Integrating Large Language Models with Multimodal Virtual Reality Interfaces to Support Collaborative Human–Robot Construction Work

Somin Park, A.M.ASCE[1]; Carol C. Menassa, F.ASCE[2]; and Vineet R. Kamat, F.ASCE[3]

Abstract: In the construction industry, where work environments are complex, unstructured and often dangerous, the implementation of human–robot collaboration (HRC) is emerging as a promising advancement. This underlines the critical need for intuitive communication interfaces that enable construction workers to collaborate seamlessly with robotic assistants. This study introduces a conversational virtual reality (VR) interface integrating multimodal interaction to enhance intuitive communication between construction workers and robots. By integrating voice and controller inputs with the robot operating system (ROS), building information modeling (BIM), and a game engine featuring a chat interface powered by a large language model (LLM), the proposed system enables intuitive and precise interaction within a VR setting. Evaluated by 12 construction workers through a drywall installation case study, the proposed system demonstrated its low workload and high intuitiveness and ease of use with succinct command inputs. The proposed multimodal interaction system suggests that such technological integration can substantially advance the integration of robotic assistants in the construction industry. **DOI: 10.1061/JCCEE5.CPENG-6106.** *© 2024 American Society of Civil Engineers.*

Author keywords: Human–robot collaboration (HRC); Human–robot interaction (HRI); Virtual reality (VR); Multimodal interaction.

Introduction

In the architecture, engineering, and construction (AEC) industry, the complex, unstructured, and often dangerous work environments have led to the increasing interest in exploring how robots can assist humans in completing the tasks (Brosque et al. 2020; Adami et al. 2021; Park et al. 2023; Wang et al. 2023). Human–robot collaboration (HRC) leverages the precision, strength, and repeatability of work allowed by robotic interfaces, blending these attributes with human workers' cognitive capability, knowledge of the craft, and adaptability to change. Given that effective communication among construction workers is important for improving labor productivity (Johari and Jha 2021), sharing essential information for correct performance within human–robot teams is equally important. Members of human teams exhibit anticipatory information-sharing initiatives to accomplish collaborative tasks. To achieve this with the integration of robotic assistants into the construction industry, easy to learn and bidirectional communication between workers and robots is necessary. The importance of the user-friendly and efficient communication with the robots is highlighted by the fact that the willingness of construction workers to engage with robotic assistants is significantly affected by their perceptions of the system's ease of use and usefulness (Park et al. 2023).

Speech communication, recognized as an easy and intuitive form of human interaction, could be the fastest and most efficient way to interact with robots (Marge et al. 2022). This mode of communication has the capacity to seamlessly transmit task-related information directly without the constraints of information loss, highlighting its potential to enhance interaction with collaborative robots. In construction, this recognition of speech's utility has prompted investigations into the application of spoken or typed natural language to improve operational efficiency (Shin and Issa 2021; Linares-Garcia et al. 2022; Park et al. 2024).

However, the reliance on speech inputs for task execution in construction environments presents significant challenges. Construction workers often come from varied backgrounds, bringing a wide range of accents into the communication process. This can lead to misrecognition of spoken commands because automatic speech recognition (ASR) systems often struggle to accurately interpret words spoken by users with heavy or uncommon accents (Saka et al. 2023). This issue is compounded by construction jargon, which includes specialized terms that may not be recognized by standard speech recognition systems developed for general use. The accurate specification of necessary task attributes (Linares-Garcia et al. 2022; Park et al. 2024) can also be affected. Such linguistic diversity undermines the practicality and convenience of speech-based interaction.

Addressing these limitations, nonverbal cues can be leveraged with verbal communication to use the respective strengths of each mode. Specifically, gestural movements of hands offer a rich array for semantics and are easily recognizable (Yongda et al. 2018). The inclusion of the hand gestures in multimodal interfaces alongside speech introduces the utility of deictic references, such as pointing at objects during conversations. This allows for more efficient verbal communications because it eliminates the need for detailed verbal descriptions (Wagner et al. 2014) that might be misunderstood due to accents or unfamiliar jargon, ultimately helping to mitigate these communication challenges.

[1]Assistant Professor, Dept. of Civil Engineering, Univ. of Texas at Arlington, Arlington, TX 76010. Email: somin.park@uta.edu

[2]Professor, Dept. of Civil and Environmental Engineering, Univ. of Michigan, Ann Arbor, MI 48109 (corresponding author). ORCID: https://orcid.org/0000-0002-2453-0386. Email: menassa@umich.edu

[3]Professor, Dept. of Civil and Environmental Engineering, Univ. of Michigan, Ann Arbor, MI 48109. ORCID: https://orcid.org/0000-0003-0788-5588. Email: vkamat@umich.edu

Note. This manuscript was submitted on April 7, 2024; approved on July 19, 2024; published online on October 28, 2024. Discussion period open until March 28, 2025; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Computing in Civil Engineering***, © ASCE, ISSN 0887-3801.

For instance, instead of providing detailed descriptions like “pick up the middle block in the row of five blocks on the right” (Paul et al. 2016) or “please pick up the Sheet 500300 and position it in the Stud 500100” (Park et al. 2024), users can issue succinct commands like “move this to that location” or “place this one there” through pointing gestures. The confluence of speech and pointing gestures offers considerable advantages in reducing cognitive load (Goldin-Meadow et al. 2001), decreasing communication errors (Lee et al. 2013), and increasing communicative efficiency (Wagner et al. 2014). Despite these benefits, the application and potential of multimodal interfaces that combine speech and hand gestures within HRC in the construction industry remain largely unexamined.

Furthermore, noise is a prevalent factor on construction sites that can adversely affect the accuracy of speech interaction, crucial for effective HRC through speech inputs (Yoon et al. 2023). Despite recent advances in ASR technology, which have enhanced the robustness of these systems against background noise through deep neural network–based models and extensive training data sets, the practical application of speech-based instructions in noisy construction environments remains a challenge. The advent of digital twins in construction, coupled with the potential for remote operations, provides an opportunity to interact with robots on construction sites from quieter and remote environments (Wang et al. 2021; Park et al. 2024). This setting potentially allows speech inputs to be delivered to robots with reduced interference from construction noise and overlapping conversations, which could significantly improve the clarity and accuracy of communication. The effectiveness of the communication could be enhanced by introducing an additional layer of verification, such as enabling both robots and operators to pose questions and review planned tasks before approving execution for robotic operations (Wang et al. 2024).

Recognizing the underexplored potential of multimodal interfaces and the potential for remote working in digital twins highlights a significant opportunity for innovation in the design of the HRC interaction systems. The existing gap in effectively integrating speech and hand gestures for HRC points toward the need for comprehensive solution to implement the interaction for construction tasks. Consequently, the objectives of this study are to (1) propose a multimodal VR interaction method for HRC in construction; (2) devise a strategy for the integration of diverse software solutions to implement the interaction method; and (3) verify the proposed method through a user study.

To this end, this paper proposes a novel multimodal interaction system that integrates voice commands and hand controller inputs within immersive virtual reality (VR) environments for HRC in construction. This multimodal interaction allows users to employ a hardware controller for pinpointing workpieces onsite while utilizing verbal commands for task specification. Moreover, it incorporates a large language model (LLM) as a virtual assistant to facilitate bidirectional communication. Further enriching the system, the integration of building information modeling (BIM) ensures retrieval of information about the workpieces for construction tasks.

The practical application of this interaction system is evaluated by construction practitioners using a case study of a drywall installation, a typical pick-and-place task. This system operates within the concepts of a digital twin framework, where the virtual world is connected to the physical world through data exchange. This connection allows for the transfer of operational data from the virtual to the physical setting, potentially influencing physical robot operations. It is important to highlight that although this study effectively demonstrates the pick and place actions, it does not encompass all possible ground actions such as moving, tilting, and gripping. Furthermore, the scope of this research is limited to the virtual environment, and it does not directly address the application of this system in real-world settings.

Related Work

Recent advancements in natural language processing (NLP) have led to the application of conversational systems, using natural language (NL) inputs, in facilitating interaction between humans and computers, including HRC. Inspired by the fact that integrating NL interaction with other input modalities can improve usability and naturalness (Grammel et al. 2010), multimodal methods for human–robot interaction (HRI) have been proposed by combined with gestures with the fact that they are important factors in conversations between humans. There have been studies to use predefined gestures with voice commands. Chen et al. (2022) developed a convolutional neural network that recognizes 16 gestures for industrial robots, such as Start, Stop, and Slow Down, aligning with corresponding short voice commands like “move inward” and “go left.”

In contrast to relying solely on predefined gestures, other studies have investigated the use of pointing gestures as a means to refer to objects or locations of interest because it provides a clear and intuitive way to convey directional information to robots (Van den Bergh et al. 2011). Yongda et al. (2018) and Constantin et al. (2022) used gestures to point to a certain direction with a finger while identifying the user intention through speech commands. Yongda et al. (2018) captured pointing gestures by Leap Motion and processed commands like “move 2 mm in this direction” to be precisely executed without the need for explicitly stating the direction verbally, such as “toward the x-axis.” To give instructions to robots such as “please bring me that thing,” Constantin et al. (2022) employed transformer model (Vaswani et al. 2017) to analyze language instructions and computer-vision techniques to detect hands and forefingers using a fixed two-dimensional (2D) camera, thereby capturing the pointed object. Although these multimodal interaction methods enhance real-world HRC by integrating speech and gestures, their applications in specific sectors like construction remain distinct.

In the construction industry, researchers have mainly explored the application of conversational systems in two domains of information retrieval (IR) and integration with VR or augmented reality (AR) technologies (Saka et al. 2023). Most of the efforts on conversational systems have focused on retrieving information from data sources such as BIM (Saka et al. 2023). These efforts have focused on enabling efficient retrieval of project information through query-answering systems. Lin et al. (2016) introduced a data retrieval and representation system for cloud BIM applications via text input. Shin and Issa (2021) developed a BIM automatic speech recognition (BIMASR) framework to convert a BIM operating environment from expert-oriented into a user-oriented. This framework allows users to manipulate BIM data using speech commands, such as changing materials within a model. Elghaish et al. (2022) proposed a data retrieval assistant that enables BIM users to perform tasks such as the creation of a room schedule through natural language commands. These studies on information retrieval underline the progress in allowing users to efficiently interact with project information via natural language queries.

With the rapid growth of artificial intelligence (AI), there has been significant advancement in LLM, which are built to be flexible and are trained on extensive data sets (Achiam et al. 2023). In the field of construction, recent studies have leveraged prompt engineering to develop customized conversation systems using

LLMs. Prieto et al. (2023) explored the applicability of LLM for automating construction schedule using natural language prompts, finding the results promising for simple use cases. Zheng and Fischer (2023) developed a dynamic prompt–based virtual assistant that interprets NL queries to improve BIM accessibility, demonstrating strong performance in intent classification and value recognition. Jang et al. (2024) proposed an LLM-BIM chaining framework that generates and modifies object classes in BIM, facilitating an interactive design detailing environment. This system effectively functioned as a design consultant that produced design details that complied with general engineering standards.

Despite these advancements, challenges such as limited domain-specific knowledge and a propensity for hallucinations are acknowledged in the use of LLM. Solutions such as incorporating human-in-the-loop approaches and developing fine-tuned LLMs have been suggested as potential solutions (Kim et al. 2024; Jang et al. 2024). Moreover, conversational systems for IR in the construction industry have demonstrated that natural language inputs can effectively manage construction project data. However, the conversational systems rely on single-mode user inputs like text or speech, highlighting the need for the advancement of multimodal interaction systems to enhance the efficiency and intuitiveness of user interactions with other agents.

Meanwhile, by leveraging visualization functionalities of VR and AR, several studies have incorporated conversational systems within VR/AR in construction. In the VR domain, studies have explored the use of virtual humans to improve educational experiences in construction. Eiris-Pereira and Gheisari (2018) used a virtual people factory (VPF) web-based application (Rossen et al. 2009) for conversational modeling. They demonstrated its application through a high-risk hazard scenario in construction, aiming to improve student communication skills. However, this approach relied solely on text inputs from users. Wen and Gheisari (2023) focused on virtual field trips for mechanical and plumbing systems, developing a conversational system by combining VPF and Google Diagflows. This allowed users to interact with objects using computer mice and make inquiries such as “is this a hot water return pipe?” through text input. Nonetheless, the method of integrating these two types of inputs was not discussed. Both systems also have limitations due to their reliance on predefined templates for NL answers, which could restrict the flexibility and adaptability of conversations.

Hussain et al. (2024) introduced a virtual training system designed to deliver knowledge effectively, using LLM as an instructor. This system does not use predefined templates, enabling users to ask about various construction hazard situations and receive tailored responses. However, the background information on risky situations provided in the prompt was general, which limited the ability to derive project-specific insights from the conversations.

In the AR domain, Chen et al. (2024) addressed construction safety compliance by proposing a visual construction safety query system. This system employs a deep learning–based vision-language model, allowing users to inquire about safety issues onsite using voice input through AR glasses. Despite its advances, they analyzed spoken words alongside image scenes rather than incorporating gestures into the interaction system. Chen et al. (2020) proposed a multimodal interaction system using human hands designed for swarm robot selection that could be applied to various industrial domains including construction. This system, facilitated through AR, enables users to issue instructions like “select the robots in this range” using speech and pointing gestures. Although this development provided the integration of software modules and hardware components for the use of AR in HRC, its applicability is limited to the selection of multi robots rather than facilitating collaborative tasks with the robots.

In addition to the aforementioned application areas, there are studies that explore the use of voice-based conversational systems to help workers perform construction tasks. One example is the work of Linares-Garcia et al. (2022), who developed a voice-based intelligent virtual assistant (VIVA) specifically designed to increase the productivity of construction workers during welding tasks. The virtual agent system was developed based on Google Actions requiring expected questions and semantic knowledge (the task steps). However, users are expected to give questions such as “After connecting PS-3 and B-8, what should I do next,” which include pieces’ tag IDs or location of the items in previous steps as part of the context addition.

Ye et al. (2023) explored the impact of ChatGPT, which is one of the LLMs, on trust in HRC assembly task. It showed that people perceived less mental load and high trust when using a GPT-enabled robot assistant compared with using fixed control commands. Park et al. (2024) introduced a framework that enables NL-based interactions with robots for pick-and-place construction operations, demonstrating its effectiveness through drywall installation tasks. This system employs deep learning–based language models, allowing it to accurately process instructions that specify targets, destinations, and placement methods. However, the scope of these two studies was limited to a single modal interaction, necessitating precise mentions of task-relevant information, such as the names or attributes of objects, in instructions. Expanding to encompass multiple modalities in communication thus promises to improve intuitiveness and usability in interaction systems for collaboration.

All the previous studies on conversational systems in construction have primarily focused on users receiving information from virtual assistants, without addressing how users should respond when the information provided is incorrect. This oversight could be particularly significant in operation-critical environments such as construction sites, where accurate and timely information is crucial for decision-making and safety.

In the construction industry, the previous studies on NL-based conversational systems for VR/AR or robotics have the following limitations:

- Lack of support for HRC in construction: Many studies have proposed interaction systems for education, safety, and other areas, rather than for robotic completion of actual construction work. There is a need for a framework that facilitates the execution of construction tasks through conversational systems in a VR environment.
- Limited user input channels to speech: Most of the studies depend on NL inputs, either through voice or text. To improve intuitiveness and efficiency in interaction, additional input channels can be added.
- Inherent limitations of conversational systems themselves: Firstly, some systems require training data to develop the conversational systems. Second, several studies rely on predefined templates for generating natural language responses. Additionally, there is research necessitating reference to previous steps in multistep tasks. Overall, all the studies position users (or construction workers) primarily as recipients of information, where the NL answers have been designed to guide users. The intention behind the generated NL responses has been to lead users, rather than enabling users to participate as collaborative partners capable of offering feedback, providing instructions or dismissing the wrong information. To harness the potential of bidirectional communication in collaboration, the design of the conversational systems with an advanced language model is needed.

Table 1. Comparison of characteristics about conversational systems to support applications on VR/AR or operations in construction

References	Application	AR/VR	User inputs		Conversation				Interaction command examples
			NL inputs	Others	A	B	C	D	
Eiris-Pereira and Gheisari (2018)	Education (hazard scenarios)	VR (desktop)	Text	N/A	Yes	Yes	N/A	Yes	Not provided
Wen and Gheisari (2023)	Education (jobsite investigation)	VR (desktop)	Text	Mouse (pointing)	Yes	Yes	N/A	Yes	Is this a hot water return pipe?/This is a supply pipe
Chen et al. (2024)	Safety inspection	AR	Voice	Image data	Yes	No	N/A	Yes	Are these workers safe?/What should they do to improve safety?
Hussain et al. (2024)	Education (safety training)	VR (desktop)	Voice	N/A	No	No	N/A	Yes	I am here to learn about safety rules for fall hazards. Can you help me out with that?
Chen et al. (2020)	HRI	AR	Voice	Hand (pointing)	Yes	N/A	N/A	N/A	Select the robots in this range
Linares-Garcia et al. (2022)	Construction tasks (steel connection)	N/A	Voice	N/A	Yes	No	Yes	Yes	After connecting PS-3 and B-8, what should I do next?
Ye et al. (2023)	HRC	VR (headset)	Voice	N/A	No	No	N/A	Yes	Give me a driller
Park et al. (2024)	HRC	VR (desktop)	Text	N/A	Yes	N/A	N/A	Yes	Please pick up the 4 by 8 drywall panel and hang it into the 500100 vertically
Proposed system	HRC	VR (headset)	Voice	Hand controller	No	No	No	No	Please pick up this and place it there

Note: A = it requires training data to develop the interaction process; B = it depends on predefined templates for NL answers; C = it necessitates referencing previous steps in multi-step tasks; D = it lacks mechanisms for human operators (or users) to verify NL answers; and N/A = not applicable.

Table 1 provides a summary of these limitations. To address these challenges, this study proposes a multimodal interaction system that facilitates conversation with collaborative robots in construction. This study integrates diverse software solutions, enhancing the efficacy and scope of HRC in the construction industry.

Technical Approach

Overview of the Proposed System

Fig. 1 presents an overall framework for a multimodal interaction system in VR, designed for easy and intuitive interaction with construction robots. The proposed system requires the integration of user interaction channels, the robot operating system (ROS), BIM, and a game engine. The game engine, which encompasses a chat interface, enables a human operator to interact with a robot using natural language commands. Users can communicate with robots using two input channels: speech and controllers. The outputs from these channels become the input for the chat interface. The game engine plays an important role in visualizing information from BIM, displaying the construction site's data, including three-dimensional (3D) geometric information and the semantic details of building materials. A robot engaged in interaction within the game engine is controlled via ROS, completing the system's loop of HRC.

Interaction Interface

Fig. 2 outlines the software integration architecture for the proposed multimodal interaction system in HRC. The diagram provides a visual representation of how different components and inputs are integrated within the system. The subsequent subsections will articulate the design of the integration strategies. Section "Integration of Speech and VR Controller Inputs" will detail the methods for integrating two distinct user inputs: speech and handheld controllers captured by the VR interface. Building material information retrieved from Rhinoceros and Grasshopper is leveraged in the integration. Following that, section "Bidirectional Communication" will concentrate on the design and flow of interaction between a human operator and a robot, specifically discussing how to implement the bidirectional communication through GPT-4 and conduct collaborative operations in ROS in the Unity game engine.

Integration of Speech and VR Controller Inputs

Fig. 3 shows how inputs from two different channels—speech and VR controllers—are combined in the interaction system. Voice

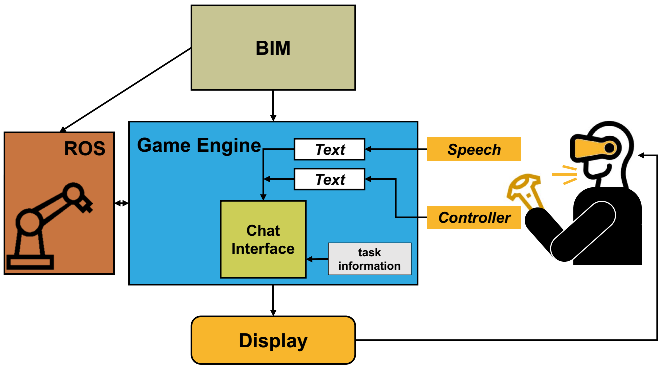


Fig. 1. Overview of the proposed system.

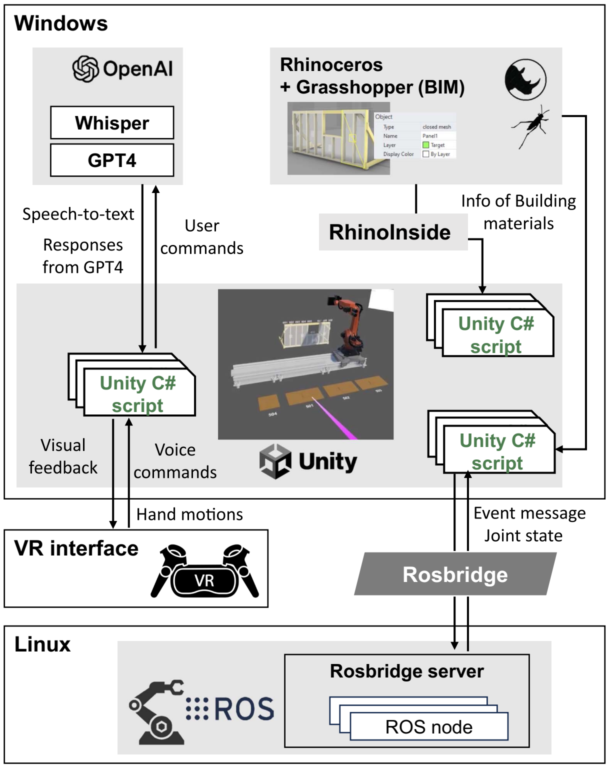


Fig. 2. Implementation of the proposed multimodal interaction system.

commands are captured by the VR headset's microphone and subsequently processed by Whisper, which is an automatic speech recognition (ASR) or speech-to-text (STT) system (OpenAI 2023). Whisper shows robustness in ASR because it was trained on very large data sets including 680,000 h of multilingual audio data (Radford et al. 2023).

When the user is engaging with the VR environment, VR controllers provide an interactive means of selecting objects within the virtual space. Selecting objects via VR controllers is streamlined using a ray interactor. The ray interactor projects a virtual beam from the controller, allowing a user to point at and select an object from a distance. To visually indicate that the object has been selected, the color of the object briefly changes to red, signaling successful engagement.

A key feature of the multimodal interaction proposed in this study is the retrieval of selected object information from BIM data during the provision of user instructions. Unity generates interactable objects from BIM through the integration of Rhino and Grasshopper applications. Utilizing Rhino.Inside (McNeel 2023), an open-source add-in, Unity is enabled to concurrently operate the two applications. Once initiated alongside Unity, Grasshopper, which provides a visual programming environment, imports BIM data into Rhino. Following this, it retrieves both the geometric and semantic information of objects from the BIM data, which is subsequently transmitted to Unity.

As shown in Fig. 4, the Grasshopper workspace uses multiple blocks to deliver the required BIM data in Rhinoceros to Unity for interaction purposes. This transfer results in the creation of 3D objects within Unity that are visually consistent with the Rhino model, encompassing color, geometry, and semantic information such as names, layers, and IDs. The IDs can be integrated into the user messages, and the names and layers are utilized to determine the interactivity of objects within Unity. A part of C# script in Fig. 4 presents how to get various types of object data from the Rhino. Although the current interaction design does not harness all the semantic data available—such as type and position—this information can be extracted and has the potential to be employed in interactions with robots.

Upon selection of an object, its ID information is retrieved from the BIM data and stored in a text format, ready for further processing or use within the system. The retrieved ID is then held in reserve until the user activates the send button on the chat interface. For instance, the placeholder ### in the sentence "The ID of the target object is ###" would be replaced with the captured ID of the selected object. This textual information, representing the selected object, is then combined with the text that has been transcribed from the user's spoken command.

To illustrate, if a user verbally commands "pick up this one" while concurrently selecting an object with the ID 127 using the controller, the consolidated command is formulated as "pick up this one. The ID of the target object is 127." This composite text, which contains inputs from both speech and controller commands, is sent to the chat interface when the user presses the send button of the chat interface. This process establishes a multimodal interaction framework, seamlessly integrating verbal commands with controller-based selections to enable effective communication and control within the VR space and reduces burden on user to find object IDs.

Bidirectional Communication

In this study, the scope of the human operator's duties in the proposed interaction system includes the following activities: issuing commands to the robots, verifying the accuracy of how these commands are interpreted by the robot equipped with the capability of GPT-4, and ultimately, supervising the execution of these tasks by the robot. The conversation with a robot in the proposed interaction system is specifically designed to facilitate the first two of these activities. The necessity for the human operator to verify the interpretation of commands arises from the potential for inaccuracies due to errors in STT conversion or misinterpretations by the chat system. Therefore, the chat system is designed to detect and rectify potential errors, thereby enhancing the overall accuracy and efficiency of task execution.

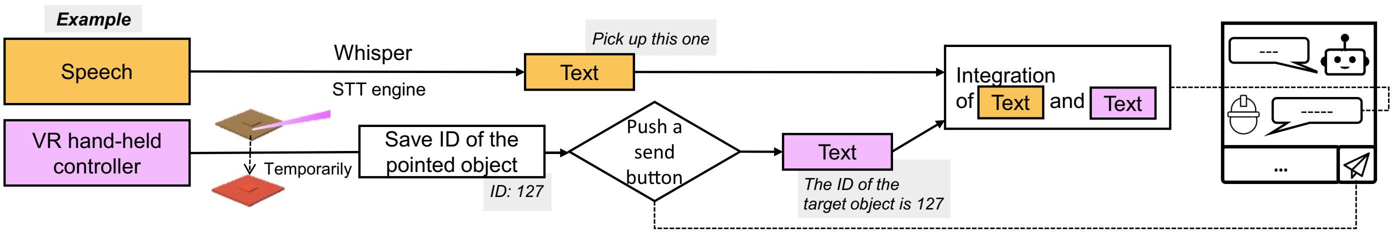
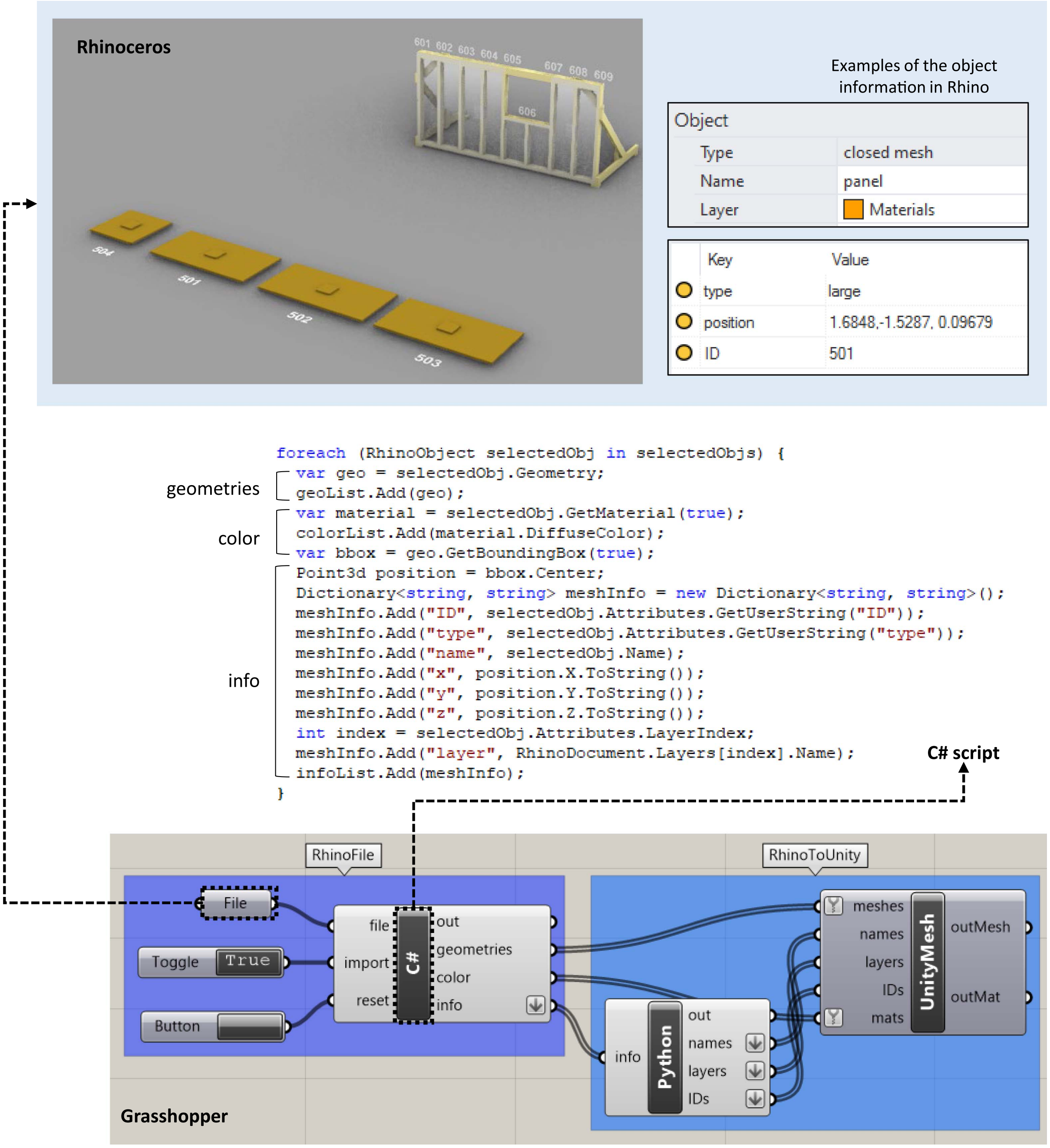


Fig. 3. Integration of inputs from voice commands and the VR controller.



Examples of the object information in Rhino

Object	
Type	closed mesh
Name	panel
Layer	🟧 Materials

	Key	Value
●	type	large
●	position	1.6848,-1.5287, 0.09679
●	ID	501

```
foreach (RhinoObject selectedObj in selectedObjs) {
  var geo = selectedObj.Geometry;
  geoList.Add(geo);
  var material = selectedObj.GetMaterial(true);
  colorList.Add(material.DiffuseColor);
  var bbox = geo.GetBoundingBox(true);
  Point3d position = bbox.Center;
  Dictionary<string, string> meshInfo = new Dictionary<string, string>();
  meshInfo.Add("ID", selectedObj.Attributes.GetUserString("ID"));
  meshInfo.Add("type", selectedObj.Attributes.GetUserString("type"));
  meshInfo.Add("name", selectedObj.Name);
  meshInfo.Add("x", position.X.ToString());
  meshInfo.Add("y", position.Y.ToString());
  meshInfo.Add("z", position.Z.ToString());
  int index = selectedObj.Attributes.LayerIndex;
  meshInfo.Add("layer", RhinoDocument.Layers[index].Name);
  infoList.Add(meshInfo);
}
```

Fig. 4. BIM data in Rhinoceros and Grasshopper workspace.

Fig. 5 describes the process flowchart from the issuing to the approval of the instructions. In this diagram, rectangles represent the actions executed by the human operator, and hexagons depict the robot's expected responses. Within the proposed conversation system, the user (human operator) issues instructions that include specific details about the task. The chat system then analyzes these instructions to identify the essential task information and seeks the user's confirmation.

The user, serving as the ultimate decision maker, assesses the chat system's interpretation of the task information. If the user agrees with the chat system's interpretation, they respond affirmatively, prompting the chat system to acknowledge with a reply of "OKAY!!!" The user then finalizes their approval by clicking an Approval button on the chat interface, which triggers the robot to begin the task. Alternatively, if the user does not concur with the chat system's interpretation or if the initial instructions were incorrect, the chat system requests the user to provide the correct information to ensure accurate task execution. This interactive process is essential for ensuring clear communication and precise task management between the human operator and the robot.

To implement this process, GPT-4 was utilized, which is a large-sized pretrained language model that demonstrates powerful capabilities to understand and generate human language (Achiam et al. 2023). Prompt engineering is a technique that utilizes natural language task specifications to design prompts for LLMs regarding the downstream task instead of directly altering or training the models (Trad and Chehab 2024). Eliminating the need for the model training with extensive data and time, this approach enables obtaining desired responses in a flexible manner and with fewer resource demands (Trad and Chehab 2024).

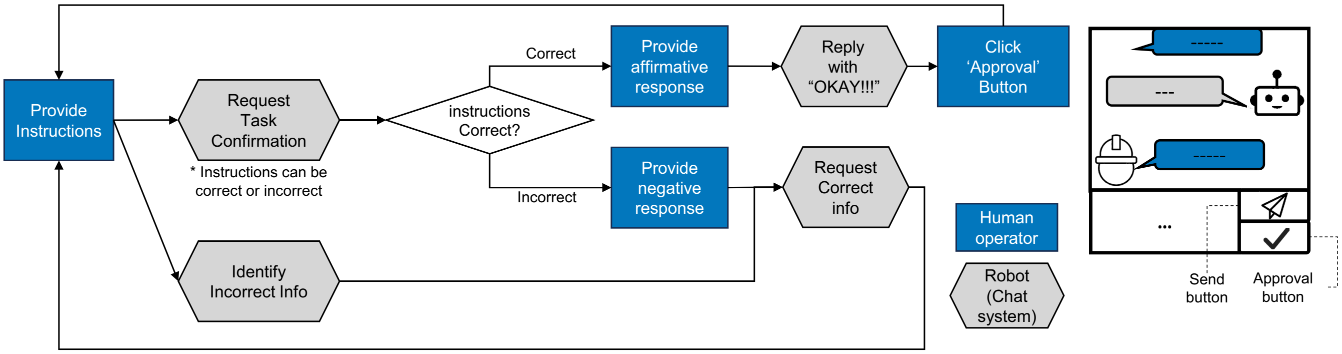


Fig. 5. Flowchart for instruction approval in the conversational system.

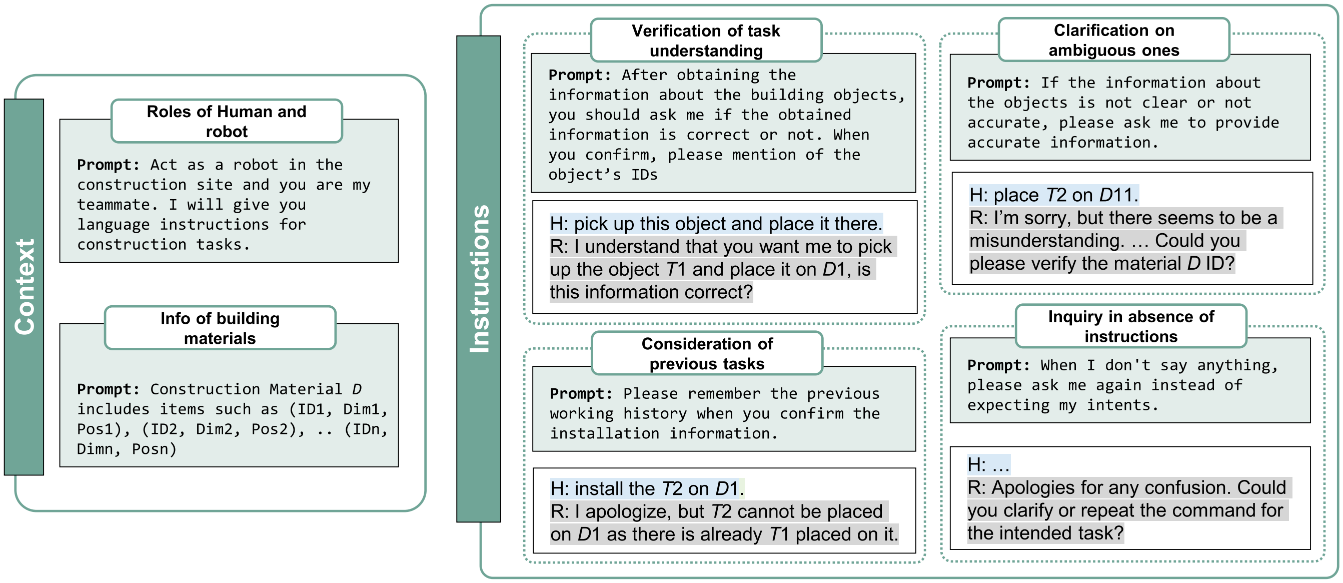
Recent studies utilizing GPT models in construction have included elements such as task description, assistant's roles, BIM information, constraints, rules, and so on into their prompts (Ye et al. 2023; Zheng and Fischer 2023). However, these prompts, although comprehensive, are not specifically detailed or tailored for HRC as required in the proposed study. To ensure the prompt is fully aligned with the needs of the proposed interaction system, this study meticulously designs the GPT prompt with components like roles, task-specific object information, and various instructions for HRC.

In this work, the prompt of the GPT contains two contexts and four types of task instructions to build effective communication between a human operator and a robot as shown in Fig. 6. The context provided in the prompt first outlines the roles of a human operator and a robot. For example, a sentence like "Act as a robot in the construction site and you are my teammate" in the prompt sets the tone for the text generated by the GPT. Following this context, task-related information is integrated into the prompt. This information typically includes the semantic details of construction materials, enabling the GPT to perform reasoning based on the commands issued by the human operator.

The prompt also includes four main instructions for the collaboration with robots, enabling the process flowchart described in Fig. 5. Each instruction, as shown in Fig. 6, is presented with corresponding examples showing expected robot responses (R) to human messages (H). The first instruction involves verification of task understanding, which includes GPT's interpretation of the operator's instructions and asks the operator's verification of this interpretation. This step ensures that the GPT's understanding aligns with the operator's intent. In addition, the IDs of the target objects should be mentioned when a robot asks for confirmation so that the human operator can clearly identify the understanding of the robot. Second, the inclusion of clarification on ambiguous instructions is vital. This step is designed so the GPT can seek further information from the user instead of making assumptions when provided with incomplete or incorrect information.

Third, the consideration of previous tasks is integrated in the prompt. This aspect is important for multitasking scenarios, encouraging the robot to take into account tasks that have been previously completed. If the workpiece already used is given to the robot as task information, the robot should have ability to recognize it based on this context. The sentence in the prompt like "please remember the previous working history when you confirm the installation information" can lead to ideal responses such as "T2 cannot be placed on D1 as there is already T1 placed on it."

Lastly, to avoid randomly generated responses that are not related to the intent, the inclusion of inquiry in the absence of instructions is essential. This is designed to prevent the GPT from making



Roles of Human and robot

Prompt: Act as a robot in the construction site and you are my teammate. I will give you language instructions for construction tasks.

Info of building materials

Prompt: Construction Material *D* includes items such as (ID1, Dim1, Pos1), (ID2, Dim2, Pos2), .. (IDn, Dimn, Posn)

Verification of task understanding

Prompt: After obtaining the information about the building objects, you should ask me if the obtained information is correct or not. When you confirm, please mention of the object's IDs

H: pick up this object and place it there.
R: I understand that you want me to pick up the object *T*1 and place it on *D*1, is this information correct?

Consideration of previous tasks

Prompt: Please remember the previous working history when you confirm the installation information.

H: install the *T*2 on *D*1.
R: I apologize, but *T*2 cannot be placed on *D*1 as there is already *T*1 placed on it.

Clarification on ambiguous ones

Prompt: If the information about the objects is not clear or not accurate, please ask me to provide accurate information.

H: place *T*2 on *D*11.
R: I'm sorry, but there seems to be a misunderstanding. … Could you please verify the material *D* ID?

Inquiry in absence of instructions

Prompt: When I don't say anything, please ask me again instead of expecting my intents.

H: …
R: Apologies for any confusion. Could you clarify or repeat the command for the intended task?

Fig. 6. Information in the prompt of GPT to manage communication scenarios with examples.

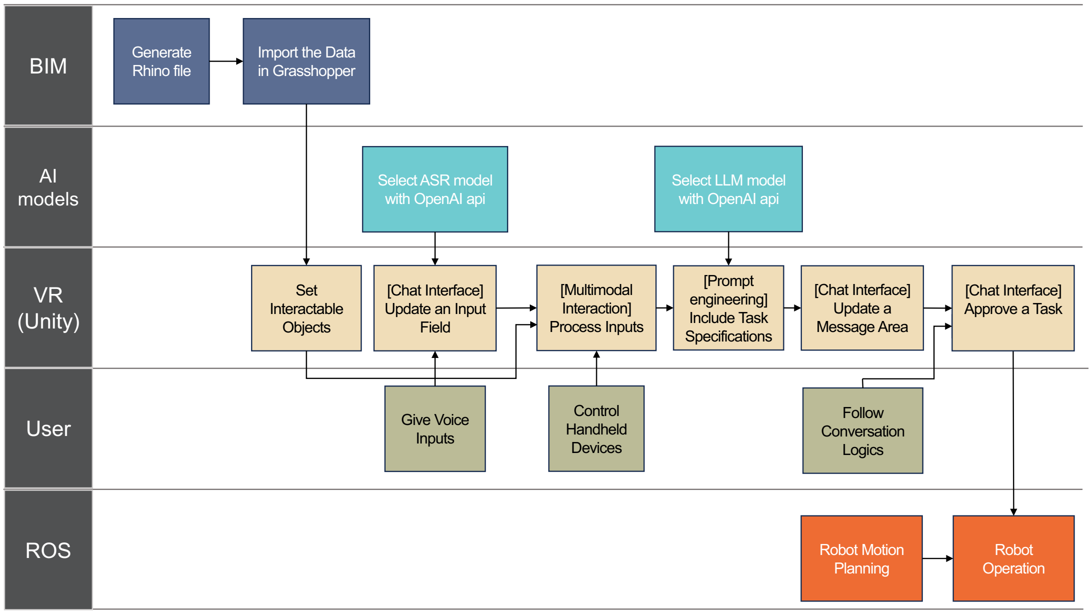

BIM
AI models
VR (Unity)
User
ROS
Generate Rhino file
Import the Data in Grasshopper
Select ASR model with OpenAI api
Select LLM model with OpenAI api
Set Interactable Objects
[Chat Interface] Update an Input Field
[Multimodal Interaction] Process Inputs
[Prompt engineering] Include Task Specifications
[Chat Interface] Update a Message Area
[Chat Interface] Approve a Task
Give Voice Inputs
Control Handheld Devices
Follow Conversation Logics
Robot Motion Planning
Robot Operation

Fig. 7. Interactive workflow diagram of the proposed system.

incorrect assumptions about human intent or providing responses that are out of context when a blank message is sent. This approach ensures that even when instructions are unclear or entirely absent, the GPT is guided to respond in a manner that is more appropriate and aligned with the intended purpose.

The proposed prompt is incorporated as an essential component within the C# script for leveraging GPT functionality in Unity. On the Unity script, an application programming interface (API) key for OpenAI is also included because it is necessary for authentication purposes in utilizing OpenAI's application. In addition, the script for GPT integration includes command lines to facilitate the integration of two input channels and selection of the GPT model. For this study, GPT-4, the latest model available, was chosen to ensure the most advanced capabilities are employed. The temperature value, which is a parameter to influence the randomness in text generation, was set to zero to minimize the variability in the responses (Zheng and Fischer 2023).

Finally, for the actualization of tasks using a robot, the ROS running on Linux OS was employed. To do this, BIM data in Rhino and task-related information are accessed in ROS through Rosbridge using the ROS# library. As mentioned in the first instruction of the prompt, the robot is required to reference the IDs of target objects when seeking confirmation from the user. Before the user confirms by pressing an approval button, the robot's response, which includes these IDs, is processed within the Unity script. During this process, a conditional statement is used to isolate and store only the IDs in a text format in the Window OS. The stored ID information is sent to ROS once a user presses an approval button of the chat interface.

Utilizing this information, robot motion planning and robot control are conducted. The motion planning for robotic movements needs to consider collision-free motion plans designing paths where both the robot and any workpiece it carries avoid collisions with other building materials or structures in the environment. This study uses the motion planning method as proposed by Wang et al. (2021), which was developed for mobile industrial arm manipulators representing a general application for construction robotics. The calculated motion plan is then mirrored in Unity, where a virtual robot executes the movements as per the plan. The virtual robot is controlled with joint state data from ROS.

In summary, the proposed system leverages several software components to enable multimodal interactions with construction robots in a virtual environment. Fig. 7 highlights the main components related to the integration of BIM, AI models, VR using Unity, and ROS. The arrows in Fig. 7 specifically represent the data flow and the progression of tasks throughout the system.

The workflow starts when users generate a Rhino file in the BIM layer, which is then imported within Grasshopper as depicted in Fig. 4. In the AI models layer, both an ASR model and a LLM are selected, and OpenAI API is provided to Unity. This setup facilitates advanced speech-based interactions with an intelligent virtual assistant that follows task-specifications depicted in Fig. 6. Through a chat interface, users can see the updated input fields and message areas, and they can finally approve a task based on the conversation logic depicted in Fig. 5. During interaction, users provide voice inputs and control handheld devices to engage in multimodal interactions (Fig. 3). The process culminates in the ROS layer, where the system performs robot motion planning and operations based on refined user inputs, demonstrating a seamless integration of digital interfaces and robotic execution aimed at enhancing operational efficiencies in technologically advanced environments.

Experimental Evaluation

Case Study

An experiment that aims to assess the effectiveness of the proposed multimodal interaction system through two primary objectives was designed. The first objective is to evaluate whether the integration of speech and hand controller inputs enhances the effectiveness of HRC over a single mode of interaction. The second objective focused on assessing the effectiveness of a bidirectional interaction system, powered by a virtual assistant, in enabling precise collaboration between humans and robots to complete construction tasks. The experiment included two different interaction parts: (1) speech-based interaction and (2) multimodal interaction (speech + handheld controller) with a focus on the task of drywall installation. The task of drywall installation, which is one of the pick-and-place operations, was selected because the operations are one of the most common robotic manipulation tasks (Cheng et al. 2021).

Twelve construction workers were recruited for the experiment and were asked to interact with a virtual robot through a VR headset, visualizing a simulated work environment as shown in Fig. 8. This environment includes a stud frame and four drywall panels, each uniquely identified by a three-digit ID.

The experiment employed a six-degrees-of-freedom (DOF) Kuka industrial robotic arm (Augsburg, Germany), which is mounted on a track. The scope of robot actions in this experiment was limited to pickup and place activities, necessitating specific information about the target object for pickup and the destination for placement. The experiment utilized two sizes of panels: three standard panels measuring 1.22 by 2.44 m (4 by 8 ft) and one uniquely sized panel measuring 1.22 by 1.22 m (4 by 4 ft). The panels represent target objects for pickup, and studs serve as destination for placement. The VR environment also featured a chat interface consisting of several elements: an input field at the bottom, a message area in the middle, a time bar at the top, and two buttons, for sending messages and approving robot tasks, on the lower right side.

Participants were tasked with installing four panels, utilizing either speech or multimodal interactions. There was no set sequence for installing the panels. In the speech interaction, workers gave installation instructions by referring to the IDs or locations of panels and studs. For example, a worker might say, "Please place Panel 504 in the second rightmost position," illustrated in Fig. 9. With multimodal interaction, workers gave instructions using demonstrative words such as "this" and "that" while pointing at objects with a handheld controller. For instance, a worker could instruct, "Please place this panel at this stud." As a result of selecting objects using the controller, sentences such as "(the ID of the target panel is 501) (the destination is the center of Stud 602)" could be added in the input message. It is also permissible in multimodal interactions to mention the panel's ID while pointing at it with the controller.

Participants were instructed to intentionally provide incorrect instructions for panel installation to evaluate the GPT model's capacity for error detection and correction within the chat application.

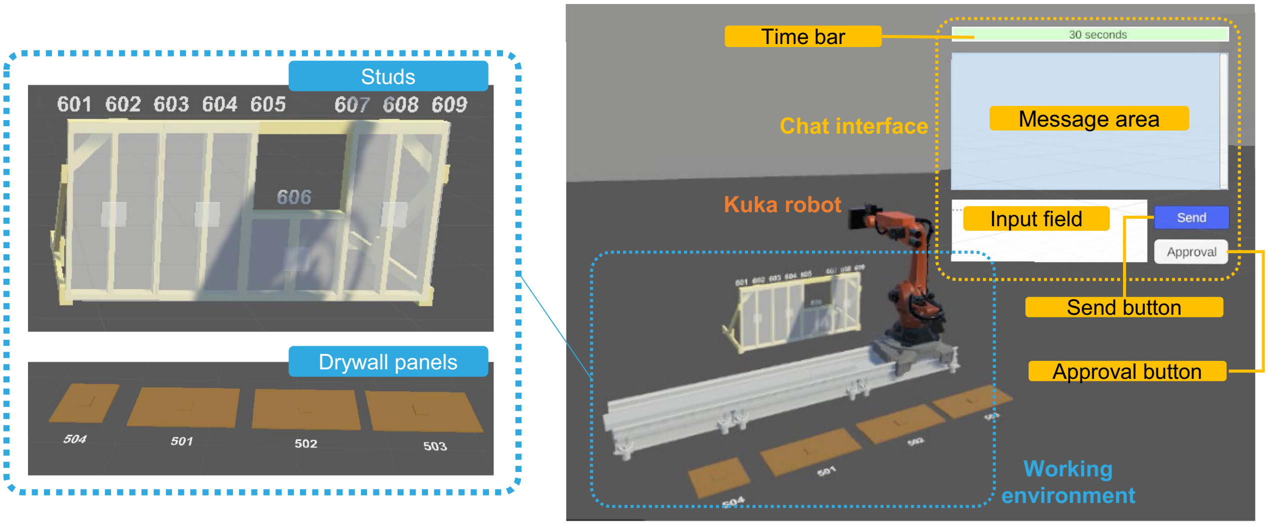
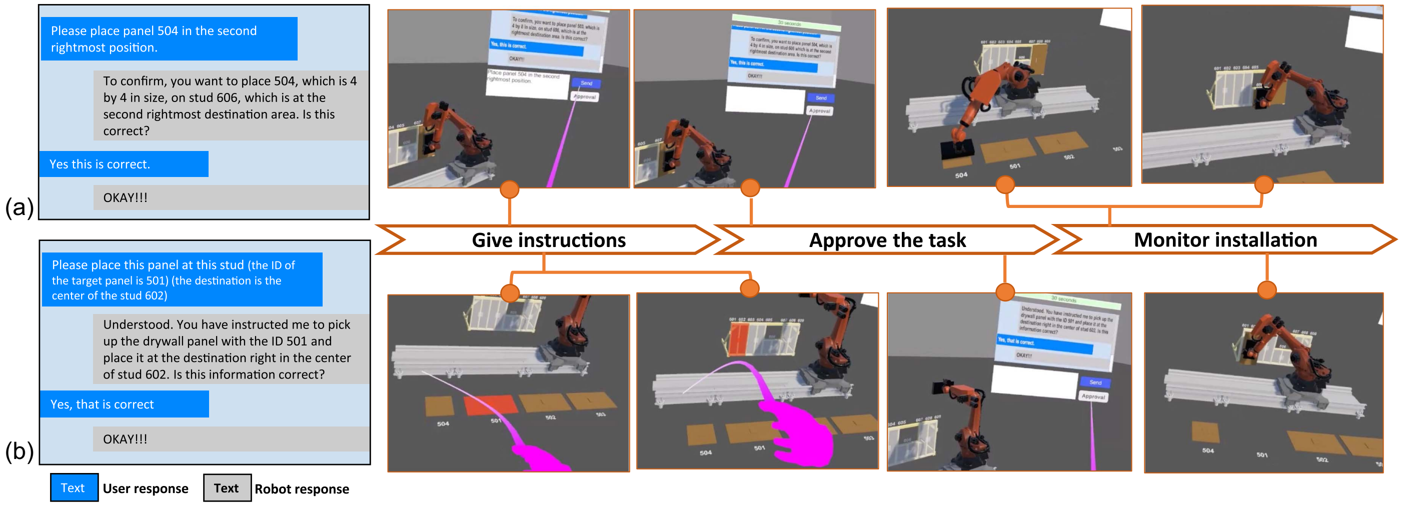


Fig. 8. VR environment in unity.



(a)

Please place panel 504 in the second rightmost position.

To confirm, you want to place 504, which is 4 by 4 in size, on stud 606, which is at the second rightmost destination area. Is this correct?

Yes this is correct.

OKAY!!!

(b)

Please place this panel at this stud (the ID of the target panel is 501) (the destination is the center of the stud 602)

Understood. You have instructed me to pick up the drywall panel with the ID 501 and place it at the destination right in the center of stud 602. Is this information correct?

Yes, that is correct

OKAY!!!

Text User response Text Robot response

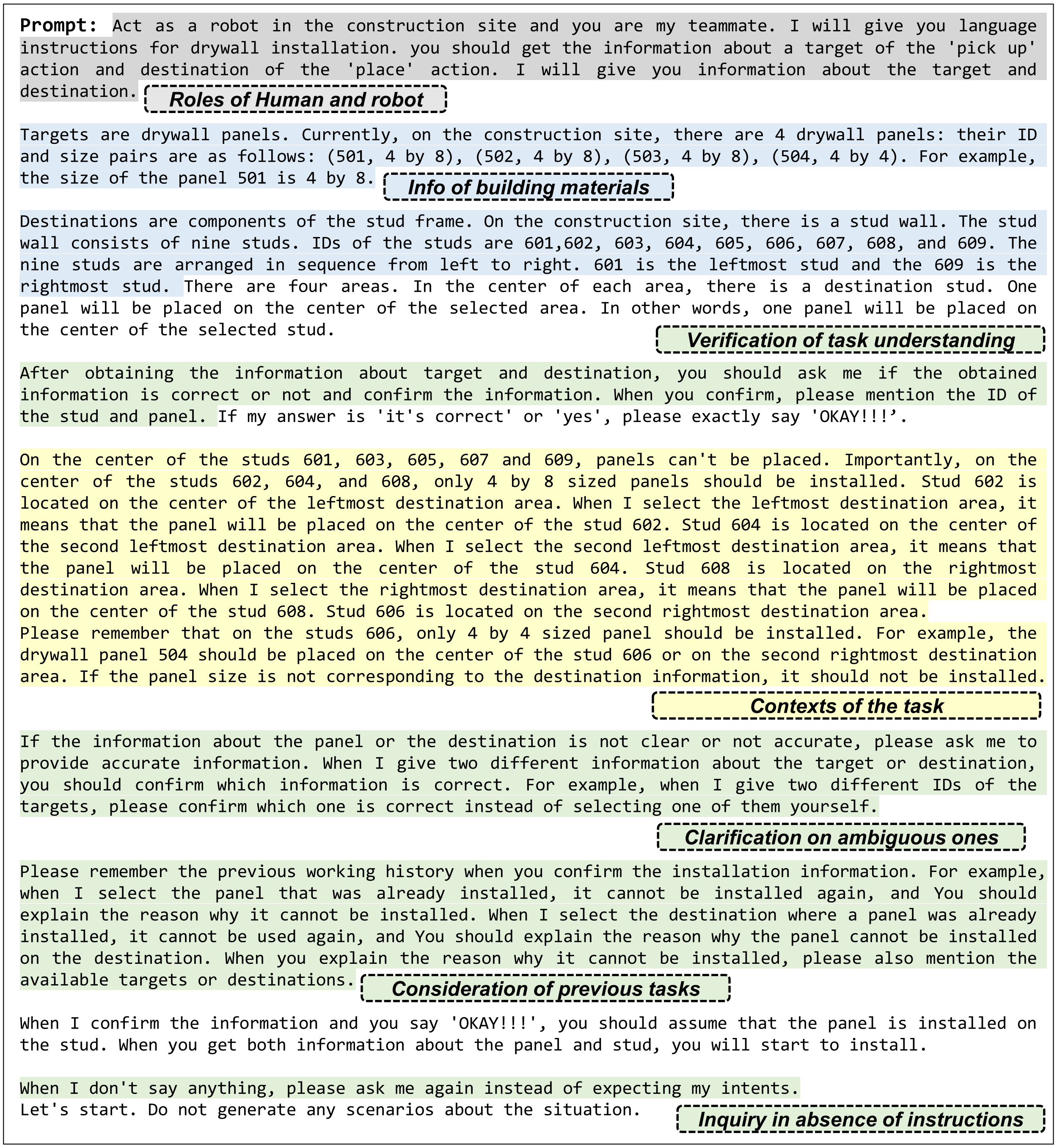
Fig. 9. Examples of HRC in VR interface: (a) verbal interaction; and (b) multimodal interaction.

They were instructed to issue two incorrect ones during each interaction phase, specifically at least two for speech interaction and two for multimodal interaction. These incorrect instructions can be categorized into four types:

- Mismatched Pairing: This involves instructing incompatible combinations, such as pairing a 4 by 8 panel with Stud 606, where only a 4 by 4 panel fits.
- Materials Not Present: This refers to instructions involving nonexistent IDs in the virtual environment, such as instructing the robot to locate a panel or stud with an ID of 106, which does not exist.
- Component Already Installed: This occurs when a participant instructs the installation of a component, such as Panel A, even though it has already been installed.
- Partial Information: This involves giving incomplete or simultaneous information about multiple targets without necessary details about the destination or object.

Participants were given the freedom to choose when and which type of incorrect instruction to use, with the only stipulation being that they must provide at least two incorrect instructions before the end of each interaction phase.

Fig. 10 illustrates the prompts input into the GPT model for the experiment. These prompts were designed to integrate both the situational context and the specific instructions elaborated in Fig. 5, along with additional task-relevant context. This included critical specifics like the dimensions and identifiers of the target objects, as well as their intended placement locations. Notably, the automatic transfer of semantic information regarding target objects



Prompt: Act as a robot in the construction site and you are my teammate. I will give you language instructions for drywall installation. you should get the information about a target of the 'pick up' action and destination of the 'place' action. I will give you information about the target and destination.

Roles of Human and robot

Targets are drywall panels. Currently, on the construction site, there are 4 drywall panels: their ID and size pairs are as follows: (501, 4 by 8), (502, 4 by 8), (503, 4 by 8), (504, 4 by 4). For example, the size of the panel 501 is 4 by 8.

Destinations are components of the stud frame. On the construction site, there is a stud wall. The stud wall consists of nine studs. IDs of the studs are 601,602, 603, 604, 605, 606, 607, 608, and 609. The nine studs are arranged in sequence from left to right. 601 is the leftmost stud and the 609 is the rightmost stud. There are four areas. In the center of each area, there is a destination stud. One panel will be placed on the center of the selected area. In other words, one panel will be placed on the center of the selected stud.

After obtaining the information about target and destination, you should ask me if the obtained information is correct or not and confirm the information. When you confirm, please mention the ID of the stud and panel. If my answer is 'it's correct' or 'yes', please exactly say 'OKAY!!!'.

On the center of the studs 601, 603, 605, 607 and 609, panels can't be placed. Importantly, on the center of the studs 602, 604, and 608, only 4 by 8 sized panels should be installed. Stud 602 is located on the center of the leftmost destination area. When I select the leftmost destination area, it means that the panel will be placed on the center of the stud 602. Stud 604 is located on the center of the second leftmost destination area. When I select the second leftmost destination area, it means that the panel will be placed on the center of the stud 604. Stud 608 is located on the rightmost destination area. When I select the rightmost destination area, it means that the panel will be placed on the center of the stud 608. Stud 606 is located on the second rightmost destination area.
Please remember that on the studs 606, only 4 by 4 sized panel should be installed. For example, the drywall panel 504 should be placed on the center of the stud 606 or on the second rightmost destination area. If the panel size is not corresponding to the destination information, it should not be installed.

If the information about the panel or the destination is not clear or not accurate, please ask me to provide accurate information. When I give two different information about the target or destination, you should confirm which information is correct. For example, when I give two different IDs of the targets, please confirm which one is correct instead of selecting one of them yourself.

Please remember the previous working history when you confirm the installation information. For example, when I select the panel that was already installed, it cannot be installed again, and You should explain the reason why it cannot be installed. When I select the destination where a panel was already installed, it cannot be used again, and You should explain the reason why the panel cannot be installed on the destination. When you explain the reason why it cannot be installed, please also mention the available targets or destinations.

When I confirm the information and you say 'OKAY!!!', you should assume that the panel is installed on the stud. When you get both information about the panel and stud, you will start to install.

When I don't say anything, please ask me again instead of expecting my intents.
Let's start. Do not generate any scenarios about the situation.

Fig. 10. Prompt for GPT for drywall installation.

from the BIM into the GPT's prompts was not part of this study. To incorporate information about the building material into the prompts, we manually entered the information in the prompt.

Participants

The requirement for participation in the experiment included people aged 21 or older who have prior experience working on construction sites. However, certain groups were excluded from participation to ensure safety. This exclusion applied to individuals who are pregnant, elderly, or those with preexisting conditions that may affect their virtual reality experience, such as vision abnormalities, psychiatric disorders, or other medical conditions. Additionally, participants were required to avoid wearing glasses when using a VR headset to ensure optimal interaction and to prevent potential discomfort related to the fit of the headset.

A total of 12 construction personnel were recruited in the state of Michigan, and Table 2 presents their demographic information. The participant demographic profile indicates uniformity in gender, with the entire group consisting of male individuals. The age distribution is moderately varied, with the majority (58.33%) within the 30–39 age range, 16.67% in the 40–49 age range, and 25% being over 49 years old. Educational levels among participants show diversity: 8.33% are high school graduates or hold a general educational development (GED) qualification, half have some college or vocational training, 16.67% hold an associate degree, and 25% have earned a bachelor's degree.

Occupational roles of the 12 participants span across the construction industry, with skilled craftsmen (carpenter and drywall finisher), foremen, superintendents (piping labor, general trades, and general), managers (project, BIM, and operations) and a smaller representation from detailers, and instructors/coordinators. Despite the current job titles not necessarily involving onsite work, all 12 participants had prior experience physically working on construction sites. The range of work experience among the participants extended from under 10 years to more than 29 years, with the average work experience in the construction industry being 19.83 years for the group.

The experiment for each participant was structured into four sessions, with a total duration of 55 min as shown in Fig. 11. During the 10-min introduction phase, participants are briefed on what to do and shown how to use voice commands and handheld controllers to interact with a robot in a virtual environment. With the explanation of how to use a chat interface, they are also introduced to how the virtual setting visualized through a VR headset looks like.

Following the introduction, participants engaged in a 10-min trial task, practicing with speech-based and multimodal (speech + controller) interactions. The participants were introduced to example instructions and their tasks. The main experiment, lasting 25 min, challenges participants to apply these interaction methods to install four drywall panels. The order in which the four panels are installed is not predetermined, allowing participants to install them in any order. Finally, participants are asked to complete a 10-min survey on Google Forms, where they provide feedback on their experience, evaluating workload, intuitiveness, ease of use, and their personal preference between the two interaction methods.

Table 2. Demographic information of 12 participants

Item	Characteristics	Frequency	Percentage
Gender	Male	12	100.00
Age (years)	30–39	7	58.33
	40–49	2	16.67
	Above 49	3	25.00
Education levels	High school graduate or GED	2	8.33
	Some college or vocational training	6	50.00
	Associate degree	1	16.67
	Bachelor's degree	3	25.00
Job titles	Skilled craftsmen	2	16.67
	Foreman	2	16.67
	Superintendents	3	25.00
	Managers	3	25.00
	Detailers	1	8.33
	Instructors/coordinators	1	8.33
Work experience (years)	0–9	3	25.00
	10–19	4	33.33
	20–29	2	16.67
	Over 29	3	25.00

Results and Discussion

Workload

The National Aeronautics and Space Administration Task Load Index (NASA-TLX) was utilized to measure workload perceived by participants (Hart 2006). As one of the most widely used instruments to assess overall subjective workload (Hoonakker et al. 2011; Li et al. 2019), NASA-TLX consists of six domains: Mental demand, Physical demand, Temporal demand, Performance, Effort, and Frustration. In this study, the Effort dimension was divided into Mental Effort and Physical Effort to capture more understanding of the effort type. Participants rated each subscale on a five-point Likert scale ranging from one (strongly low) to five (strongly high).

Figs. 12 and 13 present the assessment of the perceived workload for two interaction methods among 12 participants using the NASA-TLX scale. Fig. 12 indicates that participants' workload perception was largely consistent across most categories for both speech and multimodal interaction methods. This consistency suggests that the type of interaction method does not significantly alter the perceived workload. However, a notable variance in responses regarding mental demand was observed for speech interaction compared with multimodal interaction, implying that speech interaction may be mentally more challenging for some participants.

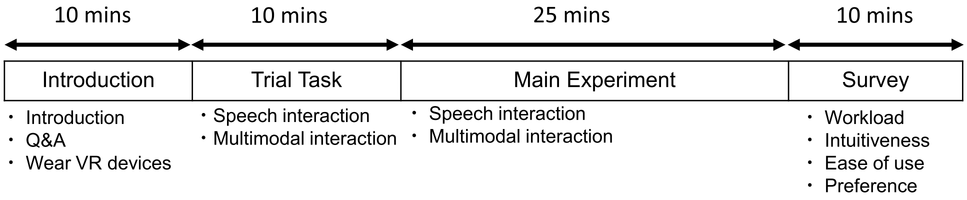


Fig. 11. Timeline of the experiment.

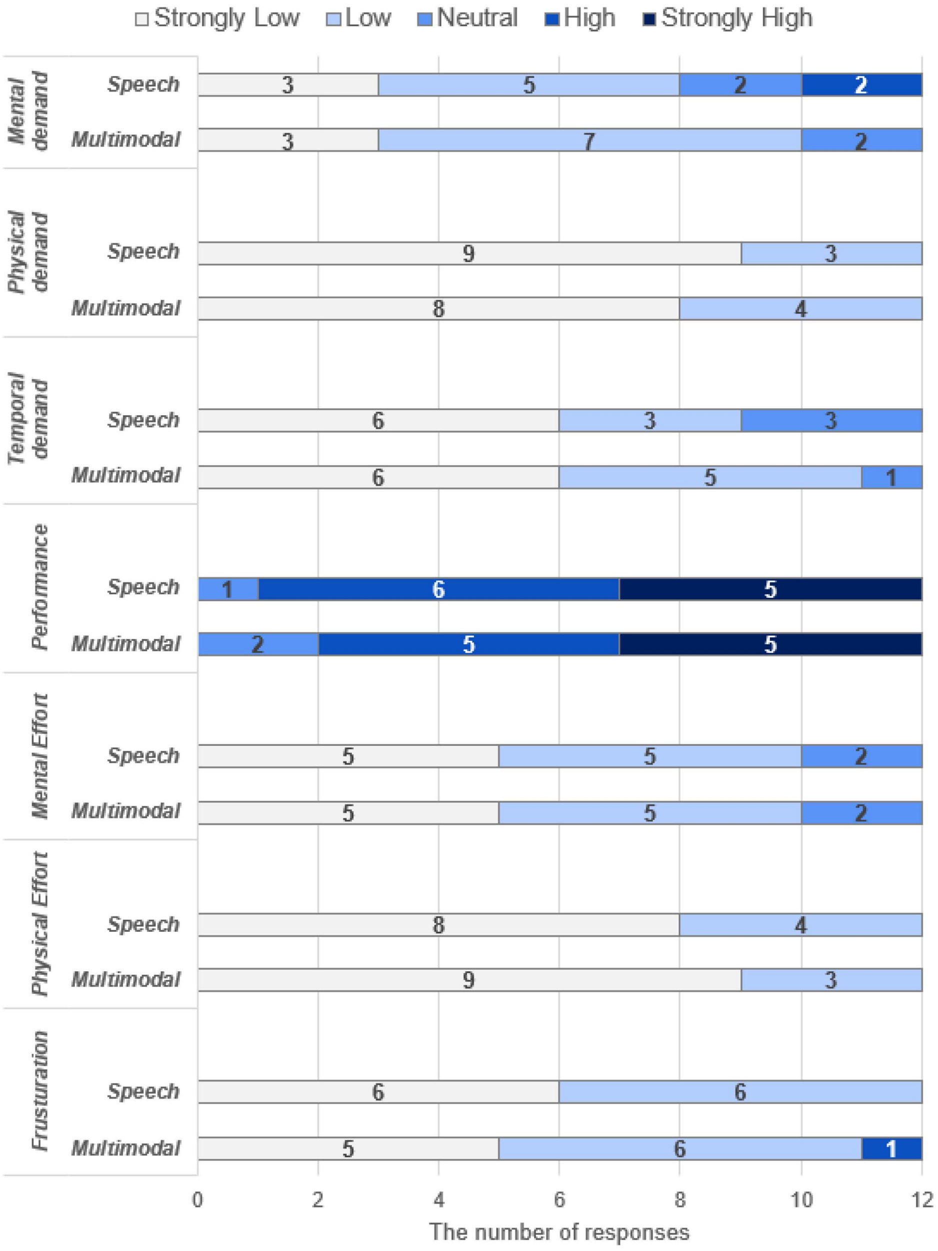
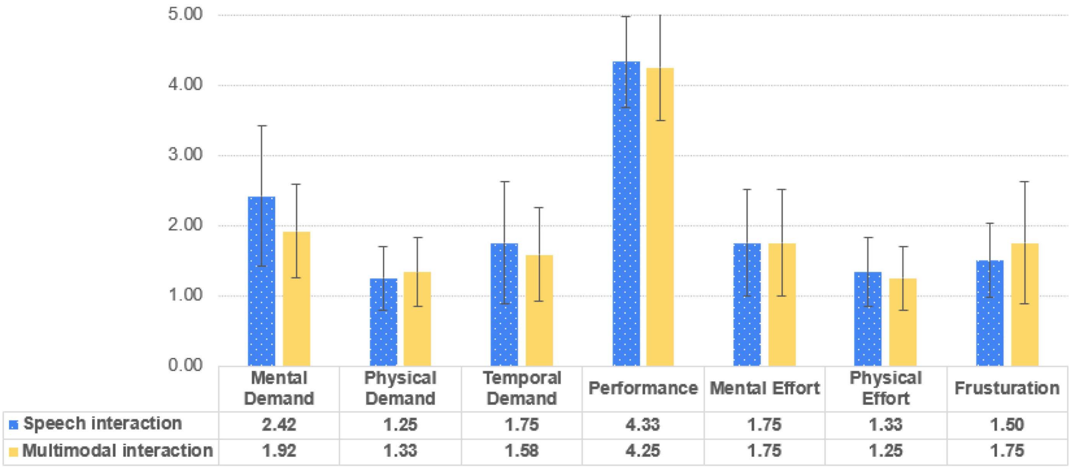


Fig. 12. Distribution of NASA TLX results.



	Mental Demand	Physical Demand	Temporal Demand	Performance	Mental Effort	Physical Effort	Frusturation
■ Speech interaction	2.42	1.25	1.75	4.33	1.75	1.33	1.50
■ Multimodal interaction	1.92	1.33	1.58	4.25	1.75	1.25	1.75

Fig. 13. Mean scores of NASA TLX results.

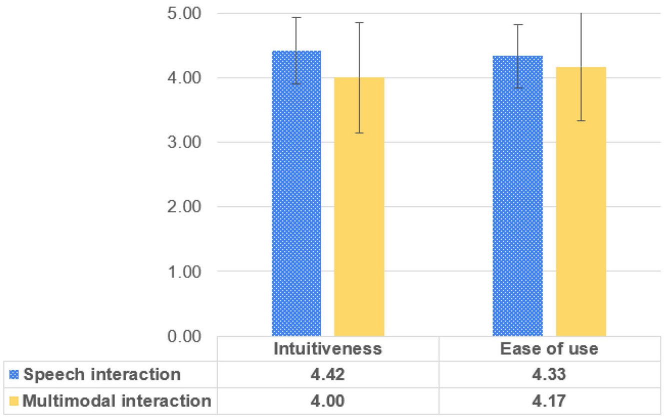


Fig. 14. Mean scores for intuitiveness and ease of use assessment.

Additionally, one high rating in the frustration category for multimodal interaction indicates possible issues within this method that could lead to user frustration.

Fig. 13 shows the mean workload scores of each dimension, including error bars that represent the standard deviation. The comparison between the two interaction methods showed a minimal difference, with the discrepancy in scores across all criteria being 0.5 or less. This finding highlights the relatively equivalent workload perception between speech and multimodal interactions among the participants in the experiment. Regarding the variability, the standard deviations for the speech interaction method ranged from 0.452 for Physical Demand to 0.996 for Mental Demand, reflecting a higher variability in perceived mental demand among participants. Conversely, the multimodal interaction method showed more consistent variability across dimensions, with most values hovering around 0.753. This consistency suggests a more uniform participant response in the multimodal setting.

Intuitiveness and Ease of Use

To assess the interaction modalities, two specific statements were used: "This interaction method was intuitive to interact with a robot" was used to evaluate intuitiveness, and "This interaction method was easy to interact with a robot" was used to measure ease of use. These items have been utilized in previous research to evaluate robot interaction methods (Nieuwenhuisen et al. 2010; Szafir et al. 2015; Fischinger et al. 2016; Campeau-Lecours et al. 2018). Participants expressed their level of agreement with each statement using a five-point Likert scale, ranging from one (strongly disagree) to five (strongly agree). Fig. 14 shows the average scores of the responses, including standard deviation error bars to illustrate variability in responses to intuitiveness and ease of use.

Although users favored speech interaction slightly more in usability, that both methods were perceived as relatively intuitive and easy to use, with average scores equal to or above four out of five. The standard deviations for speech interaction, at 0.514 for intuitiveness and 0.492 for ease of use, suggest a relatively consistent perception among users. In contrast, the multimodal interaction method displayed higher variability, with standard deviations of 0.853 for intuitiveness and 0.835 for ease of use. This variability indicates that user experiences with the multimodal interaction method were more diverse compared with speech interaction.

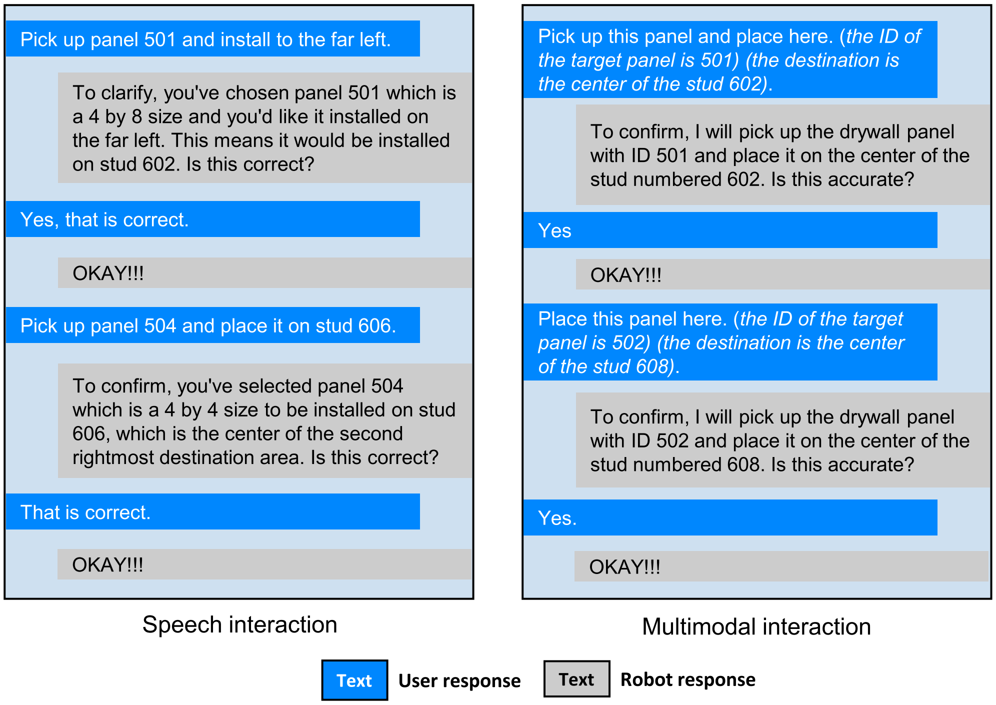
Preferences

To understand participants' interaction preferences within the VR environment, the survey included two questions. Initially, participants were asked "Which type of interaction do you prefer in the VR environment?" Subsequently, to gain insight into their choices, the question "Why do you prefer that interaction?" was posed. This led to 66.7% of participants (8 out of 12) favoring multimodal interaction with the remaining 33.3% (4 out of 12) opting for speech interaction. Participants' justifications for their preferred interaction method varied and could be systematically categorized into five themes: efficiency, accuracy, ease of use, engagement, and versatility. These categorizations are detailed in Table 3, which summarizes the participants' responses.

A notable observation is that both speech and multimodal interactions were commended for their efficiency and accuracy. Participants who preferred speech interaction highlighted its speed and precision as the reason of their choice and noted a perceived redundancy in the multimodal approach's combination of gestures and speech. In contrast, those who preferred multimodal interaction valued its higher accuracy and efficiency, noting particularly the consequent decrease in verbal miscommunication. Furthermore, multimodal interaction was also noted for its versatility because it allows for an integration of verbal and gestural communication, which could influence the more natural and engaging way to interact with the robot. The capability to use speech and handheld controller was lauded for its ease of use, contributing to the perceived engagement and versatility of the multimodal method. These factors outline the users' preferences for their respective interaction methods in a VR environment.

Table 3. User feedback on preferences for two interaction methods: speech and multimodal

Interaction method	Theme	Statements
Speech interaction	Efficiency	• Speech interaction seems faster. • In multimodal interaction, duplicating hand gestures and some speech seems a bit redundant.
	Accuracy	• Speech was more accurate. • I can be thorough with my intent during speech interaction.
Multimodal interaction	Efficiency	• It seemed quicker. • It just seemed to be more efficient and quicker.
	Accuracy	• More accurate input, less opportunity to miss-speak or be misunderstood.
	Ease of use	• I feel like it is easier for me to use my words and hands at the same time when I am working. • I believe that it is easier to be able to point at an object then to have the exact identity. • It was an easier version of communication. I am able to just point to an object and instruct the robot to perform a task.
	Engagement	• I felt more involved in giving the commands to the robot
	Versatility	• To be able to physically give instructions as well as verbal instructions.



Speech interaction

Pick up panel 501 and install to the far left.

To clarify, you've chosen panel 501 which is a 4 by 8 size and you'd like it installed on the far left. This means it would be installed on stud 602. Is this correct?

Yes, that is correct.

OKAY!!!

Pick up panel 504 and place it on stud 606.

To confirm, you've selected panel 504 which is a 4 by 4 size to be installed on stud 606, which is the center of the second rightmost destination area. Is this correct?

That is correct.

OKAY!!!

Multimodal interaction

Pick up this panel and place here. (*the ID of the target panel is 501) (the destination is the center of the stud 602)*.

To confirm, I will pick up the drywall panel with ID 501 and place it on the center of the stud numbered 602. Is this accurate?

Yes

OKAY!!!

Place this panel here. (*the ID of the target panel is 502) (the destination is the center of the stud 608)*.

To confirm, I will pick up the drywall panel with ID 502 and place it on the center of the stud numbered 608. Is this accurate?

Yes.

OKAY!!!

Fig. 15. Examples of conversation for success panel installation.

Performance of Chat Application

Regarding the drywall panel installation, the execution of robot tasks was not validated, focusing instead on assessing the effectiveness of interaction methods and the accurate interpretation and transmission of task information to the robot. This approach was based on the assumption that accurate communication would ensure error-free task execution by the robot. The interaction system design includes the Approval button, shown in Fig. 8, which requires operator confirmation before any task information is sent to the robot. This setup allows operators to address and rectify any misinterpretations by resubmitting the instructions. Throughout the experiment, the chat application displayed commendable accuracy, generating responses that were consistent with the given instructions. During the experiment, 12 participants successfully executed four panel installations, with Fig. 15 illustrating examples of conversations in two interaction ways.

There were no instances where GPT misinterpreted or responded incorrectly to correct user instructions. Notably, three participants sent blank messages to GPT by pressing the Send button without saying anything. In response to these instances, GPT followed the prompt for handling the absence of instructions, offering responses such as "How can I assist you further?" and "I am sorry, but I need your confirmation. Can you confirm that the information is correct?"

In analyzing the communication length between interaction methods, a notable difference was found in the average number of words per instruction: speech interaction commands averaged 8.27 words, with the longest command "Okay, robot, could you please pick up Panel 503 and install it at the rightmost portion of the framing?," reaching 19 words, and the shortest being "Panel 504 to Stud 606," which contains five words. Conversely, multimodal interaction commands were more concise, averaging 6.65 words. The longest command was "Okay, robot, could you please pick up Panel 503 and install it at the rightmost portion of the framing?" comprising 18 words, whereas the briefest instruction, "install this here," consisted of just three words. This distinction not only illuminates the multimodal interface's capacity to support more succinct communication but also its effectiveness in making interactions more streamlined, marking a critical enhancement in the chat interface's role in facilitating efficient and accurate HRC tasks.

Additionally, the participants intentionally issued a total of 55 incorrect instructions to test the system's capability. The chat system, powered by GPT, accurately pinpointed errors in 51 cases, achieving an accuracy rate of 92.73%. These results are described in Table 4, which details the types of incorrect instructions and the frequency with which GPT recognized and addressed the issues.

Table 4. Chat system's detection of incorrect instructions

	Speech interaction		Multimodal interaction	
Incorrect instructions	Number of cases	Issues detected	Number of cases	Issues detected
Mismatched pairing	22	22	11	10
Materials not present	4	4	1	1
Component already installed	2	1	5	3
Partial information	0	0	10	10
Total	28	27	27	24

During the speech interaction phase, the system accurately detected 27 out of 28 issues, resulting in a 97.43% accuracy rate. All cases of Mismatched Pairing and Materials Not Present were correctly identified, but one Component Already Installed case was missed. In the multimodal interaction phase, the system correctly identified 24 out of 27 issues, yielding an accuracy of 88.89%. It successfully detected all Partial Information cases but struggled with Mismatched Pairing and Component Already Installed, missing one and two cases, respectively. Future improvements could focus on enhancing detection accuracy in these challenging areas to increase overall system robustness.

Fig. 16 further exemplifies the system's reactions to incorrect instructions across four distinct categories, with the incorrect instructions indicated by dotted lines and the inputs from handheld controllers highlighted in italics. The GPT component within the chat system showcased its analytical prowess by not only identifying wrong instructions but also by proposing alternative actions where applicable in some cases. For example, when the instruction incorrectly directed to place Panel 504 on Stud 605 instead of its correct paired destination, Stud 606, the chat system inquired, "would you like to install Panel 504 on Stud 606 instead?"

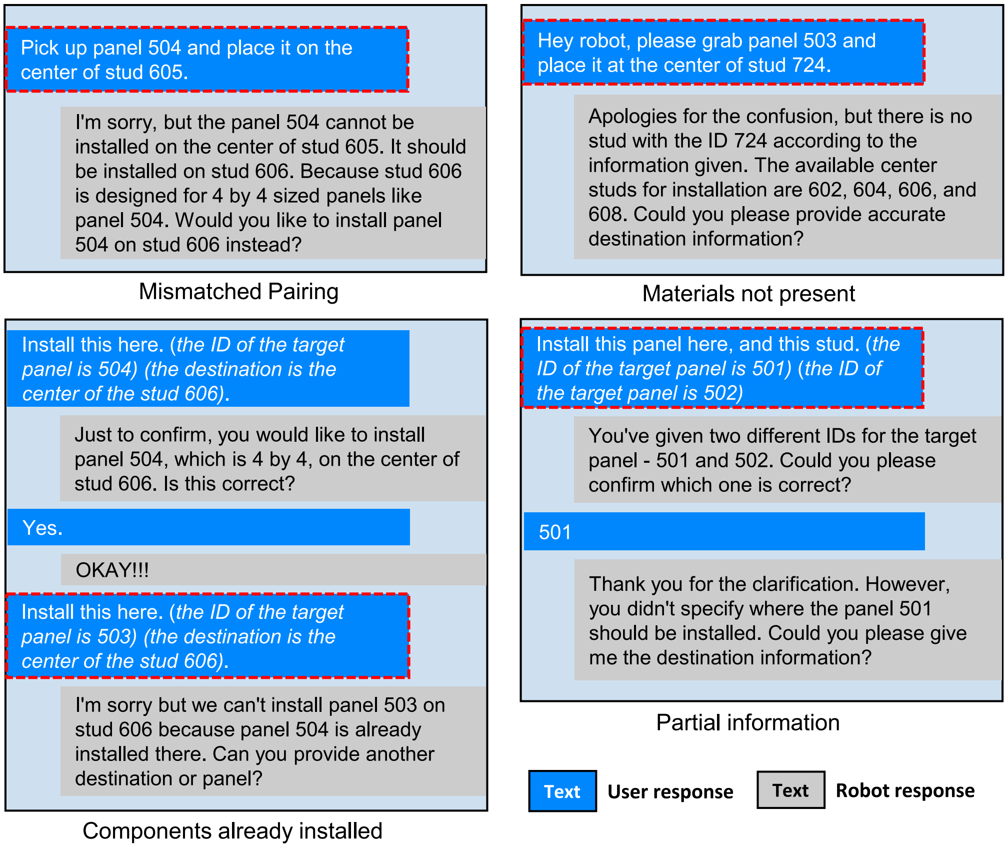
However, the system's detection was not infallible; it missed an error in one instance of mismatched pairing and in three cases where components were not reported as already installed, prompting the system to request user confirmation. The examples of these errors are shown in Fig. 17. For example, even though Panel 501 cannot be installed on Stud 608 because the Panel 503 is already installed there, it did not catch it and it just asked if its understanding is correct. Another example shows that the chat system did not catch the fact that the Panel 502 cannot be installed on the Stud 606 designed for 4 by 4-sized panels like Panel 504.

Discussion

The user study conducted as part of this study has demonstrated the potential for successful deployment of the proposed multimodal interaction system within VR interfaces, integrated with a chat application, for HRC. In a case study focused on drywall installation, construction workers reported experiencing a low workload and high usability with the multimodal system. Although the assessment of workload and usability between the two interaction methods showed similar results, this uniformity may stem from the simplicity of the task, which possibly did not challenge the participants enough to discern a significant difference. This aspect suggests a need for further exploration with more complex tasks to truly gauge the differential impact on usability and workload.

However, the findings indicated a strong preference for the multimodal approach, with two-thirds of participants selecting it over speech interaction. This preference was influenced by the multimodal system's perceived efficiency, accuracy, ease of use, emergence, and versatile communication combining verbal and gestural inputs. The preference may also stem from the multimodal systems' demonstration of more concise commands on average than those of speech interaction, indicating a streamlined communication process. Efficiency and accuracy contributed to a marked preference for speech interaction, yet they were also cited as advantages of multimodal interaction. The versatile nature of multimodal interaction contributed to ease of use and helped to deepen user engagement in HRC tasks.

The GPT-4 model exhibited a high degree of responsiveness within the chat application, showing its sophisticated reasoning capabilities. Nonetheless, it did not achieve perfect accuracy. To address this, this study incorporated an additional module that



Pick up panel 504 and place it on the center of stud 605.

I'm sorry, but the panel 504 cannot be installed on the center of stud 605. It should be installed on stud 606. Because stud 606 is designed for 4 by 4 sized panels like panel 504. Would you like to install panel 504 on stud 606 instead?

Mismatched Pairing

Hey robot, please grab panel 503 and place it at the center of stud 724.

Apologies for the confusion, but there is no stud with the ID 724 according to the information given. The available center studs for installation are 602, 604, 606, and 608. Could you please provide accurate destination information?

Materials not present

Install this here. (*the ID of the target panel is 504) (the destination is the center of the stud 606)*.

Just to confirm, you would like to install panel 504, which is 4 by 4, on the center of stud 606. Is this correct?

Yes.

OKAY!!!

Install this here. (*the ID of the target panel is 503) (the destination is the center of the stud 606)*.

I'm sorry but we can't install panel 503 on stud 606 because panel 504 is already installed there. Can you provide another destination or panel?

Components already installed

Install this panel here, and this stud. (*the ID of the target panel is 501) (the ID of the target panel is 502)*

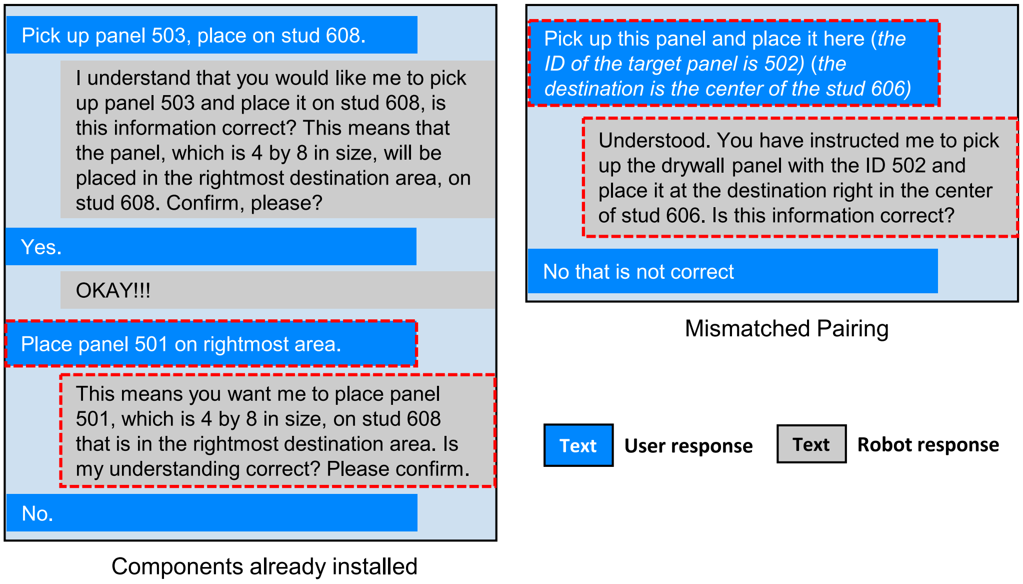
You've given two different IDs for the target panel - 501 and 502. Could you please confirm which one is correct?

501

Thank you for the clarification. However, you didn't specify where the panel 501 should be installed. Could you please give me the destination information?

Partial information

Fig. 16. Examples that correctly caught the incorrect instructions.



Pick up panel 503, place on stud 608.

I understand that you would like me to pick up panel 503 and place it on stud 608, is this information correct? This means that the panel, which is 4 by 8 in size, will be placed in the rightmost destination area, on stud 608. Confirm, please?

Yes.

OKAY!!!

Place panel 501 on rightmost area.

This means you want me to place panel 501, which is 4 by 8 in size, on stud 608 that is in the rightmost destination area. Is my understanding correct? Please confirm.

No.

Pick up this panel and place it here (*the ID of the target panel is 502) (the destination is the center of the stud 606)*

Understood. You have instructed me to pick up the drywall panel with the ID 502 and place it at the destination right in the center of stud 606. Is this information correct?

No that is not correct

Fig. 17. Examples that did not catch the incorrect instructions.

allowed the GPT-4 to query its own interpretations, thereby enabling human operators to review and correct any misinterpretations. During the experiments, the chat system demonstrated an impressive accuracy rate of 92.73%, correctly identifying errors in 51 out of 55 intentional incorrect instructions. However, the system's occasional failure to detect errors in instructions underscored the need for continued human oversight.

These findings highlight the potential of integrating intuitive multimodal interfaces with AI-driven chat systems in HRC in the construction industry. By reducing cognitive load and enhancing task accuracy, this innovative approach paves the way for more efficient, reliable, and user-centric HRC systems. These advancements are not only expected to improve operational efficiency in construction but also provide substantial support to construction workers in their daily tasks. Furthermore, the experimental results emphasize the critical role of a well-designed human–AI interface. Although AI demonstrates a capability to reduce the cognitive load of human operators by accurately interpreting and responding to most instructions, the necessity for human oversight remains. This balanced approach could potentially lead to enhanced precision in task execution and a reduction in operational errors, fostering a more efficient and reliable HRC system in construction.

Conclusion

This paper proposed a multimodal interaction system for HRC in construction, leveraging VR to enhance the interaction between human workers and robots. The proposed system integrated speech and handheld controller inputs to enable easy and intuitive communication with construction robots. It employed VR controllers to point at objects of interest and NL commands to specify tasks. Furthermore, the system integrated BIM for material data retrieval, a robotic operation system for robot control, and GPT-based chat system for bidirectional communication. The practical application of the system was demonstrated through a drywall installation task, validated by 12 construction workers. Their successful completion of the task using the multimodal interaction highlighted the system's low workload and high levels of intuitiveness and ease of use.

This study makes several key contributions to the field of HRC in construction, primarily through the proposal and implementation of a multimodal interaction system. First, the system integrates speech and handheld controller inputs with the use of BIM data within Unity. BIM data were used not only for providing visual information but also as a functional part of the operational workflow. Building components selected during construction activities were used to trigger robot actions, enhancing interaction and operational efficiency.

Second, a significant aspect of this study is the successful integration of diverse software components, including BIM, ROS, external servers for OpenAI API, and a game engine like Unity. This strategic integration not only augments the functionality of each individual component but also ensures that the entire system operates efficiently in a cohesive manner. This allows for intuitive communication and seamless collaboration between construction workers and robotic assistants, leading to more effective collaboration than could be achieved by any single technology independently. Moreover, the system's architecture is designed with scalability, enabling it to expand to support a variety of construction activities for HRC.

Additionally, the study extends domain knowledge by designing GPT prompts and proposing a conversation flow for HRC, showcasing the potential of advanced AI assistants in enhancing HRC. Next, the user study with construction workers provided in-depth qualitative and quantitative analyses on the HRC experiences, offering valuable insights into the system's operational effectiveness and user satisfaction.

However, there are several limitations that should be addressed in future research. Firstly, the case study focused on drywall installation which represents one aspect of construction pick-and-place tasks and limited to two actions of pick and place rather than various grounding actions. Future studies should expand the proposed interaction system to include a wider variety of construction activities and structures. Furthermore, the management of ground actions, including moving, tilting, and gripping within construction activities, should be systematically incorporated into future studies to enhance the applicability and robustness of the findings.

Secondly, this study has the lack of automatic integration of the BIM data to the GPT prompt. This necessitated the manual inclusion of semantic information about workpieces in the GPT prompt. Automating this process in future research could significantly enhance the chat system for HRC, leading to more advanced task management and user interaction.

Thirdly, the absence of female participants in the experiment might introduce a gender bias in both the delivery of instructions and in perceptions related to workload and usability. This limitation could potentially affect the generalizability of the study's findings across different demographic groups. In future studies, efforts will be made to recruit a more diverse participant pool to address and mitigate this issue.

Fourthly, the system to display all of object IDs around corresponding objects in the virtual environment may not be effective for more complex construction tasks that involve intricate structures and numerous materials because it might confuse operators. Future implementations might require additional modules that enable users to access only the information pertinent to specific queries or to selectively display data, enhancing effective information management and accessibility.

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request. This includes user study data, such as conversation transcripts, survey questions, and responses.

Acknowledgments

The work presented in this paper was supported financially by two United States National Science Foundation (NSF) Award Nos. 2025805 and 2128623. The support of the NSF is gratefully acknowledged.

References

Achiam, J., et al. 2023. "GPT-4 technical report." Preprint, submitted March 15, 2023. https://arxiv.org/abs/2303.08774.

Adami, P., P. B. Rodrigues, P. J. Woods, B. Becerik-Gerber, L. Soibelman, Y. Copur-Gencturk, and G. Lucas. 2021. "Effectiveness of VR-based training on improving construction workers' knowledge, skills, and safety behavior in robotic teleoperation." *Adv. Eng. Inf.* 50 (Oct): 101431. https://doi.org/10.1016/j.aei.2021.101431.

Brosque, C., E. Galbally, O. Khatib, and M. Fischer. 2020. "Human-robot collaboration in construction: Opportunities and challenges." In *Proc., 2020 Int. Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 1–8. New York: IEEE.

Campeau-Lecours, A., U. Côté-Allard, D. S. Vu, F. Routhier, B. Gosselin, and C. Gosselin. 2018. "Intuitive adaptive orientation control for enhanced human–robot interaction." *IEEE Trans. Rob.* 35 (2): 509–520. https://doi.org/10.1109/TRO.2018.2885464.

Chen, H., L. Hou, S. Wu, G. Zhang, Y. Zou, S. Moon, and M. Bhuiyan. 2024. "Augmented reality, deep learning and vision-language query system for construction worker safety." *Autom. Constr.* 157 (Jan): 105158. https://doi.org/10.1016/j.autcon.2023.105158.

Chen, H., M. C. Leu, and Z. Yin. 2022. "Real-time multi-modal human–robot collaboration using gestures and speech." *J. Manuf. Sci. Eng.* 144 (10): 101007. https://doi.org/10.1115/1.4054297.

Chen, M., P. Zhang, Z. Wu, and X. Chen. 2020. "A multichannel human-swarm robot interaction system in augmented reality." *Virtual Reality Intell. Hardware* 2 (6): 518–533. https://doi.org/10.1016/j.vrih.2020.05.006.

Cheng, S., K. Mo, and L. Shao. 2021. "Learning to regrasp by learning to place." Preprint, submitted November 17, 2021. https://doi.org/10.48550/arXiv.2109.08817.

Constantin, S., F. I. Eyiokur, D. Yaman, L. Bärmann, and A. Waibel. 2022. "Interactive multimodal robot dialog using pointing gesture recognition." In *Proc., European Conf. on Computer Vision*, 640–657. Cham, Switzerland: Springer.

Eiris-Pereira, R., and M. Gheisari. 2018. "Building intelligent virtual agents as conversational partners in digital construction sites." In *Proc., Construction Research Congress 2018: Construction Information Technology*, 200–209. Reston, VA: ASCE.

Elghaish, F., J. K. Chauhan, S. Matarneh, F. P. Rahimian, and M. R. Hosseini. 2022. "Artificial intelligence-based voice assistant for BIM data management." *Autom. Constr.* 140 (Aug): 104320. https://doi.org/10.1016/j.autcon.2022.104320.

Fischinger, D., et al. 2016. "Hobbit, a care robot supporting independent living at home: First prototype and lessons learned." *Rob. Auton. Syst.* 75 (Jan): 60–78. https://doi.org/10.1016/j.robot.2014.09.029.

Goldin-Meadow, S., H. Nusbaum, S. D. Kelly, and S. Wagner. 2001. "Explaining math: Gesturing lightens the load." *Psychol. Sci.* 12 (6): 516–522. https://doi.org/10.1111/1467-9280.00395.

Grammel, L., M. Tory, and M. A. Storey. 2010. "How information visualization novices construct visualizations." *IEEE Trans. Visual Comput. Graphics* 16 (6): 943–952. https://doi.org/10.1109/TVCG.2010.164.

Hart, S. G. 2006. "NASA-task load index (NASA-TLX); 20 years later." In Vol. 50 of *Proc., Human Factors and Ergonomics Society Annual Meeting*, 904–908. Los Angeles: SAGE.

Hoonakker, P., P. Carayon, A. P. Gurses, R. Brown, A. Khunlertkit, K. McGuire, and J. M. Walker. 2011. "Measuring workload of ICU nurses with a questionnaire survey: The NASA Task Load Index (TLX)." *IIE Trans. Healthcare Syst. Eng.* 1 (2): 131–143. https://doi.org/10.1080/19488300.2011.609524.

Hussain, R., A. Sabir, D. Y. Lee, S. F. A. Zaidi, A. Pedro, M. S. Abbas, and C. Park. 2024. "Conversational AI-based VR system to improve construction safety training of migrant workers." *Autom. Constr.* 160 (Apr): 105315. https://doi.org/10.1016/j.autcon.2024.105315.

Jang, S., G. Lee, J. Oh, J. Lee, and B. Koo. 2024. "Automated detailing of exterior walls using NADIA: Natural-language-based architectural detailing through interaction with AI." *Adv. Eng. Inf.* 61 (Aug): 102532. https://doi.org/10.1016/j.aei.2024.102532.

Johari, S., and K. N. Jha. 2021. "Exploring the relationship between construction workers' communication skills and their productivity." *J. Manage. Eng.* 37 (3): 04021009. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000904.

Kim, H., K. Mitra, R. Chen, S. Rahman, and D. Zhang. 2024. "MEGAnno+: A human-LLM collaborative annotation system." Preprint, submitted September 28, 2024. https://doi.org/10.48550/arXiv.2402.18050.

Lee, M., M. Billinghurst, W. Baek, R. Green, and W. Woo. 2013. "A usability study of multimodal input in an augmented reality environment." *Virtual Reality* 17 (Nov): 293–305. https://doi.org/10.1007/s10055-013-0230-0.

Li, J., H. Li, H. Wang, W. Umer, H. Fu, and X. Xing. 2019. "Evaluating the impact of mental fatigue on construction equipment operators' ability to detect hazards using wearable eye-tracking technology." *Autom. Constr.* 105 (Sep): 102835. https://doi.org/10.1016/j.autcon.2019.102835.

Lin, J.-R., Z.-Z. Hu, J.-P. Zhang, and F.-Q. Yu. 2016. "A natural-language-based approach to intelligent data retrieval and representation for cloud BIM." *Comput.-Aided Civ. Infrastruct. Eng.* 31 (1): 18–33. https://doi.org/10.1111/mice.12151.

Linares-Garcia, D. A., N. Roofigari-Esfahan, K. Pratt, and M. Jeon. 2022. "Voice-based intelligent virtual Agents (VIVA) to support construction worker productivity." *Autom. Constr.* 143 (Nov): 104554. https://doi.org/10.1016/j.autcon.2022.104554.

Marge, M., et al. 2022. "Spoken language interaction with robots: Recommendations for future research." *Comput. Speech Lang.* 71 (Jan): 101255. https://doi.org/10.1016/j.csl.2021.101255.

McNeel. 2023. "Rhino—Rhino.Inside." Accessed October 26, 2023. https://www.rhino3d.com/features/rhino-inside/.

Nieuwenhuisen, M., J. Gaspers, O. Tischler, and S. Behnke. 2010. "Intuitive multimodal interaction and predictable behavior for the museum tour guide robot Robotinho." In *Proc., 2010 10th IEEE-RAS Int. Conf. on Humanoid Robots*, 653–658. New York: IEEE.

OpenAI (Artificial Intelligence). 2023. "Whisper." Accessed October 25, 2023. https://openai.com/research/whisper.

Park, S., C. C. Menassa, V. R. Kamat, and J. Y. Chai. 2024. "Natural language instructions for intuitive human interaction with robotic assistants in field construction work." *Autom. Constr.* 161 (May): 105345. https://doi.org/10.1016/j.autcon.2024.105345.

Park, S., H. Yu, C. C. Menassa, and V. R. Kamat. 2023. "A comprehensive evaluation of factors influencing acceptance of robotic assistants in field construction work." *J. Manage. Eng.* 39 (3): 04023010. https://doi.org/10.1061/JMENEA.MEENG-5227.

Paul, R., J. Arkin, N. Roy, and T. M. Howard. 2016. "Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators." *Robot. Sci. Syst. Found.* 37 (10): 1–9. https://doi.org/10.15607/RSS.2016.XII.037.

Prieto, S. A., E. T. Mengiste, and B. García de Soto. 2023. "Investigating the use of ChatGPT for the scheduling of construction projects." *Buildings* 13 (4): 857. https://doi.org/10.3390/buildings13040857.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2023. "Robust speech recognition via large-scale weak supervision." In *Proc., Int. Conf. on Machine Learning*, 28492–28518. New York: Proceedings of Machine Learning Research.

Rossen, B., S. Lind, and B. Lok. 2009. "Human-centered distributed conversational modeling: Efficient modeling of robust virtual human conversations." In *Proc., 9th Int. Conf. on Intelligent Virtual Agents, IVA 2009*, 474–481. Berlin: Springer.

Saka, A. B., L. O. Oyedele, L. A. Akanbi, S. A. Ganiyu, D. W. Chan, and S. A. Bello. 2023. "Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities." *Adv. Eng. Inf.* 55 (Jan): 101869. https://doi.org/10.1016/j.aei.2022.101869.

Shin, S., and R. R. Issa. 2021. "BIMASR: Framework for voice-based BIM information retrieval." *J. Constr. Eng. Manage.* 147 (10): 04021124. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002138.

Szafir, D., B. Mutlu, and T. Fong. 2015. "Communicating directionality in flying robots." In *Proc., 10th Annual ACM/IEEE Int. Conf. on Human-Robot Interaction*, 19–26. New York: Association for Computing Machinery.

Trad, F., and A. Chehab. 2024. "Prompt engineering or fine-tuning? A case study on phishing detection with large language models." *Mach. Learn. Knowl. Extr.* 6 (1): 367–384. https://doi.org/10.3390/make6010018.

Van den Bergh, M., D. Carton, R. De Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, and M. Buss. 2011. "Real-time 3D hand gesture interaction with a robot for understanding directions from humans." In *Proc., 2011 Ro-Man*, 357–362. New York: IEEE.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. "Attention is all you need." Preprint, submitted June 12, 2017. https://arxiv.org/abs/1706.03762.

Wagner, P., Z. Malisz, and S. Kopp. 2014. "Gesture and speech in interaction: An overview." *Speech Commun.* 57 (Feb): 209–232. https://doi.org/10.1016/j.specom.2013.09.008.

Wang, X., C.-J. Liang, C. C. Menassa, and V. R. Kamat. 2021. "Interactive and immersive process-level digital twin for collaborative human–robot construction work." *J. Comput. Civ. Eng.* 35 (6): 04021023. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000988.

Wang, X., S. Wang, C. C. Menassa, V. R. Kamat, and W. McGee. 2023. "Automatic high-level motion sequencing methods for enabling multi-tasking construction robots." *Autom. Constr.* 155 (Nov): 105071. https://doi.org/10.1016/j.autcon.2023.105071.

Wang, X., H. Yu, W. McGee, C. C. Menassa, and V. R. Kamat. 2024. "Enabling building information model-driven human-robot collaborative construction workflows with closed-loop digital twins." *Comput. Ind.* 161 (Oct): 104112. https://doi.org/10.1016/j.compind.2024.104112.

Wen, J., and M. Gheisari. 2023. "*iVisit-communicate* for AEC education: Using virtual humans to practice communication skills in 360-degree virtual field trips." *J. Comput. Civ. Eng.* 37 (3): 04023008. https://doi.org/10.1061/JCCEE5.CPENG-5165.

Ye, Y., H. You, and J. Du. 2023. "Improved trust in human-robot collaboration with ChatGPT." *IEEE Access* 11 (Jun): 55748–55754. https://doi.org/10.1109/ACCESS.2023.3282111.

Yongda, D., L. Fang, and X. Huang. 2018. "Research on multimodal human-robot interaction based on speech and gesture." *Comput. Electr. Eng.* 72 (Nov): 443–454. https://doi.org/10.1016/j.compeleceng.2018.09.014.

Yoon, S., Y. Kim, M. Park, and C. R. Ahn. 2023. "Effects of spatial characteristics on the human–robot communication using deictic gesture in construction." *J. Constr. Eng. Manage.* 149 (7): 04023049. https://doi.org/10.1061/JCEMD4.COENG-12997.

Zheng, J., and M. Fischer. 2023. "Dynamic prompt-based virtual assistant framework for BIM information search." *Autom. Constr.* 155 (Nov): 105067. https://doi.org/10.1016/j.autcon.2023.105067.