Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24886

Design and Development of Embodied AI Based on Foundation Models

Bozhi Yang

School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
yang_bozhi6432@mail.ustc.edu.cn

Abstract. The disembodied intelligence technology represented by large language models is the mainstream of today's AI development. However, traditional large models are limited by the embodiment problem, and cannot reach the human level in reasoning and common sense understanding. Embodied intelligence provides solutions to this problem by integrating environment understanding, intelligent interaction, cognitive reasoning, and planning execution. This paper reviews the historical development of two pathways in artificial intelligence. Through case studies, it examines the technical approaches for integrating embodied intelligence with foundation models in scene understanding and action execution steps, exploring how their convergence advances the development of artificial general intelligence. By synthesizing existing literature, the feasibility and advantages of combining embodied intelligence with foundation models are validated. Current limitations and challenges in existing models are analyzed, and targeted development directions and optimization strategies for this field are subsequently proposed. The foundation models-based embodied intelligence significantly enhances the adaptability and generalization ability of AI in open environments through the perception-cognition-execution closed loop, and promotes the development of Artificial General Intelligence.

Keywords: Large model, foundation model, embodied intelligence, robot.

1. Introduction

In recent years, Artificial Intelligence technology represented by large language models (LLMs) in the field of Natural Language Processing (NLP) has been advancing by leaps and bounds, with disruptive results such as GPT-3 [1], GPT-4 [2], DeepSeek, and so on. These large models can fluently communicate and dialogue with humans, answer human questions, and even help humans perform tasks such as writing articles, editing code, and drawing. Among the various large models that have appeared in recent years, there are some pre-trained large models utilizing self-supervised learning that show excellent generality [3]. They can be the starting point for building more specific and specialized large models and have become the basis for building AI ecosystems. They have acquired a new name - foundation models.

However, the performance of foundation models in terms of logical reasoning is not yet up to the level of humans, and it is still subject to all kinds of common-sense errors or reasoning mistakes.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24886

Moreover, the interaction between foundation models and humans is limited to text, pictures, and so on. They cannot interact with the complex external environment through vision, hearing and other modalities. They cannot feel the external environment and its changes in real time, and thus cannot give more appropriate answers.

Because of this, some researchers have noticed another direction of AI: Embodied intelligence. As a cutting-edge field at the intersection of AI and robotics, embodied intelligence emphasizes the autonomous learning and evolution of intelligence through the dynamic interaction between body and environment. The core of embodied intelligence research lies in the in-depth integration of perception, action and cognition. Researchers in the related direction believe that true intelligence not only relies on pre-programmed algorithms and data training but also needs to perceive, learn, and adapt through the dynamic interaction between the intelligent body and the external environment.

With the development of large models, embodied intelligence based on foundation models has become a research hotspot in recent years, and various attempts to combine the two have emerged. Unlike previous robots trained with specialized task data, foundation models are often trained with a large amount of text-image data from the Internet, which often contains a large number of topics or application scenarios. In addition, the rich modal perception of embodied intelligence can provide more accurate training data for foundation models, which in turn can provide a more comprehensive and in-depth understanding of foundation models based on these data. The language organization ability of foundation models provides embodied intelligence with the ability to communicate with humans without barriers, which enables the embodied intelligence to understand and comprehend human needs and thoughts and perform actions according to the requirements. Thus embodied intelligence becomes more autonomous and task-adaptive, and can handle complex tasks and carry out long-term planning [4].

Combining embodied intelligence with foundation models can help enhance the cognitive and reasoning abilities of intelligence, and strengthen the perception and training of large language models, realize the fusion of the ‘brain’ and ‘body’ of AI. This will enable more natural human-computer interaction, solve the problem of AI adaptability in more complex and open environments, and promote the research of Artificial General Intelligence.

This paper organizes relevant technologies into a unified framework to provide a reference for future researchers interested in this field, addressing the abundance of existing research outputs that lack systematic classification. Drawing upon key research achievements in combining embodied intelligence with foundation models over the past three years, the concepts and development trajectories of both foundation models and embodied intelligence are explained. Representative techniques for integrating foundation models with embodied intelligence will be introduced, alongside an analysis of the challenges faced by this integration. Targeted projections of future development are also proposed.

2. Definition

Alan M. Turing, the father of computer science, in his 1950 paper Computing Machinery and Intelligence, first suggested two directions in which computing machines might develop. The first direction was for machines to learn abstract skills, such as playing chess, and the second direction was to equip a robot with the best possible sensory organs and then teach it human knowledge so that it could feel, learn, and think like a human [5]. The first direction provides ideas for the development of disembodied intelligence, where cognition is an abstract symbolic computational process independent of the body [6,7]. The second direction develops into embodied intelligence,

DOI: 10.54254/2755-2721/2025.PO24886

where cognition relies on the experience of feeling and interacting with the environment in a body that possesses specific sensorimotor abilities [8,9].

Since AI technology ushered in the deep learning revolution and the rise of big models in the 2010s, disembodied intelligence has become the mainstream direction of today's AI technology. On this path, researchers have proposed many groundbreaking techniques, such as the Perceptron Model, multilayer neural networks, and the Transformer [10], which laid the foundation for the emergence of large models. In addition to the various types of large language models mentioned above, AlphaFold [11], BloombergGPT [12], and other large models used for specialized big models for biology, finance and other fields also appear. Early definitions and descriptions of large models in academia and industry were very confusing, lacking systematic classification, while the wild growth of technology also brought potential ethical and other challenges to the development of large models. Therefore, in 2021, more than a hundred researchers from Stanford University, MIT and other institutes, with their keen insight into the technological revolution, pioneered the concept of "Foundation Models" in their pioneering paper On the Opportunities and Risks of Foundation Models. The paper defines Foundation Models as models that are trained on broad data at scale and are adaptable to a wide range of downstream tasks, i.e., they are large-scale, pre-trained, generic and foundational, which are usually based on Transformer [13]. More importantly, the article not only introduced this concept but also refreshed the cognitive coordinates of researchers in the field of artificial intelligence around the world. Since then, the research, development, governance, and application of large models have been integrated into the new paradigm defined by the foundation models, which is moving toward the right, controllable, and low-risk direction in an open and collaborative environment. As Rishi Bommasani, the first author of the paper, mentions in the paper and subsequent interviews, the foundation models are becoming the underlying infrastructure of digital civilization, and humans must learn how to build and master them [13,14].

To be more rigorous, foundation models are included in large models, specifically referring to those that have the four key elements mentioned above. GPT-3, GPT-4, DeepSeek, and Claude, LLaMA, etc., are both large language models and foundation models.

However, foundation models are flawed. They are limited by the problem of embodiment in the real world. There is growing evidence that foundation models are not yet capable of human-level causal reasoning and general knowledge of the physical world, and are unable to interact with the external environment. Although it is possible to use massive amounts of data in training large models, allowing them to understand the current knowledge and state of the physical world, the high data and training costs and the shortcomings of being unable to detect changes in a timely manner make subsequent developments in this direction unsustainable. Early related work made decisions by combining LLM outputs with robot strategies, but this intelligence can only be perceived by virtue of textual inputs and is incapable of tasks that require processing information about the geometry of the scene [15].

On the other hand, although the concept of embodied intelligence has been proposed for more than half a century, due to the limitations of early technologies, embodied intelligence remained at the conceptual stage until the early part of this century. It was only after the AI technology revolution mentioned above, when computer vision, natural language processing, cognitive science, large models and other fields witnessed rapid development, that embodied intelligence was able to break away from the concepts on paper and become a reality.

As an intersection research field between robotics and artificial intelligence, embodied intelligence gives a good idea to solve the difficulties faced by the foundation model: it provides a systematic solution integrating environment understanding, intelligent interaction, cognitive

DOI: 10.54254/2755-2721/2025.PO24886

reasoning, and planning execution. Firstly, the equipment used by embodied intelligence for environment sensing, such as various kinds of high-precision sensors, can collect external environment data in real-time and input them into foundation models. This not only eliminates the need to import a large amount of fixed and outdated environment data into the foundation models during training but also gives foundation models the ability to adapt to changes in the environment. Secondly, the strong reasoning and learning ability of foundation models can enable the embodied intelligence to quickly recognize and understand the external environment and its changes, and help the embodied intelligence to quickly and accurately plan the steps of the task; thirdly, through the help of robotic arms and other equipment, the embodied intelligence can turn the planning of the foundation models into an action sequence that can be carried out by itself, instead of just giving answers in the form of words or pictures.

The embodied intelligence combined with the foundation models will have both the perception and execution capabilities of robots and the reasoning and planning capabilities of large models, which will profoundly change the development path of AI. In recent years, the technologies represented by the Vision-Language Model (VLM) and Vision-Language-Action Model (VLA) have been developing rapidly, and have become the main force in the development of embodied foundation models.

3. Technical routes

The current technologies for embodied foundation models are diverse, but in summary, they all follow the perception-cognition-action loop. The technology of embodied intelligence is roughly divided into four parts: scene perception, data guidance, embodied reasoning, and action execution. Scene perception means embodied intelligence perceives and understands the external environment in various ways. Data guidance refers to the embodied intelligence learning embodied reasoning methods and interactive action generation through data training. Embodied reasoning means embodied intelligence uses the foundation model for reasoning and output. Action execution refers to the embodied intelligence converting the output of foundation models into an executable sequence of actions and completing the task.

In terms of data guidance, difficulties still exist in current research. For the embodied reasoning part, the method of calling existing foundation models is mainly used. Therefore, these are not the focus of this section. Most of the newly emerged embodied foundation models have made breakthroughs in scene perception and action execution, so this part will focus on these two directions.

3.1. Scene perception

As repeatedly emphasized in this paper before, a good embodied foundation model must be able to perceive its surroundings clearly and comprehensively, which is a basic ability of embodied intelligence. For humans, about 80% of external information is input to the brain through vision [16]. Humans can acquire environmental details and perform 3D modelling through binocularity and binocular parallax, so as to complete a series of tasks such as distance judgment and route planning. Analogously to humans, the most basic environment perception capability required for the embodied intelligence base model is vision. Both VLM and VLA have vision modules.

The scene perception technique used in mainstream embodied foundation models is multi-modal grounding. The essence of multi-modal grounding is to establish semantic mapping relationships between different modal data, with the core goal to semantically or spatially align information (e.g.,

DOI: 10.54254/2755-2721/2025.PO24886

text, image, and audio) from different modalities to achieve accurate cross-modal correlation and comprehension [17]. Currently, this technique mainly recognizes visual information in the form of images. A typical example is Google's embodied vision-language model PaLM-E. PaLM-E is based on the pre-trained PaLM foundation model and the ViT visual model, to which a multimodal encoder is added, enabling the model to recognise "multi-modal sentences" containing natural language and inserted images in user inputs. ". Unlike previous VLMs, PaLM-E's multi-modal sentences support dynamic picture insertion. Users can add picture input at any position of the sentence during input, breaking the limitation of earlier research in this area and making the input more consistent with natural human language habits.According to the researcher's results, PaLM-E demonstrates remarkable transferability on diverse datasets and outperforms the baseline model in the embodied task test [18]. As an excellent embodied foundation model, PaLM-E has now become a reference and a new benchmark for the development of subsequent embodied models. For example, the design of RT-2 includes PaLM-E [19].

However, the visual input of traditional VLMs still relies on 2D pictures, but the world in which humans live is three-dimensional. Even though it is possible to reveal the position and distance relationships of objects in the environment by means of multi-angle pictures, there are still many details of the 3D world that cannot be given in a 2D picture. To attain a level close to that of humans, embodied foundation models must adopt 3D modeling technology for environmental perception. In this field, researchers from Peking University designed LEO, the first embodied multitasking, multi-modal generalist intelligence in the 3D world, in 2024.LEO can process both 2D and 3D environmental information. For 2D information, it uses pre-trained processing to obtain visual features from the first viewpoint; for 3D information, it first extracts the features of scene objects, and then models spatial location relationships with spatial encoders to obtain 3D scene features. In addition, the researchers also propose a self-service data generation method based on scene graph and LLM to ensure the quality of the generated information. The emergence of LEO fills the lack of multimodal models in 3D spatial understanding and interaction, and is significantly better than previous methods in 3D visual language understanding tasks [20].

3.2. Action execution

As another part of the interaction between embodied intelligence and the outside world, how to translate the tasks given by foundation models into correct and recognizable action commands is one of the key topics in this field. Compared with the scene understanding direction, the action execution direction has a later start. There are many current research results in this direction, and some of the more outstanding ones are Autoregressive techniques represented by RT-2 [19] and OpenVLA [21]. In this paper, two impressive action execution techniques will be selected to introduce.

In 2024, researchers proposed a generic robot strategy model π0, whose core idea is to add an 'action expert' to a pre-trained visual language model. The whole model is divided into two parts: a VLM and an action expert. The VLM is responsible for processing the visual and natural language inputs while the action expert is responsible for generating actions. The action expert interacts attentively with the VLM at each layer of its network. The model employs a flow matching method based on diffusion, which enables π0 to generate more high-frequency action sequences - meaning that the intelligence's actions will become smoother and more refined. Test results show that π0 outperforms previous VLAs on all types of tasks, and performs equally well in the learning a new task test [22].

Embodied Multi-Modal Agent (EMMA), on the other hand, adopts a completely new way of thinking. In this model, researchers design a visual-textual parallel environment and give the same

DOI: 10.54254/2755-2721/2025.PO24886

embodied task to both environments to generate action results. This model incorporates insights distilled from LLM experts in parallel textual worlds — specifically, a method of enhancing actions through error analysis. Through this approach, EMMA was able to fine-tune the VLM in the visual world of the same task and adapt itself quickly to the dynamic environment. The article innovatively proposes the cross-modal imitation learning algorithm DAgger-DPO, which allows EMMA to generalize to a wide range of new tasks without further guidance from the LLM.EMMA represents the first VLM agent approaching human-level performance (91% success rate) on complex vision tasks, demonstrating the critical role of LLM feedback in enhancing embodied agents' generalization capabilities [23].

4. Challenges

The development of embodied intelligence has not been entirely hurdle-free. In fact, the research related to the combination of embodied intelligence and foundation models is still in a relatively early stage, which not only means that this field has a broad development prospect and many development directions, but also means that many problems need to be solved urgently. There are many challenges in the field of embodied intelligence, the most important of which are the following four directions.

4.1. Physical-semantic alignment problem

The primary challenge confronting embodied intelligence is mapping abstract semantics from foundation models into precise executable physical actions, which is termed the physical-semantic alignment problem. This entails not merely translating model instructions into successful action sequences, but achieving granular motion refinement. For instance: When a robot is tasked to deliver a near-full glass of water, how should the manipulator move to stably lift it? How must the robot navigate to prevent spillage? When grasping deformable objects, what force control minimizes shape distortion?

Addressing this requires the embodied model to establish parameter-level correspondence with physical entities, actions, and environmental variables. Crucially, the model must comprehend real-world physics. Current techniques remain inadequate for such cross-modal alignment, rendering this a core challenge in embodied intelligence development.

4.2. Real-time and computing resource bottleneck

Real-time responsiveness remains a persistent challenge for foundation models. Current foundation models incur substantial latency during computation and inference phases. For embodied agents, these models additionally require extensive environmental data processing and action generation, often exceeding inference latency itself [24]. Cumulative delays may extend end-to-end response intervals to second-level durations, whereas real-world applications, such as emergency obstacle avoidance in autonomous vehicles, demand millisecond-level thresholds.

Computational constraints present another critical bottleneck. Contemporary large models impose significant compute requirements during both training and inference. As embodied agents transition from laboratories to deployment, miniaturization and hardware simplification become imperative. This raises fundamental trade-offs: reducing onboard compute resources to achieve portability may severely compromise functional performance. While cloud-based inference offloading exists, its viability remains contingent on network availability and thus lacks universal applicability.

DOI: 10.54254/2755-2721/2025.PO24886

4.3. Data scarcity

Training datasets for embodied foundation models suffer from high acquisition costs and limited generalizability.

Unlike large language models that leverage widely available books, research papers, and web content, embodied foundation models require multi-modal datasets incorporating environmental contexts, action sequences, and sensory modalities (visual, auditory, haptic). Such data cannot be easily sourced like text or images—it demands real-time collection through robotic platforms in simulated environments, resulting in prohibitively expensive data collection.

Furthermore, while basic underlying data in mainstream VLMs and VLAs demonstrate partial interoperability, task-specific datasets necessitate customized optimization or redesign. The absence of shareable, standardized datasets forces researchers to expend substantial effort on data conversion, adaptation, and compatibility tuning rather than model training.

4.4. Sim-to-real gap

Most existing foundation model datasets are acquired and trained within restricted simulated environments, achieving high task completion rates in idealized settings. However, simulation modeling incurs substantial costs and cannot exhaustively cover all environmental variables. This results in embodied agents lacking exposure to complex real-world data during training, hindering direct deployment in open-world scenarios.

The embodied foundation models discussed in technical routes section demonstrate strong task performance and transferability. Nevertheless, researchers must recognize these impressive results remain simulation-bound—however sophisticated the virtual environments. While robots excel in simulated kitchens, real-world deployment introduces unforeseen challenges like spilled liquids or interacting children. Whether embodied agents can reliably operate in such dynamic open-world contexts remains an unresolved challenge.

5. Prospects

Despite numerous challenges, research on foundation model-based embodied intelligence undeniably remains a pivotal direction in artificial intelligence—likely evolving into the dominant research paradigm in the near term. In response to the four challenges proposed in the previous section, this paper proposes several promising pathways for advancing embodied foundation models, offering reference points for interested researchers.

5.1. Foundation model upgrade

Current embodied foundation models exhibit limited task diversity. While demonstrating proficiency in similar or identical tasks, their generalization capabilities deteriorate significantly when confronted with complex or novel task types. Furthermore, hardware components like specialized manipulators constrain transferability across robotic platforms. Consequently, core research objectives focus on developing adaptable foundation models capable of cross-scenario generalization—tailored to diverse robotic morphologies, tasks, and environments. Primary technical pathways include:

Multi-modal Fusion: Integrating heterogeneous sensory data to construct unified cross-modal representations.

DOI: 10.54254/2755-2721/2025.PO24886

3D Geometric Perception: Employing real-time 3D reconstruction for dynamic physical world modeling.

Notably, the Beijing Academy of AI's RoboBrain achieves cross-scenario multitask deployment and cross-ontology collaboration, adapting to varied environments and devices [25].

5.2. Sim-to-real transfer

This direction directly confronts the Sim-to-Real Gap challenge, fundamentally enabling embodied agents to execute real-world tasks solely through simulation training without prior physical environment exposure. Current methodologies enhance simulated training environments and datasets by integrating realistic interaction data, incorporating critical physical parameters like friction coefficients, and constructing high-fidelity simulations for accelerated learning

At ICRA 2024, a competition-focused team presented a seminal paper addressing sensing-actuation discrepancies during simulation-to-robot transfer for the Sim2Real Challenge. Their approach—combining Structured Multi-modal Mapping Strategy (SMMS) visual policies with Dynamic Feedback (DF) servo controllers—demonstrated significant performance gains, marking substantive progress in this domain [26].

5.3. Human-agent natural interaction

The ultimate target of artificial intelligence should emulate human-like qualities. Beyond perception, reasoning, and action capabilities, humans possess rich emotional intelligence and exist as social beings within interconnected communities. This research aims to equip agents with anthropomorphic interaction abilities—including emotion recognition, natural language dialogue, and social behavior generation. Current focal directions encompass:

Multi-modal affect recognition: Integrating visual, vocal, and physiological signals to detect human emotional states and trigger contextually appropriate responses

Context-aware conversation: Leveraging foundation models' contextual understanding to analyze users' real-time emotions and intentions for dynamic response generation

While contemporary foundation models already incorporate multi-modal affect recognition frameworks, these technologies hold transformative potential for future embodied agents. Ultimately, such advancements will enable embodied intelligence to achieve authentic human-like social presence in real-world interactions.

6. Conclusion

Embodied intelligence, as a pivotal frontier in AI's evolution, demonstrates transformative potential beyond traditional AI through its tight perceptual coupling with physical environments. This paper has articulated the synergistic advantages of integrating foundation models with embodied cognition, surveyed cutting-edge advancements in embodied large models, analyzed critical challenges and emerging opportunities in the field. This paradigm shift from pure computational models toward physical-world interaction marks AI's maturation into an environmentally-grounded discipline. Despite significant hurdles, embodied intelligence exhibits vast application prospects. Future technological progression will enable embodied agents to undertake increasingly complex roles across diverse domains, evolving toward more comprehensive and robust implementations. Meanwhile, breakthroughs in embodied intelligence research will accelerate progress toward Artificial General Intelligence. Such systems may soon become integral components of human

DOI: 10.54254/2755-2721/2025.PO24886

society—whether in disembodied or embodied domains, AI will profoundly permeate virtually every facet of human existence.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. Language Models are Few-Shot Learners. Conference on Neural Information Processing Systems (NeurIPS), 2020, arXiv: 2005.14165.

[2] OpenAI. GPT-4 Technical Report. 2023, arXiv: 2303.08774.

[3] Li Songyuan, Zhu Xiangwei, Li Xi. Survey of embodied agent in context of foundation model. Journal of Zhejiang University (Engineering Science), Volume 59, 2025

[4] Wang Wensheng, Tan Ning, Huang Kai, Zhang Yunong, Zheng Weishi, Sun Fuchun. Embodied intelligence systems based on large models: A survey. Acta Automatica Sinica, Volume 51, 2025.

[5] Alan M. Turing. A. Computing machinery and intelligence. Mind, 1950.

[6] Francisco J. Varela, Evan Thompson, Eleanor Rosch. The Embodied Mind: Cognitive Science and Human Experience. MIT Press, 1991.

[7] Rolf Pfeifer, Josh Bongard. How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press (Bradford Books), 2006

[8] Margaret Wilson. Six views of embodied cognition. Psychonomic Bulletin & Review, Volume 9, 2002

[9] Andy Clark. Being There: Putting Brain, Body, and World Together Again. MIT Press, 1997

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. Conference on Neural Information Processing Systems (NeurIPS), 2017, arXiv: 1706.03762.

[11] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, et al. Highly accurate protein structure prediction with AlphaFold. Nature, Volume 586, 2021.

[12] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, et al. BloombergGPT: A Large Language Model for Finance. 2023, arXiv: 2303.17564.

[13] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, et al. On the Opportunities and Risks of Foundation Models. Stanford Institute for Human-Centered Artificial Intelligence (HAI), 2021, arXiv: 2108.07258.

[14] Rishi Bommasani, Stanford HAI Interview, April 28, 2022.

[15] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. 2022, arXiv: 2204.01691.

[16] Brad Mehlenbacher. Instruction and technology: Designs for everyday learning. MIT Press, 2010.

[17] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, arXiv: 1606.01847

[18] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, et al. PaLM-E: An Embodied Multimodal Language Model. International Conference on Machine Learning (ICML), 2023, arXiv: 2303.03378.

[19] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. Google Deepmind, 2023, arXiv: 2307.15818.

[20] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li. An Embodied Generalist Agent in 3D World. International Conference on Machine Learning (ICML), 2024, arXiv: 2311.12871.

[21] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, et al. OpenVLA: An Open-Source Vision-Language-Action Model. Conference on Robot Learning (CoRL), 2024, arXiv: 2406.09246.

[22] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. π0: A Vision-Language-Action Flow Model for General Robot Control. Physical Intelligence, 2024, arXiv: 2410.24164.

[23] Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, Yuhui Shi, et al. Embodied Multi-Modal Agent trained by an LLM from a Parallel TextWorld. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024, arXiv: 2311.16714.

[24] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, et al. EmbodiedBench: Comprehensive Benchmarking Multi-modal Large Language Models for Vision-Driven Agents. 2025, arXiv: 2502.09560.

DOI: 10.54254/2755-2721/2025.PO24886

[25] Yuheng Ji, Huajie Tan, Jiayu Shi1, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, et al. RoboBrain: A Unified Brain Model for Robotic Manipulation from Abstract to Concrete. 2025, arXiv: 2502.21257.
[26] Ming Yang, Hongyu Cao, Lixuan Zhao, Chenrui Zhang, Yaran Chen. Robotic Sim-to-Real Transfer for Long-Horizon Pick-and-Place Tasks in the Robotic Sim2Real Competition. 2025, arXiv: 2503.11012.