

PACA: A Personal Autonomous Construction Assistant to Facilitate Human-Robot Interaction

Samuel A. Prieto, *Member, IEEE*, Borja García de Soto, *Member, IEEE*

Abstract—Integrating robots into construction workflows is challenging due to the complex and dynamic nature of construction sites, and the diverse workforce, many of whom lack formal training in robotics. To address these challenges, we introduce the Personal Autonomous Construction Assistant (PACA), a multimodal natural Large Language Model (mmLLM)-based interface designed to facilitate seamless communication between workers and robots in construction environments. PACA incorporates an Automatic Speech Recognition (ASR) engine to process spoken instructions in different languages, utilizing a Large Language Model (LLM) to interpret user intent and generate appropriate responses. When necessary, the LLM translates the command into robot actions, incorporating image-based reasoning to enhance decision-making. The system also features a Text-to-Speech (TTS) engine, providing spoken feedback to ensure fully bidirectional and accessible interaction with construction workers.

I. INTRODUCTION

The construction industry has traditionally been slow to adopt automation compared to other sectors such as manufacturing or logistics [1]. While robotic solutions have been successfully integrated into controlled environments, construction sites present unique challenges that make robotic adoption particularly difficult. These environments are highly dynamic, often cluttered, and subject to unpredictable changes [2], requiring robots to operate under conditions that are significantly different from structured indoor settings. Moreover, construction tasks are frequently executed by a diverse workforce, with varying levels of technical expertise, which introduces an additional barrier to seamless human-robot collaboration.

Recent advancements in robotics and artificial intelligence (AI) have sparked growing interest in the use of autonomous systems for construction. Robots have been proposed for tasks such as bricklaying [3], material transport [4], and structural inspection [5], aiming to improve efficiency, safety, and cost-effectiveness. However, a major limitation of existing approaches is the lack of intuitive and accessible interfaces for human-robot interaction [6]. Traditional robot control mechanisms, such as programming interfaces or tablet-based inputs, are often impractical for on-site workers who may not have the technical background to operate them effectively. The need for an intuitive, user-friendly interface that bridges the communication gap between construction workers and robotic assistants is evident.

This paper provides an insight into a Personal Autonomous

Construction Assistant (PACA) designed to address these challenges by introducing a multimodal natural language-based communication framework for human-robot interaction. At its core, PACA integrates a Large Language Model (LLM) with an Automatic Speech Recognition (ASR) system and a Text-to-Speech (TTS) engine, enabling workers to interact with robots through spoken commands in their native language. The system processes user input, determines the appropriate robot action, and generates spoken feedback, ensuring a fully bidirectional interaction. Additionally, the multimodal capabilities of PACA allow the robot to incorporate visual information into its decision-making process, further enhancing its ability to operate in dynamic construction environments.

The significance of PACA extends beyond convenience, it represents a crucial step to facilitate the integration of robots in construction sites. By enabling natural language interactions, PACA lowers the barrier to entry for workers who may lack formal training in robotics, making robotic workforce more accessible and practical in real-world construction settings. Furthermore, the ability to support multiple languages facilitates the collaboration of multinational teams with automated systems, reducing potential miscommunication and improving overall site efficiency.

This paper presents an overview of PACA, outlining its methodology, implementation, and expected impact on the construction industry. The remainder of the paper is structured as follows: Section 2 provides an overview of the state of the art in construction automation and human-robot interaction. Section 3 describes the methodology behind PACA's mmLLM-based workflow, detailing the key components of the system. Section 4 discusses the implementation specifics, including the robotic platform, language models, and multimodal processing capabilities. Finally, Section 5 presents conclusions and outlines future research directions to further enhance PACA's capabilities.

II. STATE OF THE ART

The adoption of automation in construction has significantly gained momentum due to its potential to alleviate the persistent issues of labor shortages, low productivity, and safety risks prevalent in construction environments. Unlike structured manufacturing facilities,

construction sites present dynamic, cluttered, and often hazardous conditions, complicating the direct adoption of robotic systems developed for controlled systems. Despite their proven benefits, existing robotic systems in construction are limited by their rigidity and inability to effectively adapt to unpredictable environments without extensive human supervision. The critical factor limiting wider adoption is the complexity of human-robot interactions (HRI) within these dynamic environments.

Emerging research underscores the importance of intuitive human-robot interaction (HRI) and collaboration (HRC) methods as crucial for safe and effective operation in complex construction scenarios [7]. Recent studies advocate for the integration of natural language processing (NLP) and multimodal approaches to facilitate more accessible and user-friendly robot interactions. For instance, the integration of Large Language Models (LLMs) with virtual reality (VR) interfaces, as demonstrated by Park et al. [8], has allowed construction workers to interact intuitively with robotic systems through simple voice commands and gestural inputs, effectively reducing cognitive load and communication errors.

The rapid advancements in NLP, particularly the adoption of LLMs, have revolutionized human-robot interaction capabilities. LLMs can interpret nuanced user instructions, translate natural language into structured robot commands, and provide human-understandable feedback, significantly lowering barriers for workers without specialized robotics training. The ROSGPT framework by Koubaa et al. [9] exemplifies this potential, enabling robots to interpret and execute instructions based on natural speech interactions across diverse robotic platforms.

Furthermore, multimodal AI technologies that fuse voice, gesture, and image-based inputs offer enhanced context-awareness and robust interaction capabilities, especially in unstructured and noisy environments typical of construction sites. Lai et al. [10] demonstrated a multimodal system combining voice commands and gestures with an LLM to precisely control robot manipulators, effectively addressing the ambiguity and limitations inherent in single-modality interfaces. Despite these advances, multimodal AI integration remains limited by real-time processing requirements, and the susceptibility to errors from LLM hallucinations or misinterpretations of user intent.

Further innovative paradigms such as brainwave-driven HRC have emerged, highlighting the importance of adaptive and context-aware robotic systems. EEG-based frameworks, as proposed by Liu et al. [11], use cognitive load assessments to dynamically adjust robotic actions, enhancing worker safety and comfort. However, such advanced methods face practical limitations, including the intrusiveness of EEG sensors, signal processing complexity, and limited practical scalability on active construction sites.

In summary, current research underscores the importance of intuitive, adaptable, and robust interfaces for construction robotics, emphasizing natural communication modalities and

multimodal interactions. Despite significant progress, existing systems generally lack seamless multilingual support, robust real-time decision-making capabilities, and comprehensive evaluations in real-world construction settings. Addressing these challenges through integrated multimodal AI is essential to fully realize the potential of robotics within the construction industry, laying the groundwork for broader adoption and enhanced collaboration between human workers and robotic systems.

III. METHODOLOGY

To facilitate human-robot interaction in construction environments, PACA follows a multimodal processing workflow designed to interpret user commands, generate appropriate responses, and execute robotic actions when necessary. The system integrates Automatic Speech Recognition (ASR), a Large Language Model (LLM), and a Text-to-Speech (TTS) engine to facilitate bidirectional communication in multiple languages. Additionally, PACA leverages visual inputs when required, allowing the robot to analyze images as part of its reasoning process. Fig. 1 provides an overview of the PACA workflow, illustrating the key components and data flow within the system.

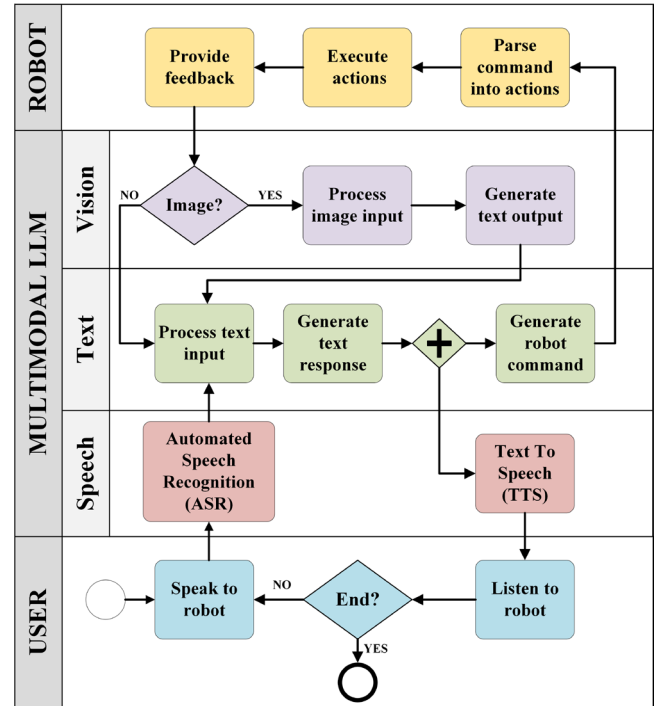


Figure 1. BPMN diagram with key elements of the process

A. User Interaction and Multimodal Input

Efficient interaction between construction workers and PACA is facilitated by a multimodal input system that captures and interprets both auditory and visual cues. To account for the complexity and noise inherent to construction environments, PACA does not continuously process audio data. Instead, users must explicitly signal the system to initiate

data processing. This signaling method ensures reliability and efficiency by avoiding unintended activation or misinterpretation of irrelevant environmental audio. Upon activation, PACA begins actively processing user input.

When activated, the system utilizes Automatic Speech Recognition (ASR) to convert spoken commands into text, enabling further processing. Notably, the system is designed to provide responses in the same language as spoken by the user. Therefore, if the command is issued in English, PACA responds in English, while commands given in other supported languages elicit responses in those same languages. This multilingual functionality ensures effective communication and usability among diverse teams.

In addition to auditory input, PACA integrates visual data as part of its multimodal capabilities. Users can also initiate interactions by providing visual inputs, which are processed alongside spoken commands when required. This combined multimodal input capability significantly enhances context-awareness and interaction precision.

After processing inputs through the multimodal Large Language Model (mmLLM), PACA communicates its understanding, clarifies instructions if needed, or confirms tasks through spoken feedback, utilizing a Text-to-Speech (TTS) engine. This ensures clear, bidirectional communication between the system and the user, critical for successful deployment in dynamic and multilingual construction sites.

B. Multimodal LLM and Decision-Making

The decision-making process within PACA is primarily managed by a mmLLM, which serves as the cognitive center of the system, responsible for understanding, reasoning, and providing coherent responses based on user interactions. At its core, the mmLLM utilizes a carefully constructed system prompt that provides foundational awareness and context regarding the tasks and commands it can execute. This system prompt can be tailored to varying degrees of specificity depending on the use case or scenario at hand, enabling flexibility and adaptability in different construction situations.

Within this prompt, two distinct layers of information are clearly delineated. The first part contains background and contextual descriptions of general tasks the robot is expected to perform, offering a comprehensive overview of the environment and typical interactions it might encounter. The second portion details Application Programming Interface (API)-specific information, explicitly defining the structured commands available for robot operation, such as moving to predefined locations, interacting physically with objects or tools, and executing sensor-driven tasks. This explicit separation ensures clarity, allowing the mmLLM to appropriately map user requests to precise and executable robot commands.

When a user initiates interaction, spoken inputs captured via Automatic Speech Recognition (ASR) and visual inputs acquired upon the user's cue are jointly provided to the mmLLM. Utilizing its multimodal capabilities, the mmLLM

processes and interprets these inputs, understanding user intent in the context of previous interactions to maintain continuity throughout the dialogue. This persistent contextual awareness is crucial, ensuring that users are not burdened with repeatedly providing the same information and allowing a seamless conversational flow, which significantly improves usability on active construction sites.

Based on this multimodal contextual analysis, the mmLLM generates coherent and contextually relevant outputs. These outputs serve two fundamental roles, first, guiding the conversation with the user by providing clear spoken responses or queries for further clarification, and second, translating user intents into structured, actionable instructions ready to be executed by the robot. The structured nature of these outputs ensures seamless integration with subsequent robot control processes, promoting effective and efficient human-robot collaboration in the inherently dynamic environment of construction sites.

C. Robot Command Parsing and Execution

The final stage in the PACA methodology involves translating high-level instructions generated by the mmLLM into actionable commands that can be executed by the robot. These high-level instructions are inherently abstract and structured according to a predefined Application Programming Interface (API), clearly outlining the tasks the robot is capable of performing, such as navigating to a specific location, interacting with objects in the environment, or performing targeted inspections.

This translation from high-level instructions to low-level robotic commands is accomplished through a dedicated parsing module that directly maps each API-defined instruction into corresponding robot-specific actions. This translation layer depends significantly on the robotic platform and its internal control protocols, so the implementation specifics will vary depending on the hardware and control middleware used.

Additionally, the translation mechanism is bidirectional. Not only does it convert mmLLM outputs into executable robot commands, but it also captures robot-generated cues from the physical environment to inform the mmLLM. For instance, when the robot reaches a predefined location, completes a task, or detects particular conditions in the environment, it can trigger corresponding cues or feedback signals back to the mmLLM. These environmental cues prompt the mmLLM to initiate follow-up interactions with the user, ensuring context-aware, dynamic, and adaptive communication between the user and PACA.

IV. IMPLEMENTATION

To test the developed methodology, PACA was deployed on a Boston Dynamics Spot robot equipped with a Spot Arm, enabling both mobility and physical interaction with the environment. The robot is fitted with an NVIDIA Jetson Orin NX 16GB, which serves as the onboard processing unit responsible for managing communication between Spot and

the cloud-based OpenAI model. The Jetson module allows PACA to send and receive requests to OpenAI servers via an internet connection, ensuring efficient processing of natural language and multimodal inputs while minimizing response latency.

For audio-based interaction, the system integrates a wireless audio receiver to capture the user's speech and a speaker for verbal responses generated through the Text-to-Speech (TTS) engine. Visual input is processed through an RGB camera mounted on the Spot Arm, allowing the robot to capture images when requested by the user. Additionally, an Insta360 X4 camera is attached to Spot for 360-degree data collection, enabling broader environmental awareness.

The overall hardware configuration of PACA is illustrated in Fig. 2. A demonstration of PACA in action can be viewed in [12].

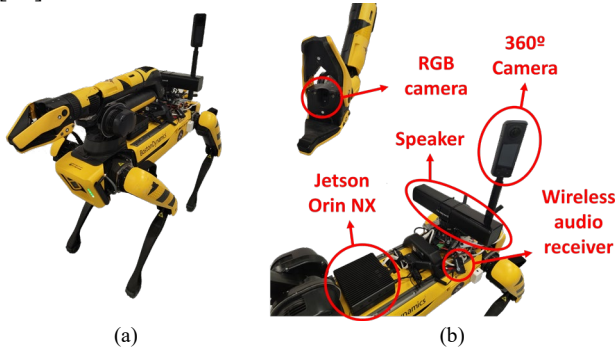


Figure 2. (a) General view of the robotic platform PACA, and (b) details of the payload used by PACA

A. User Interaction and Multimodal Input

To ensure reliable interaction in the challenging conditions of a construction site, PACA is designed to be cued into an active listening state through a dedicated remote interaction device. This device, wirelessly connected to the robot, allows the worker to initiate communication via a physical button press, ensuring a robust and unambiguous signal. Unlike voice-activated wake words or gesture-based cues, which can be masked by background noise or environmental clutter, the physical button provides a consistent and reliable method for engaging with the system. The interaction device includes multiple buttons corresponding to different functions, such as starting a conversation or commanding the robot to take a picture for image-based reasoning.

To facilitate clear speech recognition in noisy environments, the worker's hardhat is equipped with a built-in microphone, ensuring that audio commands are captured with minimal interference. Speech input is processed using OpenAI's whisper-1 model for Automatic Speech Recognition (ASR), accurately transcribing the user's commands into text before being passed to the mmLLM.

For verbal feedback, PACA employs OpenAI's tts-1 model for Text-to-Speech (TTS), generating clear and natural-sounding responses. To minimize latency in real-time interactions, the system streams the TTS output as it is generated, rather than waiting for the full response to be processed before playback. This approach significantly

reduces response lag, ensuring fluid and natural interaction between the user and the robot.

B. Multimodal LLM and Decision-Making

PACA utilizes the OpenAI's gpt-4-turbo model, which, at the time of this study, is the fastest multimodal model available from OpenAI. The mmLLM is responsible for processing user commands, analyzing visual inputs when necessary, and generating appropriate responses to facilitate human-robot collaboration.

For this case study, the system prompt provided to the mmLLM instructs PACA to assist workers by fetching the tools they request. If a worker needs a tool to complete a task, such as opening a box, they can command PACA to retrieve an appropriate tool. Upon receiving the request, PACA is instructed by the mmLLM to navigate to a predefined inventory location where tools are stored. Once there, PACA interacts with the inventory manager, relaying the worker's request.

At the inventory, the mmLLM guides PACA in prompting the inventory manager to present the requested tool. When the manager indicates a tool is ready, PACA captures an image and processes it through the mmLLM to evaluate whether the tool is appropriate for the requested task. If the tool matches the worker's requirements, PACA proceeds to pick it up and transport it back to the worker. If the tool is deemed unsuitable, PACA will prompt the inventory manager for an alternative.

All of these decision-making processes, from interpreting the worker's request to verifying the suitability of the tool and coordinating interactions between multiple users, are handled by the mmLLM. By maintaining contextual awareness throughout the process, PACA ensures a smooth and efficient workflow, reducing the cognitive burden on both the worker and the inventory manager.

C. Robot Command Parsing and Execution

PACA's ability to execute physical tasks is facilitated through the Boston Dynamics Python API, which enables direct control of the robot's mobility, manipulation, and sensing capabilities. The high-level commands generated by the mmLLM, such as "go to inventory," "go to worker's location," "take a picture," or "pick up tool," are mapped to corresponding low-level API functions that allow the robot to perform these actions.

The translation of commands occurs in a structured manner, where each high-level instruction issued by the mmLLM is parsed and converted into a sequence of executable actions using the Spot API. For instance, when PACA receives the command to "go to inventory," the system translates this into a navigation request that utilizes Spot's built-in localization and autonomous movement capabilities. Similarly, a "take a picture" command triggers the camera module to capture an image, while "pick up tool" translates into a series of precise manipulator movements coordinated through the API.

By leveraging the Boston Dynamics Python API, PACA

ensures seamless integration between high-level decision-making and low-level robotic execution, allowing for reliable task performance in dynamic construction environments.

CONCLUSIONS AND FUTURE WORK

This study presented PACA, a multimodal language-driven robotic assistant designed to facilitate human-robot collaboration in construction environments. By integrating a mmLLM with a mobile robotic platform, PACA enables natural and intuitive communication, allowing workers to issue commands using voice input while also incorporating visual processing for enhanced decision-making. Through a structured methodology, we demonstrated how PACA can assist workers in retrieving tools, autonomously navigating a construction site, and interacting with inventory personnel to fulfill user requests.

Despite its promising capabilities, the current implementation presents certain limitations. One of the primary constraints is its reliance on cloud-based services. This dependency necessitates a stable internet connection, which may not always be feasible in construction environments, and also introduces latency in system responses. Reducing this dependency by exploring local deployment options or leveraging more optimized edge-computing models could significantly enhance PACA's performance and reliability.


The current implementation focuses on a straightforward tool-fetching task, which, while useful, does not fully explore the potential of PACA in more complex scenarios. Future work should aim to develop a more advanced use case in which the robot not only assists with tool retrieval but also performs site inspections, identifies potential safety hazards, and provides real-time feedback to workers.

ACKNOWLEDGMENT

This research was partially supported by different Centers at NYUAD. In particular, the Center for Sand Hazards and Opportunities for Resilience, Energy, and Sustainability (SHORES) funded by Tamkeen under the NYUAD Research Institute Award CG013, the Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001, and the Center for Artificial Intelligence and Robotics (CAIR), funded by Tamkeen under the NYUAD Research Institute Award CG010. Part of this research benefited from the resources in the Core Technology Platform (CTP) at New York University Abu Dhabi (NYUAD), particularly the CTP's Kinesis Lab.

REFERENCES

- [1] D. Trujillo and E. Holt, "Barriers to Automation and Robotics in Construction," presented at the Associated Schools of Construction Proceedings of the 56th Annual International Conference, pp. 257–247. doi: 10.29007/1shp.
- [2] J. M. Davila Delgado *et al.*, "Robotics and automated systems in construction: Understanding industry-specific challenges for adoption," *J. Build. Eng.*, vol. 26, p. 100868, Nov. 2019, doi: 10.1016/j.job.2019.100868.
- [3] R. Dindorf and P. Woś, "Innovative solution of mobile robotic unit for bricklaying automation," *J. Civ. Eng. Transp.*, vol. 4, no. 4, Art. no. 4, Dec. 2022, doi: 10.24136/tren.2022.014.
- [4] G. Yang *et al.*, "Safe and Efficient Motion Planning for Material Transportation Robots considering Intention Prediction of Obstacles," in *2023 IEEE International Conference on Mechatronics and*

- Automation (ICMA)*, Aug. 2023, pp. 1179–1184. doi: 10.1109/ICMA57826.2023.10216040.
- [5] W. Tang and M. R. Jahanshahi, "AUTONOMOUS ROBOTIC INSPECTION BASED ON ACTIVE VISION AND DEEP REINFORCEMENT LEARNING," in *Proceedings of the 14th International Workshop on Structural Health Monitoring*, Destech Publications, Inc., Sep. 2023. doi: 10.12783/shm2023/36973.
- [6] S. Park, X. Wang, C. C. Menassa, V. R. Kamat, and J. Y. Chai, "Natural language instructions for intuitive human interaction with robotic assistants in field construction work," *Autom. Constr.*, vol. 161, p. 105345, May 2024, doi: 10.1016/j.autcon.2024.105345.
- [7] J. Yu, Q. Chen, S. Prieto, and B. García de Soto, *Human-Robot Partnership: An Overarching Consideration for Interaction and Collaboration*. 2024. doi: 10.22260/ISARC2024/0160.
- [8] S. Park, C. C. Menassa, and V. R. Kamat, "Integrating Large Language Models with Multimodal Virtual Reality Interfaces to Support Collaborative Human–Robot Construction Work," Oct. 2024, doi: 10.1061/JCCEE5.CPENG-6106.
- [9] A. Koubaa, A. Ammar, and W. Boulila, "Next-generation human-robot interaction with ChatGPT and robot operating system," *Softw. Pract. Exp.*, vol. 55, no. 2, pp. 355–382, 2025, doi: 10.1002/spe.3377.
- [10] Y. Lai, S. Yuan, Y. Nassar, M. Fan, T. Weber, and M. Rättsch, "NVP-HRI: Zero shot natural voice and posture-based human–robot interaction via large language model," *Expert Syst. Appl.*, vol. 268, p. 126360, Apr. 2025, doi: 10.1016/j.eswa.2024.126360.
- [11] Y. Liu, M. Habibnezhad, and H. Jebelli, "Brainwave-driven human-robot collaboration in construction," *Autom. Constr.*, vol. 124, p. 103556, Apr. 2021, doi: 10.1016/j.autcon.2021.103556.
- [12] SMART@NYUAD, *PACA: The Personal Autonomous Construction Assistant* , (Feb. 14, 2025). Accessed: Mar. 18, 2025. [Online Video]. Available: https://www.youtube.com/watch?v=warxBvJo_Zk