

Advancing general robotic manipulation with multimodal foundation models: An embodied AI paradigm

Shifeng HUANG^{1,2,3}, He WANG^{4,5}, Xing ZHOU^{2,3}, Wenkai CHEN⁶,
Haibin YANG^{2,3} & Jianwei ZHANG^{6*}

¹ Department of Mechanical and Aerospace Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China

² Foshan Institute of Intelligent Equipment Technology, Foshan 528234, China

³ Foshan Huashu Robot Co., Ltd., Foshan 528234, China

⁴ Center on Frontiers of Computing Studies (CFCS), School of Computer Science, Peking University, Beijing 100871, China

⁵ Beijing Academy of Artificial Intelligence, Beijing 100084, China

⁶ Department of Informatics, Universität Hamburg, Hamburg 20146, Germany

Received November 4, 2024; accepted March 3, 2025; published online April 8, 2025

Citation: Huang S F, Wang H, Zhou X, et al. Advancing general robotic manipulation with multimodal foundation models: An embodied AI paradigm. *Sci China Tech Sci*, 2025, 68(5): 1560301, <https://doi.org/10.1007/s11431-024-2910-3>

Can current robotic technologies truly replicate the full scope and intricacies of human labour? In practice, the adoption of robots remains limited, especially in open, unstructured environments commonly encountered in everyday scenarios such as services, healthcare, agriculture, construction, and numerous other fields. From the perspective of general robotic manipulation, the challenges arise from three factors.

(1) High operational barriers: human operators are obliged to master specialized robotic programming languages and gain a deep understanding of the tasks at hand. These tasks need to be broken down into action-level robotic programs, which results in high labour costs.

(2) Limited autonomous task execution: robots lack the capability to independently plan and execute actions required to achieve the target tasks. This limitation renders them unsuitable for deployment in open, unstructured environments that demand sophisticated interaction and seamless collaboration with humans.

(3) Poor manipulation skills: robots struggle with many tasks, particularly those requiring skills for dexterous manipulation. For instance, in shoe factories, workers are still necessary to manually thread shoelaces. Firstly, robots face significant challenges in acquiring the specialized and complex skills needed for specific manipulations. Secondly, their execution of such skills tends to be rigid and inefficient, lacking the smoothness and finesse of human performance.

Recently, the rise of embodied AI (referring to AI systems with a physical embodiment) [1] appears to offer a promising pathway for developing general robotic manipulation. In this context, the robot seems to possess an intelligent brain, demonstrating a high degree of autonomous planning capability. An intriguing observation among practitioners is that the emer-

gence of embodied AI almost coincided with major breakthroughs in generative AI, exemplified by the release of ChatGPT in late 2022, a landmark achievement in the development of Large Language Models (LLMs) [2]. This convergence of innovations has thus sparked thought-provoking questions.

(1) Why have LLMs catalyzed the development of embodied AI?

(2) Are large-scale “language” models alone sufficient to achieve the vision of general robotic manipulation we aspire to realize?

Thus, it prompts us to reflect on the essence of embodied AI and to assess the capabilities and limitations of LLMs in addressing the challenges of general robotic manipulation. Our perspective on this issue is as follows.

By leveraging ultra-large models enabled by scaling laws [3], LLMs exhibit remarkable performance in common sense knowledge and reasoning and show proficiency in contextual understanding and continuous dialogue. Robots are commonly regarded as mere “automated devices” rather than “intelligent agents”. Recent advancements indicate that LLMs could equip robots with a brain-like system. LLMs facilitate natural language interaction between operators and robots, enabling robots to effectively comprehend human intentions and make decisions. Therefore, LLMs have the potential to overcome the bottleneck of “high operational barriers” and ease the challenges of “limited autonomous task execution”.

Embodied AI, however, posits that intelligence arises from the dynamic interactions between an agent’s embodiment (physical body) and its real-world environment, evolving continuously over the course of its lifetime. Therefore, the fundamental distinction between embodied AI and traditional AI (that is, disembodied AI) lies in the presence of physical embodiments in embodied AI. Robotic manipulation, a fundamental topic within this paradigm, goes beyond language-based feedback modalities to embrace robotic proprioception, tactile sensing, force/torque feedback, vision, and more. These rich modalities are crucial for the identification of an object’s affor-

* Corresponding author (email: jianwei.zhang@uni-hamburg.de)

dances [4] and the formation of intelligence. For example, when we grasp an object, we use tactile feedback to assess the object's texture and weight, while visual input aids in determining its shape and orientation. These combined modalities allow us to understand how the object can be manipulated. Clearly, LLMs (language alone) cannot meet the requirements of embodied AI, as they lack the action-oriented capabilities necessary for robotic manipulation. Instead, we need more powerful foundation models that integrate the various modalities a robot perceives, such as vision-language models (VLMs) [5] and vision-language-action models (VLAs) [6], which are collectively termed multimodal foundation models (MFMs).

General robotic manipulation can be advanced by MFMs within the embodied AI paradigm. Nowadays, the technological framework suitable for embodied AI remains an open issue and can generally be categorized into two types: (1) end-to-end (E2E) pipelines [7] and (2) modular (or hierarchical) pipelines [8]. Below, we discuss each approach.

An E2E pipeline maps raw sensor data directly to actions without intermediate hand-crafted modules. This approach avoids information distortion during transmission between modules, and reduces cumulative errors and latency. Furthermore, an E2E pipeline can more effectively capture human skills, because it is often difficult to accurately model abstract actions using analytical methods. Despite the promising potential of E2E pipelines, their limitations should be recognized [9]. A fully end-to-end pipeline operates like a black-box model, lacking interpretability and safety guarantees. Its performance heavily depends on the availability of large-scale, high-quality datasets during the training phase, and achieving high generalization requires a large-scale neural network. This dependency on extensive data and complex models can lead to significant challenges in terms of data annotation costs, computational resources, and the risk of overfitting. Additionally, the lack of interpretability makes it hard to validate the model's decisions, which is detrimental to safety-critical applications. Overall, we

should remain optimistic about the potential of E2E pipelines, but we must also approach the current challenges with caution.

A more deployable and cost-effective solution for most robot manufacturers is to adopt a modular pipeline, in which modules such as perception, planning, and control are independently designed. In this article, we present a potential solution for the modular pipeline, characterized by “cerebrum-cerebellum” integration, complex task decomposition, and composition of primitive skills, as illustrated in Figure 1. MFMs act like the cerebrum, enabling multimodal perception, reasoning, and decision-making, while also fulfilling the requirements of natural human-robot interaction (nHRI). In this context, human users can convey tasks through various forms, including speech, text, and images, and the robot is capable of understanding the underlying intent behind the received information. In addition, MFMs decompose tasks, especially complex and long-horizon ones, into simpler subtasks in a top-down manner (task planning). Each subtask can then be mapped to a combination of skills. At this level, the granularity of skills is still relatively coarse and can be further broken down into primitive skills [10]. Real-life tasks are virtually limitless, and as the environment changes over time, they often require the acquisition of unforeseen skills. Therefore, policy learning for primitive skills, like using reinforcement learning (RL) or imitation learning (IL) to master these skills, should be designed in a continuous learning paradigm to avoid catastrophic forgetting of previously learned skills. We recommend using lightweight models for policy learning to facilitate training and ensure a real-time response. Furthermore, the proposed policy learning architecture (cerebellum) is hybrid. The analytical-based method can serve as an auxiliary component for data-driven policy learning. This hybrid architecture reduces the complexity of policy learning while offering better physical interpretability. Action sequences for tasks can be generated using MFMs that combine skills from the bottom-up (action planning). Robot manipulation capabilities

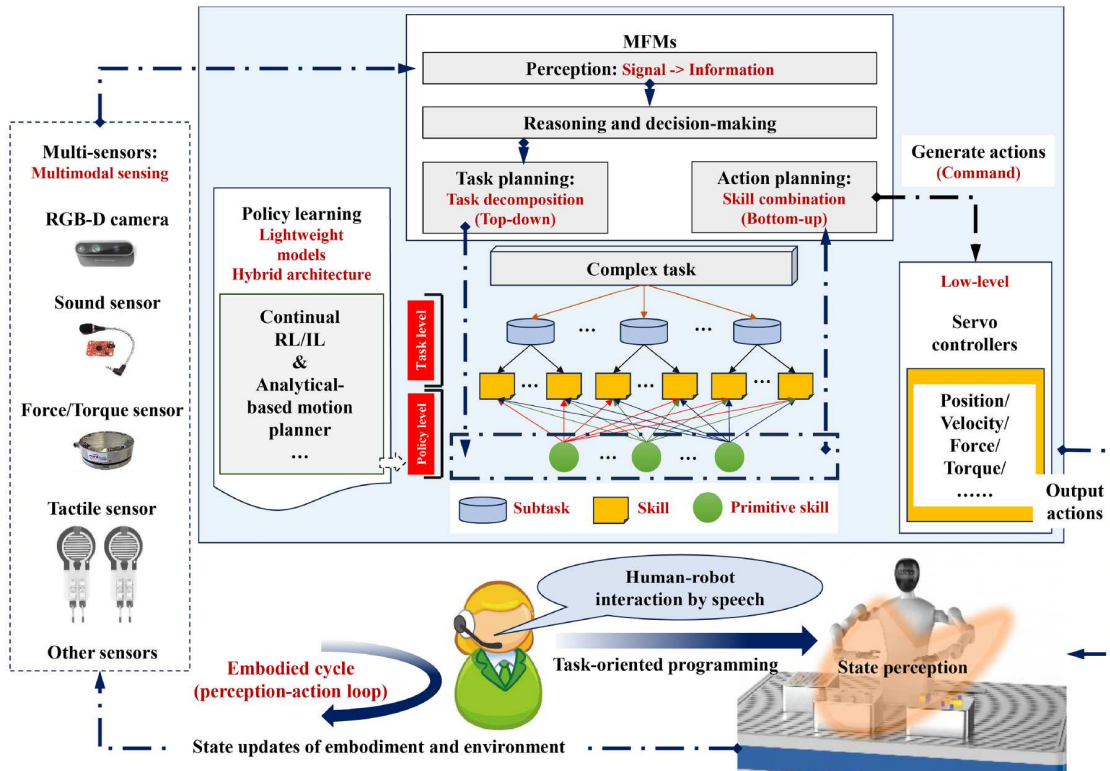


Figure 1 A framework for the modular pipeline: “cerebellum-cerebrum” integration, decomposition of complex tasks, and composition of primitive skills.

are expected to continuously improve within a perception-action loop.

In fact, the principles underlying task decomposition and the combination of primitive skills are inspired by human behaviors. When faced with a long-horizon task, humans naturally break it down into smaller subtasks. A crucial aspect of human intelligence lies in the capacity to generalize from one instance to another, allowing us to apply the acquired knowledge flexibly across diverse scenarios. This remarkable generalization ability originates from a profound mastery of the fundamental primitive skills essential for those manipulations. The descriptions of primitive skills, the design principles of MFMs applicable to robotic manipulation, the mechanisms for decomposing complex tasks, and the approaches to skill combination are detailed in our supplementary file.

Finally, to promote the progress of embodied AI in the realm of general robotic manipulation, the following aspects should be emphasized: (1) Developing high-performance AI chips and embodied AI-oriented network architectures to enable the deployment and real-time computation of large models on terminal platforms. (2) Exploring the cognitive evolution and collaboration of multi-embodied AI in social contexts, forming safe, causal, and trustworthy world models. (3) Achieving improved tokenization-based multimodal integration, developing interpretable AI theories for transparent decision-making, and aiming to mitigate MFM hallucination. (4) Leveraging both real-world and synthetic data to create high-quality datasets for various sectors. (5) Building domain-specific models with scalable functionality and long-term memory to support vertical applications. (6) Improving the efficiency of learning complex skills via zero-shot sim-to-real transfer, and tackling the challenges of lifelong skill development in real-world environments.

Acknowledgements This work was supported by the Guangdong Provincial Science and Technology Program (Grant No. 2023A0505030003).

Supporting Information The supporting information is available online at tech.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

- 1 Li C S, Zhang R H, Wong J, et al. Behavior-1k: A benchmark for embodied AI with 1000 everyday activities and realistic simulation. In: Proceedings of the 6th Conference on Robot Learning (CoRL 2022). Auckland: PMLR, 2023. 80–93
- 2 Kalyan K S. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Nat Language Process J*, 2024, 6: 100048
- 3 Antonello R, Vaidya A, Huth A. Scaling laws for language encoding models in fMRI. In: Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS). New Orleans: MIT Press, 2023. 21895–21907
- 4 Rai A, Buettner K, Kovashka A. Strategies to leverage foundational model knowledge in object affordance grounding. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle: IEEE, 2024. 1714–1723
- 5 Zhang J Y, Huang J X, Jin S, et al. Vision-language models for vision tasks: A survey. *IEEE Trans Pattern Anal Mach Intell*, 2024, 46: 5625–5644
- 6 Zitkovich B, Yu T H, Xu S C, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Proceedings of the 7th Conference on Robot Learning (CoRL 2023). Atlanta: PMLR, 2023. 2165–2183
- 7 Chib P S, Singh P. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Trans Intell Veh*, 2024, 9: 103–118
- 8 Chitta S, Jones E, Ciocarlie M, et al. Mobile manipulation in unstructured environments: Perception, planning, and execution. *IEEE Robot Automat Mag*, 2012, 19: 58–71
- 9 Chen L, Wu P H, Chitta K, et al. End-to-end autonomous driving: Challenges and frontiers. *IEEE Trans Pattern Anal Mach Intell*, 2024, 46: 10164–10183
- 10 Hiranaka A, Hwang M, Lee S, et al. Primitive skill-based robot learning from human evaluative feedback. In: Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Detroit: IEEE, 2023. 7817–7824