

Multi-Model Large Language Models for Human-Robot Interaction

Amarnath Awasthi¹, Abhinav Singh², Aastha Gupta³, Aditya Prajapati⁴

Assistant Professor, Axis Institute of Higher Education¹

DOI: <https://doi.org/10.47772/IJRISS.2026.100900212>

Received: 09 September 2026; Accepted: 14 September 2026; Published: 06 October 2026

ABSTRACT

Human-robot interaction (HRI) becomes challenging when robots must interpret human instructions expressed through multiple forms of communication. People often combine speech, pointing gestures, visual cues, and contextual information, and these signals may be incomplete, ambiguous, or inconsistent. Multimodal large language models (MLLMs) can combine such inputs to support more natural interaction, but unreliable interpretations and hallucinated information may still lead to incorrect robotic decisions. This paper proposes a reliability-focused multimodal framework for MLLM-based HRI that integrates voice, pointing gestures, visual perception, and contextual cues for human intention and object understanding. The framework processes information from each modality and combines the resulting evidence within the MLLM reasoning stage. It then examines whether the modalities support the same interpretation, identifies conflicts, estimates uncertainty, and verifies the predicted interpretation against available visual and sensor evidence. When the evidence provides sufficient support for a safe decision, the system selects the appropriate response. When the information is ambiguous, contradictory, or poorly supported, the system asks the user for clarification instead of making an unsupported assumption. Contextual and emotional cues are used as additional evidence to improve interpretation rather than as definitive measures of a person's internal state. The proposed framework is evaluated through controlled HRI scenarios involving ambiguous instructions, conflicting speech and gestures, contextual changes, and hallucination-related cases. Evaluation considers intention accuracy, conflict-detection accuracy, hallucination rate, clarification effectiveness, response time, task-completion rate, and unsafe-action rate. By combining multimodal reasoning with uncertainty estimation, verification, and clarification, the proposed framework aims to improve the reliability, safety, and trustworthiness of MLLM-based human-robot interaction.

Keywords— Human-Robot Interaction, Multimodal Large Language Models, Human Intention Understanding, Multimodal Fusion, Conflict Resolution, Hallucination, Uncertainty, Clarification.

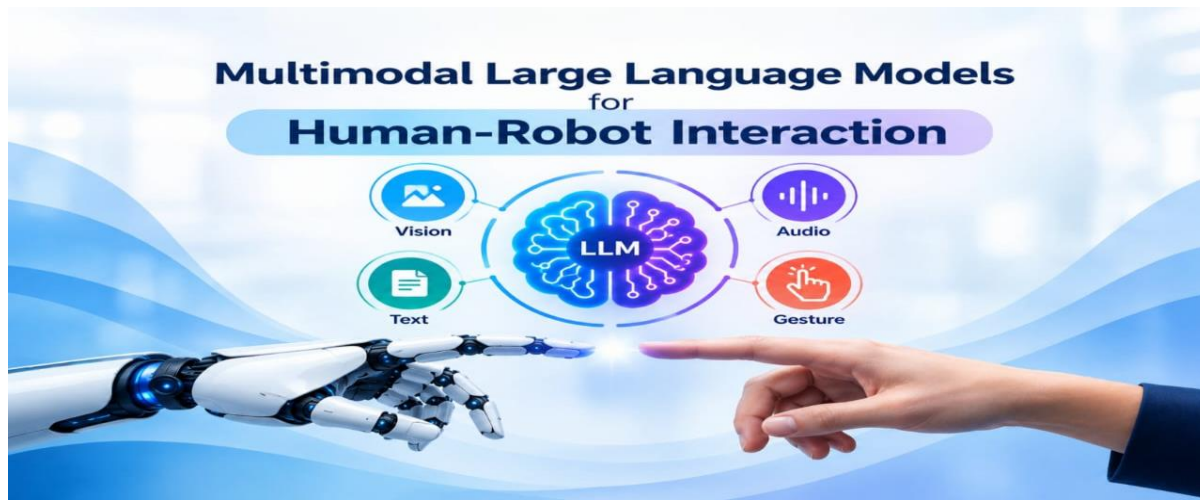
INTRODUCTION

The integration of Large Language Models (LLMs) into robotic systems has opened up new avenues for human-robot communication. Traditionally, robots rely on a suite of individually designed modules, each with its own distinct function - one for deciphering user intentions, another for planning tasks, and a third for generating corresponding actions. In contrast, LLM-based approaches can decipher natural, spoken instructions and adapt them into tangible actions, thereby reducing the need for painstakingly programmed rules and bespoke pipelines. The development of multimodal LLMs has significantly expanded these capabilities, enabling the simultaneous processing of information from diverse sources, such as text, images, audio, and sensor data from the robot itself. This allows a robot to seamlessly integrate spoken instructions with visual cues from its surroundings, thereby identifying relevant objects, understanding its environment, and determining an optimal course of action.

Notably, some recent approaches have incorporated collections of pre-defined skills, actions, or demonstrations, which can be leveraged to construct novel solutions for unfamiliar tasks without the need for exhaustive programming. This not only streamlines the development process but also enables robots to perform with greater ease and fluency, particularly during interactive scenarios. The enhanced coordination between a robot's physical movements and verbal responses is yet another benefit of these integrated approaches.

Researchers have also explored the potential of combining LLMs with vision-language models and speech-recognition systems to create more natural, conversational interactions between humans and robots. Initial results

from these systems are promising, with one notable example achieving an accuracy rate of approximately 87.5% in understanding spoken commands and a success rate of around 86.3% in completing corresponding tasks. While these advancements are significant, they also underscore the limitations of current LLM-powered robots, which remain prone to errors and inconsistencies.



One of the primary challenges hindering the deployment of these robots in real-world settings is their ability to interpret contextual information, adapt to diverse user preferences, and respond to complex, unscripted situations. In real-world environments, robots must contend with unexpected scenarios, human unpredictability, and a multitude of unseen variables, all of which can compromise their performance and reliability. Therefore, it is crucial to continue refining the **robustness, safety, adaptability, privacy, and reliability** of LLM-based robotic systems to ensure seamless, trustworthy interaction in everyday human-robot applications.

Background / Conceptual Foundations

Large Language Models

Large Language Models (LLMs) have expanded the capabilities of intelligent systems by enabling natural-language understanding, contextual interpretation, and instruction-based reasoning. In robotics, these capabilities allow users to communicate tasks using flexible language rather than relying exclusively on predefined commands. However, language alone is often insufficient for interaction in physical environments because robotic systems must also understand the objects, people, and situations associated with an instruction. Therefore, the use of language models in HRI requires their linguistic reasoning to be connected with environmental information.

Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) extend language-based reasoning by incorporating information from multiple modalities, such as text and visual observations. This enables a model to relate a linguistic instruction to information contained in a visual scene rather than interpreting the instruction independently. Embodied multimodal models such as PaLM-E demonstrate the integration of language with visual and sensor information for reasoning about physical environments. This capability is particularly relevant to HRI because human communication frequently combines verbal and non-verbal information.

For the proposed research, multimodal reasoning provides the foundation for combining speech, pointing gestures, visual observations, and contextual information. Instead of treating each modality as an isolated source, the system considers their relationships to construct a more complete interpretation of the user's instruction.

Vision-Language-Action Models

Vision-Language-Action (VLA) models extend multimodal understanding toward robotic action by connecting visual and linguistic information with executable behavior. This approach is important for embodied systems

because a robot must translate an interpreted instruction into an appropriate response within a physical environment. However, errors in perception or interpretation can propagate into subsequent decisions. An incorrectly identified object or misunderstood instruction may therefore result in an inappropriate action. Consequently, reliable HRI requires mechanisms that evaluate whether an interpretation is sufficiently supported before it is converted into action.

Multimodal Perception in HRI

Human communication in HRI can involve speech, gestures, body movements, visual information, and contextual cues. Multimodal perception enables robots to combine these heterogeneous information sources to obtain a richer representation of an interaction. Research in multimodal HRI highlights the importance of integrating asynchronous data streams while dealing with practical issues such as noise, incomplete observations, and variations in sensing conditions.

In the proposed framework, speech provides the linguistic description of a request, while pointing gestures provide additional spatial or referential information. Visual perception identifies objects and their locations within the scene. Contextual information can further assist interpretation when the user's instruction is incomplete or ambiguous. Combining these sources allows the system to relate what the user says and indicates to what is actually observed in the environment.

Human Intention Understanding

Human intention understanding is a central requirement of HRI because a robot must determine the user's intended goal before selecting an appropriate response. Human instructions are not always explicit; they may depend on the surrounding environment, previous interaction, gestures, or contextual information. As a result, determining intention from a single modality can be insufficient in situations involving incomplete or ambiguous instructions.

Multimodal intention understanding addresses this limitation by considering evidence from several sources simultaneously. For example, the statement “give me that one” does not explicitly identify an object through language alone. A pointing gesture can provide the missing reference, while visual perception can determine which objects are available in the indicated region. Thus, intention understanding in the proposed framework depends on the combined interpretation of language, gesture, visual evidence, and relevant context.

Pointing and Deictic Gestures

Pointing is an important deictic form of communication because it allows a person to indicate an object, location, or direction without explicitly naming it. In HRI, pointing can complement speech by providing spatial information that may be absent from the verbal instruction. However, gesture interpretation can be affected by factors such as viewpoint, occlusion, inaccurate pointing direction, and the presence of multiple nearby objects. Therefore, pointing should be interpreted in relation to speech and visual evidence rather than being treated as an independent command.

Multimodal Conflict, Uncertainty, and Reliability

Multimodal inputs may provide inconsistent information during natural interaction. For example, a user may verbally refer to one object while pointing toward another, or the visual system may fail to detect the object mentioned in speech. Such situations create uncertainty about the user's intended meaning and require the system to determine whether the available evidence is sufficient for a reliable decision. Recent work on multimodal perception and decision-making identifies robustness, sensor noise, adaptation, and safety as important challenges for HRI systems.

Reliability is also affected by hallucination, in which an MLLM produces information that is inconsistent with the available visual evidence. Multimodal hallucinations can include incorrect claims about the existence,

attributes, or relationships of objects. Such errors are particularly important in HRI because unsupported model interpretations may influence subsequent robotic decisions.

These challenges establish the conceptual basis of the proposed reliability-focused framework. The system first combines multimodal evidence to understand the user's intention and identify the relevant object or action. It then examines whether the modalities provide a consistent interpretation and whether the interpretation is supported by available visual or sensor evidence. When the evidence is sufficiently reliable, the system can proceed toward an appropriate response. When the information remains ambiguous, contradictory, or unsupported, the system requests clarification rather than relying on an uncertain assumption. This approach connects multimodal understanding with uncertainty handling, verification, and safer decision-making, which forms the central foundation of the proposed methodology.

Related Work

Research on large language models (LLMs) in robotics has developed rapidly over the past few years. Earlier studies mainly used language models to convert human instructions into executable robot plans. Later research connected language models with visual and sensory information, followed by approaches in which models could directly connect perception with robot actions. This progression provides an important background for understanding the remaining challenges in human-centred Human–Robot Interaction (HRI).

Early research mainly treated an LLM as a bridge between human language and robot planning. A human could provide an instruction in natural language, and the model could transform that instruction into a sequence of smaller actions. However, generating a sequence of actions did not guarantee that the robot could actually perform them. Therefore, researchers introduced feasibility checks based on the robot's available skills and physical capabilities. Other approaches represented the generated plans as structured programs or executable code. This made the robot's decisions easier to inspect, test, and modify. As a result, research gradually moved from simply asking whether LLMs could generate robot plans toward ensuring that those plans were feasible, understandable, and suitable for physical environments.

A major limitation of language-only systems is that language does not always provide enough information about the surrounding environment. To address this issue, researchers began connecting language models with sensory information such as images and robot-state data. PaLM-E is an important example of this direction, as it incorporates visual and sensor information into the model so that the system can reason about both the environment and the robot. This approach is particularly relevant to HRI because human instructions are frequently dependent on the surrounding situation. For example, a statement such as "put it there" cannot be properly interpreted without knowing which object and location the person is referring to. Later approaches also explored the use of short-term memory so that the system could retain recently observed information and use it when interpreting subsequent instructions.

The next stage involved connecting language and perception directly with robot actions. Vision-Language-Action (VLA) models were developed to allow a model to associate visual information and natural-language instructions with physical robot actions. RT-2 demonstrated that knowledge learned from large-scale datasets could contribute to robotic manipulation and help robots respond to tasks that were not explicitly included in their robot-specific training data. The Open X-Embodiment initiative extended this direction by bringing together data from different robot platforms and environments. The underlying idea was that learning from multiple embodiments could improve the ability of a model to generalize across different robots and tasks. Open-source models such as OpenVLA and Octo have further supported research in this area by making VLA-based experimentation more accessible. Although these models have achieved promising results in physical manipulation, handling tasks that require social understanding, contextual interpretation, or human intention remains more challenging.

Multimodal Large Language Models (MLLMs) extend these developments by allowing multiple forms of information to be considered together. In HRI, a robot may receive speech, gestures, gaze, facial expressions, and visual information at the same time. Processing these signals jointly can help the robot understand relationships that may be missed when every modality is handled independently. For example, when a person

points toward an object and says "that one," the spoken statement alone does not identify the object. The robot needs to connect the person's gesture with the objects visible in the environment. However, general-purpose MLLMs may still be less precise than specialized models for fine-grained tasks such as accurate gaze or gesture estimation. Therefore, hybrid systems are often useful, where dedicated perception models provide detailed observations and an MLLM performs higher-level reasoning over those observations.

Another important research direction is human intention understanding. In natural interaction, people do not always express their requirements clearly. They may use incomplete instructions, gestures, gaze, or information from the surrounding environment to communicate what they want. For example, when several objects are present, the instruction "give me that object" is ambiguous unless the robot also considers visual and contextual information. Existing approaches use combinations of speech, video, gaze direction, hand movement, body posture, object location, and interaction history to estimate human intentions. MLLMs can potentially combine these sources and use their learned knowledge to interpret human activities. However, recognizing an action that is already happening is different from predicting what a person will do next. Future intention prediction requires consideration of previous actions, current context, possible goals, and environmental conditions, making it a more difficult and still developing research problem.

Human emotion understanding is another important area for socially capable robots. Traditional affective computing systems have used facial expressions, voice characteristics, body movements, and other behavioural signals to estimate emotional states. MLLMs provide an opportunity to consider several of these signals together. A robot may use a person's words, tone, facial expression, and body language when selecting an appropriate response. However, emotion recognition remains challenging because people express emotions differently, and cultural and individual differences can affect these expressions. Emotions can also be mixed or unclear, making them difficult to determine from a single observation. In addition, models trained on controlled datasets may not behave equally well during natural interactions. Therefore, emotional information should be treated as supporting evidence rather than as a definitive representation of a person's internal state.

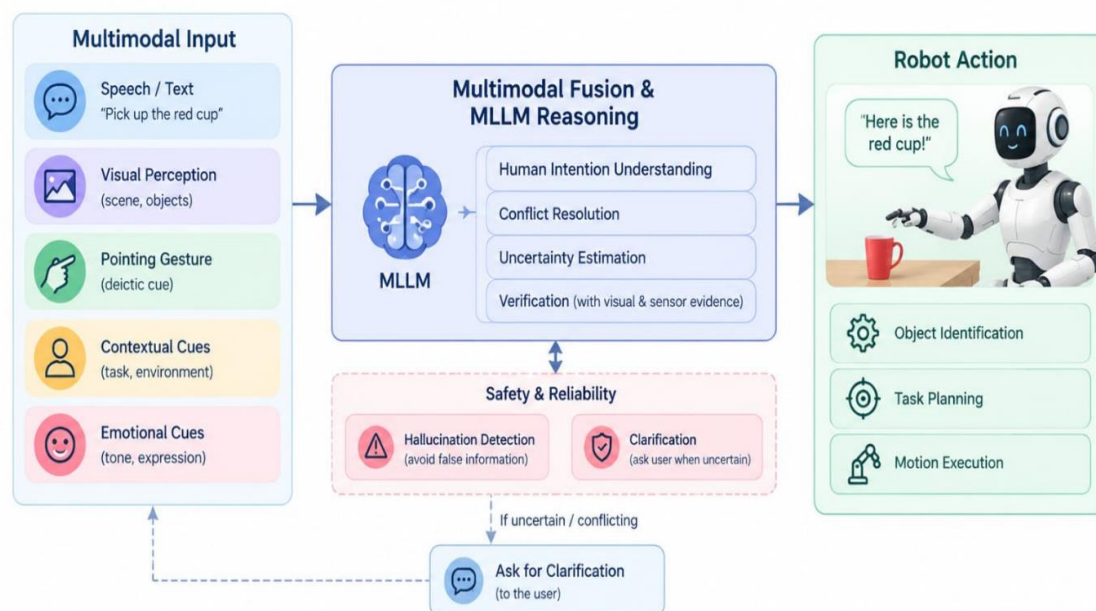
Multimodal conflict resolution is closely connected to both intention and emotion understanding. In real interactions, different modalities may provide contradictory information. A person might verbally indicate one location while pointing toward another, or their tone and facial expression may differ from the literal meaning of their words. Existing systems commonly handle such situations either by asking the person for clarification or by assigning predefined importance to different modalities. Although these strategies can be useful, neither is ideal in every situation. Repeated clarification can interrupt the interaction, while fixed modality priorities may fail when the reliability of a particular signal changes with context. This creates a need for adaptive multimodal reasoning in which the system evaluates available evidence, considers context, detects conflicts, and estimates confidence before deciding how to respond.

Reliability is also a major concern when MLLMs are connected to physical robots. One important problem is hallucination, where a model produces information that is not supported by the available evidence. In robotics, such errors can be more serious than ordinary text-based mistakes. For example, if a model incorrectly assumes that an object exists or misunderstands the environment, the resulting decision may cause the robot to perform an inappropriate action. Existing research has explored methods such as grounding model outputs in sensor observations, checking generated information before execution, and measuring hallucination under specific conditions. However, reliable evaluation of hallucination in real interactive HRI environments remains an important challenge, particularly because evaluation methods differ between studies.

Overall, existing research has made substantial progress in connecting language, perception, reasoning, and robot action. However, several human-centred challenges remain. Reliable integration of voice, gestures, vision, context, and emotional information is still difficult, particularly when the available signals are incomplete or contradictory. Future HRI systems therefore need to move beyond simple multimodal fusion toward adaptive reasoning that can estimate human intention, use emotional and contextual cues, recognize uncertainty, and avoid unsupported conclusions. Addressing these issues can contribute to the development of MLLM-based robots that interact with humans in a more reliable, natural, and context-aware manner.

METHODOLOGY

This study proposes a reliability-focused multimodal framework for Human–Robot Interaction (HRI) that integrates speech, pointing gestures, visual information, and contextual cues within an MLLM-based reasoning process. The methodology is designed to improve human intention and object understanding while reducing errors caused by conflicting inputs, uncertainty, and hallucinated information.



Multimodal Input and Perception

The system receives spoken commands, pointing gestures, visual scenes, and contextual information through appropriate sensing components. Speech is converted into text, while vision identifies relevant objects and their spatial information. Gesture processing determines pointing direction and meaningful movements. These modalities are initially processed separately to retain their individual evidence.

Multimodal Fusion and MLLM Reasoning

The extracted information is combined within the MLLM for joint reasoning. The model analyses relationships between speech, gestures, visual observations, and contextual information to construct a unified interpretation. Previous interaction context is also considered when the current instruction is incomplete or ambiguous.

Intention and Conflict Understanding

The MLLM determines the user's probable intention and identifies the relevant object or action. It then compares evidence across modalities to detect inconsistencies. For example, if speech refers to one object while the pointing gesture indicates another, the system identifies the conflict and evaluates the reliability of the available evidence.

Reliability, Verification, and Decision

Before executing an action, the predicted interpretation is evaluated using confidence estimation and available visual or sensor evidence. Unsupported objects, incorrect locations, or inconsistent interpretations are flagged for verification. If sufficient evidence supports the decision, the system proceeds with the appropriate response. If uncertainty or conflict remains high, it requests clarification instead of making an unsupported assumption.

Evaluation

The framework is evaluated through controlled HRI scenarios involving ambiguous instructions, conflicting speech and gestures, contextual variations, and hallucination-related cases. Performance is assessed using **intention accuracy, conflict-detection accuracy, hallucination rate, clarification effectiveness, response time, task-completion rate, and unsafe-action rate**. This evaluation measures both interpretation accuracy and the system's ability to recognize and handle uncertainty.

Experimental Setup

The proposed framework is not a complete robotic system but rather a foundation for a system. Therefore, it is evaluated in controlled interaction scenarios and in accordance with pre-defined evaluation metrics. The system aims to investigate the ability of the framework to grasp human intentions in different modalities, manage conflicting information, incorporate emotional and contextual aspects, and improve the precision of MLLM-based decisions.

Experimental Environment

The evaluation environment is composed of a camera and a microphone as the main input devices. The camera captures the environment, objects, their position, pointing gestures, and human gestures. The microphone picks up the user's voice, and a speech-processing module converts the speech into text.

The extracted visual, speech and gesture information is then inputted into the Multimodal Large Language Model (MLLM) for integrated reasoning. The MLLM takes into account the relationships between the various modalities to understand the user's intended goal. The system produces a prediction or a clarification request, depending on the evidence it has. This configuration enables the reasoning process of the framework to be tested without relying on physical robot hardware.

Evaluation Scenarios

Four representative scenarios are identified to test the key capabilities of the proposed framework.

Scenario 1 – Human Intention Understanding:

When there are multiple objects in the scene, the user gives a verbal instruction and a pointing gesture. For example, the user could point to an object and say, "Give me that one." The framework integrates speech, gesture and visual information to recognise the object and select the action.

Scenario 2 – Multimodal Conflict Resolution:

The user gives contradictory information in various modalities, e.g., verbally pointing to one object and pointing to another. The framework assesses the evidence and context available, not a predetermined priority for one modality. If the conflict is not resolved with sufficient clarity, the system does not make an unsupported decision but rather asks for clarification.

In this scenario, the student is able to recognise and respond to emotions and context.

The same instruction is given in different emotional or situational contexts, e.g., a neutral request and an urgent request. The framework takes into account the tone of the voice, facial expressions and context when deciding on an appropriate response. Unless there is evidence to the contrary, the task intention should be the same.

Scenario 3 – Hallucination and Reliability:

The framework is tested with scenes where the MLLM might produce an interpretation that is not supported by the visual evidence, e.g., describing an object that is not present in the scene. The verification stage is a

comparison of the model's interpretation with the perceptual information available prior to the selection of a final action. Unsupported interpretations are therefore not considered actions but rather uncertain.

Evaluation Metrics

The framework is assessed using the following indicators:

Intention Accuracy: The accuracy of the object, goal, or intention that the system identifies.

Conflict Resolution Accuracy: Evaluates the accuracy of processing conflicting multimodal inputs.

Context Sensitivity: Evaluates the system's ability to react to relevant emotional and situational cues while keeping the intended task.

Hallucination Rate: The number of interpretations that are not supported by the perceptual information.

Successful Action Rate: Percentage of times the final system decision matches the user's intended action.

Clarification Rate: The percentage of times the system needs more information than it currently has when the evidence is not enough or conflicting.

Response Time: Time taken to process the multimodal input and produce a final decision or a clarification request.

These metrics together give a structured way to assess the accuracy of the final decision and the framework's capacity to identify uncertainty, deal with conflicting information, and prevent unsupported interpretations before an action is chosen.

Expected results and discussion

The proposed framework is expected to make Human–Robot Interaction (HRI) more reliable by enabling the robot to utilize various types of information simultaneously. The system takes into account speech, gestures, visual information, context, and emotional cues before determining the response. The expected results are mainly related to human intention understanding, multimodal conflict resolution, hallucination and reliability, and emotion and context in interaction.

Human Intention Understanding

The proposed framework is expected to improve the robot's understanding of human intentions through speech, gestures, visual information, and context. This is useful when verbal instructions are not provided or are unclear. For example, when a user points at an object and says “that one,” the pointing direction, scene information, and speech can be used to determine the object the user is pointing at.

If the instruction is not clear, additional information can be obtained from previous interactions and the context. If there is still not enough evidence, the framework can recognize the uncertainty and request clarification instead of making a decision without sufficient evidence. Thus, the expected improvement is more accurate intention understanding with safer handling of ambiguous situations.

Multimodal Conflict Resolution

The framework is also designed to deal with cases involving conflicting communication signals. For instance, a user can refer to one object while pointing to another object. In such cases, using only one modality can result in an incorrect action.

The proposed approach is to review the evidence and take into account the context in which it is found to choose an interpretation. If the evidence is clearly in favor of one option, the system can proceed with that interpretation. If the conflict cannot be resolved reliably, it can ask the user for clarification. This should minimize the chances

of errors due to conflicting information and eliminate unnecessary interruptions. The aim is not to increase the number of clarifications requested, but to request clarification when there is a lack of sufficient evidence in the system.

Hallucination and Reliability

One of the expected benefits is better control over hallucination errors. Sometimes, an MLLM will generate an interpretation that is not consistent with the current scene. This is especially significant in HRI, as the robot might choose the wrong physical action based on this error.

The proposed verification stage aims to check the model's understanding of the perceptual information before it is executed. For example, if the MLLM identifies an object that is not verified in the scene, the system should not take any action based on the information it generates, but rather identify the problem. This process should minimize the impact of unsupported model outputs. It should not, however, be taken as a promise that there will be no hallucinations at all.

Emotion and Context Sensitivity

The framework is intended to increase the robot's sensitivity to the context in which an instruction is given. The tone of voice, facial expression, body language, and the context of the interaction can provide further clues as to how the interaction should be handled.

For instance, the same instruction can be delivered in a regular or urgent voice. This difference can be utilized by the system when choosing an appropriate response style. Meanwhile, emotional information is used as supporting evidence and not as a definite measurement of a person's internal emotional state. This helps to avoid the emotional interpretation from interfering with the task or intention unnecessarily.

OVERALL DISCUSSION

The overall proposed framework is expected to make HRI more reliable by bringing together various sources of human communication into the same decision process. The system does not process speech, gestures, or visual information separately; rather, it considers all of these signals together and uses context to understand what is happening. This can assist the robot in handling ambiguous instructions, conflicting signals, and unsupported model responses before taking an action.

The anticipated assessment should provide a more comprehensive understanding of the framework's grasp of human intent, its ability to manage multimodal conflicts, its ability to manage errors related to hallucination, and its ability to adjust its responses to various scenarios. However, some challenges will remain, such as unclear speech, difficult-to-observe gestures, limited visual information, or situations in which the user's intent cannot be inferred from the available evidence.

The results are not experimental results, but rather expected results, since the proposed framework has not yet been tested with a large number of users or on a physical robot. The framework will need to be further tested in real-world HRI settings to determine its effectiveness in real-world situations.

Limitations

The proposed framework has some limitations. The study uses controlled evaluation scenarios rather than extensive testing with a physical robot in real-world environments. Its performance under unpredictable conditions therefore requires further validation. The framework also depends on the quality of speech recognition, gesture detection, object detection, and sensor data. Noise, poor lighting, occlusion, and unclear gestures may affect the system's interpretation. Human intention may remain uncertain when instructions are incomplete or change during interaction. Although the framework includes uncertainty estimation and clarification, some ambiguous situations may remain unresolved. Finally, verification can reduce hallucinations

but cannot completely prevent them when sensory information is incomplete or inaccurate. Future studies should test the framework with physical robots, diverse users, and more complex interaction settings.

Future Work

Future work will focus on validating the proposed framework through physical robot experiments and longer real-world interactions. Testing with diverse users and environmental conditions will help determine how effectively the system handles background noise, poor lighting, occlusion, unclear gestures, and changing instructions.

Further research can improve confidence estimation and develop more adaptive methods for determining the reliability of individual modalities. Larger and more varied interaction scenarios can also be used to evaluate conflict resolution, hallucination mitigation, and clarification effectiveness more thoroughly. Integrating stronger perception and verification components may further improve the system's ability to distinguish reliable observations from uncertain model interpretations. These developments could support practical applications of multimodal HRI in environments such as healthcare, education, workplaces, and assisted living.

CONCLUSION

This study presented a reliability-focused multimodal framework for improving human–robot interaction using multimodal large language models (MLLMs). The proposed approach combines voice, pointing gestures, visual perception, and contextual information to support more accurate human intention and object understanding. Rather than relying on a single modality, the framework considers evidence from multiple sources and examines whether they provide a consistent interpretation.

The framework also incorporates uncertainty estimation and verification to reduce the effects of hallucinated or unsupported information. A key component is the clarification mechanism, which allows the system to request additional information when the available evidence is ambiguous or conflicting. This helps prevent uncertain interpretations from directly leading to inappropriate actions. Overall, the framework provides a structured approach to making MLLM-based HRI more reliable, safe, and suitable for situations in which human instructions cannot be understood from speech alone.

REFERENCES

1. D. Drieset al., “PaLM-E: An Embodied Multimodal Language Model,” Proceedings of the 40th International Conference on Machine Learning, vol. 202, pp. 8469–8488, 2023.
2. B. Ichteret al., “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances,” Proceedings of the 6th Conference on Robot Learning, vol. 205, pp. 287–318, 2023.
3. W. Huang et al., “Inner Monologue: Embodied Reasoning through Planning with Language Models,” Proceedings of the 6th Conference on Robot Learning, vol. 205, pp. 1769–1782, 2023.
4. B. Zitkovich et al., “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control,” Proceedings of the 7th Conference on Robot Learning, vol. 229, pp. 2165–2183, 2023.
5. S. Trick, D. Koert, J. Peters, and C. Rothkopf, “Multimodal Uncertainty Reduction for Intention Recognition in Human-Robot Interaction,” arXiv preprint arXiv:1907.02426, 2019.
6. X. Zhao, H. Li, T. Miao, X. Zhu, Z. Wei, and A. Song, “Learning Multimodal Confidence for Intention Recognition in Human-Robot Interaction,” arXiv preprint arXiv:2405.14116, 2024.
7. S. Duncan, S. Alambeigi, and J. Pryor, “A Survey of Multimodal Perception Methods for Human–Robot Interaction in Social Environments,” ACM Transactions on Human-Robot Interaction, vol. 13, no. 4, 2024.
8. Y. Lai et al., “Natural Multimodal Fusion-Based Human–Robot Interaction: Application With Voice and Deictic Posture via Large Language Model,” IEEE Robotics and Automation Magazine, vol. 33, no. 2, pp. 29–38, 2026.
9. M. Abokiet al., “Multimodal Human–Robot Interaction Using Human Pose Estimation and Local Large Language Models,” Advanced Robotics Research, 2026, doi: 10.1002/adrr.202500175.

10. S. Yeon, M. Lee, J. Bae, and S. Lee, "Resolving Ambiguity in Pointing Gestures Using Contextual Reasoning from Large Language Models," *Computer Modeling in Engineering & Sciences*, vol. 147, no. 1, 2026, doi: 10.32604/cmes.2026.079954.
11. C. Li et al., "Do MLLMs Understand Pointing? Benchmarking and Enhancing Referential Reasoning in Egocentric Vision," *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.
12. T. Schreiter et al., "Multimodal Interaction and Intention Communication for Industrial Robots," *arXiv preprint arXiv:2502.17971*, 2025.
13. G. A. Abbo, S. Lenaerts, and T. Belpaeme, "Multimodal Large Language Models for Real-Time Situated Reasoning," *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1161–1163, 2026.
14. S. Spezialetti, G. Placidi, and M. Rossi, "Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives," *Frontiers in Robotics and AI*, vol. 7, 532279, 2020.
15. C. Bai et al., "Hallucination of Multimodal Large Language Models: A Survey," *arXiv preprint arXiv:2404.18930*, 2024.