



MULTIMODAL PERCEPTION IS THE FOUNDATION OF HUMANOID INTELLIGENCE

Abstract

This paper explains that multimodal perception is essential for humanoid intelligence, which involves not just sensing but also reasoning, decision-making, and action in human-like robots. While unimodal systems rely on a single input and struggle in complex environments, multimodal systems combine vision, sound, language, and other data to achieve better understanding and adaptability. The study highlights that intelligent humanoid behavior emerges from the integration of perception, reasoning, and action, making multimodal perception the foundation for advanced, autonomous AI systems.

1. Introduction

Humanoid intelligence refers to the decision-making ability of an embodied AI agent that perceives, comprehends, reasons, and interacts with its environment through ongoing sensory input. A system that depends solely on a single sensor or modality lacks the contextual awareness necessary for precise and dependable decision-making, as crucial environmental data may be absent. In contrast, multimodal intelligence combines various sensory inputs such as vision, speech, touch, and environmental signals, facilitating a more thorough understanding of the surrounding world. The significance of humanoid intelligence is swiftly growing across multiple fields, including healthcare, domestic assistance, disaster rescue operations, industrial manufacturing, and service sectors. As humanoid systems become more embedded in everyday life and industrial settings, the demand for strong multimodal perception and reasoning becomes increasingly vital. Recent progress in large-scale vision-language frameworks, like GPT-4V, has showcased the capabilities of multimodal AI systems. These frameworks can process and comprehend both visual and textual information at the same time, enabling AI systems to interpret complex environments more efficiently. Such advancements mark a significant leap toward intelligent humanoid systems that can perceive and reason in ways that are more akin to human cognition. Over the decades of robotics research, humanoid robots have gradually transitioned from basic mechanical systems to more advanced platforms. While many humanoid robots are designed to physically resemble humans, they have historically faced challenges in executing tasks that humans perform naturally and efficiently.

However, the rise of generative AI, large language models (LLMs), and multimodal learning systems has greatly enhanced the capacity of humanoids to learn, adapt, and carry out complex

tasks. Currently, humanoid robots are being utilized as assistants in various sectors including manufacturing, hospitality, customer service, and other practical applications. This swift advancement indicates a significant shift in technology: firstly, moving from traditional AI-based robotics to humanoid-focused intelligence, and secondly, transitioning from standard task-oriented robots to sophisticated adaptive humanoid systems. Recent advancements in deep learning and generative AI, especially the emergence of multimodal large language models (MLLMs) that can merge text, vision, speech, and contextual comprehension, have propelled the evolution of humanoid robotics. These developments are enhancing humanoid AI systems beyond mere human-like appearances by facilitating multimodal perception, interactive communication, adaptive behaviors, and real-time decision-making. Contemporary humanoid systems are progressively crafted to integrate human-like embodiment with expressive features such as facial expressions, gestures, emotional engagement, and context-sensitive responses driven by multimodal sensory integration.

2. Keywords

Humanoid Intelligence, Multimodal, Unimodal, Language Learning Model (LLM), Vision language model (VLM), Vision Language action (VLA).

3. What Humanoid Intelligence Structurally Requires

Before assessing any architecture, we must define the target. [Cao \(2024\)](#) proposes six evolutionary stages of humanoid development. The diagram below illustrates this progression and its perceptual demands at each layer.

Six Stages of Humanoid Development (Adapted from [Cao, 2024](#))



Fig no: 1 Stages of Humanoid Development

Six stages of humanoid development and their perceptual requirements (Cao, 2024). Stages 2–6 are inherently multimodal.

The key observation from above stages. Stage 1 — constructing a human-looking physical body — can be achieved without multimodal perception. Every subsequent stage requires at least two sensory channels working in concert. Humanoid intelligence, located at Stage 6, is the synthesis of all prior stages. It therefore requires all their perceptual resources simultaneously.

For a system to achieve humanoid intelligence, it must be capable of simultaneously processing and fusing multiple sensory modalities into a coherent world model that supports real-time

decision-making and physical action. This is precisely the definition of multimodal perception. Humanoid intelligence requires multimodal perception by structural necessity.

4. Why Unimodal perception cannot achieve Humanoid Intelligence

4.1 The Unimodal Architecture and Its Inherent Ceiling

A unimodal learning system processes a single input type through an encoder and model to produce an output. Its architecture is straightforward and efficient within its domain.

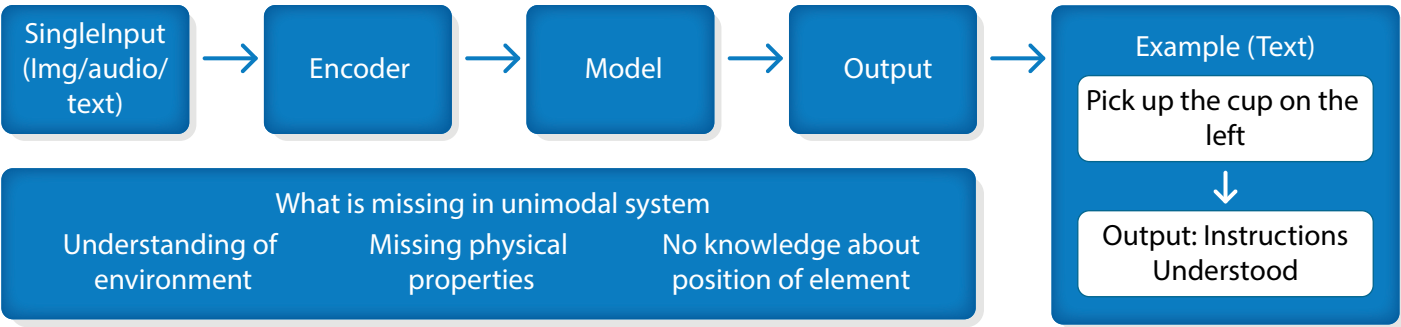


Fig no: 2 Unimodal Pipeline

With this we can be able to achieve the unimodal related task accurately as it focus only on required topic it performs very well in narrow tasks such as image classification, speech transcription, or text generation. However, Zhao et al. (2025) observe that achieving accurate perception and robust decision-making in human-robot interaction remains deeply challenging precisely because human behavior is inherently complex, dynamic, and variable — qualities that no single sensor can adequately capture.

4.2 Three Categorical Failures of Unimodal Systems

The failures of unimodal systems in humanoid contexts are not marginal, they are categorical. Wang (2025) identifies the core problem like when a humanoid process the image without any context it will give wrong prediction and it leads to failure.

Failure 1 - Ungrounded Language: When an LLM is trained only as a language only system can process the text with high accuracy but cannot give instructions to physical objects. Without vision

and depth sensing, a command such as “hand me over the glass of water on the top of table” where it doesn’t provide no actionable information about position, distance and object identity the instruction can be understood but not deliverable.

Failure 2 - Unresolvable dependency: A single sensor provides a single perspective. Take an example of X-ray diagnosis where it can understand the report, but we cannot provide our feedback or our context to by this it will lead to incorrect output.

Failure 3 - Insufficient Multimodal Intent Recognition: Human communication itself is multimodal. When a person in the middle of the speech pointing an object and says “pass me the object” here the single modality cannot understand neither “that” nor speech alone itself. Zhao et al. (2025) highlighted in their survey: during HRI, people simultaneously use body language, speech and facial signals as part of a single integrated communicative act. By removing any one of these channels from the robot’s perception does not simply reduce accuracy it removes entire categories of information.

Dimension	Unimodal System	Multimodal System
Input channels	Single (text / image / audio)	Multiple (vision + language + touch + and more)
World model completeness	one perspective only	More than one perspective
Language grounding	Abstract — no physical reference	Grounded — language linked to visual scene

Table 1. Comparison of unimodal and multimodal systems across dimensions critical to humanoid intelligence

5. The AI Architecture Evolution as Evidence for Multimodal Necessity

The Progression of AI models from language only to vision language to vision language action It is a record of researchers discovering,

at each stage, that the previous architecture was categorically insufficient for the next class of tasks that humanoid operation requires. Each step is developed to resolve the limitations that occur in previous steps.
Evolution towards humanoid intelligence

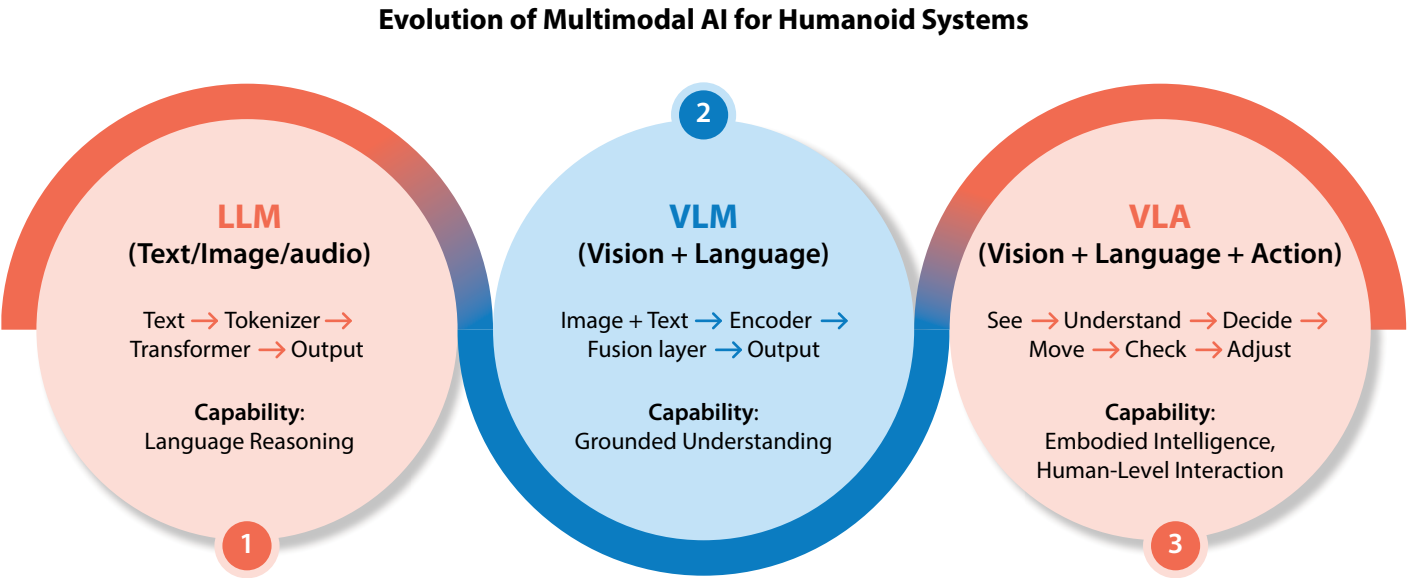


Fig no:3 Evolution of Humanoid Intelligence

Here we can understand the evolution of Humanoid Intelligence from LLM to VLM to VLA

5.1 Large Language Models:

LLM are trained under a large-scale text dataset where we can generate text by predicting tokens and will receive the output.

For humanoid robot command itself is not enough because the robot can understand the command but cannot navigate to the respective place with the help of command it needs some external push this is not an limitation in LLM it is an architectural gap.

5.2 Vision-Language Models:

VLM resolves the issue of grounding gap by fusing Vision and Language this will resolve such as visual questioning and understanding the situation and more. With the help of VLM we will be able to perform the tasks which LLM isn't able to do.

[Zhao et al. \(2025\)](#) specifically tells large-scale vision-language frameworks as a main driver of the recent advances in robotic decision-making capability. The addition of vision to language enables robots to locate objects and ground verbal commands in observed physical reality. Each of these capabilities is impossible without the second modality.

5.3 Vision–Language–Action (VLA) Models

A Vision Language Action (VLA) model combines vision, language, and action so that a robot can see the environment, understand human commands, and perform physical actions. In simple terms, a robot uses its camera to see, its AI model to understand instructions, and its motors to act in the real world.[Cao\(2024\)](#)

Flow of VLA: See → Understand → Decide → Move → Check → Adjust

In VLAs, directly informs decision-making, and actions generate feedback that updates future perceptions. This loop interaction is important for humanoid robots. This shows why multimodal is a foundational layer that enables humanoid intelligence.

6. Why Multimodality is Foundational, Not Optional:

This is the central claim of this paper. The main difference between optional and foundational comes when the absence makes the target impossible, not a less efficient or unreachable to the goal. This section builds that argument in four steps.

6.1 The Foundation Test: What Does 'Foundation' Actually Mean?

A capability is considered foundational when the system becomes impossible without it. In humanoid intelligence, multimodal systems is not simply an optional improvement or it is not an

performance enhancement it is necessary. Removing it makes the target unachievable, and no alternative can replace its functionality. With all this understanding we can prove why multimodal is an foundation.

6.2 Condition 1 — Without Multimodal Perception, Humanoid Intelligence Cannot Function

Consider what a humanoid robot must do to operate in an unstructured human environment like locating the objects in world map while parallel understanding verbal instructions about those objects. It must adjust grip force in real time based on the command which user provides. None of these tasks are possible with one sense. [Zhao et al. \(2025\)](#) confirm in their survey due to limitations of unimodal they aim to fuse the different information from different sensors to handle complex cases.

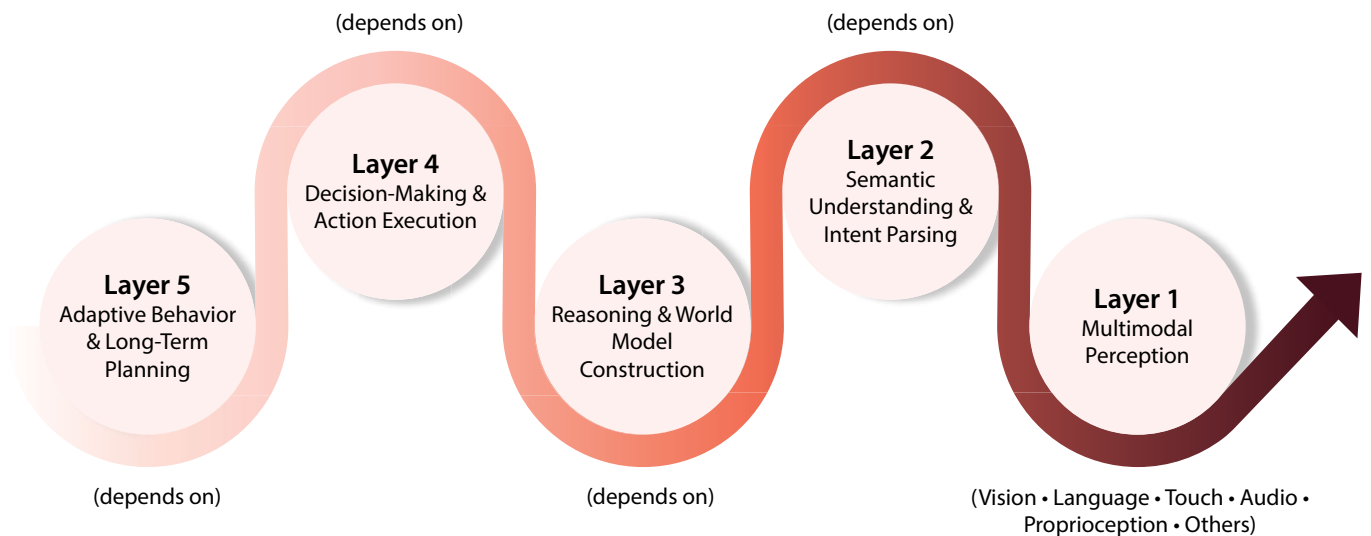
A system with vision but no language understanding cannot receive verbal instructions. A system with language but no vision cannot locate physical referents. A system with but no haptic feedback cannot regulate grip.

6.3 Condition 2 — No Unimodal Substitute Can Replace Multimodal Perception

Yes. Even if we use an very powerful language model we cannot replace the multimodal perception. A model cannot reason about information it has never received. A vision cannot understand the sound of the object. A language only system cannot determine where an object is on world map. These are not the limitations. These are the current unimodal architecture design. [Wang \(2025\)](#) confirm in their survey the system that lack cross modal alignment cannot produce semantic representations of environments that present information across multiple channels parallelly. For humanoid robots this absence is disqualifying.

6.4 Condition 3 — Every Capability of Humanoid Intelligence Depends on Multimodal Perception First

This condition proves that why multimodal is foundation. Multimodal perception satisfies this condition because it is the input layer — and intelligence begins with input.



According to [Zhao et al.](#), the removal of the foundational perception layer results in a complete failure of the understanding and decision-making process. In such scenarios, decision-making is not only diminished in effectiveness but also becomes fundamentally unreliable, which leads to the breakdown of higher-order capabilities such as reasoning, contextual understanding, and adaptive action.

Unimodal perception has been crucial in earlier AI and robotics research and continues to underpin many current applications. However, when it comes to humanoid intelligence, unimodal systems encounter significant limitations because humanoid systems necessitate highly interactive behavior, swift decision-making, and contextual understanding derived from various sensory inputs. Due to their restricted information-processing capabilities, unimodal systems struggle to simultaneously interpret visual data, contextual meanings, environmental signals, and cues from human interactions effectively. Consequently, they frequently experience diminished accuracy, inadequate adaptability, and unreliable performance in intricate real-world situations.

Moreover, human communication and interaction with the environment inherently involve multiple signals, including language, emotion, gestures, intentions, actions, and surrounding contextual information. A single modality cannot adequately capture or interpret these intricate human-environment interactions, especially in Human-Robot Interaction (HRI) contexts.

To address these challenges, recent research has increasingly concentrated on multimodal perception, which amalgamates information from various sensory sources such as vision, audio, touch, environmental signals, and other contextual inputs. Through the integration of multimodal data, humanoid robots can enhance their understanding of spoken commands, interpret their environments more effectively, reason about tasks, and translate

that understanding into suitable physical actions via motor systems and control mechanisms. This integration establishes a seamless link among perception, reasoning, and action, allowing robots to engage more naturally and effectively with humans and their surrounding environments.

Furthermore, multimodal perception greatly enhances environmental awareness, contextual comprehension, adaptability, and decision-making abilities in humanoid systems. Consequently, it improves the capacity of humanoid robots to work alongside humans in daily tasks and ever-changing settings. From this viewpoint, multimodal perception is not merely designed to overcome the shortcomings of unimodal systems; rather, it is an essential prerequisite for attaining humanoid intelligence itself. The evidence and discussions provided earlier illustrate that multimodal intelligence is not a mere enhancement but the fundamental basis upon which intelligent humanoid behavior is constructed.

7. Real-World Scenario (HealthCare):

7.1 Unimodal:

Now a days we are using this unimodal learning behavior in every corner where we need single behavior needed like HealthCare. We use in terms of diagnosis the patient's health with the help of reports, x-ray and more. The main drawback is it will understand only the reports if it shows good it will give context as patient is fine, but the patient may not be fine due to some other symptoms. Here the image is used not in a doctor's context or patient context.

Eg: Ada Health and more.

7.2 Multimodal:

As we see the issue with unimodal perception, we can still resolve them with multimodal systems. We provide in a way where a doctor or patient provides the image, symptoms and more context about the condition our decision becomes more reliable, but we have limitations too in this. Fusion complexity if the combining the models are wrong the output becomes worse. Data alignment issues.

Eg: ChatGPT, Gemini and more.

8. Open Challenges

8.1 Adaptation

It is important for a robot to continual learning because every day every minute the environment will get changed so It need to grasp every change like it was trained on a even surface but when come's to reality surface may get changed. If it is working in manufacturing unit, the instrument shape differs from one to other, so it needs to be trained continuously. [Wang \(2025\)](#)

8.2 Synchronization Issues

Currently, humanoid robots lose critical information between sensor readings and data fusion. Before the data from multiple sensors is properly combined, its system receives incomplete data and inconsistent inputs which leads to miscommunication between sensors. This causes incorrect actions and unexpected behavior.

9. Future Enhancements

Multimodal integration will be central to this progress. Future will seamlessly fuse inputs from multiple sensors into a unified perception, enabling the robot more accurately than expected.

Coming next, we can build good architecture to decrease sensor failure and noise. On Intelligence side, continual learning will allow humanoids to experience across tasks and environments without forgetting pretrained data.

10. Conclusion

From this viewpoint, multimodal perception is not just a facet of humanoid intelligence; it serves as its essential foundation. Intelligence does not arise solely from language processing, logical reasoning, or planning abilities. Instead, it emerges from the capacity to perceive the environment coherently through various sensory modalities, establish meaning through contextual understanding and experience, and execute actions based on integrated perception and reasoning.

Humanoid intelligence is cultivated through the ongoing integration of perception, reasoning, and action, rather than depending on any single capability in isolation. Much like human cognition, where behavior and decision-making rely on multisensory integration, intelligent humanoid systems must also be rooted in multimodal perception to operate effectively in intricate real-world settings.

The healthcare case study further illustrates the notable distinction between systems functioning with a single sensory modality and those employing multiple sensory modalities. Multimodal systems offer enhanced contextual awareness, adaptability, reliability, and decision-making abilities compared to unimodal systems.

Consequently, multimodal intelligence is not merely an optional enhancement for humanoid systems; it is the fundamental prerequisite for humanoid intelligence itself. In the absence of multimodal perception and integration, genuine humanoid intelligence cannot be realized.



References

- [1] A. Pal, "LLM, VLM, and VLA," Medium, Feb. 20, 2025. [Online]. Available: <https://medium.com/@arpipal2/llm-vlm-and-vla-d758b91479eb>.
- [2] L. Wang, "Multi-scenario reasoning: Unlocking cognitive autonomy in humanoid robots for multimodal understanding," *TechRxiv*, 2025. Arxiv: [2412.20429] [Multi-Scenario Reasoning: Unlocking Cognitive Autonomy in Humanoid Robots for Multimodal Understanding](https://arxiv.org/abs/2412.20429)
- [3] OpenAI, "GPT 4V(ision) system card," 2023. [Online]. Available: <https://openai.com/index/gpt-4v-system-card/>
- [4] L. Cao, "Humanoid robots and humanoid AI: Review, perspectives and directions," *ACM Computing Surveys*, 2024. <https://arxiv.org/abs/2405.15775>
- [5] W. Zhao, K. Gangaraju, and F. Yuan, "Multimodal perception-driven decision-making for human-robot interaction: a survey," *Frontiers in Robotics and AI*, vol. 12, Aug. 2025. <https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2025.1604472/full>

About the Authors



Charan Kumar Yadav Mondem is an Emerging Tech Engineer, specializing in humanoid robotics, research, and patent development. His work focuses on innovations at the intersection of artificial intelligence and physical machines.

He is actively engaged in research initiatives and patent-driven innovation, contributing to advancements in next-generation robotic technologies.

For more information, contact askus@infosys.com

Infosys
Navigate your next

© 2026 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.