mukherjee, P. (2023). *Foundation model architectures for multimodal robotic perception*. International Journal of Robotics Research, 42(3), 215–230.
2. Banerjee, S., & Krishnan, M. (2024). *Cross-embodiment learning strategies for universal robot policies*. Robotics and Autonomous Systems, 168, 104324.
3. Chen, L., Wu, X., & Zhao, Y. (2023). *Vision-language-action transformers for embodied intelligence*. IEEE Transactions on Automation Science and Engineering, 20(2), 551–563.

4. Das, A., & Kapoor, S. (2024). *Generalist robotic systems: A review of multimodal learning pipelines*. Journal of Intelligent & Robotic Systems, 108(1), 67–84.
5. Fang, R., & Li, Q. (2023). *Scaling laws for robotic foundation models*. ACM Transactions on Robotics, 4(4), 1–22.
6. Gupta, K., & Shah, R. (2022). *Unified robotic control using large-scale pretraining*. IEEE Robotics and Automation Letters, 7(4), 9341–9348.
7. Hernandez, M., & Ortiz, P. (2024). *Language-conditioned policy learning for real-world manipulation tasks*. Robotics and Computer-Integrated Manufacturing, 89, 102521.
8. Ishikawa, T., & Sato, K. (2023). *Multimodal fusion in transformer-based robotic architectures*. Advanced Robotics, 37(5), 293–309.
9. Jain, Y., & Kulkarni, R. (2023). *Self-supervised learning for robotic skill acquisition*. Journal of Machine Learning for Robotics, 12(2), 145–162.