Foundational Models and The Pursuit of General Robotics

Aki Bhabad
10/9/2025
University of North Carolina at Chapel Hill

Key terms for ease of understanding (bolded):

Reasoning: The model's ability to decompose goals, infer preconditions, and chain actions (often expressed in natural language plans or intermediate steps).

General robotics: Embodied systems that can perform varied, human-like tasks in unstructured environments, not just fixed routines (contrasts with industrial arms or autonomous vehicles).

Embodied: Coupled to a physical body and sensors (e.g., cameras, depth, touch), producing actions that change the world and require feedback.

Unstructured environments: Open-world settings with variable objects, layouts, lighting, and noise where task conditions are not tightly controlled.

Reinforcement Learning (RL): Learning policies via trial-and-error to maximize reward; powerful but data-hungry and risky on real hardware.

Imitation Learning (IL): Learning from human demonstrations; sample-efficient but can overfit to demonstrated contexts.

Teleoperation: A human directly controls a robot to collect demonstrations or perform tasks (often used to build IL datasets or correct policies).

Abstract

Large language models (LLMs) and vision-language models (VLMs) are reshaping robotics by giving machines the ability to reason, plan, and communicate through natural language. Their success has renewed the long-standing pursuit of general robotics – robots capable of performing diverse, human-like tasks in unstructured environments. This literature review synthesizes how foundation models are being integrated into robotics and what still prevents their full embodiment. This review addresses the question: To what extent can large language models and vision-language models contribute to the development of general-purpose robotics, and what barriers still limit their embodiment? Specifically, it aims to: (1) evaluate how LLMs and VLMs extend reasoning and communication within robotics; (2) identify the major technical bottlenecks—grounding, limited embodied data, and the simulation-to-reality gap; and (3) assess emerging hybrid approaches that integrate foundation-model reasoning with physical interaction. Across recent studies – including SayCan, RT-2, DROID, Open X-Embodiment, FP3, Embodied-R1, and SightGAN – the evidence shows that while LLMs provide a strong cognitive foundation, they remain largely detached from physical experience. True general robotics will

not emerge from scaling parameters but from scaling embodied interaction—robots that learn not only from data, but from the world they move in.

Introduction

Over the past two years, large language models (LLMs) have redefined the landscape of artificial intelligence. From conversational agents like ChatGPT to enterprise-level autonomous software engineer agents, these systems have demonstrated an unseen ability to reason, plan, and interact using natural language. Similarly, vision-language models (VLMs) (the class of models that power multimodal-AI, such as interpreting images and spatial scenes) extend these capabilities by linking textual reasoning with visual understanding. Together, these architectures are often referred to as foundational models, since they are trained on massive, diverse datasets that enable broad generalization rather than narrow specialization. These models put forth the ability of artificial reasoning, which is now a crucial component of intelligent systems. Their success in natural language processing and vision has brought an urgent question within robotics research: Are these foundation models providing a path toward general-purpose robots capable of reasoning and acting in the real world, and what else is needed to reach this autonomy?

This review therefore examines the extent to which foundation models meaningfully advance the pursuit of general robotics. It has three aims: (1) to evaluate what cognitive capabilities LLMs and VLMs currently contribute to robotic reasoning; (2) to analyze the major bottlenecks that prevent full embodiment; and (3) to assess the effectiveness of emerging hybrid approaches that attempt to connect abstract reasoning with physical interaction.

Historically, robotics has evolved through a sequence of specifically specialized frameworks. Industrial robot arms dominate manufacturing floors, autonomous vehicles navigate complex road systems, and household devices such as Roombas or warehouse sorters execute narrow routines (JoinColossus 2024). Each system is defined by fixed environmental assumptions and heavily engineered control loops. The long-standing ambition to progress from these specialized domains to more open-ended domains is known as **general robotics**, which remains mainly unachieved. This includes humanoid or multi-modal robots designed to generalize knowledge; exemplified by projects like PaLM-E, RT-2, and SayCan, which integrate LLMs into perception and control pipelines to connect reasoning with physical action (Billard et al.; SayCan 2022).

Now, in extremely recent work: models like FP3 (Yang et al., 2025) and Embodied-R1 (Yuan et al., 2025) are starting to link reasoning with 3D spatial understanding and embodied pointing, while other studies focus on tactile sensing (Azulay et al., 2023) and human-robot collaboration (Liu et al., 2024) to bring robots closer to real-world adaptability. Together, these efforts suggest

that progress toward general robotics depends on coupling cognitive intelligence with sensory grounding and real-world interaction.

Before the emergence of foundation models, most robotic learning followed two primary approaches: **reinforcement learning (RL)** and **imitation learning (IL)** (Xiao 2023). RL agents learned policies by trial and error to maximize reward, but training was often inefficient and unsafe for physical systems (Tang 2025). Imitation learning, by contrast, relied on human demonstrations to teach robots specific behaviors, producing stability at the cost of limited generalization. Both frameworks struggled to scale beyond carefully modeled environments, exposing a persistent "software bottleneck" in robotics: an inability to link cognitive reasoning with embodied control. The emergence of LLMs and VLMs revived optimism that this barrier could be reduced. Recently, researchers have begun combining these reasoning models with new cutting-edge sensory and spatial learning systems to make progress toward true embodiment.

This literature review examines this possibility. It surveys recent research on both how foundational models and emerging embodied systems are converging toward general robotics. Which organizes the discussion around three related themes: (1) the cognitive potential of LLMs and VLMs in enabling intelligent reasoning; (2) the technical barriers that limit the adoption of general robotics and (3) the hybrid approaches that integrate reasoning, perception, and physical learning. Together, these themes assess how close robotics is to achieving embodied general intelligence, and what barriers remain.

1.0 The Promise of LLMs and VLMs in Robotics

The integration of large language models (LLMs) and vision-language models (VLMs) into robotics has opened a new frontier for reasoning, planning, and communication between humans and machines. Traditional robotic systems have relied on narrowly defined programming or reinforcement learning algorithms that limit their adaptability in complex environments. These falter in complex, changing environments (the real world). Foundation models such as PaLM-SayCan and RT-2 marked a shift from hard-coded routines to flexible, instruction-based behavior grounded in language. By translating words into actions, these systems demonstrated that LLMs and VLMs can function both as a planning interface and as a medium of reasoning – a major cognitive leap for robotics.

Yet, these early systems also exposed an important limitation: language understanding alone cannot guarantee successful physical execution. The following studies, SayCan and RT-2, demonstrate both the promise and the fragility of using foundation-model reasoning in real-world robotic contexts.

1.1 SayCan - LLM's first introduction to Robotics

The SayCan framework (Ahn et al., 2022) provided one of the first empirical demonstrations of integrating LLM reasoning into robot control. Built on Google's PaLM language model, SayCan enabled a mobile manipulator robot to interpret and execute multi-step natural language commands such as "throw away the Coke can and bring me a sponge." Its experimental results highlight both the potential and the limitations of this approach.

The model achieved 100% plan success on simple single-action commands, such as tasks that directly specify one skill. For example, "pick up the apple" – where the instruction corresponds neatly to a pre-trained behavior. It also achieved 93% success for verb-based tasks, such as "restock the rice chips on the far counter," which require the model to infer the appropriate object and context for an action rather than having it explicitly stated. However, performance declined for more complex, context-dependent queries. In embodiment tasks (commands that test the robot's understanding of the current physical state of the world or of itself, such as locating an item or assessing whether a table is clean), the success rate dropped to 64%. Performance fell further in long-horizon reasoning tasks, which involve chaining several dependent actions across time. For instance, "throw away the Coke can, fetch a sponge, and wipe the table", where cumulative reasoning and environmental feedback are required for completion, achieved only 47% success (SayCan 2022). These results demonstrate that while LLMs can effectively generate logical, high-level action sequences in language, they struggle when that reasoning is paired with an embodiment system that has to adapt to continuous environmental feedback or unmodeled physical variability.

1.2 RT-2 - VLM introduction to Robotics

Building off SayCan, RT-2 (Vemprala et al., 2023) extends language reasoning into the visual domain. By incorporating large web-based vision-language pretraining, RT-2 enables robots to recognize uncommon objects and semantically plan tasks beyond their original training distribution. The model demonstrates that natural language can function as an intermediate "chain-of-thought" layer, allowing the robot to verbally reason through its actions before execution. However, the RT-2 team explicitly notes that such pretraining "does not enable the robot to perform new motions by virtue of including this additional experience," revealing that semantic reasoning does not automatically translate into motor or physical generalization (RT-2 2023). Furthermore, the system's largest model, *RT-2-PaLI-X-55B*, containing 55 billion parameters-can only operate at 1–3 Hz when run through a cloud-based multi-TPU service. This dependency on external infrastructure (the cloud) highlights the scalability and latency challenges that currently prevent such models from being deployed in real-time, real-world scenarios (RT-2 2023).

1.3 LLMs as a Cognitive Interface

Across both studies, researchers emphasize that LLMs facilitate far more intuitive human-robot interaction. As Billard et al. (2025) observe, language-based interfaces allow users to communicate complex instructions in natural human language, enabling robots to respond and adapt more fluidly within conversational contexts (Billard 2025). This aligns with the broader motivation of general robotics: to create embodied systems that can reason abstractly yet act concretely, guided by intuitive communication rather than pre-coded scripts.

1.4 Summary: The Cognitive Promise and Its Limits

Current literature thus portrays LLMs and VLMs as a powerful cognitive foundation for the next generation of robots. They extend robotic intelligence from perception and control to reasoning and dialogue. Yet, these same studies also expose their most significant limitations: the absence of physical grounding, the inability to invent new motions, and the prohibitive computational demands of operating large-scale models in embodied settings.

The promise of foundation models, therefore, lies in their conceptual breadth: but their real-world translation remains incomplete with the only recent breakthrough in LLMs and VLMs. The next section explores the bottlenecks that continue to separate reasoning from real-world embodiment: grounding, data scarcity, and the gap between simulation and reality.

2.0 Persistent Bottlenecks: Grounding, Data, and Simulation

While LLMs and VLMs have introduced new reasoning capabilities into robotics, the core technical bottlenecks that limited earlier approaches continue to constrain progress. The literature consistently identifies three interlocking challenges:

(1) the grounding gap between symbolic reasoning and physical perception,
(2) the data bottleneck caused by the scarcity and cost of embodied experience, and
(3) the persistent simulation-to-reality divide that prevents seamless transfer from virtual to physical performance.

Together, these obstacles explain why foundation-model intelligence has yet to produce robust, general-purpose robots.

2.1 The grounding gap

A foundational issue in bridging language reasoning with physical action is the grounding problem: the inability of models trained on text or images to anchor their understanding in real-world sensory experience.

As early LLM-based frameworks such as SayCan demonstrated, robots can logically plan sequences of actions but cannot reliably perceive or adapt to the physical effects of those actions (SayCan 2022). Later analyses confirm that this limitation stems from the nature of multimodal training itself.

According to the *Firoozi 2023* study on vision-language models, current multimodal systems "can analyze 2D images, but they lack a connection to the 3D world, which encompasses 3D spatial relationships, 3D planning, 3D affordances, and more" due to the "scarcity of 3D data paired with language descriptions"(Firoozi et al. 2023). Without this embodied grounding, even the most advanced LLMs remain detached from the causal structure of physical environments. This is a critical gap that separates linguistic reasoning from embodied performance – a mind without a body.

This gap now motivates many of the hybrid approaches explored later in Section 3, where spatial and tactile data are used to give models a physical sense of context.

2.2 The data bottleneck

The second and most widely cited barrier concerns the quantity and diversity of robotic data. Unlike internet text or image corpora, real-world robot data must be physically collected through trials, teleoperation, or simulation.

Tang et al. (2025) note that "experience on a real physical system is tedious to obtain, expensive, and often hard to reproduce," since each roll-out consumes significant time and resources. Reinforcement learning experiments reveal that even minor model inaccuracies compound over time, causing simulated robots to diverge quickly from real-world performance (Tang 2025). Similarly, this survey emphasizes that it is both "inefficient and unsafe for RL agents to collect trial-and-error samples directly in the physical world," underscoring the practical limits of data collection.

Recent studies in imitation learning reinforce this conclusion. Lin et al. (2024) demonstrate that the diversity of environments and objects is far more important than the absolute number of demonstrations: performance rises with new contexts but plateaus once variation declines. Their scaling-law analysis shows that generalization "scales approximately as a power law with the number of training objects and environments," meaning data breadth, not volume, influences success (Lin 2024). This explains why many foundation models trained on homogeneous lab data fail to transfer knowledge across robots or settings. As McCarthy et al. (2025) summarizes, robotics faces a "chicken-and-egg problem: data cannot be easily collected

due to limited robot capabilities, and capabilities cannot easily be improved due to the lack of data" (McCarthy 2025).

2.3 The simulation-to-reality divide

To compensate for scarce real-world data, many researchers rely on simulated environments. Yet, the gap between simulation and reality remains one of the most persistent and well-documented barriers in robotics.

Billard et al. (2025) describe the "sim-to-real" discrepancy as one of the central obstacles in robotics: simulated physics engines fail to capture the variability of real-world contact forces, deformable surfaces, and environmental noise (Billard 2025). Tang (2025) likewise observed that as "small model errors accumulate, the simulated robot can quickly diverge from the real-world system," producing policies that perform well in simulation but collapse under real conditions (Allen 2019). These mismatches make it nearly impossible to train foundation models that both generalize semantically and act reliably across physical domains.

Recent work seeks to narrow this gap by improving the realism of synthetic sensory data. Azulay et al. (2023) propose SightGAN, a *generative adversarial model* that converts simulated tactile images into photorealistic, physically consistent ones.

By enhancing the fidelity of touch-based feedback, SightGAN allows robots trained in simulation to generalize better to physical sensors – a step toward making virtual training data "feel" more like the real world. While still experimental, this line of research shows how refining simulation quality, rather than merely scaling model size, can meaningfully advance embodied learning.

2.4 Summary

Collectively, these studies reveal that the software bottleneck in robotics predates LLMs/VLMs and continues to define their limitations.

Foundation models have not yet overcome the scarcity of embodied data or the fragility of simulated learning - which is what robotics required. Their linguistic reasoning excels in abstract planning but falters without grounded perception and real-world feedback. As a result, progress toward general robotics depends less on scaling model size and more on scaling diversity of embodied experience. Until robots can learn from rich, multimodal, and physically grounded datasets, the reasoning power of LLMs will remain largely disembodied: a mind without a body.

This next section examines hybrid approaches that directly target these constraints: expanding embodied data, sharing skills across robot bodies, and grounding language with 3-D perception, touch, and human feedback.

3.0 Emerging Hybrid Solutions

Despite the enduring limitations of grounding, data, and simulation, recent work in robotics suggests a gradual shift toward hybrid embodied learning. New systems combine the high-level abstraction of foundation models with the low-level adaptability of physical learning, forming what is described as “embodied foundation models”.

These hybrid approaches aim to close the gap identified earlier. Linking language, perception, and control through shared, multimodal experience. Four main strategies define this new frontier:

(1) scaling real-world data diversity
(2) sharing knowledge across robot bodies
(3) embedding spatial and tactile grounding
(4) ensuring transparency and safety in multimodal reasoning.

3.1 Scaling embodied data through real-world diversity

One of the most direct responses to the data bottleneck has been the large-scale collection of diverse real-world robot trajectories.

The DROID dataset (Khazatsky et al., 2024) exemplifies this trend. Compiling 76,000 trajectories across 564 scenes, 86 tasks, and 52 buildings, DROID demonstrates that exposing models to varied environments markedly improves “performance, robustness, and generalization ability.” Its creators argue that diversity in both physical setting and interaction type, rather than sheer volume of repetitions, is the “central ingredient for training such generalizable policies.” However, they also note that collecting such data is uniquely difficult: robot manipulation datasets “cannot be easily scraped from the internet,” and moving robots outside of controlled labs introduces “logistical and safety challenges.” DROID thus represents both progress and constraint, proof that real-world embodiment enhances learning, but also evidence of how costly and complex such embodiment remains.

3.2 Learning from non-robotic multimodal data

Another strategy for expanding robot experience focuses on synthetic or proxy data, training robots from large collections of human behavior rather than robot demonstrations.

McCarthy et al. (2025) argue that deep learning “techniques and scaling laws” developed for text and video “offer a path towards more general-purpose robot capabilities,” provided models can bridge the “chicken-and-egg” gap between data collection and capability growth. Their study proposes using internet video, which captures humans performing physical actions, as a proxy for robot learning data. By aligning visual sequences with action semantics, robots

could theoretically learn affordances and motion patterns before physical training. This method, though very speculative, represents a creative strategy for synthetic embodiment, where robots could learn from observation at a planetary scale.

3.3 Cross-robot learning and Collaborative foundation models

Another major step forward comes from collaborative, cross-robot training efforts such as Open X-Embodiment (2023). This project assembled data from 22 distinct robot types across 21 research institutions, demonstrating 527 skills spanning over 160,000 tasks. The resulting model, RT-X, shows "positive transfer" across robots, meaning that knowledge learned by one system can improve the performance of others. This finding parallels the scaling laws of language models: pooling experiences across agents yields better generalization than isolated learning. In essence, Open X-Embodiment treats diverse robot bodies as analogous to varied human experiences: distinct embodiments feeding into a single foundation model. The success of RT-X suggests that general robotics may arise not from one supermodel trained in isolation, but from shared learning across many embodied agents.

Recent work also focuses on giving robots a better understanding of space itself. FP3 (Yang et al., 2025) introduces what the authors call a 3D foundation policy – a model that learns from both normal camera images and depth data (RGB-D), helping robots understand the size, shape, and position of objects around them. Unlike earlier 2D vision-language systems such as RT-2, FP3 doesn't just recognize what an object is; it also understands where it is in three-dimensional space and how it can be moved or grasped. In tests on everyday manipulation tasks, FP3 reached over 90% success even with objects and shapes it had never seen before, roughly doubling the generalization performance of previous models that used only visual and text inputs.

The authors describe this as an important step toward spatial generalization, scaling not just the amount of data robots learn from, but the richness of that data. However, FP3 still relies on carefully collected 3D scans and depth sensors, showing that truly universal spatial understanding will require larger, more varied 3D datasets.

Together, RT-X and FP3 show that expanding what robots experience, across different bodies, environments, and spatial settings (the training data), may be just as important as expanding the size of their neural models (the LLMs).

3.4 Grounded and Collaborative Embodiment

Beyond data scale, researchers are improving how reasoning connects to the physical world. The Embodied-R1 model (Yuan et al., 2025) pairs vision-language reasoning with reinforced pointing behaviors that tie language to concrete 3-D locations. Despite being far smaller than RT-2 with about 3 billion parameters, it achieved 87.5 % real-world success on the XArm platform, showing that better grounding can outweigh sheer model size.

Touch sensing is another path toward embodiment. Azulay et al. (2023) developed SightGAN, which converts synthetic tactile images into realistic ones, allowing robots trained in simulation to transfer more smoothly to real sensors. By making simulated “touch” feel real, SightGAN directly reduces the sim-to-real gap identified earlier.

Human-robot collaboration also plays a growing role. Liu et al. (2024) showed that adding human feedback loops to LLM-based manipulation systems helps with complex, long-horizon reasoning that pure code-generation approaches struggle with. Letting humans guide corrections teaches robots to adapt in changing situations.

Together, these approaches suggest that future progress will depend on grounded, multimodal, and collaborative learning. With robots that can reason with language, perceive in 3D, sense through touch, and work with humans in real time.

3.6 Summary

Together, these studies mark the first tangible movement toward embodied generalization.

[3.1] DROID and Open X-Embodiment show that pooling diverse experiences strengthens generalization.

[3.2] McCarthy et al. show that large-scale human and internet multimodal data can theoretically supplement limited robot experience, displaying the extent of which the industry is exploring creative and innovative solutions.

[3.3] FP3 and Embodied-R1 prove that grounding in 3-D space improves real-world control even in smaller models.

[3.4] SightGAN and Liu et al. demonstrate that tactile realism and human feedback are closing the simulation gap.

Despite this progress, none of these systems fully merges language-based reasoning with autonomous, adaptive motor control. Hybrid embodied models therefore represent a transitional stage, beginning to bridge disembodied intelligence and truly general robotics, but that bridge is not yet complete. The next challenge will be scaling these systems beyond research labs into durable, trustworthy autonomy in everyday environments.

Discussion and Synthesis

Across the reviewed literature, a clear narrative emerges: LLMs and foundation models have revolutionized the cognitive dimension of robotics, but embodiment, data, and real-world generalization remain the defining barriers between simulated intelligence and physical autonomy.

Section 1 showed that LLMs and VLMs introduced a cognitive substrate for robots. They allowed machines to plan, reason, and communicate through language, giving rise to what some researchers call "robotic thought." Yet these systems remained largely disembodied: they could describe and predict actions but not feel or adapt to their consequences.

Section 2 revealed why this abstraction fails in practice. The grounding gap prevents models from connecting symbols to sensations; the data bottleneck limits how much embodied experience robots can acquire; and the simulation-to-reality divide ensures that performance in controlled settings rarely transfers to the messy dynamics of the physical world. In short, LLM-based reasoning expanded what robots can imagine, but not what they can do.

Section 3 then traced how current research is beginning to close these divides through hybrid approaches that combine reasoning with embodiment. Large, diverse datasets such as DROID and Open X-Embodiment address data scarcity by pooling physical experience across platforms. Synthetic-data efforts like McCarthy et al. (2025) push this further, suggesting that robots might pre-train on human demonstrations and internet video before ever touching the real world. Spatially grounded models such as FP3 and Embodied-R1 translate abstract goals into 3-D understanding and motor action, while SightGAN and Liu et al. integrate tactile realism and human feedback to reduce the simulation gap.

Taken together, these developments suggest a conceptual shift in how researchers pursue general robotics. Rather than trying to scale a single all-in-one model, the field is moving toward multimodal ecosystems – networks of models and sensors that share experience across different bodies and data domains. This distributed view of intelligence reframes embodiment as a shared property of systems, not just an attribute of an individual robot.

At the same time, the literature emphasizes that technical progress must be matched by transparency and accountability. As Xiao et al. (2023) and Billard et al. (2025) argue, the more autonomous these systems become, the more essential it is that their decision processes remain interpretable, fair, and aligned with human values. Ethical reliability is therefore emerging as the fourth pillar of embodiment – joining reasoning, perception, and control.

In essence, the trajectory of robotics now mirrors that of language models a decade ago: progress will be less on larger parameters and more on richer interaction. The next stage of research will need to focus on scaling experience rather than scale alone. Thus developing robots that not only reason about the world but also learn how to interact with it through living in it.

Conclusion

This literature review set out to address the question: To what extent are large language models (LLMs) a solution to general robotics, and what barriers prevent general robotic's mainstream adoption? The evidence across contemporary robotics research makes the answer increasingly clear: Foundational models represent a crucial step toward general robotics, but not its completion. They provide the reasoning, communication, and abstraction capabilities that robotics has historically lacked, but without physical grounding, diverse embodied data, and scalable infrastructure, these cognitive advances cannot yet translate into fully autonomous general-purpose robots.

Recent hybrid frameworks show that this convergence is already underway. Datasets like DROID and Open X-Embodiment are expanding real-world experience; McCarthy et al. demonstrate that human and internet data can supplement limited robot trials; and models such as FP3, Embodied-R1, and SightGAN reveal how spatial and tactile grounding can close the simulation gap. Together, these projects redefine progress in robotics not as scaling model size, but as scaling experience – the variety, richness, and interconnectedness of what robots can learn from.

Still, the road to truly general robotics remains long. No current system combines autonomous reasoning, multimodal perception, and adaptive motor control without human intervention. The future therefore lies not in a single breakthrough model but in an ecosystem of embodied foundation systems that learn collectively, act safely, and explain their decisions transparently. As ethical and technical boundaries blur, the question is no longer whether robots can think, but whether they can perform responsibly and effectively within the physical world.

In essence, LLMs have given robots a mind; embodiment research is now teaching them to move with understanding. General robotics will emerge when these two forms of intelligence, cognitive and physical, become one continuous process of learning from the world itself.

Future research should prioritize three directions that consistently emerge across the reviewed literature. First, robotics requires far larger and more diverse 3-D, tactile, and multimodal datasets to improve grounding and spatial understanding. Second, models must become computationally efficient enough to run on-device rather than depend on cloud inference, reducing latency and enabling real-time control. Third, collaborative learning frameworks – such as cross-robot skill sharing and human-in-the-loop correction systems – should be expanded to support safer, more adaptive behavior. These avenues represent concrete steps toward merging foundation-model reasoning with reliable physical autonomy.

References

1. Ahn M, Brohan A, Chebotar Y, et al. 2022. Do as I can, not as I say: Grounding language in robotic affordances. arXiv [Internet]. Available from: https://say-can.github.io/assets/palm_saycan.pdf
2. Allen K, Barto A, Lillicrap T, Peters J, et al. 2019. Deep reinforcement learning for robotics: A survey of real-world successes. Columbia University [Internet]. Available from: https://www.cs.columbia.edu/~allen/S19/rl_robotics_survey.pdf
3. Azulay O, Mizrahi A, Curtis N, Sintov A. 2023. Augmenting tactile simulators with real-like and zero-shot capabilities. arXiv [Preprint]. arXiv:2309.10409v1. Available from: https://arxiv.org/pdf/2309.10409.pdf
4. Billard A, Albu-Schaeffer A, Beetz M, et al. 2025. A roadmap for AI in robotics. arXiv [Internet]. Available from: https://arxiv.org/pdf/2507.19975
5. Firoozi R, Tucker J, Tian S, Majumdar A, Sun J, Liu W, Zhu Y, Song S, Kapoor A, Hausman K, Driess D, Wu J, Lu C, Schwager M. 2023. Foundation models in robotics: Applications, challenges, and the future. arXiv [Internet]. Available from: https://doi.org/10.48550/arXiv.2312.07843
6. JoinColossus. 2024. When will robots go mainstream? [Internet]. Available from: https://joincolossus.com/article/when-will-robots-go-mainstream/
7. Khazatsky A, Jing M, Choi H, et al. 2024. DROID: A large-scale in-the-wild robot manipulation dataset. Proceedings of Robotics: Science and Systems (RSS) [Internet]. Available from: https://www.roboticsproceedings.org/rss20/p120.pdf
8. Lin Z, Chen T, Levine S, Finn C. 2024. Data scaling laws in imitation learning. arXiv [Internet]. Available from: https://arxiv.org/pdf/2410.18647
9. Liu H, Zhu Y, Kato K, Tsukahara A, Kondo I, Aoyama T, Hasegawa Y. 2024. Enhancing LLM-based robot manipulation through human-robot collaboration. IEEE Robotics and Automation Letters. doi:10.1109/LRA.2024.3415931
10. McCarthy J, Gao S, Xu R, et al. 2025. Learning from internet video for robotic generalization. Journal of Artificial Intelligence Research [Internet]. Available from: https://arxiv.org/abs/2404.19664
11. Open X-Embodiment Collaboration. 2023. Open X-Embodiment: Robotic learning datasets and RT-X models. arXiv [Internet]. Available from: https://arxiv.org/abs/2310.08864
12. Tang C, Xiong Y, Kumar R, et al. 2025. Deep reinforcement learning for robotics: A survey of real-world successes. Annual Review of Control, Robotics, and Autonomous Systems [Internet]. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev-control-030323-022510

13. Vemprala S, Zeng A, Florence P, Zeng E, et al. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv [Internet]. Available from: https://arxiv.org/pdf/2307.15818
14. Xiao T, Chen X, Huang Z, et al. 2023. Robot learning in the era of foundation models: A survey. arXiv [Internet]. Available from: https://arxiv.org/abs/2311.14379
15. Yang R, Chen G, Wen C, Gao Y. 2025. FP3: A 3D foundation policy for robotic manipulation. arXiv [Preprint]. arXiv:2503.08950. Available from: https://arxiv.org/pdf/2503.08950
16. Yuan Y, Cui H, Huang Y, Chen Y, Ni F, Dong Z, Li P, Zheng Y, Hao J. 2025. Embodied-R1: Reinforced embodied reasoning for general robotic manipulation. arXiv [Preprint]. arXiv:2508.13998v1. Available from: https://arxiv.org/pdf/2508.13998