# Multimodal Humanoid Robotic Interaction System

Jinhao Zheng

School of Mechanical Engineering, Tongji University, Shanghai, China

2453136@tongji.edu.cn

**Abstract.** The interactive system of multimodal humanoid robots is a pivotal technology for achieving robotic intelligence. This paper presents a systematic review of the key cutting-edge technologies, difficulties, challenges, and development prospects pertaining to such interactive systems. Firstly, the paper enumerates the core technologies in this field, including the application of visual-language models in robot training, long-context multimodal command navigation, learning to imitate human movements, force and torque sensing control, and multimodal intent fusion technology. Insights into the current challenges faced by the field are also derived from these technologies, such as complex system integration, weak environmental adaptability, insufficient safety, and limitations in hardware performance. Finally, the paper looks ahead to the future development of multimodal humanoid robots, proposing that the in-depth integration of large models with embodied intelligence, multimodal interaction, and the coordinated development of software and hardware are the keys to realizing robotic intelligence. The paper argues that the interactive system of multimodal humanoid robots will significantly enhance the level of robotic intelligence, find wide application in various fields, and greatly drive the development of productivity.

**Keywords:** Multimodal interaction; humanoid robot; vision-language model; human-robot collaboration; embodied intelligence.

## 1. Introduction

Although more than a decade has elapsed since the first modern humanoids were unveiled, progress has remained incremental, constrained by the ceilings of contemporary materials, energy storage, and control theory. Yet scholarly fascination and industrial investment continue to swell, propelling a steady cadence of incremental breakthroughs.With the continuous advancement of artificial intelligence, the development of humanoid robots has also accelerated, rendering them increasingly intelligent. Intelligent robots are poised to provide increasingly sophisticated assistance to humanity; consequently, the pursuit of robotic intelligence has become a focal domain within the scientific and technological community. The multimodal humanoid robotic interaction system constitutes a pivotal enabling technology for the realization of advanced robotic intelligence. Indeed, by equipping humanoid platforms with multimodal interaction capabilities, the system enables robots communicate with humans in a human-like manner, thereby allowing users to convey their intentions effortlessly and ensuring that assigned tasks are executed with high fidelity. Accordingly, this paper concentrates exclusively on the multimodal humanoid–robot interaction system as its central technological focus.

Substantial progress has recently been achieved within the scientific community in this domain. For instance, [1] presents a pragmatic integration of off-the-shelf vision–language models for robotic grasping, empirically validating the feasibility of “out-of-the-box” manipulation in unstructured domestic environments, thereby substantiating the practical viability of domestic service robots [1]. Concurrently, [2] introduces the first Vision-Language-Action (VLA) model capable of navigating in response to complex, multimodal instructions (e.g., “go to where I was snacking last night”), substantially elevating the naturalness of human–robot interaction [2]. Furthermore, [3] proposes a method that enables a humanoid robot to reproduce a repertoire of dexterous manipulation skills—such as typing or playing tennis—after only one hour of human motion demonstration, dramatically enhancing human–robot interactivity and allowing any user to act as the robot’s “instructor” [3]. These advances collectively demonstrate that a multimodal

humanoid–robot interaction architecture is technically attainable and constitute a decisive milestone toward fully intelligent robotic systems.

This paper presents a systematic review. Accordingly, the remainder of this paper is organized as follows. First, this paper surveys and critically analyzes the state-of-the-art technologies in the field, with particular emphasis on the key enabling techniques underpinning multimodal interaction systems. Next, this paper distills the prevailing technical limitations and delineates prospective directions for refinement and future development. Finally, this paper presents a forward-looking perspective on the realization of these technologies. This tripartite structure constitutes the principal argumentative trajectory of the present study.

## 2. Methods

### 2.1. Integrating Vision–Language Models into Multimodal Robotics

Vision-language models are a mature AI technology; porting them to robotics makes machines both more multimodal and more intelligent. [1] embeds such a model in robot training and proposes an open-knowledge framework that almost doubles task-success rates. By associating images with natural language the way humans do, the robot gains the flexibility and efficiency needed for truly intelligent operation [1]. Early robots followed bare-bones commands and worked only in pre-mapped zones where every object had a fixed label. Once a vision-language model is wired in, the machine can train itself: told to “get an apple,” it pairs the word with any apple it sees, then plans and executes the grasp. From that moment on, an apple on any counter, in any light, is still an apple—and the robot still knows what to do. That kind of open-world competence is what will finally make robots useful outside the lab and inside everyday life.

### 2.2. Long-Context Multimodal Instruction Navigation with Vision-Language Models

Current AI-based VLM models are equipped with long-context processing capabilities, namely a memory function, which enables them to continue discussing topics that users have brought up in previous conversations. Applying this capability to robots will significantly enhance their level of intelligence [2]. This article presents an excellent case in point. The research focuses on addressing the challenge of robots' long-context multimodal instruction understanding. Among its key components, the "Mobility VLA" navigation strategy consists of long-context VLMs, endowing the robot with environmental comprehension and common-sense reasoning abilities. It is a vision-language model capable of processing long video sequences [2]. For instance, if a user asks the robot to find a book left on the sofa the previous night, the model can integrate previously recognized images with the textual description provided by the user to locate the desired item. Unlike traditional robots that only respond to simple instructions, robots integrated with long-context vision-language models can connect their current visual observations with past memories and complex linguistic inputs. Their intelligence is even sufficient to assist humans in completing all household chores, representing a substantial leap in capability.

### 2.3. The Imitation of Human Movements by Robots

In a multimodal humanoid robot interaction system, motion imitation serves as a crucial interaction mode. Just like humans, when learning a new dance move or a new sport, it refers to and imitate others' movements, which enables us to master new actions more quickly [3]. This article demonstrates how a research team from Stanford University has enabled robots to imitate human movements. They employed algorithms to allow robots to recognize human actions and map human joint movements to their own. Furthermore, the study indicates that using a relatively small amount of demonstration data enables robots to learn a series of dexterous movements, significantly enhancing their learning efficiency [3]. Therefore, when robots need to perform a new task and learn new movements, they no longer have to go to great lengths to write control codes as did for traditional robots. Instead, only need to perform the corresponding actions, and the robots can

learn—making the process as simple as teaching a real human being. This also allows robots to interact better with humans and serve as more capable human companions.

### 2.4. Multimodal Control of Robots Achieved by the Combined Use of Force, Torque, and Tactile Information

In the technology of multimodal humanoid robots, the robot's perception of force and touch is also a crucial component. While visual and language perception have equipped robots with considerable capabilities, they still fall short of enabling robots to complete tasks accurately in certain specialized scenarios. Take picking up a piece of jelly as an example: this is trivial for humans, as it is possible to control the force applied to pick it up without breaking it. For robots, however, this is difficult to achieve due to their lack of tactile sensing capabilities. Therefore, endowing robots with tactile perception is a vital aspect of multimodal humanoid robot development [4]. This article introduces just such a technology. In the study, researchers proposed a reinforcement learning model that integrates force, torque, and tactile information, which significantly enhances the robot's adaptability to objects of varying stiffness [4]. This method improves the stability of the robot's motion execution and strengthens its ability to control object deformation. This enables robots to move beyond mere brute force and instead apply different levels of force to different objects, just like humans do—making them more efficient and safer.

### 2.5. Application of Multimodal Intent Fusion Methods to Multimodal Robots

One of the ultimate goals of multimodal humanoid robots is to understand human intentions and thereby better assist humans. A robot that can grasp user intentions can provide more accurate and seamless services, flexibly adapt to user preferences, and thus achieve intelligent human-robot interaction [5]. This article presents precisely such an approach. In the study, the researchers proposed a Human-Robot Interaction Multimodal Intent Fusion (HIL-MIF) algorithm [5]. This algorithm integrates multimodal information from gaze, gestures, and language, introduces weight factors to evaluate the importance of each modality, and dynamically adjusts these weight factors based on user feedback. In doing so, it achieves high-precision intent understanding and enhances the personalization and reliability of human-robot interaction. This significantly improves the intelligence of robots, allowing them to truly become human companions—because it enables robots to comprehend human intentions and adjust their task-execution strategies according to different users' preferences. In turn, this facilitates more complex human-robot interaction, enabling robots to truly "understand" human commands.

## 3. Discussion

### 3.1. Challenges

The research on multimodal humanoid robot interaction systems is highly challenging. First, there is its complexity: such systems need to integrate multiple modalities including vision, language, memory, and touch, yet these modules often operate independently. Designing a system that enables the collaborative work of these multimodalities—while avoiding information conflicts and delays—thus presents immense research difficulties, with extremely high requirements for both algorithms and computing power.

Furthermore, even if system-level issues are resolved, ensuring the robot’s robustness when applied to real-world scenarios remains problematic. The real world features complex and ever-changing environments, and the information transmitted to the robot is subject to multiple interferences; whether the robot can maintain system stability when facing various unexpected situations is also a critical challenge.

Safeguarding the robot’s safety is another vital concern. Since multimodal interactive humanoid robots need to interact with humans, it is essential to ensure that the robot does not harm humans and that the user’s safety is protected during interaction.

Additionally, hardware development constitutes one of the bottlenecks in the advancement of multimodal humanoid robots. For instance, vision and touch both require high-performance sensors, and the performance ceiling of these sensors ultimately determines the robot's performance limits.

Data scarcity also constitutes a significant challenge. The training of multimodal interaction models often requires massive volumes of data, even annotated data. However, unlike large models that can source data from the Internet, such extensive real-world data is notoriously difficult to acquire, rendering the training process extremely challenging.

Lastly, the generalization capability of the model also poses a crucial challenge. Tasks in the real world are highly variable, especially when applied in environments beyond factories, where scenarios are uncertain and factors such as object positions change constantly. Consequently, robots exhibit insufficient generalization ability when transitioning across tasks and scenarios, and may even suffer from overfitting. This presents a formidable challenge to the practical application of robots.

### 3.2. Future prospects

Multimodal interactive humanoid robots hold exceptionally broad prospects for development. As artificial intelligence technology continues to make breakthroughs, it will greatly facilitate the in-depth integration of large models and embodied intelligence [6-9]—specifically, the application of multimodal interaction technology to humanoid robots. This integration will endow robots with cognitive and decision-making capabilities comparable to those of humans. This enables robots to achieve genuine situational understanding and autonomous decision-making

The future development of robots will focus on intelligence, and the multimodal humanoid robot interaction system is undoubtedly the key technology to enhance robots' intelligence level [10]. Equipped with multimodal interaction technology, robots will be better able to execute human commands, learn to perform various tasks with human-like agility, and comprehend human intentions. By introducing models such as attention mechanisms and graph neural networks, robots will be enabled to achieve cross-modal semantic alignment for reasoning. This, in turn, enhances the accuracy of intent understanding, thereby making human-robot interaction more natural.

With the continuous advancement of electronic technology, the performance of hardware such as flexible electronics and neuromorphic computing will gradually improve, which will significantly enhance robots' perceptual capabilities and action execution abilities. The improvement in hardware also enables more efficient algorithm implementation; the coordinated development of software and hardware allows robots to achieve lightweight algorithm design and edge computing deployment. This enables robots to maintain performance in constrained environments and accelerates their popularization in practical scenarios.

Once robots acquire such capabilities, they can be applied across diverse fields. Smart manufacturing, medical care, and household services, among others, will all become arenas where robots can fully demonstrate their capabilities. Humans will significantly reduce their workload, free themselves from trivial matters, and focus on more important tasks—even highly dangerous tasks can be assigned to robots. This will propel human society to a new level.

## 4. Conclusion

This paper offers a comprehensive review of interactive systems for multimodal humanoid robots. It provides detailed coverage of cutting-edge technologies in this domain, including visual-language perception models, robots' long-text memory capabilities, imitation learning of human movements, and force/torque sensing technologies. The advancement of these technologies has significantly driven the progress of multimodal humanoid robots, bringing about an unprecedented enhancement in their level of intelligence. While certain difficulties and challenges remain—such as complex system integration, poor environmental adaptability, and insufficient safety—these issues will be gradually addressed with the continuous development of large models

and embodied intelligence technologies. It is believed that in the near future, multimodal humanoid robot technology will be fully realized and widely applied across various fields. By then, robots will become indispensable partners to humans and propel the productivity of human society to new heights.

## References


[1] Liu P, Orru Y, Vakil J, Paxton C, Shafiullah NM, Pinto L. Ok-robot: What really matters in integrating open-knowledge models for robotics. arXiv preprint arXiv:2401.12202. 2024 Jan 22.

[2] Chiang HT, Xu Z, Fu Z, Jacob MG, Zhang T, Lee TW, Yu W, Schenck C, Rendleman D, Shah D, Xia F. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. arXiv preprint arXiv:2407.07775. 2024 Jul 10.

[3] Fu Z, Zhao Q, Wu Q, Wetzstein G, Finn C. Humanplus: Humanoid shadowing and imitation from humans. arXiv preprint arXiv:2406.10454. 2024 Jun 15.

[4] Zhou Q, Feng B, Li B, Liu C, Chen Y, Bi Y. Enhancing Human-Robot Collaborative Transportation of Deformable Objects using Multi-modal Reinforcement Learning and Adaptive Admittance Control. Advanced Engineering Informatics. 2026 Jan 1;69:103905.

[5] Jiang S, Zhang W, Liu J, Li H, Wang Z, Zhou Y, He B. A novel human-in-the-loop multimodal intention fusion method for human-robot interaction. IEEE Transactions on Automation Science and Engineering. 2025 Jun 17.

[6] Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, Du Y. A survey of large language models. arXiv preprint arXiv:2303.18223. 2023 Mar 31;1(2).

[7] Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008 Sep 1;95(3):759-71.

[8] Cangelosi A, Bongard J, Fischer MH, Nolfi S. Embodied intelligence. InSpringer handbook of computational intelligence 2015 (pp. 697-714). Berlin, Heidelberg: Springer Berlin Heidelberg.

[9] Gupta A, Savarese S, Ganguli S, Fei-Fei L. Embodied intelligence via learning and evolution. Nature communications. 2021 Oct 6;12(1):5721.

[10] Stiefelhagen R, Ekenel HK, Fugen C, Gieselmann P, Holzapfel H, Kraft F, Nickel K, Voit M, Waibel A. Enabling multimodal human–robot interaction for the karlsruhe humanoid robot. IEEE Transactions on Robotics. 2007 Oct 8;23(5):840-51.