Crimson Publishers
Wings to the Research

Mini Review

The Future of Generative AI in Robotics

John Atkinson*

AI-Empowered, Santiago, Chile

ISSN: 2832-4463

***Corresponding author:** John Atkinson, AI-Empowered, Santiago, Chile

Submission: May 30, 2025
Published: July 17, 2025

Volume 4- Issue 5

How to cite this article: John Atkinson*. The Future of Generative AI in Robotics. COJ Rob Artificial Intel. 4(5). COJRA. 000596. 2025.
DOI: 10.31031/COJRA.2025.04.000596

Abstract

Generative Artificial Intelligence (GenAI) is fundamentally reshaping robotics, moving the field beyond rigid, pre-programmed systems toward flexible, adaptive, and creative machines. Traditional robotics has long relied on precise control systems, detailed planning, and narrow task definitions, but GenAI through technologies such as large vision-language models, diffusion models, and imitation learning enables robots to learn from demonstrations, natural language, and online data. These advances are further amplified by collaborative efforts like Open X-Embodiment, which pool data from diverse robots to build scalable, generalist AI models. Despite these breakthroughs, significant challenges remain before robots can be fully integrated into everyday life. Issues such as safety, interpretability, data efficiency, and real-time performance continue to limit deployment in high- stakes or consumer-facing contexts. Moreover, robots still lack the general- purpose commonsense needed for complex, multi-step tasks in unstructured environments. Nonetheless, the future of robotics is being rapidly transformed by GenAI, with promising directions including open-ended skill acquisition, personalized user interactions, and integration with emerging technologies. Accordingly, this review discusses recent research, challenges and applications of GenAI and robotics and its impact in real-life applications.

Keywords: Generative artificial intelligence; Intelligent robotics; Large vision language models; LLM-driven robotics

Introduction

Generative Artificial Intelligence (GenAI) has made significant strides in recent years, particularly in natural language processing, image synthesis, and multimodal learning. Its integration with robotics an area traditionally dominated by deterministic control systems, perception algorithms, and classical planning signals a paradigm shift toward more adaptable, data-driven, and creative robotic systems. Recent advancements in GenAI are reshaping the field of robotics, opening new possibilities for learning, adaptation, and interaction. Unlike traditional robotics, which often relies on pre-programmed behaviors and narrowly defined tasks, GenAI enables robots to reason, imagine, and create solutions in dynamic environments [1]. This shift marks a critical evolution in how robots are designed and deployed, with the potential to impact industries ranging from manufacturing and healthcare to education and the creative arts. Through the integration of large vision language models, multimodal learning, and imitation from demonstrations, robots can now learn new tasks continuously from diverse sources including online content, user instructions, and their own experiences [2,3]. This approach dramatically reduces the need for manual programming and expands the range of environments in which robots can operate e effectively.

GenAI also improves human-robot interaction by making communication more natural and intuitive. Robots equipped with generative models can interpret spoken or written commands, understand context, and generate appropriate responses or actions. Moreover, the creative potential of GenAI is unlocking new roles for robots in art, design, and entertainment. To support these sophisticated functions, researchers are increasingly exploring new computational architectures, AI techniques and collected robotics datasets. They make the deployment of advanced robotic systems more practical and scalable. Together, these developments suggest a future in which GenAI is a foundational element of next-generation robotics. This review explores the current trajectory and future prospects of GenAI in robotics.

Review

Tasks that are cognitively or physically trivial for humans often pose substantial challenges for robotic systems, whereas tasks that demand sustained precision or endurance are relatively straightforward for machines to perform [4]. For instance, a robot can play chess or maintain a fixed grip on an object indefinitely with high reliability. In contrast, tasks such as tying shoelaces, intercepting a moving object, or engaging in natural language dialogue require sophisticated perceptual, motor, and cognitive integration. These challenges stem from several key limitations:

1. Imprecise motor control and coordination

2. Constrained perceptual understanding due to dependency on limited-resolution sensor data, and

3. An absence of intuitive physical reasoning, which humans typically develop through embodied experience.

Traditionally, roboticists have addressed these limitations through model- based control and explicit motion planning. This approach typically involves the use of vision systems to detect and classify objects and environments, followed by the construction of detailed predictive models to estimate the consequences of specific motor commands. Based on these models, planners generate highly deterministic action sequences, which are rigorously tested and incrementally refined in controlled laboratory settings to ensure robustness and repeatability [5]. This approach has its limits. Robots trained like this are strictly choreographed to work in one specific setting. Compared with other elds, such as computer vision, robotics has been in the dark ages. But that might not be the case for much longer, because the eld is experiencing a big shake-up. Thanks to new approaches such as GenAI, the focus is now shifting from feats of physical dexterity to building general-purpose robot brains in the form of deep neural networks. Much as the human brain is adaptable and can control different aspects of the human body, these networks can be adapted to work in different robots and different scenarios.

Recent technological trends in GenAI and robotics are being driven by the integration of advanced generative models, particularly foundation models that combine multiple modalities such as vision, language, and motor control [6]. These large-scale, general-purpose models enable robots to generalize across diverse tasks and environments, representing a major step toward true embodied intelligence. Additionally, diffusion models are emerging as powerful tools for robotic planning, capable of generating high-quality action sequences, adaptive policies, and even full simulations providing greater flexibility and robustness in decision-making [7,8].

Another key area is simulation-to-real transfer, where generative models play a critical role in narrowing the gap between virtual training and real-world deployment. By generating realistic textures, physics behaviors, and sensor noise, these models make it easier to transfer skills learned in simulation to physical robots. At the same time, researchers are working on embodied agents that incorporate memory and reasoning capabilities, aiming to create robots that can understand context, recall relevant experiences, and reason symbolically. These trends collectively point toward a future of more autonomous, intelligent, and adaptable robotic systems. Thus, instead of the traditional painstaking planning and training, deep learning and neural networks have been used to create systems that learn from their environment on the go and adjust their behavior accordingly.

At the same time, the emergence of low-cost hardware such as commercially available components and affordable robotic platforms like Stretch has significantly lowered the barrier to entry for conducting large-scale robotic experimentation. In general, current research leverages artificial intelligence and Generative AI (GenAI) to train robotic systems via two state-of-the-art techniques [9]:

A. **Reinforcement learning (RL):** it allows systems to improve through trial and error, so the robotic system can adapt its movements in new environments. It can be used learning to create a robotic system that can do extreme tasks (i.e., parkour) with minimal pre-programming. This approach is inspired by human navigation in which Humans receive information about the surrounding world from their eyes, and this helps them instinctively place one foot in front of the other to get around in an appropriate way. Thus, a robot can use a camera to look ahead. The robot was then able to memorize what was in front of it for long enough to guide its leg placement. The robot learned about the world in real time, without internal maps, and adjusted their behavior accordingly [9,10].

B. **Imitation learning:** a model learns to perform tasks by, for example, imitating the actions of a human tele-operating a robot or using a VR head- set to collect data on a robot. This technique has recently become more popular with robots that do manipulation tasks. By pairing this technique with GenAI methods such as Large Language Models (LLM), GANs (Generative Adversarial Networks), Transformers and Diffusion models, researchers have been able to quickly teach robots to do many new tasks. This may extend the technology propelling GenAI from the realm of text, images, and videos into the domain of robot movements [11,12].

A common approach begins with human teleoperation, where a human operator manually controls the robot to demonstrate target behaviors. These demonstrations serve as foundational data for training, which is subsequently leveraged by generative AI (GenAI) techniques such as diffusion models to enable the robot to learn complex skills autonomously from the provided data [5]. For instance, researchers have successfully trained robots to perform over 200 distinct tasks, including fine motor activities such as peeling vegetables and pouring liquids, with ongoing efforts aimed at scaling this capability to over 1,000 skills by year-end [13]. In parallel, industry efforts have advanced the development of multimodal robotic foundation models. A notable example is

Covariant's RFM-1, which integrates diverse input modalities text, images, video, robot command sequences, and sensor measurements to facilitate flexible task specification and execution.

GenAI models not only enhance a robot's ability to interpret complex multimodal instructions but also enable the generation of contextual visual representations (e.g., task-related images or video simulations). A recent development by Stanford researchers, ALOHA (Affordable Low-cost Open-source Hardware for teleoperation), demonstrated that a robot could learn to perform tasks such as cooking shrimp using as few as 20 human demonstrations, supplemented by data from unrelated tasks (e.g., removing a paper towel or tape) [14]. These findings indicate that GenAI enables cross-task generalization, where training on a specific task can improve performance on others through shared representational learning and transferable skill acquisition.

Recent advancements suggest that GenAI has the potential to render many conventional robotics methodologies obsolete. This evolution is timely, as the robotics field despite decades of rigorous algorithmic development and system engineering continues to face significant limitations in core areas such as perception, motion planning, reasoning, grasping, manipulation, and human robot interaction, particularly when operating in unstructured, dynamic environments characteristic of the human world [15]. Deep learning-based approaches are increasingly demonstrating competitive performance relative to traditional, model-based techniques in both control and sensorimotor processing tasks. In particular, large language models (LLMs), when trained on sufficiently diverse and large-scale datasets, exhibit a compelling capacity to generalize across a wide range of tasks and situational contexts, offering a promising new paradigm for robotic autonomy and adaptability [16,17].

However, gathering training data for robots is costly and slow. Some estimates show that to reach a similar amount of data available for Natural Language Processing (NLP), from streams of images and text produced by internet users, robotics training data needs to scale up by a factor of 27 million. A recent community effort named Open X-Embodiment has produced a dataset of 22 robots, 527 skills and 160,266 tasks, which seems a sizeable start. However, the feasibility of ever gathering sufficient data to develop a general-purpose robotics model is questionable.

The complexity of real-world human robot interactions requires exceptionally high standards of reliability and robustness. While zero-shot performance rates of 50% to 75% may be considered notable achievements under controlled laboratory conditions, such performance levels remain insufficient for safety-critical or human-facing deployment scenarios. Beyond quantitative benchmarks, concerns related to the reliability and trustworthiness of general-purpose robotic models present significant challenges. Unlike language-based systems (e.g., ChatGPT or Gemini), where occasional factual inaccuracies or hallucinations may be tolerable, physical robotic systems operating in human environments must adhere to strict safety and dependability constraints. Consequently, robotics must continue to integrate models grounded in physical reasoning and embodied understanding of the environment.

To address these challenges, researchers have begun exploring the integration of Large Vision-Language Models (LVLMs) into robotic systems [18,19]. Early research suggests that LVLMs significantly enhance capabilities in scene understanding, human robot interaction, and high-level action planning. Models such as GPT-4 and Gemini, having been trained on internet-scale multimodal data, exhibit a form of emergent commonsense knowledge that can potentially be leveraged for robotic reasoning and decision-making in open-world environments [20]. However, this commonsense representation remains fundamentally different from human-like under- standing and continues to raise questions about reliability and interpretability. Nevertheless, the semantic priors embedded within LVLMs particularly regarding everyday objects, actions, and interactions offer a promising foundation for advancing robotic perception and interaction in complex, dynamic settings.

Nonetheless, significant challenges remain in addressing the complexities associated with operating in dynamic, unstructured environments. How robots can physically interact with their environment will depend on their bodies, and a next step is highlighted in the `SayCan' project [21], in which the PaLM model is grounded in the affordances of real-world mobile robots into two primary components:

a) **LLM:** it uses language models such as GPT-4 that understands and generates natural language. This model is good at understanding contextual nuances, inferring implicit intents, and generating actionable plans based on natural language inputs (aka. prompts).

b) **Action model:** it performs semantic grounding by translating natural language commands into executable low-level robotic actions. It evaluates the operational feasibility of candidate actions, ranks them according to task-specific and environmental context, and manages their sequential execution within the robot's control architecture.

A related research direction is to develop LVLMs with an advanced, physical commonsense understanding of the world. An essential ingredient is curated data collection of examples from videos for a better understanding of physical properties of objects and physical effects in manipulating them [22]. Designing robotic systems that can safely and reliably work in the real world remains a challenging issue, but GenAI is injecting the field with fresh ideas.

Other effort such as Open X-Embodiment [23] aims at collaboratively developing generalist AI models for robots (aka. RT-X models), that can learn and adapt to various robots, tasks, and environments. It involves creating a large, open-source dataset of real robot trajectories, and providing standardized data formats and model checkpoints for research. The goal is to move beyond training separate models for each robot and task to enable robots to leverage experience from diverse sources. The initiative has been able to partner with 34 research labs and about 150 researchers to collect data from 22 different robots. The resulting dataset consists of robots demonstrating 527 skills, such as picking, pushing, and moving. The initiative sought to establish a robot internet by aggregating robotic data from laboratories worldwide, thereby enabling access to larger, more scalable, and diverse datasets for the research community. This effort parallels the deep learning breakthrough catalyzed by the introduction of ImageNet, a large-scale online image dataset that significantly advanced computer vision and laid the foundation for modern generative AI. In this context, researchers developed two implementations of a robotic model named RT-X: one designed for local deployment on individual laboratory infrastructure, and another accessible remotely via web-based interfaces, facilitating distributed experimentation and collaboration.

The larger, web-accessible model was pretrained with internet data to develop a 'visual commonsense', or a baseline understanding of the world, from LLMs and image models. When the RT-X model was ran on many different robotics platforms, robots were observed to learn skills 50% more successfully than in the systems each individual lab was developing. These large robotic dataset and GenAI which are able to analyze image and language data, might offer robots important hints as to how the surrounding world works. These models provide high-level semantic representations of the world, which can support robotic systems in tasks involving reasoning, inference, and visual understanding. In order to evaluate this capability, researchers deployed a robot pre-trained on a large multimodal model and instructed it to identify a specific person's image. Despite the absence of explicit training data containing images of the individual, the robot successfully localized the target image, leveraging its web-scale, multimodal knowledge to infer its identity through contextual and semantic associations.

Novel LVLMs has been introduced for robots using the previous approach, RT-2 This model gets its general understanding of the world from online text and images it has been trained on, as well as its own interactions in the real world. It translates that data into robotic actions. Each robot has a slightly different way of translating English into action.

While rapid advancements in robotic systems are advancing, significant challenges remain before they can be viably deployed in real-world, consumer-facing environments. Current platforms exhibit limited dexterity and reliability, making it difficult to justify their high cost for everyday users. Moreover, these systems generally lack robust commonsense reasoning capabilities, which constrains their ability to perform multitask operations or adapt to unstructured scenarios. Progress is still needed to transition from basic manipulation tasks such as object grasping and placement to more complex, goal-directed activities involving sequential and context-aware actions. For instance, tasks like reassembling a board game, packaging its components, and returning it to a designated storage location exemplify the level of functional autonomy yet to be achieved. Accordingly, several applications could be useful in the near future, including:

A. **Motion and trajectory generation:** Generative models like Variational Autoencoders (VAEs), GANs, and diffusion models are increasingly used to generate plausible movement trajectories for complex robotic systems.

B. **Grasp and manipulation planning:** Generative models can create synthetic grasp configurations or infer manipulation strategies in high-dimensional spaces, often outperforming traditional planning methods in unstructured environments.

C. **Scene understanding and simulation:** GenAI can produce synthetic environments and simulate sensor data, which is useful for training robots in virtual worlds before deployment.

D. **Language-to-action translation:** LLMs combined with generative policies allow robots to interpret and act on natural language commands, enabling more intuitive human-robot interaction.

E. **Design and prototyping:** Generative design tools assist in the physical design of robotic components by creating novel, optimized shapes or mechanical architectures.

As a consequence, generative models face several critical challenges that limit their deployment in real-world robotics. Safety and reliability remain major concerns, as these models are inherently stochastic and can produce unpredictable or unsafe outputs, which is particularly problematic in high-stakes domains like healthcare or manufacturing. Additionally, data efficiency is a barrier as training such models typically requires large-scale datasets that are costly and impractical to obtain in physical environments; research is ongoing in self- supervised and few-shot learning to address this. Another demanding issue is interpretability it is often unclear why a generative model made a particular decision, complicating debugging and eroding user trust, especially in settings that demand human-robot collaboration. Finally, the real-time performance of generative models poses a challenge due to their high computational demands, motivating efforts to optimize them for efficient inference on edge devices.

Based on the recent advances, future directions in robotics will be increasingly shaped by the integration of GenAI, paving the way for more adaptive, creative, and intelligent machines. One key trend is open-ended skill acquisition, where robots continually learn new tasks through interaction, web-based information, and human demonstrations, moving beyond pre-programmed behavior. This adaptability also supports the emergence of creative robotics, allowing machines to contribute to fields like art, architecture, and music. Additionally, generative AI enables personalized robotics, where systems tailor their actions to individual user preferences especially impactful in domestic and healthcare settings.

Conclusion

GenAI is poised to revolutionize robotics by enabling systems that are not only reactive but also imaginative, adaptive, and creative. While significant challenges remain in safety, interpretability, and efficiency, the convergence of generative modeling and robotics opens the door to more intelligent, versatile, and collaborative machines. A major breakthrough is in open-ended learning, where robots leverage generative models to acquire new skills from human demonstrations, natural language, and large-scale internet data moving away from rigid, pre-programmed instructions. This allows robots to generalize across tasks, adapt in real time, and handle more complex, unstructured scenarios. Generative AI also enhances human-robot interaction by allowing robots to interpret intent, generate natural language responses, and refine their behavior through continuous feedback. Furthermore, GenAI is pushing robotics into creative and personalized domains. Robots can now participate in artistic, architectural, and musical endeavors, suggesting a future where machines become collaborators in creative industries. In personal settings, generative models enable robots to tailor their behavior to individual users, especially valuable in assistive healthcare and home automation. Overall, GenAI will transform robotics from task-specific tools into adaptive, intelligent partners capable of evolving with human needs.

References

1. Prystawski B, Goodman N, Li MY (2023) Why think step-by-step? reasoning emerges from the locality of experience. ArXiv 3107: 70926-70947.
2. Zeng F, Gan W, Wang Y, Liu N, Yu PS (2023) Large language models for robotics: A survey.
3. Peng A, Sucholutsky I, Li BZ, Sumers TR, Griffiths TL, et al. (2024) Learning with language-guided state abstractions. ArXiv.
4. Walter C (2005) You, robot. Hans moravec of carnegie mellon university aspires for robots to be humanity's successors. Sci Am 292(1): 23-23A.
5. Park J, O'Brien J, Cai C, Morris M, Liang P, et al. (2023) Generative agents: Interactive simulacra of human behavior. ArXiv.
6. Heikkila M (2024) Is robotics about to have its own ChatGPT moment? MIT Technology Review.
7. Zhang K, Yun P, Cen J, Cai J, Zhu D, et al. (2025) Generative artificial intelligence in robotic manipulation: A survey. ArXiv.
8. Wong L, Mao J, Sharma P, Siegel ZS, Feng J, et al. (2023) Learning adaptive planning representations with natural language guidance. ArXiv.
9. Cheng X, Shi K, Agarwal A, Pathak D (2023) Extreme parkour with legged robots. ArXiv.
10. Kalashnikov D, Varley J, Chebotar Y, Swanson B, Jonschkowski R, et al. (2021) Mt-opt: Continuous multi-task robotic reinforcement learning at scale. ArXiv.
11. Alto V (2023) Modern generative AI with ChatGPT and OpenAI models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4. Paxckt, Birmingham, England, pp. 286.
12. Liu H, Zhu Y, Kato K, Tsukahara A, Kondo I, et al. (2024) Enhancing the LLM-based robot manipulation through human-robot collaboration. IEEE Robotics and Automation Letters 9(8): 6904-6911.
13. Ravichandran Z, Cladera F, Hughes J, Murali V, Hsieh MA, et al. (2025) Deploying foundation model-enabled air and ground robots in the field: Challenges and opportunities. ArXiv.
14. Team A, Aldaco J, Armstrong T, Baruch R, Bingham J, et al. (2024) Aloha 2: An enhanced low-cost hardware for bimanual teleoperation. ArXiv.
15. Chen S, Xiao A, Hsu D (2024) Llm-state: Open world state representation for long-horizon task planning with large language model. ArXiv.
16. Ge Y, Hua W, Ji J, Tan J, Xu S, et al. (2022) Language models: Past, present, and future. Communications of the ACM 65(7): 56-63.
17. Atknson-Abutridy J (2025) Large language models: Concepts, techniques and applications. Taylor & Francis, CRC Press, Boca Raton, Florida, USA, pp. 184.
18. Honerkamp D, Bijchner M, Despinoy F, Welschehold T, Valada A (2024) Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. IEEE Robotics and Automation Letters 9(10): 8298-8305.
19. Yu S, Lin K, Xiao A, Duan J, Soh H (2024) Octopi: Object property reasoning with large tactile-language models. ArXiv.
20. Bu Q, Li H, Chen L, Cai J, Zeng J, et al. (2025) Towards synergistic, generalized, and efficient dual-system for robotic manipulation. ArXiv.
21. Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, et al. (2022) Do as I can, not as I say: Grounding language in robotic affordances. ArXiv.
22. McCarthy R, Tan DCH, Schmidt D, Acero F, Herr N, et al. (2024) Towards generalist robot learning from internet video: A survey. ArXiv.
23. Collaboration E, O'Neill A, Rehman A, Gupta A, Maddukuri A, et al. (2025) Open x-embodiment: Robotic learning datasets and RT-X models. ArXiv.