RESEARCH-ARTICLE

GenComUI: Exploring Generative Visual Aids as Medium to Support Task-Oriented Human-Robot Communication

YATE GE, Tongji University, Shanghai, China

MEIYING LI, Tongji University, Shanghai, China

XIPENG HUANG, Tongji University, Shanghai, China

YUANDA HU, Tongji University, Shanghai, China

QI WANG, Tongji University, Shanghai, China

XIAOHUA SUN, Southern University of Science and Technology, Shenzhen, Guangdong, China

View all

Published: 25 April 2025

Citation in BibTeX format

CHI 2025: CHI Conference on Human Factors in Computing Systems
April 26 - May 1, 2025
Yokohama, Japan

Conference Sponsors:
SIGCHI

https://doi.org/10.1145/3706598.3714238
ISBN: 9798400713941

GenComUI: Exploring Generative Visual Aids as Medium to Support Task-Oriented Human-Robot Communication

Yate Ge
College of Design and Innovation
Tongji University
Shanghai, China
geyate@tongji.edu.cn

Meiying Li
College of Design and Innovation
Tongji University
Shanghai, China
mzlzyca@tongji.edu.cn

Xipeng Huang
College of Design and Innovation
Tongji University
Shanghai, China
2333319@tongji.edu.cn

Yuanda Hu
College of Design and Innovation
Tongji University
Shanghai, China
ydhu@tongji.edu.cn

Qi Wang
College of Design and Innovation
Tongji University
Shanghai, China
qiwangdesign@tongji.edu.cn

Xiaohua Sun
School of Design
Southern University of Science and Technology
Shenzhen, Guangdong, China
sunxh@sustech.edu.cn

Weiwei Guo*
College of Design and Innovation
Tongji University
Shanghai, China
weiweiguo@tongji.edu.cn

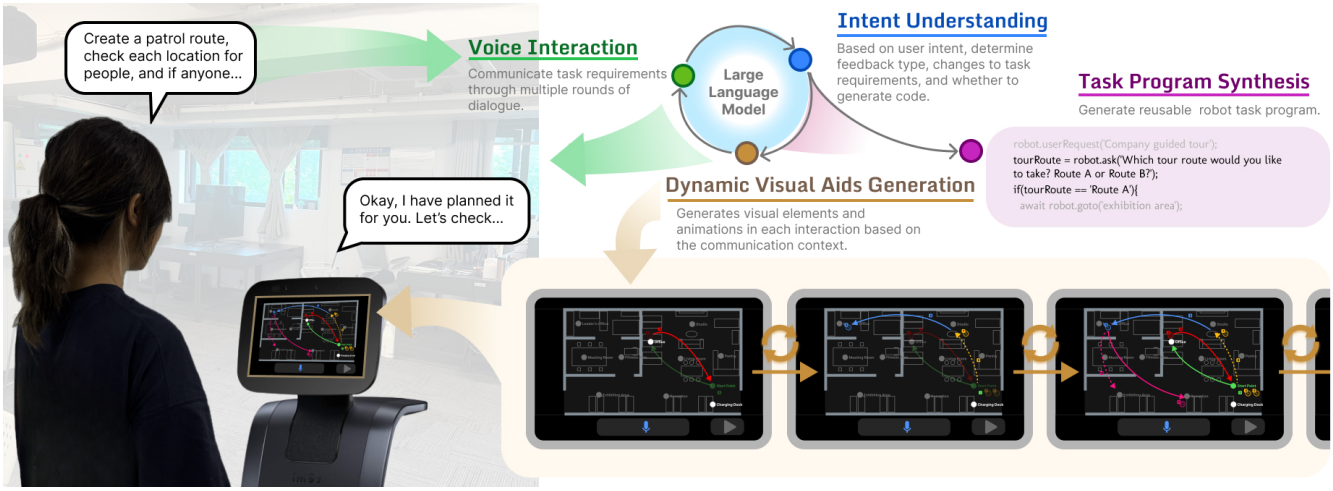

Figure 1: Overview of GenComUI: A system integrating voice interaction, user intent understanding, generative visual aids, and code generation. The system dynamically updates visual aids based on task communication context, combining multimodal feedback through graphical animations and voice output to confirm and refine task requirements with users.

*Corresponding Author.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714238

Abstract

This work investigates the integration of generative visual aids in human-robot task communication. We developed GenComUI, a system powered by large language models (LLMs) that dynamically generates contextual visual aids—such as map annotations, path indicators, and animations—to support verbal task communication and facilitate the generation of customized task programs for the robot. This system was informed by a formative study that examined how humans use external visual tools to assist verbal communication in spatial tasks. To evaluate its effectiveness, we conducted

a user experiment (n = 20) comparing GenComUI with a voice-only baseline. The results demonstrate that generative visual aids, through both qualitative and quantitative analysis, enhance verbal task communication by providing continuous visual feedback, thus promoting natural and effective human-robot communication. Additionally, the study offers a set of design implications, emphasizing how dynamically generated visual aids can serve as an effective communication medium in human-robot interaction. These findings underscore the potential of generative visual aids to inform the design of more intuitive and effective human-robot communication, particularly for complex communication scenarios in human-robot interaction and LLM-based end-user development.

CCS Concepts

• **Human-centered computing** → *Human computer interaction (HCI)*; *Interactive systems and tools*; *Natural language interfaces*; • **Computing methodologies** → Artificial intelligence.

Keywords

Human-Robot Interaction, Robot Programming, Service Robots, Conversational Interaction, Large Language Models, Generative UI

ACM Reference Format:

Yate Ge, Meiying Li, Xipeng Huang, Yuanda Hu, Qi Wang, Xiaohua Sun, and Weiwei Guo. 2025. GenComUI: Exploring Generative Visual Aids as Medium to Support Task-Oriented Human-Robot Communication. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3706598.3714238

1 Introduction

In service robotics, verbal communication allows non-expert users to naturally and intuitively express their needs when interacting with robots, enabling broad applications across home services, education, healthcare, and retail [8, 16, 32, 49, 55, 66]. Previous studies have demonstrated a growing interest in using verbal communication for robot programming, allowing users to specify commands, define goals, or create simple programs that align with their needs [3, 14, 25, 53, 63, 64]. Particularly with the advancement of large language models (LLMs) [70], traditional end-user robot programming can evolve into a collaborative and iterative process [37]. Users can now convey complex intentions and specify desired program outcomes through multi-turn, iterative communication while LLMs produce detailed specifications [21, 56].

Although LLMs have significantly improved human-robot natural language interactions [38] and code generation [33], robot programming via verbal communication remains constrained in task-oriented scenarios due to its unstructured and ambiguous nature, often resulting in the abstraction matching problem [41]. Moreover, speech-based communication encounters challenges such as speech recognition errors and being constrained to programming commands that are not easy to describe verbally, making it less effective than text-based communication for handling complex tasks [3, 62]. These challenges hinder task-oriented human-robot communication, especially when describing the robot's spatial movements and the logical relationships of task execution [60]. In End-User Development (EUD), multiple representations [2] have been utilized to support program representations across varying levels of abstraction, aligning with users' program authoring requirements. These representations complement one another, enabling users to harness their strengths to enhance comprehension, modification, and execution of programs.

To address the challenges in verbal programming, it is essential to consider the unique characteristics of human-robot voice interactions. Specifically, this involves designing interaction methods that reduce cognitive effort, ensure intuitive use, and promote natural human-robot interaction [6], while simultaneously enabling users to articulate and customize robotic tasks through verbal communication. Achieving this requires the integration of effective visual representation techniques and adaptive interaction strategies that facilitate seamless and comprehensible task specification.

In human-to-human communication, visual aids are widely used to support verbal communication by enhancing comprehension, retention, and engagement [15, 22, 35]. Visual aid tools, such as images, charts, graphs, diagrams, and animations, can complement or supplement words in both written and spoken texts, clarifying complex information [15, 35] or improving understanding in specific contexts [1, 42, 51]. In the field of Human-Computer Interaction, dynamic UI methods are employed to dynamically modify interfaces at runtime in response to interaction contexts or user requests [4, 30, 57, 59]. However, there is limited research on how dynamic user interfaces can be leveraged to implement visual aids in verbal interactions. Additionally, the understanding and creative capabilities of LLMs for interacting with multimodal interfaces can be utilized to generate multimodal interaction interfaces and enable dynamic multimodal interactions [21, 46, 58]. LLMs have the potential to dynamically generate interfaces during interactions based on the interaction context. This makes them particularly promising for aligning dynamic interfaces with user needs. Specifically, LLMs can integrate intent recognition and multimodal interaction generation, making them a compelling method for achieving more natural multimodal interactions that seamlessly combine visual UIs with verbal communication.

Motivated by these studies, this work explores the untapped potential of leveraging visual aids, drawing inspiration from how humans use them to support verbal communication, as mediating tools to enhance verbal task communication between humans and robots.

We developed GenComUI, a system powered by LLMs that facilitates voice-based robot task customization through dynamic visual aids generation (Figure 1). The system seamlessly integrates voice interaction, task program generation, and context-aware dynamic visual content generation to enhance human-robot communication. To guide the design of this system, we conducted an observational study of human behavior while communicating spatial guidance tasks using paper maps and pens. Insights from this formative study informed the development of a visual aids module designed to provide timely and progressive feedback, interpret user task intentions, and integrate visual aids with speech to improve clarity and intuitiveness. Drawing inspiration from commonly used visual elements in human communication, such as arrows, labels, and symbols, the system improves its ability to convey complex task information

in a natural and effective manner. These design considerations enabled GenComUI to leverage dynamic and context-aware visual aids to mediate and improve verbal task communication, addressing challenges inherent to traditional voice-based human-robot interaction.

To evaluate the effectiveness of generative visual aids in facilitating task communication for end-user robot programming and users' perception of GenComUI, we conducted a within-subjects user study (n = 20). We compared the baseline system with our full system using quantitative data to investigate the impact of generative visual aids on task-oriented communication. Additionally, we used qualitative methods to collect user perceptions of GenComUI and their views on how our generative UI design supports task communication. Our findings indicate that GenComUI, through its dynamically generated graphical interface supporting voice interaction, significantly improved user efficiency and accuracy in complex task communication, and enhanced user understanding and trust in the robot's capabilities, but showed less pronounced advantages in simple tasks. Based on these findings, we discuss design implications for dynamically generated visual aids as a communication medium to improve task communication in human-robot interactions.

In summary, this paper makes the following contributions:

- A formative study that reveals how humans use external visual media to facilitate task communication.
- GenComUI, a proof-of-concept system that dynamically generates visual aids on the robot's screen based on communication context to support feedback and confirmation in task communication.
- A within-subjects user study with 20 participants evaluating the impact of GenComUI on task-oriented communication and user experience.
- A set of design implications informed by findings from the user study, offering insights into how dynamically generated visual aids can serve as an effective communication medium in human-robot interaction.

2 Related Work

2.1 Task-oriented Human-Robot Communication

Task-oriented communication between humans and robots is becoming increasingly important, especially in the context of end-user robot programming. Natural language programming allows non-expert users to specify robot tasks through verbal instructions [14, 31, 39, 48, 63]. This approach enables users to communicate complex task requirements and specify reusable programs through multi-turn dialogues [25, 61, 63], making it a fundamental aspect of end-user robot programming [3, 40].

The integration of artificial intelligence with natural language programming is considered a key method for future household robots and intelligent agents to provide personalized services [19]. Through natural dialogue, this approach enables untrained users without programming knowledge to define reusable robot programs that align with their practical needs. Recent advancements in large language models (LLMs) [65] have further accelerated developments in this field. Several studies have explored how LLM capabilities can support end-users in defining robot tasks using natural language [18, 20]. However, specifying robot tasks based on natural language descriptions of desired program outcomes still poses challenges due to the abstraction gap between natural language and program code [41]. Recent advancements in LLM-based end-user programming have investigated structured and visualized "intermediate-level" representations to address the abstraction gap [21, 41].

LLM-based end-user development (EUD) transforms programming into a collaborative and iterative communication process [19, 37]. Effective communication between users and robots often requires a continuous cycle of "intention expression → result feedback → intention adjustment" to complete task specification [23]. This process typically integrates multiple modalities to provide user feedback and represent program or task context [18, 29, 69], or supports users in expressing intentions through multimodal interactions [53]. Furthermore, human-like multimodal interactions have been explored to facilitate task communication between users and robots [27, 28].

In light of these advancements and the challenges inherent in LLM-based EUD systems, there is an opportunity to enhance verbal programming. This motivates us to draw inspiration from human-to-human verbal communication and explore natural, intuitive interactions that leverage large language models to improve the usability of verbal programming.

2.2 Visual Aids in Human-Robot Communication

Screens on robots serve an important function in enhancing communication by displaying facial expressions and complementing voice interactions [13, 24, 68]. Beyond expressive functions, screens also facilitate the communication of complex messages, yet their integration with verbal communication remains an area requiring further exploration.

Currently, the way touch screens are used in service robots has led to their perception as mere "screen bearers", diminishing the sense of rich interaction with an autonomous agent. In most cases, they are employed as a means to circumvent the current limitations of full speech interaction while also providing an effective way to enable complex interactions [8]. However, balancing this with other interaction modalities is important to maintain a rich interaction experience. The consistency between voice interactions and graphical user interfaces is crucial for improving system usability and user satisfaction [50], making it easier for users to understand and operate the system. Research indicates that providing graphical feedback during dialogues significantly enhances user comprehension of the robot's responses and intentions [52].

Visual media can support task communication through various forms, including robot-mounted screens, external display devices, and extended reality (XR) interfaces [12, 12, 44]. Additionally, external objects and gestures can serve as references to assist in task communication, enhancing naturalness and comprehension [27, 28]. Some researchers have explored projection techniques that allow robots to visualize task information on physical objects in the

environment [5]. Compared to single-modality interactions, interactions with robots that exhibit natural multimodal expression provide a more engaging and intuitive communication experience [11]. Drawing inspiration from human-to-human interaction patterns has proven effective in refining robot communication strategies [28]. Since human-AI agent communication is inherently dialogic, requiring iterative exchanges to refine intent expression, the selection of interaction modalities should be contextually adapted to different scenarios [23].

2.3 LLM-driven Dynamic UI Generation

Recent work has demonstrated that large language models can serve as intermediaries for multimodal interfaces, enabling both understanding and generation beyond purely natural language content. For example, in graphical user interfaces, LLMs can interpret the semantics and structure of UIs [17, 67] and generate UI designs [36, 43], demonstrating their potential in processing and generating non-linguistic visual representations.

Furthermore, the world knowledge and in-context learning capabilities of LLMs [70] enable the dynamic generation of multimodal interactions tailored to the context on the fly. For instance, GenEM [46] utilizes LLMs to flexibly generate and adapt robot expressive behaviors based on natural language instructions and user preferences. Similarly, SiSCo [58] showcases the ability of LLMs to synthesize both natural language and visual signals adaptively for efficient collaboration. In the context of end-user programming, Cocobo [21] illustrates how LLMs can dynamically translate between natural language and visual programming representations.

Unlike traditional rule-based or template-based approaches, these LLM-based methods show the potential for adaptive contextual understanding and immediate generation across multiple modalities. This motivates our exploration of leveraging LLMs to dynamically generate visual aids for task-oriented communication between humans and robots.

3 Formative study

To better understand how people utilize visual tools to enhance verbal task-oriented communication and to inform the design of visual aid methods for human-robot communication in spatial robot task customization, we conducted a formative study.

In this study, participants were paired for task-based communication, with one member of each pair provided with paper and a pen to facilitate visual communication. The objective was to observe and analyze how individuals integrate natural language and visual tools in their interactions, ultimately informing the design of visual aids modules for human-robot task communication.

By collecting and analyzing data, we hope to explore:

- In what situations do people use external visual tools to aid communication?
- How visual and natural language dialogue are combined to facilitate communication?
- The types of visual elements that users use to assist communication.

3.1 Participants

We recruited 8 participants (5 female). Each participant was paired with one of the two researchers, forming a total of 8 groups. The recruits consisted of students and teachers from the university. The experiment for each group lasted between 10 and 15 minutes.

3.2 Procedure/Task

Each participant was assigned to a group where they took on the role of maintenance staff (Role B), while the researcher acted as the manager (Role A). Participants were informed of the specific experimental procedures and provided with printed task maps and colored pens.

The scenario was set in a designated college space within the school, where the manager (Role A) verbally assigns tasks to the maintenance staff (Role B). The maintenance staff uses a paper map and a pen to communicate with the manager, ensuring they accurately understand the manager's intentions and provide a specific execution plan. Two tasks were designed: an exhibition reception task and a patrol task. The detailed scripts can be found in the appendix A.

Before the experiment began, Role A reviewed a script and noted down the tasks they needed to assign to Role B. During the actual communication, Role A was not permitted to refer to the script, and their objective was to ensure the tasks were accurately conveyed to Role B. Role B was responsible for confirming their understanding through questioning and paraphrasing, as well as formulating a concrete execution plan. To facilitate task confirmation, Role B was provided with five different colored pens to annotate pre-printed maps, with no restrictions on paper usage. Each experimental session was recorded on video, capturing participants as they completed the two tasks. The experiment concluded once Role A confirmed that Role B had accurately understood the assigned tasks. Examples of participants' annotations on the paper maps can be seen in Figure 2. After the experiment, participants were asked about the specific situations in which they used pen and paper, how these tools assisted in articulating their intentions, and what specific benefits they provided.

3.3 Analysis

We analyzed the experiment recordings and synthesized the interview findings, identifying the specific needs and contexts in which users employed visual aids. Based on the experimental data, all participants utilized visual aids by marking or drawing on the map (8/8). Additionally, the majority of participants (7/8) reported that visual aids were beneficial in articulating their intentions.

3.3.1 ***T1. When do people use visual modalities to communicate?*** First, during the interaction between Role A and Role B, Role B would draw while listening, responding with phrases like *"OK"* or *"understood"* to indicate comprehension of the request. Several participants noted that *"taking notes while listening helps with memory"* (P1, P2, P3, P4, P7). Second, during the final plan confirmation, Role B would always use the previously drawn content to describe the task requirements while summarizing them to Role A. Some participants repeatedly marked certain points to emphasize key aspects during their summary (P6).

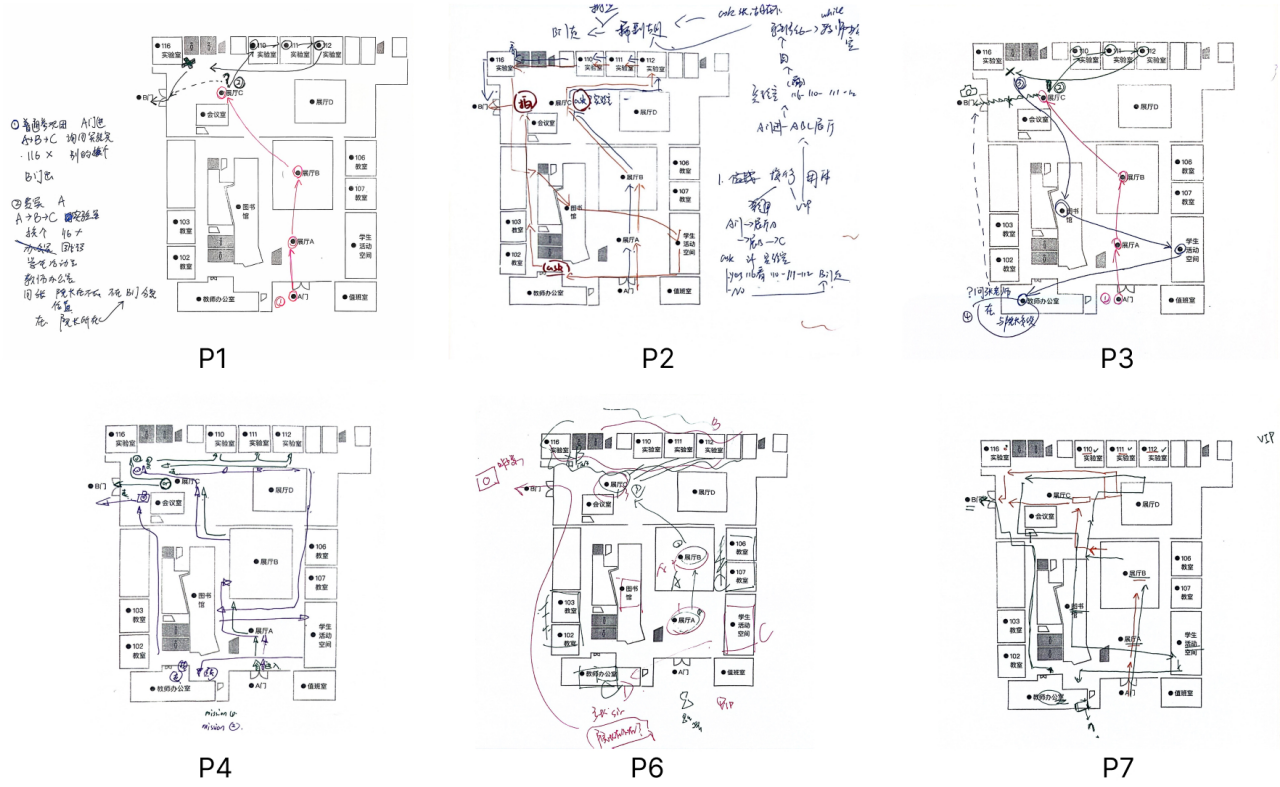

Figure 2: Sample maps drawn by participants during the formative study, showing how they used visual elements to represent and communicate spatial tasks through annotations, paths, and markers.

Participants believed that drawing and marking helped clarify their intentions. They observed that visual aids in communication *"helping to communicate more clearly"* and *"improving the detail and accuracy of the communication"* (P4, P6). Additionally, participants mentioned that *"drawing allows the other person to see their ideas, which facilitates clearer information conveyance"* (P2). Observations from the researchers showed that participants who were more proficient in visual marking demonstrated higher accuracy and fluency in summarizing tasks.

3.3.2 ***T2. How are visual and natural language dialogue combined to facilitate communication?*** During the task communication, participants frequently referred to the map. When confirming tasks with Role A, Role B would point to their annotations, explaining the task in alignment with their drawings and markings. In cases of logical branches, such as *"asking whether the principal is present in the college,"* participants would explain both the *"present"* and *"absent"* branches. If the branch led to a different location, such as *"if the principal is not in the college, exit through the main gate to end the tour,"* participants would simultaneously point to the main gate on the map while explaining.

When reconfirming the task requirements, Role B would add supplementary drawings after consulting with Role A. Participants mentioned that *"seeing the content on paper facilitates easier organization of thoughts"* (P4, P6, P7), emphasizing that synchronization between verbal communication and visual aids was key to facilitating understanding.

3.3.3 ***T3. What types of visual elements do users use to assist communication?*** The content annotated by participants often aligned closely with the specific aspects of the task. For aspects of the task involving sequential steps, such as *"first go to the classroom, then to the tool room, and finally to the library to complete the patrol task,"* participants would record the sequence using numbers, letters, or other markers. Particularly when the task involved spatial transitions, participants would use arrows to indicate the starting and ending points of these transitions. Participants mentioned that *"recording instructions and describing routes require pen and paper"* (P1, P3), and that *"making handwritten annotations helps in understanding the order of locations"* (P5, P7).

Participants often sketched logical branches (P1, P2, P3, P4, P6, P7), stating that *"some obvious logic is easy to represent on paper"* (P4). For specific events, such as *"checking whether the appliances are turned off in the activity space,"* participants would use distinct icons or brief text to describe them. Some participants would note down information at particular spots on the paper. For different scenarios, such as *"regular groups and VIP groups,"* participants used different colored pens to distinguish between them.

Type	Visual Elements	Representation
Sequence	Line with arrow	Show the sequence of movement or path in the map Potential task sequence
	Circle annotation	Mark location Highlight information
	Symbol	Numeric symbol
Logic	Line with arrow	Point to a logical branch
	Text label	Presenting content directed by a logical branch
Annotation	Line with arrow	Point the annotation information to the annotated object
	Text label	Annotation on the space Label information that is not directly related to the location
	Symbol	Event annotation
Global	Color	Point the annotation information to the annotated object. Highlight information
	Note	Add note to record task requirements

Table 1: Visual Elements summarized from the formative study

3.4 Summary

Based on our formative study observations, we categorized the visual elements used in task communication into four main types: sequence, logic, annotation, and global elements. For each type, we identified specific visual elements (such as lines with arrows, circle annotations, text labels, symbols, colors, and notes) and their representational purposes. For example, sequence-related elements were used to show movement paths and task orders, while logic elements helped represent conditional branches. Annotation elements served to mark locations and add contextual information, and global elements like color coding helped highlight and organize information across different aspects of the task. The complete categorization of these visual elements and their specific uses is shown in Table 1.

3.5 Design Considerations

Drawing from how humans naturally employ visual aids in communication, particularly their patterns of using visual elements to clarify complex tasks, we formalize the following design considerations for our system:

[DC1] Provide continuous and progressive visual feedback throughout the communication process to support step-by-step task understanding and verification, ultimately facilitating the successful completion of task communication.
[DC2] Facilitate memory and task comprehension by interpreting user task intentions to plan and organize feedback, including the use of visual aids and speech output, ensuring users stay aware of complex task sequences and spatial relationships.
[DC3] Enable effective robot-to-human communication by organically integrating visual aids and speech, allowing the system to convey information more comprehensively and intuitively.
[DC4] Leverage visual elements commonly used in human communication (e.g., arrows, labels, symbols) and their associated usage patterns to represent tasks, ensuring clarity and natural alignment with user expectations.

4 GenComUI System

Based on the design considerations distilled from the formative study in section 3.5, we propose GenComUI (Figure 1), an LLM-based robot EUD system that incorporates generative visual aids. The system is implemented on the Temi V2 robot[1], a mobile robotic platform equipped with a touchscreen. The system consists of four core modules designed to enable verbal robot programming through iterative, multi-turn communication:

- **Voice Interaction Module**: Handles speech-to-text and text-to-speech conversion, enabling bidirectional voice communication between users and the robot.
- **User Intention Understanding Module**: Analyzes user input and dialogue context to understand communication progress, tracks task specifications, and plans appropriate responses, including visual aids generation and code updates.
- **Generative Visual Aids Module**: Generates visual interface elements and animations on spatial maps according to visual aid requirements from the Intention Understanding Module.
- **Task Program Synthesis and Deployment Module**: Generates and deploys executable robot code based on user specifications, with built-in testing capabilities for iterative refinement.

The system leverages LLMs with structured output[2] capabilities to handle complex interaction logic by generating multiple coordinated outputs in response to context, including planning and dynamically generating visual aids. For further details on the underlying mechanisms of these modules, please refer to Section 4.2- 4.4.

[1]https://www.robotemi.com/
[2]https://platform.openai.com/docs/guides/structured-outputs

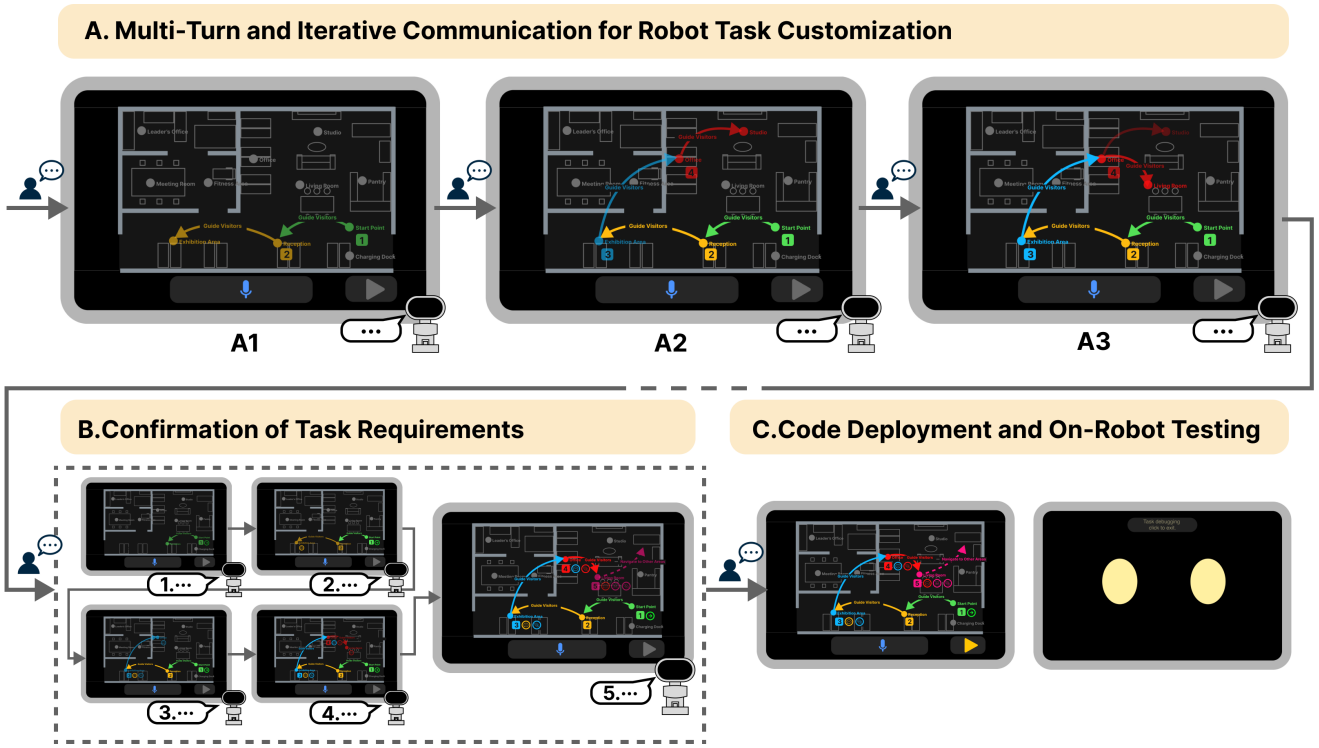

Figure 3: The communication process in GenComUI: (A) Multi-turn and iterative communication for robot task customization, showing progressive visual feedback as users specify and modify task requirements; (B) Task requirement confirmation phase with step-by-step visual and verbal verification; (C) Code deployment and on-robot testing phase enabling real-world validation of the specified task.

4.1 Example Usage Scenario

Here we present a sample usage scenario in an office setting to demonstrate how GenComUI and its generative visual aids support multi-turn verbal task communication between humans and robots. The main visual interface of GenComUI is shown in Figure 3.

Lily, a secretary at a tech company, is responsible for coordinating meetings and managing schedules. While she wants to leverage a robot assistant for her daily tasks, she faces two challenges: the robot's predefined functions are too rigid for her dynamic needs, and she lacks programming expertise to customize robot behaviors. GenComUI addresses these challenges by enabling natural dialogue-based task specification.

In this scenario, Lily wants to create a "visitor reception" task where the robot notifies and guides participants to meetings. After launching GenComUI, she initiates the dialogue through the **Voice Interaction Module** by tapping the voice button. She says: *"Hello Temi, I would like to develop a visitor reception service."*

The **User Intention Understanding Module** processes her input and generates an appropriate response: "Okay, the robot will lead the visitor to the reception area first, then go to the work display area. Do you have any other requirements?" Simultaneously, the **Generative Visual Aids Module** visualizes the task flow using connected lines and fade-in animations to highlight the newly added requirements (Figure 3-A1) (**DC2**).

Lily further instructs the robot to then lead visitors to the staff office area and creative studio. The robot acknowledges this and updates the visual aids accordingly (Figure 3-A2).

When Lily modifies the task by saying, "I want to modify it. After the staff area, lead them to the drawing room," the robot understands the user's intent to modify the task steps. The system updates the visualization using fade-out animations for removed elements and fade-in for new ones (Figure 3-A3).

After several rounds of communication, aided by step-by-step visual feedback (**DC1**), Lily indicates she has finished specifying the task. The robot enters confirmation mode (Figure 3-B), generating synchronized visual and voice presentations of the complete task steps (**DC3**): "Step one, activate the service with the keyword 'visitor reception', then the robot will lead the visitor to the reception area," "Step two, robots lead visitors to the exhibition area," and so on.

Upon Lily's confirmation, the **Task Program Synthesis and Deployment Module** generates and deploys the program code. The screen displays the complete task flow with icons, text, and connecting arrows (**DC4**). After deployment, Lily tests the task through the activated "Test" button (Figure 3-C). During testing, she identifies the need for the robot to return to the reception area after completing the task. She easily makes this modification by exiting test mode and communicating the additional requirement.

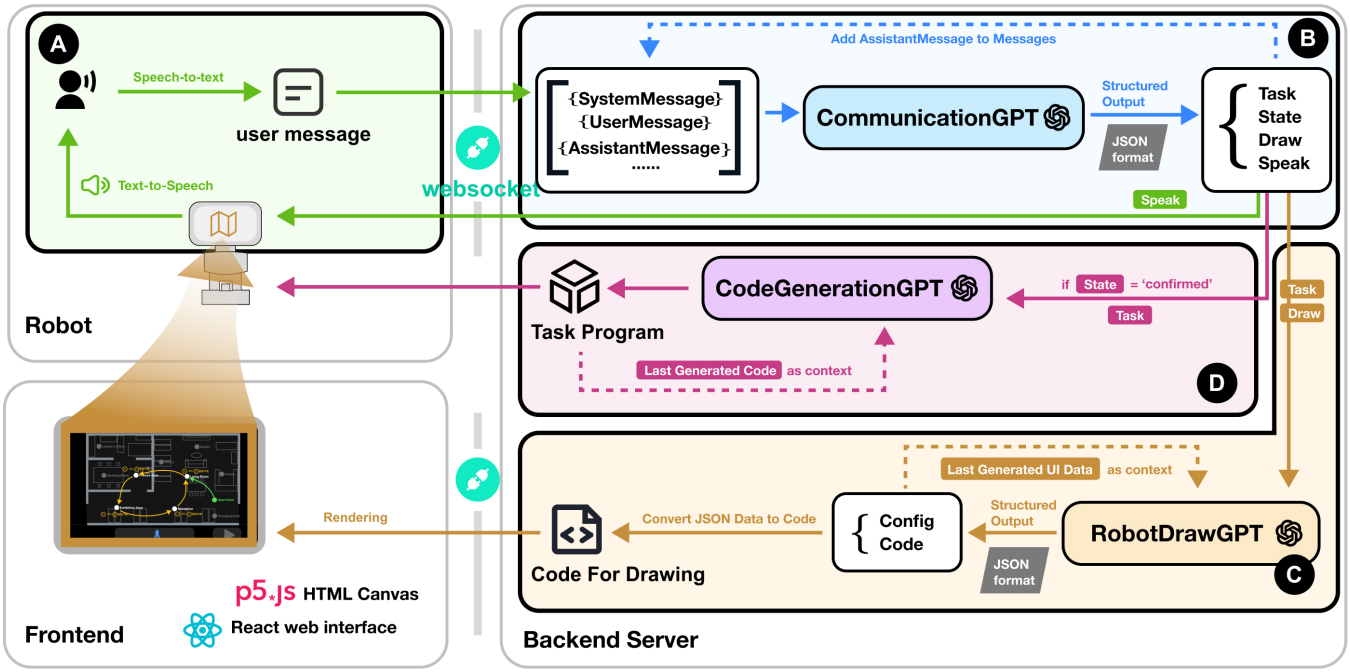

Figure 4: GenComUI system architecture: (A) Voice Interaction Module enabling bidirectional voice communication between users and robot; (B) User Intention Understanding Module analyzing user input and dialogue context to generate structured outputs; (C) Generative Visual Aids Module creating dynamic visual interface elements and animations on a spatial map; (D) Task Program Synthesis and Deployment Module for generating and executing robot code.

Through this process, Lily successfully creates a customized robot task that matches her specific needs, demonstrating how GenComUI enables non-programmers to naturally specify and refine robot behaviors through multi-modal interaction in an on-the-fly style[60].

4.2 Human-to-Robot: Intention Understanding from User Speech

The **Intention Understanding Module** aims to obtain clear task specifications through multi-turn interactions by interpreting user intent, understanding requirements, and tracking interaction progress.

To achieve this, the module leverages the **CommunicationGPT** model with *chain-of-thought reasoning* and *few-shot learning techniques* (detailed in Appendix B.1). The model takes system prompts and dialogue history as input and produces structured outputs containing:

- **Task**: Current robot task step descriptions based on the ongoing communication
- **State**: Communication progress tracking to guide the interaction flow
- **Speak**: Robot's verbal response content
- **Draw**: Visual aids type for rendering

These structured outputs are then processed by the backend to generate appropriate robot behaviors, including verbal responses, task-related code updates, and visual aid presentations. See Figure 4.B for the detailed workflow, and Figure 3 for the application of visual aids in the communication process.

4.3 Robot-to-Human: Visual Aids Generation and Presentation

The **Generative Visual Aids Module** dynamically generates graphical interface content during interactions based on the context of task communication using **RobotDrawGPT** (see Appendix B.2). Based on the structured outputs from the **User Intention Understanding Module**, the system supports three presentation modes for visual-aided robot-to-human communication:

The *Feedback* mode highlights modified parts of the task flow, using fade-in animations to emphasize newly added steps and fade-out animations for deleted steps. In *Confirm* mode, the system synchronizes task steps with corresponding graphical animations and voice descriptions, highlighting each visual element as its associated step is narrated (Figure 3-B1). The *None* mode simply displays the current task flow without any animations.

The visual content is composed of two main types of elements. Location markers include icons representing robot actions, task numbers, and configurable location colors. These are connected by arrows that feature customizable colors and styles (solid or dashed lines), along with task descriptions. The arrows can establish various types of connections: between two specific locations, from one location to any location, or from any location to a specific destination.

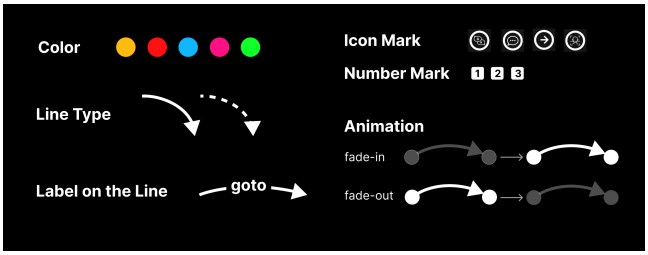

Figure 5: Visual design elements and animation specifications in GenComUI: Color palette for different task components, line types (solid and dashed) for path representation, text labels for spatial behavior description, icon and number markers for location identification, task steps and robot behaviors, and animation effects (fade-in/fade-out) for visual feedback.

To maintain visual continuity, each rendering process updates only the elements related to the user's refined intent while keeping other parts unchanged. As shown in Figure 4, during the requirement communication process, the system generates different drawing behaviors according to the user's intent. The detailed styles and configurable elements are shown in Figure 5, and for implementation details, please refer to **RobotDrawGPT**'s prompt in Appendix B.2.

4.4 Task Program Synthesis and Deployment

Table 2: Robot Commands

Robot Command	Description
userRequest: *WakeWord* →	Activate via *WakeWord*
goto: *Place* →	Move to *Place*
say: *Speech* →	Say the contents of *Speech*
ask: *Speech* →	Ask the contents of *Speech*
humanDetection →	Detect a person in front of the robot

GenComUI creates and modifies robot task programs using an LLM based on the task flow input, system prompt (see Appendix B.3), and, if previously generated, existing code and task flow to inform the generation process.

After final user confirmation, the **Task Program Synthesis and Deployment Module** leverages **CodeGenerationGPT** (see Appendix B.3) to generate executable robot code based on the confirmed task specifications. The module synthesizes JavaScript programs that orchestrate the robot's basic commands (Table 2) according to the specified task flow.

Once code generation is complete, the system deploys the program to the robot's runtime environment and activates the test functionality. Users can then enter test mode to validate the program's behavior using the specified wake word. The testing interface provides an exit option that returns users to the customization interface for iterative refinement if needed.

4.5 Implementation Detail

GenComUI uses the open-source program *temi-woz-android*[3] to establish communication between the robot and the backend server via WebSocket[4]. This enables the backend server to control Temi's behavior by sending WebSocket commands and receiving callback messages. The backend, developed using Node.js, handles interaction events, manages robot behavior, and makes calls to the OpenAI API. The robot task programs synthesized by the system are written in JavaScript and executed on the backend server. The system uses *GPT-4o-2024-08-06* through the OpenAI API[5], which supports structured output[6] and provides response times and reasoning capabilities that meet the requirements of this project. We set the temperature parameter to 0 in all API calls. The frontend, built with the React framework and `p5.js`[7], dynamically inserts and renders the visual aids code on Temi's Screen.

5 User Study

The role of visual aids in facilitating task-oriented human-robot communication lacks sufficient empirical exploration in HCI research. Using GenComUI as a research tool, we aimed to probe two key questions: **(RQ1)** How do users perceive and experience visual aids during task-oriented communication with robots? and **(RQ2)** What design implications can be derived for integrating generative visual aids in human-robot communication systems? To investigate these research questions, we designed a comparative study that systematically examined how visual aids influence human-robot communication. Through a mixed-method approach comparing GenComUI with a baseline system without visual feedback, we sought to understand the specific role and impact of visual aids in task-oriented robot programming scenarios.

5.1 Baseline System for Comparsion

To gain deep insights [26] into how visual aids influence task-oriented human-robot communication, we designed a baseline system that retained all functionalities of GenComUI but omitted the generative visual aids module. While the baseline system displayed only static robot expressions on screen, we provided participants with a paper map identical to the one shown in GenComUI's interface for reference. This controlled design ensured that both systems had comparable response times and behaviors, allowing us to specifically examine the role of visual aids in human-robot task-oriented communication. By controlling this single variable, we could focus on understanding how visual aids shape the communication process and user experience during task-oriented interactions.

5.2 Participants

This study recruited 20 participants (12 females, 8 males, aged 18–60, M = 26.5, SD = 8.38) via an online questionnaire, ensuring a diverse range of backgrounds in robot interaction, LLM familiarity, and programming expertise. Half of the participants (10/20) had no prior experience with robots, while the others had experience with

[3] https://github.com/tongji-cdi/temi-woz-android
[4] https://developer.mozilla.org/en-US/docs/Web/API/WebSockets_API
[5] https://api.openai.com/v1/chat/completions
[6] https://openai.com/index/introducing-structured-outputs-in-the-api/
[7] https://p5js.org/

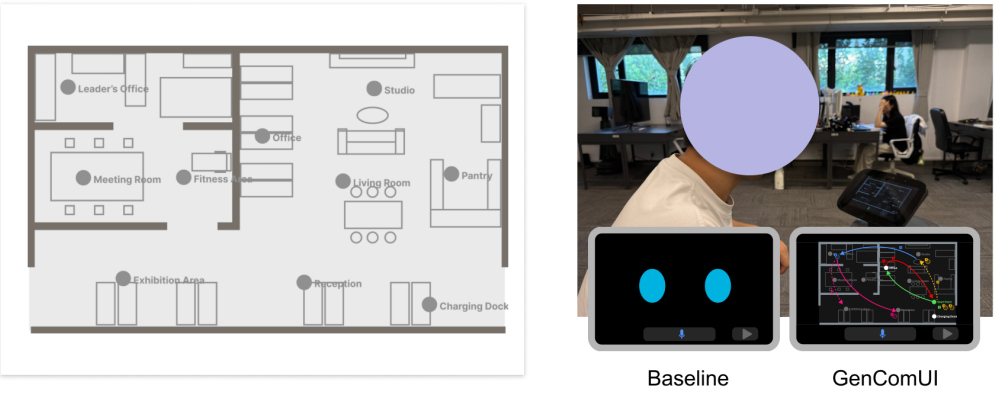

Figure 6: Experimental setup for system comparison: (Left) Paper floor plan provided to participants during baseline testing; (Right) Study environment showing interface comparison between baseline system (displaying a facial expression) and GenComUI (showing generative visual aids).

home service (8/20) or industrial robots (2/20). Most participants frequently used LLMs (13/20), while some were occasional users (3/20) and others had never used LLMs (4/20). In terms of programming expertise, participants ranged from professionals (3/20) to those with basic skills (7/20) and no experience (10/20).

5.3 Setup

The experiment was conducted in a simulated office environment as shown in Figure 6. The setup consisted of a Temi robot, a computer running the backend system, video and audio recording equipment for data collection, and printed task guidelines for participants. For the baseline condition, participants were provided with a physical map identical to the one displayed in GenComUI's interface. Two researchers were present throughout each session: one facilitating the experiment and another conducting behavioral observations.

5.4 Task

We designed four spatial programming tasks that required participants to create robot programs involving navigation and actions in an office environment. The tasks were divided into two complexity levels based on the minimum number of required robot commands, ensuring that tasks of the same complexity were interchangeable.

Low-complexity Tasks (minimum 4 robot commands each):

- **Task L1**: Office Patrolling and Employee Notification
  Program the robot to patrol specific office areas and notify designated employees
- **Task L2**: Employee Guidance and Area Introduction
  Program the robot to guide an employee through office areas while providing area descriptions

High-complexity Tasks (minimum 8 robot commands each):

- **Task H1**: Visitor Guidance at Reception
  Program the robot to receive visitors and guide them through multiple office locations
- **Task H2**: Employee Gathering and Preparation Work
  Program the robot to locate multiple employees and coordinate a meeting preparation sequence

Each participant completed two tasks with each system: one low-complexity task and one high-complexity task. To reduce task-specific biases and improve the generalizability of our findings, the two tasks of the same complexity level were designed to be interchangeable. Task assignment was randomized through a lottery system by the researcher, with the remaining two tasks assigned to the other system condition. This approach ensured a balanced design while examining system performance across diverse cases. See Appendix C.1 for detailed task descriptions.

5.5 Procedure

The experiment followed a within-subjects design where participants experienced both systems. The researcher used a computerized randomization program to determine each participant's system order, which ultimately achieved a balanced distribution (10:10) across conditions. The procedure for each system condition consisted of three phases:

Training Phase (15 minutes): Participants watched a system-specific tutorial video and received hands-on guidance from researchers on how to operate the system. For the baseline system, participants were provided with a paper map; for GenComUI, the map was displayed on the robot's screen.

Task Execution Phase (30 minutes): Participants completed two tasks with each system: one low-complexity task and one high-complexity task. For each task, participants communicated their programming intentions through voice commands until the robot indicated understanding by requesting a task trigger word. Upon execution, participants could interrupt and modify the program if needed, with the task concluding only when participants perceived that the robot had successfully executed the intended program.

Assessment Phase: After completing tasks with each system, participants filled out questionnaires.

Following the completion of both conditions, participants engaged in a 20-minute semi-structured interview.

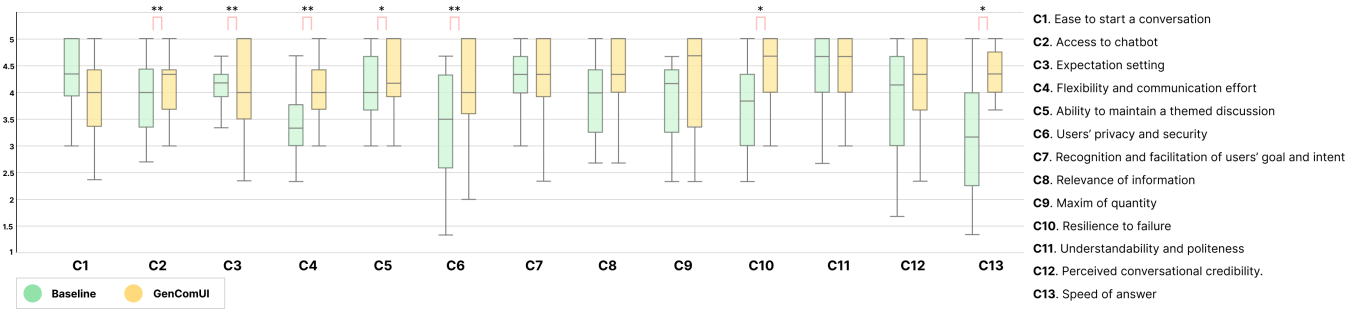

Figure 7: The result of Chatbot Usability Scale[9]. This figure shows the usability evaluation of the two systems using a 5-point Bot Usability Scale, with select questions renumbered and displayed. (*: p < .050, **: p < .010)

5.6 Measurements

5.6.1 Questionnaire. To evaluate how generative visual aids affect human-robot task-oriented communication, we employed three complementary measurement scales. The Chatbot Usability Scale [9] was used to assess the dialogue-based interaction quality, focusing on how visual aids influence task communication effectiveness and user experience. We incorporated the Godspeed questionnaire [7] to evaluate whether visual aids enhanced users' perception of the robot as an intelligent communication partner during task programming. Additionally, the System Usability Scale (SUS) [34] provided a standardized measure of overall system usability. This combination of metrics helped us comprehensively understand the impact of visual aids on task-oriented communication and derive design implications. For data analysis, we conducted statistical comparisons between the GenComUI and baseline conditions. We applied paired t-tests for normally distributed variables and Wilcoxon signed-rank tests for non-normally distributed variables to analyze the questionnaire responses.

5.6.2 Interview. We conducted semi-structured interviews to gather comprehensive user feedback. The interviews began with questions about users' operational experiences with both systems, asking them to compare and identify their preferred system with the rationale, which helped us understand users' overall perception of the systems. We then explored how users leveraged the visual aids to assist in completing natural language programming tasks. Following this, we focused on the GenComUI interface design, investigating users' perceptions and expectations of our interface features. Each interview lasted approximately 20 minutes and was audio-recorded for subsequent analysis. The complete set of interview questions is provided in Appendix C.2. We followed Boyatzis's [10] thematic analysis guidelines. Two researchers independently coded the transcripts, developed initial codebooks, and identified preliminary themes. Through iterative discussions, the researchers refined the codes and themes until reaching consensus after several rounds of review.

5.6.3 Task Performance Metrics. To evaluate task performance, we collected quantitative data through backend logs during each session. These logs captured key metrics including task completion time, success rates, and the frequency of repeated communications for modification in each user task. We performed paired t-tests and Wilcoxon signed-rank tests (for non-normally distributed data) to compare these metrics between conditions.

6 Findings

6.1 Quantitative Results

6.1.1 Task Completion and Observation Report. All participants (20/20) successfully communicated tasks to the robot, achieving correct task execution across both systems. Statistical analysis revealed no significant difference in task completion time between GenComUI and the baseline system ($p > 0.05$; Baseline: Mdn = 0:05:26, Std = 0:03:16; GenComUI: Mdn = 0:06:47, Std = 0:02:52). The number of voice dialogue turns (Baseline: Mdn = 11, Std = 4; GenComUI: Mdn = 11, Std = 4) also showed no significant difference between systems.

6.1.2 Chatbot Usability Scale Analysis. The questionnaire consists of 42 questions divided into 14 sections, such as "ease of starting conversation" and "access to chatbot". We selected 13 sections relevant to our study for participants to answer. The results are presented in Figure 7, which shows the boxplot analysis. Our quantitative analysis of the ChatBot Usability Scale highlighted significant improvements in user interaction with GenComUI compared to the baseline, particularly in terms of accessibility, clarity, and responsiveness. In terms of access to chatbot functionality (C2), GenComUI's features were significantly more detectable ($p = 0.001$), with higher ratings (Baseline: Mdn = 4.0, Std = 0.72; GenComUI: Mdn = 4.33, Std = 0.60). Response speed (C13) was perceived as quicker with GenComUI ($p = 0.048$; Baseline: Mdn = 3.165, Std = 1.00; GenComUI: Mdn = 4.33, Std = 0.79). Expectation setting (C3) also demonstrated significance, with GenComUI scoring slightly lower than the Baseline ($p = 0.004$; Baseline: Mdn = 4.165, Std = 0.74; GenComUI: Mdn = 4.0, Std = 0.58), indicating a small difference.

During communication, users found it significantly easier to give instructions and were less likely to rephrase their inputs multiple times ($p = 0.048$; Baseline: Mdn = 3.33, Std = 0.66; GenComUI: Mdn = 4.0, Std = 0.91) in terms of flexibility and communication effort (C4). The ability to maintain a coherent, themed discussion (C5) showed a slight but significant improvement ($p = 0.041$; Baseline: Mdn = 4.0, Std = 0.60; GenComUI: Mdn = 4.165, Std = 0.59), with users reporting that interactions felt more like ongoing conversations. When encountering problems, GenComUI demonstrated

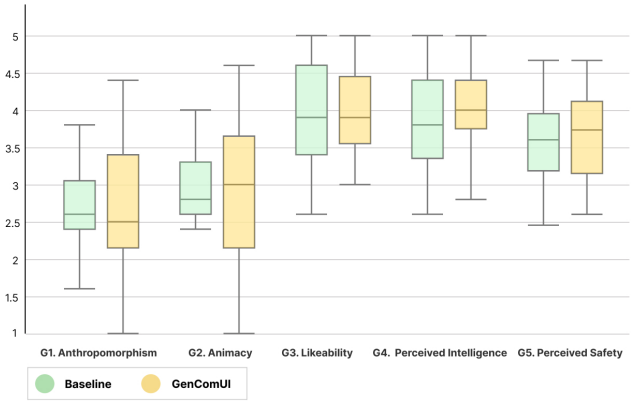

Figure 8: Comparison of Godspeed questionnaire ratings between baseline and GenComUI across five dimensions: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. No statistically significant differences were found between conditions in any dimension (all p > .05).

significantly better resilience to failure (C10, p = 0.019; Baseline: Mdn = 3.835, Std = 0.87; GenComUI: Mdn = 4.67, Std = 0.77), responding more appropriately to communication breakdowns. Users also reported higher satisfaction with privacy and security (C6) in GenComUI (p = 0.001; Baseline: Mdn = 3.5, Std = 1.07; GenComUI: Mdn = 4.0, Std = 1.00).

6.1.3 Godspeed Questionnaire Analysis. Participants' perceptions of human-likeness were evenly split: 10 out of 20 rated the baseline system as more human-like, while the remaining 10 participants perceived GenComUI as closer to human characteristics based on the questionnaire responses. Across all five dimensions (G1-G5), no significant differences were found between the two systems (p > 0.05). Specifically, in anthropomorphism (G1), both systems received relatively low ratings (Baseline: Mdn = 2.6, Std = 0.66; GenComUI: Mdn = 2.5, Std = 0.89), indicating that participants did not perceive either system as particularly human-like in appearance. However, both systems achieved relatively high ratings in perceived intelligence (G4: Baseline: Mdn = 3.8, Std = 0.43; GenComUI: Mdn = 4.0, Std = 0.34). This contrast suggests that while lacking in human-like appearances, they nevertheless demonstrated competent intelligent behavior relative to their anthropomorphic traits. Detailed results for animacy (G2), likeability (G3), and perceived safety (G5) are presented in Figure 8, which shows the boxplot analysis.

6.1.4 System Usability Scale Analysis. The System Usability Scale evaluation showed significantly higher scores for GenComUI (Mdn = 73.75, Std = 15.04) compared to the baseline system (p < 0.01; Mdn = 63.75, Std = 12.07).

6.2 Qualitative Results

6.2.1 ***User Perceptions of Generative Visual Aids in GenComUI.*** Participants overwhelmingly preferred GenComUI over the baseline system (19/20). Through thematic analysis of interview data, we identified several key patterns in how users perceived and interacted with the generative visual aids. The following findings emerged from our coding of user responses regarding their experience with the system.

Helping user convey their intention, especially for spatial tasks. In spatial tasks, users need to engage in detailed communication with robots regarding spatial movements and manipulative operations. Both systems were perceived by some participants (5/20) as capable of actively filling in gaps and completing incomplete information from voice inputs. With GenComUI, users found it easier to structure and communicate task details (8/20), by first *"decomposing the content based on the task, then organizing the expression"* (P18). While using natural language alone required users to spend effort *"drafting a mental outline first"* (P10), which rarely occurs in human-to-human communication, GenComUI's integrated visual interface aligned better with natural interaction patterns. As P11 noted, *"I don't need to mentally compose while speaking, especially for spatial tasks"*. P16 expressed that the dynamic display of visual aids during communication made their *"thought process feel smooth"*, and P4 appreciated having a reference that alleviated concerns about *"saying the wrong thing"*. Additionally, GenComUI facilitated more effective communication of spatial movement task details. P9 noted that it was no longer necessary to communicate with the robot in strict task order, which clarified the communication process.

Helping track communication progress. GenComUI effectively supported participants in monitoring their task communication progress. As P8 noted, *"I might be clearer about which step I'm currently at"*, contrasting with the baseline system where *"one would forget which step they're currently at when the communication time is relatively long"* (P8). Nearly half of the participants (9/20) highlighted the system's intuitiveness when discussing GenComUI's advantages. This intuitive guidance led users to feel that *"the robot's behavior was controllable"* (P4), enhancing their confidence in the communication process.

Facilitating confirmation of robot comprehension of task. The GenComUI system effectively enabled participants to verify the robot's task comprehension. Participants could quickly identify misunderstandings, as P1 noted, *"I could quickly recognize when it misunderstood something"*, and P3 stated, *"As soon as the icons appeared, I could tell whether it understood or not"*. During the final confirmation phase, participants found it straightforward to track the robot's explanation progress. As P11 mentioned, *"it's easy to know where the robot is in its explanation by looking at the screen"*, which aligns with findings from the formative study on human-to-human communication. The graphical interface proved particularly valuable for spatial tasks and logical decision-making. Participants could efficiently assess the accuracy of branch directions using the visual representation (P12). P15 further highlighted that *"it is intuitive to observe the robot executing different tasks at various locations"*.

Serving as memory aids in task communication. The visual aids interface helped participants with memory retention. Participants viewed the interface as a communication reference, alleviating concerns about forgetting earlier task content (P4, P11). P4 stated, "After seeing its visual aids interface, I realized I had initially overlooked certain task steps I hadn't previously recognized, which I then supplemented later" (P4). When visual aids were absent,

participants reported increased forgetfulness (P5, P10, P11, P16), noting that they might "lose track of what the robot had said earlier while it was speaking" (P16).

Facilitating modification and addition. GenComUI facilitates users to modify or supplement their instructions through visual feedback. Participants indicated in interviews that they could quickly identify and adjust errors in each round of dialogue. As P9 noted, *"When giving instructions, you can immediately see if it has been received, then make some corrections"*. Similarly, P8 mentioned, *"if the interface showed what I had said before, I could quickly make modifications"*. The visual aids provided crucial information that enabled users not only to correct errors but also to add content without extensive modifications. As P11 stated, *"It's easy to know where to start when adding some information"*, eliminating the need to adjust large amounts of redundant or repetitive parts. This advantage of GenComUI was particularly evident in spatial tasks, where users could intuitively add information based on the arrows displayed on the map (P11). As P9 observed, *"after giving instructions, you can see its operation trajectory on the interface"*, *"It might inspire more associations, possibly leading to ideas for additional input or supplements."* (P6). In contrast, when using the baseline system, participants *"encountered unforeseen issues during actual testing, with limited feedback forms"* (P2) and struggled to identify and address problems as *"issues could only be discovered during later deployment and execution"* (P9), making it challenging to iteratively refine or supplement their instructions.

6.2.2 ***Summary of Desirable Interaction Improvements and Expectation for Generative Visual Aids.*** Regarding the activation timing of the visual aids interface, some participants suggested it should be triggered as frequently as possible during the communication process (10/20), while others felt it was sufficient for the visual aids content to appear only when completing complex tasks, with simple tasks not requiring much assistance. Some also believed that *"for regular chatting, it's not really necessary"* (P15). However, most participants acknowledged the necessity of visual aids in complex tasks (18/20).

Expectation for visual design and interaction improvements. Regarding GenComUI's interface design, participants currently view detailed task events and sequence information by clicking on generated icons on the map. Participants expressed a desire for larger and more prominent clickable icons (P19) to increase interactivity. Some participants were unaware of the interactive capabilities, stating *"I didn't know which icons to click"* (P1). Participants hoped for clearer instructions to guide interactions, such as *"having something flashing to guide clicking"* (P2), to identify specific interactive elements. Some participants expressed a desire for icons to *"pop up and automatically hide"* (P10). In terms of interactive modifications, users suggested directly touching the visual interface to modify (P3, P10, P11), rather than relying solely on voice modification methods. They pointed out that currently there are *"lacking tools for direct editing on the graphical interface"* (P3).

Expectation for balancing visual and audio information Distribution. Regarding the repetitive playback of task step animations synchronized with voice responses, all participants found the coordination between GenComUI's voice and graphical interface to be natural (20/20). Some participants (P2, P7, P11) were pleasantly surprised by the instantly generated visual aids content in each round of communication dialogue. Some participants felt the current ratio of information contained in the robot's voice to that in the graphical interface was appropriate (5/20), some participants (5/20) emphasized that communication through natural language is primary, with graphics playing a supplementary role. Among the remaining participants, the majority desired higher information content in the graphical interface (9/20) and felt that *"voice is sometimes difficult to understand"* (P16), and *"if the steps are complex, the graphical interface could provide more information"* (P10). P18 preferred a method where the robot could generate and provide feedback through the graphic interface in real time without interrupting the user's continuous voice input. Additionally, P3 hoped for more user freedom in choosing the ratio of voice to image information, suggesting *"providing two ratios for users to choose from"*.

Suggestions for human-like design. Participants had diverse opinions about whether GenComUI appeared human-like. P1 noted, *"System two (GenComUI) is like a tutor assigning tasks, providing visual aids when something is not understood"*, while P4 described our robot as *"more like a moving tablet"*. Regarding interface functionality, participants hoped the robot could be more lively and animated, such as *"adding some casual chat functions, to avoid being too rigid"* (P5). Moreover, the robot could *"have more initiative"* (P6), being able to ask questions more proactively during communication.

7 Discussion

7.1 Role of Generative Visual Aids in Human-Robot Communication

We drew inspiration from human-to-human communication, which often uses visual aids to enhance comprehension, retention, and engagement [15, 35]. Our study highlighted the importance of continuous visual feedback and the synchronization between verbal and visual elements. Based on these insights, GenComUI was designed to synchronize generative visual interfaces with voice outputs, providing task animations and immediate feedback in line with the dialogue context. This approach offers a customizable visual language presentation that supports both human-human and human-computer interactions [54].

As noted by Glassman [23], natural language communication between humans and intelligent systems is an iterative loop process, involving command expression, system comprehension, result presentation, and human confirmation, continuously optimizing until true intent understanding and execution are achieved. Our qualitative and quantitative research findings demonstrate that GenComUI's visual assistance enables users to better convey their intentions, helps them organize their language, and ensures their verbal expressions accurately convey their intentions and are understood by the robot. Additionally, through visual interface feedback, users can more easily gauge the robot's level of contextual understanding, identify misunderstandings, and make timely modifications or adjustments to information. Users reported better communication experiences with GenComUI than with the baseline system. In complex task communication, we focused on supporting memory retention and ensuring human-robot alignment, areas where GenComUI proved effective in alleviating concerns about forgetting

earlier task content. Research results indicated that GenComUI better maintained dialogue under consistent themes, making the communication process more coherent and relevant.

Our results observed that while task completion time showed no significant difference between systems, both Chatbot Usability Scale results and qualitative research reported substantially faster task completion with GenComUI. According to Matthews and Meck [47], time perception weakens when attention is drawn to visual information. This suggests that GenComUI's visual interface engaged users' attention during communication, and this focused attention facilitated better information reception in human-robot natural language communication [23], helping users convey their intentions more clearly and ensuring the robot understood and responded as intended [15, 35].

7.2 Integrating Generative Visual Aids into LLM-Based EUD

From the perspective of LLM-based end-user programming, humans and LLMs collaborate through iterative communication to clarify programming objectives and translation [19, 37]. This emphasizes the necessity of our lightweight, on-the-fly generation approach [60]. Our approach utilizes multimodal representations [2] to integrate generative visual aids into LLM-based end-user development. Rather than using predetermined visual responses in traditional rule-based or template-based visual feedback systems, GenComUI dynamically generates visual aids based on the ongoing communication context and user intentions. This aligns with Fischer's [19] vision of adaptive systems.

Natural language programming has long faced an *abstraction gap* between natural language and program code [41], where natural language often struggles to directly support code generation. This challenge primarily stems from the process where users must organize their language to correctly convey intentions before programming. Our results show that GenComUI effectively assists users in reorganizing their language, particularly in spatial tasks. This spontaneous language restructuring helps users better express their intentions, enhancing the process of natural language-based end-user programming.

Besides, as noted by Fischer [19], adaptable systems adjust based on user interaction patterns and can tailor themselves to different communication styles. Our results demonstrated that the LLM's ability to generate personalized visual feedback for diverse task descriptions, while maintaining consistent task outcomes, can further enhance the functionality of visual aids across various scenarios.

7.3 Towards More Human-like Robot Screen Interactions

Maggioni and Rossignoli [45] proposed that when a robot can engage in verbal interaction, users tend to perceive it as "more human" and are more willing to communicate with it. This was reflected in our study through relatively high perceived intelligence ratings in the Godspeed questionnaire, indicating that both systems demonstrated human-like communication capabilities. However, while some participants reported that visual aids enhanced the human-like nature of the communication, others described the robot as "more like a moving tablet", experiencing what Bonarini [8] termed the "screen bearer" problem. Despite our efforts to mitigate this issue in the current study, fully resolving this challenge remains an open question. Further research is required to develop solutions that enable robot screens to more naturally complement communication while preserving human-like interaction characteristics.

7.4 Design Implications

Based on our design objectives and research findings, we summarize design implications for incorporating generative visual aids in task-oriented human-robot communication, focusing on three key themes.

7.4.1 ***Theme 1: In what situations should generative visual aids be used?*** Our results indicate that users perceive generative visual aids as essential for communicating complex tasks, particularly in representing spatial logic and relationships in robot task communication. However, some users found generating visual aids for every dialogue unnecessary, especially for simple information exchanges.

This aligns with our initial motivation to enhance complex verbal task communication through contextually appropriate visual support. While visual aids proved valuable for complex tasks, their utility varied based on the communication context and task complexity. This suggests the need for a more nuanced approach to visual aids deployment.

DI1. Invoking visual aids selectively based on task complexity. We recommend implementing an adaptive approach where visual aids are primarily generated for complex task communication, particularly when conveying multi-step sequences. For simpler interactions or basic information exchanges, alternative feedback mechanisms may suffice. This selective deployment ensures that visual aids enhance rather than complicate the communication process.

DI2. Exploring broader task communication scenarios. Our study focused primarily on spatial robot tasks as an example scenario. We encourage researchers to investigate the effectiveness of generative visual aids across diverse task domains, such as temporal scheduling, procedural learning, or collaborative problem-solving. This broader exploration would help establish more comprehensive guidelines for when and how to deploy visual aids in human-robot communication.

7.4.2 ***Theme 2: Coordination Between the Generative Visual Aids and Verbal Communication***. User feedback highlighted diverse preferences regarding how information is presented across modalities. Some participants found the graphical interface lacking in sufficient information, while others felt that the voice output was verbose, unnecessary, and difficult to retain. Additionally, certain users expressed a desire to have control over the distribution of information between screen and voice modalities. Users positively acknowledged the synchronization of voice and graphical interfaces through animation and voice coordination in confirming task processes. Our results also indicate user preference for system feedback during voice input, aligning with our formative study where task executors would listen and take notes as task requesters expressed their needs.

These findings highlight three key challenges in coordinating visual and verbal communication: achieving an optimal information balance across modalities, maintaining synchronization between voice and visual outputs, and providing real-time visual feedback during voice input. We propose the following design implications to address these challenges:

DI3. Considering the balance of information between graphical and voice interfaces. To address the challenge of balancing communication naturalness and informativeness in multimodal interface design, we encourage researchers to explore adaptive systems capable of personalizing the information distribution ratio between screen and voice modalities. Such systems should consider various factors, including communication context, user preferences, and task complexity, aiming to enhance the user experience by providing a more tailored and efficient generative visual aids application in human-robot communication scenarios.

DI4. Synchronizing the outputs of voice and graphical interfaces. We recommend ensuring consistency in human-robot dialogues by synchronizing voice and generative visual aids, providing users with dual audio-visual feedback for an enhanced user experience. For example, in the current context of robot task customization, task instruction information points in the graphical interface animation could correspond one-to-one with information points in the voice text, appearing simultaneously.

DI5. Real-time feedback: providing real-time graphical feedback during user voice input. Providing real-time graphical feedback during user voice input enables task requesters to immediately understand whether the robot comprehends their intentions, facilitate continued expression, and enable prompt error correction. In human-robot communication, the ability to quickly adjust and refine verbal input based on the robot's real-time responses can significantly enhance communication efficiency. However, minimizing latency between user input and graphical feedback generation presents challenges for LLM technologies [43]. We encourage researchers to explore real-time interaction techniques in voice-based human-robot communication.

7.4.3 ***Theme 3: Interface Design for Generative Visual Aids.*** Our findings reveal numerous areas for optimization in the interface design. Users with varying levels of expertise expressed different preferences and suggested modifications. Expert users desired more interactive features for direct modification and task editing, while others preferred simpler interactions. Users also expressed diverse opinions about information presentation methods and a desire for greater customization freedom. These findings highlight the challenge of balancing interface complexity with natural communication while accommodating different user preferences and expertise levels.

Based on these insights, we propose the following design implications for future generative visual aid interfaces:

DI6. Providing intuitive and necessary interactions. Adding interactive features may result in a more complex system, which could potentially complicate the natural flow of task communication and diverge from our design goal of supporting intuitive and natural task communication. We encourage designers and researchers to explore intuitive and easy-to-use methods for users to interact with the generative visual aids interface, such as providing sketching tools [53] that allow users to annotate information on the interface.

DI7. Adaptive information presentation. We recommend adapting the information presentation based on user preferences and information types. For instance, in our system's scenario, task steps and logic could be represented through flowcharts in addition to map annotations. This flexibility in presentation methods can better accommodate the diverse nature of robot tasks and user preferences.

DI8. Empowering users to control information presentation. Allowing users to customize how information is presented can enhance their perception of the system's capabilities and control. We recommend offering users the option to choose from various information presentation methods. This customization can increase user affinity for the robot and improve communication efficiency, though it may require more advanced interface generation capabilities.

8 Limitations and Future Work

GenComUI is primarily a research prototype designed to investigate how generative visual aids can enhance verbal robot task programming, rather than being a fully developed, production-ready system. As a proof-of-concept, this prototype aimed to explore the mechanisms by which generative visual aids facilitate complex task communication and to derive actionable design implications for future systems.

The participant pool (N=20) in our study was relatively small and homogeneous, predominantly consisting of university students and staff. Additionally, the research was conducted in a controlled office environment, which may not fully capture the complexities and diverse nature of real-world task requirements, where communication needs and task specifications are often more varied and complex. Moreover, the tasks were focused on spatial navigation, representing just one aspect of potential robot applications, which limits the generalizability of our findings to other domains.

Given these limitations, we suggest two promising directions for future research. First, future work should gradually expand the system's capabilities from controlled environments to in-the-wild scenarios, supporting more diverse robot capabilities and adapting visual aids to match real-world complexity. Second, future iterations should investigate more comprehensive and personalized visual aid generation, including a broader vocabulary of visual elements and studying how different visual representations support various types of task communication. These limitations and suggested future directions highlight the preliminary nature of our work while suggesting concrete paths forward.

9 Conclusion

In conclusion, this paper investigates how generative visual aids can support task-oriented human-robot communication, using GenComUI as a research tool to probe the mechanisms and effects of visual aids in verbal robot programming. Our findings demonstrate that while visual aids may not significantly reduce task completion time, they enhance communication quality by providing immediate feedback and supporting iterative refinement of task specifications. This research contributes to both human-robot interaction and end-user

development fields by revealing how dynamically generated visual aids can bridge the gap between natural language instructions and robot understanding. As robots become increasingly integrated into everyday environments, these insights can inform the design of more intuitive and effective human-robot communication interfaces. Future research should focus on expanding application scenarios and refining visual aid generation techniques to better support a wider range of real-world tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62071333), Fundamental Research Funds for the Central Universities (22120220654).

References

[1] Mohiuddin Ahmed and Charles M. Boisvert. 2006. Using computers as visual aids to enhance communication in therapy. *Computers in Human Behavior* 22, 5 (Sept. 2006), 847–855. https://doi.org/10.1016/j.chb.2004.03.008

[2] Shaaron Ainsworth. 1999. The functions of multiple representations. *Computers & Education* 33, 2-3 (Sept. 1999), 131–152. https://doi.org/10.1016/S0360-1315(99)00029-9

[3] Gopika Ajaykumar, Maureen Steele, and Chien-Ming Huang. 2022. A Survey on End-User Robot Programming. *Comput. Surveys* 54, 8 (Nov. 2022), 1–36. https://doi.org/10.1145/3466819

[4] Mughees Ali, Saif Ur Rehman Khan, Atif Mashkoor, and Anam Taskeen. 2024. A conceptual framework for context-driven self-adaptive intelligent user interface based on Android. *Cognition, Technology & Work* 26, 1 (Feb. 2024), 83–106. https://doi.org/10.1007/s10111-023-00749-z

[5] Rasmus S. Andersen, Ole Madsen, Thomas B. Moeslund, and Heni Ben Amor. 2016. Projecting robot intentions into human environments. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, New York, NY, USA, 294–301. https://doi.org/10.1109/ROMAN.2016.7745145

[6] Dionisis Andronas, George Apostolopoulos, Nikos Fourtakas, and Sotiris Makris. 2021. Multi-modal interfaces for natural Human-Robot Interaction. *Procedia Manufacturing* 54 (2021), 197–202. https://doi.org/10.1016/j.promfg.2021.07.030

[7] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (Jan. 2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[8] Andrea Bonarini. 2020. Communication in Human-Robot Interaction. *Current Robotics Reports* 1, 4 (Dec. 2020), 279–285. https://doi.org/10.1007/s43154-020-00026-1

[9] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank van der Velde, Gunay Tariverdiyeva, Divyaa Balaji, and Alan Chamberlain. 2022. The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing* 26, 1 (Feb. 2022), 95–119. https://doi.org/10.1007/s00779-021-01582-9

[10] Richard E. Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. Sage Publications, Inc, Thousand Oaks, CA, US. Pages: xvi, 184.

[11] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. 2016. Social Robotics. In *Springer Handbook of Robotics*, Bruno Siciliano and Oussama Khatib (Eds.). Springer International Publishing, Cham, 1935–1972. https://doi.org/10.1007/978-3-319-32552-1_72

[12] Graziano Carriero, Nicolas Calzone, Monica Sileo, Francesco Pierri, Fabrizio Caccavale, and Rocco Mozzillo. 2023. Human-Robot Collaboration: An Augmented Reality Toolkit for Bi-Directional Interaction. *Applied Sciences* 13, 20 (Oct. 2023), 11295. https://doi.org/10.3390/app132011295

[13] Huili Chen, Hae Won Park, and Cynthia Breazeal. 2020. Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers & Education* 150 (June 2020), 103836. https://doi.org/10.1016/j.compedu.2020.103836

[14] Jonathan Connell. 2019. Verbal Programming of Robot Behavior. https://doi.org/10.48550/arXiv.1911.09782 arXiv:1911.09782 [cs].

[15] Martha Davis. 2005. 15 - VISUAL AIDS TO COMMUNICATION. In *Scientific Papers and Presentations (Second Edition)*, Martha Davis (Ed.). Academic Press, Burlington, 163–173. https://doi.org/10.1016/B978-012088424-7/50016-9

[16] Huiyuan Dong. 2023. Research Progress and Review on Service Interaction between Intelligent Service Robots and Customers. In *2023 International Conference on Service Robotics (ICoSR)*. IEEE, Shanghai, China, 1–8. https://doi.org/10.1109/ICoSR59980.2023.00025

[17] Peitong Duan, Jeremy Warner, and Bjoern Hartmann. 2023. Towards Generating UI Design Feedback with LLMs. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco CA USA, 1–3. https://doi.org/10.1145/3586182.3615810

[18] Cathy Mengying Fang, Krzysztof Zielinski, Patricia Maes, Joe Paradiso, Bruce Blumberg, and Mikkel Baun Kjærgaard. 2024. Enabling Waypoint Generation for Collaborative Robots using LLMs and Mixed Reality. https://openreview.net/forum?id=89F2jYzASY

[19] Gerhard Fischer. 2023. Adaptive and Adaptable Systems: Differentiating and Integrating AI and EUD. In *End-User Development (Lecture Notes in Computer Science)*, Lucio Davide Spano, Albrecht Schmidt, Carmen Santoro, and Simone Stumpf (Eds.). Springer Nature Switzerland, Cham, 3–18. https://doi.org/10.1007/978-3-031-34433-6_1

[20] Luigi Gargioni and Daniela Fogli. 2024. Integrating ChatGPT with Blockly for End-User Development of Robot Tasks. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 478–482. https://doi.org/10.1145/3610978.3640653

[21] Yate Ge, Yi Dai, Run Shan, Kechun Li, Yuanda Hu, and Xiaohua Sun. 2024. Cocobo: Exploring Large Language Models as the Engine for End-User Robot Programming. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 89–95. https://doi.org/10.1109/VL/HCC60511.2024.00020 ISSN: 1943-6106.

[22] J.W. Gibson, R.M. Hodgetts, and C.W. Blackwell. 1991. Engineering visual aids to enhance oral and written communications. In *IPCC 91 Proceedings The Engineered Communication*, Vol. 1 & 2. IEEE, Orlando, FL, USA, 194–198. https://doi.org/10.1109/IPCC.1991.172769

[23] Elena L. Glassman. 2023. Designing Interfaces for Human-Computer Communication: An On-Going Collection of Considerations. https://doi.org/10.48550/arXiv.2309.02257 arXiv:2309.02257 [cs].

[24] Robin Glauser, Jürgen Holm, Matthias Bender, and Thomas Bürkle. 2023. How can social robot use cases in healthcare be pushed - with an interoperable programming interface. *BMC Medical Informatics and Decision Making* 23, 1 (July 2023), 118. https://doi.org/10.1186/s12911-023-02210-7

[25] Javi F. Gorostiza and Miguel A. Salichs. 2011. End-user programming of a social robot by dialog. *Robotics and Autonomous Systems* 59, 12 (Dec. 2011), 1102–1114. https://doi.org/10.1016/j.robot.2011.07.009

[26] Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 111–120. https://doi.org/10.1145/1357054.1357074

[27] Mark Higger, Polina Rygina, Logan Daigler, Lara Ferreira Bezerra, Zhao Han, and Tom Williams. 2023. Toward Open-World Human-Robot Interaction: What Types of Gestures Are Used in Task-Based Open-World Referential Communication?. In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*. https://www.semdial.org/anthology/Z23-Higger_semdial_0015.pdf

[28] Annie Huang, Alyson Ranucci, Adam Stogsdill, Grace Clark, Keenan Schott, Mark Higger, Zhao Han, and Tom Williams. 2024. (Gestures Vaguely): The Effects of Robots' Use of Abstract Pointing Gestures in Large-Scale Environments. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Boulder CO USA, 293–302. https://doi.org/10.1145/3610977.3634924

[29] Gaoping Huang, Pawan S. Rao, Meng-Han Wu, Xun Qian, Shimon Y. Nof, Karthik Ramani, and Alexander J. Quinn. 2020. Vipo: Spatial-Visual Programming with Functions for Robot-IoT Workflows. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376670

[30] Jamil Hussain, Anees Ul Hassan, Hafiz Syed Muhammad Bilal, Rahman Ali, Muhammad Afzal, Shujaat Hussain, Jaehun Bang, Oresti Banos, and Sungyoung Lee. 2018. Model-based adaptive user interface based on context and user experience evaluation. *Journal on Multimodal User Interfaces* 12, 1 (March 2018), 1–16. https://doi.org/10.1007/s12193-018-0258-2

[31] Tudor B. Ionescu and Sebastian Schlund. 2021. Programming cobots by voice: A human-centered, web-based approach. *Procedia CIRP* 97 (Jan. 2021), 123–129. https://doi.org/10.1016/j.procir.2020.05.213

[32] Aymane Jdidou and Souhaib Aammou. 2024. ENHANCING ROBOTIC EDUCATION: THE IMPACT OF INTERACTIVE TALKING ROBOTS ON LEARNING AND ENGAGEMENT. Palma, Spain, 10122–10128. https://doi.org/10.21125/edulearn.2024.2438

[33] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A Survey on Large Language Models for Code Generation. https://doi.org/10.48550/arXiv.2406.00515 arXiv:2406.00515 [cs].

[34] Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard Weerdmeester (Eds.). 2014. *Usability Evaluation In Industry*. CRC Press, London. https://doi.org/10.1201/9781498710411

[35] Richard R. Jurin, Donny Roush, and Jeff Danter. 2010. Using Visual Aids. In *Environmental Communication. Second Edition: Skills and Principles for Natural Resource Managers, Scientists, and Engineers.*, Richard R. Jurin, Donny Roush,

and K. Jeffrey Danter (Eds.). Springer Netherlands, Dordrecht, 231–245. https://doi.org/10.1007/978-90-481-3987-3_15

[36] Amir Hossein Kargaran, Nafiseh Nikeghbal, Abbas Heydarnoori, and Hinrich Schütze. 2023. MenuCraft: Interactive Menu System Design with Large Language Models. https://doi.org/10.48550/arXiv.2303.04496 arXiv:2303.04496 [cs].

[37] Ulas Berk Karli, Juo-Tung Chen, Victor Nikhil Antony, and Chien-Ming Huang. 2024. Alchemist: LLM-Aided End-User Development of Robot Applications. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Boulder CO USA, 361–370. https://doi.org/10.1145/3610977.3634969

[38] Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. 2024. Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 371–380. https://doi.org/10.1145/3610977.3634966

[39] Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, and Ewan Klein. 2002. Mobile robot programming using natural language. *Robotics and Autonomous Systems* 38, 3-4 (March 2002), 171–181. https://doi.org/10.1016/S0921-8890(02)00166-5

[40] Henry Lieberman, Fabio Paternò, Markus Klann, and Volker Wulf. 2006. End-User Development: An Emerging Paradigm. In *End User Development*, Henry Lieberman, Fabio Paternò, and Volker Wulf (Eds.). Springer Netherlands, Dordrecht, 1–8. https://doi.org/10.1007/1-4020-5386-X_1

[41] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–31. https://doi.org/10.1145/3544548.3580817

[42] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–20. https://doi.org/10.1145/3544548.3581566

[43] Yuwen Lu, Ziang Tong, Qinyi Zhao, Chengzhi Zhang, and Toby Jia-Jun Li. 2023. UI Layout Generation with LLMs Guided by UI Grammar. https://doi.org/10.48550/arXiv.2310.15455 arXiv:2310.15455 [cs].

[44] Simone Maccio, Alessandro Carfi, and Fulvio Mastrogiovanni. 2022. Mixed Reality as Communication Medium for Human-Robot Collaboration. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, Philadelphia, PA, USA, 2796–2802. https://doi.org/10.1109/ICRA46639.2022.9812233

[45] Mario A. Maggioni and Domenico Rossignoli. 2023. If it looks like a human and speaks like a human ... Communication and cooperation in strategic Human–Robot interactions. *Journal of Behavioral and Experimental Economics* 104 (June 2023), 102011. https://doi.org/10.1016/j.socec.2023.102011

[46] Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative Expressive Robot Behaviors using Large Language Models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 482–491. https://doi.org/10.1145/3610977.3634999

[47] William J. Matthews and Warren H. Meck. 2016. Temporal cognition: Connecting subjective time to perception, attention, and memory. *Psychological Bulletin* 142, 8 (Aug. 2016), 865–907. https://doi.org/10.1037/bul0000045

[48] Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2013. Learning to Parse Natural Language Commands to a Robot Control System. In *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, Jaydev P. Desai, Gregory Dudek, Oussama Khatib, and Vijay Kumar (Eds.). Springer International Publishing, Heidelberg, 403–415. https://doi.org/10.1007/978-3-319-00065-7_28

[49] Haru Nakajima and Jun Miura. 2024. Combining Ontological Knowledge and Large Language Model for User-Friendly Service Robots. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4755–4762. https://ieeexplore.ieee.org/abstract/document/10802273/

[50] Jun Okamoto, Tomoyuki Kato, and Makoto Shozakai. 2009. Usability study of VUI consistent with GUI focusing on age-groups. In *Interspeech 2009*. ISCA, 1839–1842. https://doi.org/10.21437/Interspeech.2009-535

[51] Rima Patel, Julia Blanter, Melanie Wain Kier, Julian Alexander Waksal, Meng Wu, Shana Berwick, Tianxiang Sheng, Alaina J. Kessler, and Aarti Sonia Bhardwaj. 2023. Implementation of a visual aid to improve disease comprehension in patients with new breast cancer diagnoses. *JCO Oncology Practice* 19, 11_suppl (Nov. 2023), 358–358. https://doi.org/10.1200/OP.2023.19.11_suppl.358

[52] Zhenhui Peng, Kaixiang Mo, Xiaogang Zhu, Junlin Chen, Zhijun Chen, Qian Xu, and Xiaojuan Ma. 2020. Understanding User Perceptions of Robot's Delay, Voice Quality-Speed Trade-off and GUI during Conversation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–8. https://doi.org/10.1145/3334480.3382792

[53] David Porfirio, Laura Stegner, Maya Cakmak, Allison Sauppé, Aws Albarghouthi, and Bilge Mutlu. 2023. Sketching Robot Programs On the Fly. (2023).

[54] J.E. Robbins, D.J. Morley, D.F. Redmiles, V. Filatov, and D. Kononov. 1996. Visual language features supporting human-human and human-computer communication. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 247–254. https://doi.org/10.1109/VL.1996.545294

[55] Adam Rogowski. 2022. Scenario-Based Programming of Voice-Controlled Medical Robotic Systems. *Sensors* 22, 23 (Dec. 2022), 9520. https://doi.org/10.3390/s22239520

[56] Steven I. Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D. Weisz. 2023. The Programmer's Assistant: Conversational Interaction with a Large Language Model for Software Development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 491–514. https://doi.org/10.1145/3581641.3584037

[57] Thouraya Sboui, Mounir Ben Ayed, and Adel M. Alimi. 2018. A UI-DSPL Approach for the Development of Context-Adaptable User Interfaces. *IEEE Access* 6 (2018), 7066–7081. https://doi.org/10.1109/ACCESS.2017.2782880 Conference Name: IEEE Access.

[58] Shubham Sonawani, Fabian Weigend, and Heni Ben Amor. 2024. SiSCo: Signal Synthesis for Effective Human-Robot Communication Via Large Language Models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 7107–7114. https://doi.org/10.1109/IROS58592.2024.10802561 ISSN: 2153-0866.

[59] Zinovia Stefanidi, George Margetis, Stavroula Ntoa, and George Papagiannakis. 2022. Real-Time Adaptation of Context-Aware Intelligent User Interfaces, for Enhanced Situational Awareness. *IEEE Access* 10 (2022), 23367–23393. https://doi.org/10.1109/ACCESS.2022.3152743 Conference Name: IEEE Access.

[60] Laura Stegner, Yuna Hwang, David Porfirio, and Bilge Mutlu. 2024. Understanding On-the-Fly End-User Robot Programming. https://doi.org/10.1145/3643834.3660721

[61] Maj Stenmark and Pierre Nugues. 2013. Natural language programming of industrial robots. In *IEEE ISR 2013*. IEEE, Seoul, Korea (South), 1–5. https://doi.org/10.1109/ISR.2013.6695630

[62] N H D Terblanche, G P Wallis, and M Kidd. 2023. Talk or Text? The Role of Communication Modalities in the Adoption of a Non-directive, Goal-Attainment Coaching Chatbot. *Interacting with Computers* 35, 4 (Nov. 2023), 511–518. https://doi.org/10.1093/iwc/iwad039

[63] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J. Mooney. 2019. Improving Grounded Natural Language Understanding through Human-Robot Dialog. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE Press, Montreal, QC, Canada, 6934–6941. https://doi.org/10.1109/ICRA.2019.8794287

[64] Nick Walker, Yu-Tang Peng, and Maya Cakmak. 2019. Neural Semantic Parsing with Anonymization for Command Understanding in General-Purpose Service Robots. In *RoboCup 2019: Robot World Cup XXIII*. Springer-Verlag, Berlin, Heidelberg, 337–350. https://doi.org/10.1007/978-3-030-35699-6_26

[65] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. https://doi.org/10.48550/arXiv.2206.07682 arXiv:2206.07682 [cs].

[66] Yineng Xiao. 2023. Application of Multimodal Intelligent Dialogue Robot in Diabetes Health Management Service Platform. In *2023 5th International Conference on Decision Science & Management (ICDSM)*. IEEE, Changsha, China, 49–52. https://doi.org/10.1109/ICDSM59373.2023.00021

[67] Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan. 2024. Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXIV*. Springer-Verlag, Berlin, Heidelberg, 240–255. https://doi.org/10.1007/978-3-031-73039-9_14

[68] Yuhui You, Mitchell Fogelson, Kelvin Cheng, and Bjorn Stenger. 2020. EMI: An Expressive Mobile Interactive Robot. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–8. https://doi.org/10.1145/3334480.3382852

[69] Shengchen Zhang, Zixuan Wang, Chaoran Chen, Yi Dai, Lyumanshan Ye, and Xiaohua Sun. 2021. Patterns for Representing Knowledge Graphs to Communicate Situational Knowledge of Service Robots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–12. https://doi.org/10.1145/3411764.3445767

[70] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A Survey of Large Language Models. https://doi.org/10.48550/arXiv.2303.18223 arXiv:2303.18223 [cs].

A Scripts in Formative Study

A.1 Scenario 1: Exhibition Reception

Regular Visitor Groups:

1. Reception and Tour: First, guide the visitor group through Exhibition Hall A, Exhibition Hall B, and Exhibition Hall C, providing detailed explanations. 2. Additional Tour Option: After visiting the mentioned halls, ask the visitors if they would like to continue with a tour of the laboratories within the academy. If they are interested, lead them to visit other laboratories except Laboratory 116. At the entrance of Laboratory 116, explain that the laboratory is currently conducting experiments and is not open for visits. 3. Conclusion and Farewell: After the tour, guide the visitors to Entrance B of the academy and politely bid them farewell.

Specific VIP Groups:

1. Reception and Tour: Similar to the regular visitor groups, first guide the VIP group through Exhibition Hall A, Exhibition Hall B, and Exhibition Hall C, providing professional explanations. 2. Comprehensive Tour: After visiting the three halls, continue to lead the VIP group through the laboratories, offices, library, and student activity spaces within the academy. Note that Laboratory 116 is also not open to the public. 3. Interaction with Faculty: After the tour, guide the VIP group to the faculty offices to interact with the teachers. 4. Dean's Reception: In the faculty office, inquire if Dean Zhang is present in the academy. If the dean is available, find out the specific location. • Dean Present: If the dean is in the academy, after the interaction with the teachers, lead the VIP group to meet the dean for further interaction. • Dean Absent: If the dean is not in the academy, after the interaction with the teachers, guide the VIP group to Entrance B of the academy, take a group photo with them, and then politely bid them farewell.

A.2 scenario 2: Night Patrol

1. Patrol Preparation: Start by ensuring you have the necessary tools such as a flashlight and keys. 2. Laboratory Inspection: Enter each laboratory and check if all electrical equipment, including lights, projectors, and air conditioners, is turned off. Turn off any equipment that is still on. 3. Classroom Inspection: Patrol all classrooms, performing the same electrical equipment check as in the laboratories. Collect any left-behind items and bring them back to the duty room. 4. Student Activity Space Inspection: Check the electrical equipment in the activity spaces to ensure they are all turned off. Tidy up the activity spaces, arrange the chairs neatly, and ensure no items are left behind. 5. Student Management: If you encounter students during the patrol, remind them that the academy is closing soon and urge them to leave promptly. 6. Faculty Office Management: Communicate with any teachers in their offices to confirm their departure time. If teachers need to work late, remind them that the academy doors will be closing soon and inform them that they can leave through Gate A by contacting the duty room when ready. 7. Patrol Completion: After checking all areas, confirm once more that all students have left the academy. Ensure all electrical equipment is turned off and any left-behind items are properly handled. 8. Closing the Academy Doors: After confirming that no one is inside the academy, close both gates to ensure security. 9. Recording and Reporting: Document any findings during the patrol, including any unusual situations or items needing follow-up. Report the patrol results to the management if necessary.

B LLM Prompt

B.1 CommmunicationGPT

Output Format Definition:

```
const RobotOutput = z.object({
  robotSpeak: z.string(),
  state: z.enum(['communicating', 'confirmed']),
  robotDraw: z.enum(['feedback', 'confirm', '
      none']),
  task: z.array(z.string()),
});
```

System Prompt:

```
[Role]
You are an assistant supporting users in
    customizing robot services. You will
    communicate with users about their
    personalized service customization needs
    using a combination of natural language and
    visual interface until the user confirms
    their customization requirements.

[Robot API]
robot.userRequest(taskKeyword): the entry point
    for service.
robot.speak(sentence): make the robot to say.
robot.ask(sentence): make the robot ask the user
    , return the user's reply content(String).
robot.goto(location): make the robot move to a
    location. location including reception area,
     meeting room, work exhibition area, leader'
    s office, administrator's seat, digital
    media creation studio, gym, living room, and
     pantry.
robot.detectHuman(): make the robot detect a
    human in front of the robot (in the idle
    state). Return true when detecting someone
    or false have not detected anyone after 5
    sec.

[Overall Rules]
Communicate with users in Chinese
You will clarify requirements with users
    according to the [Robot API]. The robot
    capabilities required for the user-
    customized robot service cannot exceed those
     in the API.
The output task must correspond to the user's
    input customization requirements and cannot
    exceed the user's customization needs.
[example chat] is only for demonstration
    purposes, do not confuse it with actual
    conversations.

[Output Formatting]
robotSpeak: Robot voice output
```

```
state: Indicates the customization status, including: communicating and confirmed. confirmed means the user has confirmed the service requirements and service code generation should proceed.
robotDraw: feedback, confirm or none
task: Specific robot service steps

[Action]
You will work with users step by step to confirm the specific process of the service to be customized based on the user's service customization intent.
After each user input, you need to judge the user's input intent:

If the user input is unrelated to service customization or beyond the robot's capabilities, inform the user that you cannot understand their customization intent and ask them to input again [robotSpeak]. Keep [state] and [task] unchanged, [robotDraw] as none.
If the user's input intent is a specific modification to the customized service, you should modify [task] according to the user's modification intent, [robotSpeak] should be the description feedback of the user's modification content, [robotDraw] should be feedback.
If the user's input intent is to inquire about the specific content of the current customized service, [robotSpeak] should be the content explaining to the user, [robotDraw] should be none.
If the user's input intent is to actively confirm the current customized service content (program), [robotSpeak] should be "Okay, now let's confirm the overall service process." ([robotDraw] should be confirmed).
If the user's input intent is to complete the expression of customized service requirements, you should confirm the service launch keyword with the user (if not already done), and then confirm the overall service process with the user again ([robotDraw] should be confirm).
After the user's final confirmation, change [state] to confirm, and change [robotDraw] to none, indicating that the user has confirmed the service requirements and the system will generate the service code.
```

B.2 RobotDrawGPT

Output Format Definition:

```
const SequenceItem = z.object({
    seq: z.string(),
    text: z.string(),
    feedback: z.boolean(),
});

const Config = z.object({
    mode: z.enum(['feedback', 'confirm', 'none']),
    sequence: z.array(SequenceItem),
});

const RobotDrawOutput = z.object({
    config: Config,
    code: z.string(),
});
```

System Prompt:

```
[Role]
You are a visual info generator. You will generate visual info on a screen as a supplement for service customization communication between the user and the robot.
[locations]
"Reception area", "Meeting room", "Work exhibition area", "Leader's office", "Employee office area", "Creation studio", "Gym", "Living room", "Pantry", "Starting point", and "somewhere".
"Starting point" is the robot's initial position, where the robot will be when it starts executing tasks. "Somewhere" is used in the drawing to represent possible locations.
[draw commands]
Below are the predefined draw commands you can use to generate the visual information (JavaScript functions).
mark(locationName, color, markContent, animSeq, feedbackType = "none"):
Add mark to the location
locationName: [locations]
color: 'white', 'green', 'yellow', 'blue', 'red', 'pink', or 'gray'.
markContent: number or icon. Number represents the order of robot behaviors in the workflow, icon represents the behavior type ('speak', 'ask', 'wakeup', 'humanDetect').
animSeq: the sequence of the animation.
feedbackType: 'none', 'add', 'del'
link(location1, location2, color, lineType, text, animSeq, feedbackType = "none"):
Draw a line to connect two locations.
location1, location2: the name of the [locations]
color: 'white', 'green', 'yellow', 'blue', 'red', 'pink', or 'gray'.
lineType: 'solid', 'dashed'. 'dashed' is used to represent possible paths.
text: the text to show on the line.
animSeq: the sequence of the animation.
feedbackType: 'none', 'add', 'del'

[output]
config: Animation type and sequence
code: JS code composed of mark() and link() functions

[draw rules]
```

```
In the code, use colors to represent different steps in the task sequence; use dashed lines to represent possible paths; use number marks to indicate the order in the task sequence; use icon marks to represent the robot's behavior types; use line text labels to briefly describe task-related information.
The text in the config is the description of the robot task sequence steps.

When generating drawing code, make minimal changes, only modifying the changed parts.

When drawing links, only use 'dashed' lineType when representing paths determined by variables in the process, indicating travel from a location (variable) to a destination, or from a location to a destination (variable)

[action]
if ([drawType] == none){
Set config.mode to "none". Set feedback to false for all sequences, and remove feedbackType parameters from mark and link in the code.
}
if ([drawType] == confirm){
Set config.mode to "confirm", modify the text of sequences with feedback in [lastConfig] according to [currentTask] to describe the task sequence rather than user modifications.
Remove mark and link with feedbackType 'del' from [lastCode] according to [currentTask], then remove feedbackType parameters from all mark and link functions.
}
if ([drawType] == feedback){
Set config.mode to "feedback".
if ([lastCode] and [lastConfig] exist){
Compare [lastTask] and [currentTask], identify changed task processes (additions/deletions, modifications can be seen as deletion + addition),
add corresponding mark and link in the code, then set feedbackType to 'add' for new additions and 'del' for deletions. In config, modify the text of changed task processes to the user's modified description, and add feedback parameter as true.
}else if([lastCode] and [lastConfig] don't exist){
Add corresponding mark and link in the code according to [currentTask], then set feedbackType to 'add' for new additions. In config, modify the text of task processes to the user's modified description, and add feedback parameter as true.
}
}

[example]
${example_draw}
```

B.3 CodeGenerationGPT

System Prompt:

```
You are a robot program generator. You will modify the code according to the changes in user's requirements and the original code.
This code is used to implement a customized service for a service robot, using the following robot API:
robot.speak(sentence): Makes the robot say the content in 'sentence'. This function returns a promise, so it can be awaited.
robot.ask(sentence): Makes the robot ask the user and return the user's reply. This function returns a promise with the reply content as its value.
robot.goto(location): Makes the robot move to a location according to a pre-defined location name. This function returns a promise when the robot arrives. The defined locations currently include:
Reception area, Meeting room, Work exhibition area, Leader's office, Employee office area, Creation studio, Gym, Living room and Pantry.
robot.detectHuman(): This function returns a promise which resolves to true when a human is detected. It resolves to false if no human is detected after a 5-second delay.
robot.userRequest(task): The "task" parameter is the keyword used to initiate the robot service. This function returns a promise when the user inputs this task keyword.
The code you generate is for a social service bot in the lab, and when generating the code, you need to consider:
You can only use the functions provided above to control the robot, do not call other APIs.
The code should start with the robot.userRequest() function. The generated code serves as a service of the robot. Users can call this service at any time by using specific task keywords.
Communicate with users in Chinese.
When calling robot.speak() or robot.ask() functions, do not use function methods and expressions that are difficult for the user to understand, such as join.
Keep the code as short as possible.
There is no need to include any content outside of the code, nor is there a requirement to explain the code. If the complete code cannot be output in one response, it can be divided into multiple outputs.
Please generate the answer in the format of Node.js.
original code: ${code}
old user requirements: ${quireOld}
new user requirements: ${quireNew}
Your output of new code:
```

C User Study

C.1 Task Descriptions

(1) **Low-complexity Task L1: Office Patrolling and Employee Notification**
In this task, the robot is assigned to patrol the office after working hours and notify employees who remain in the office to leave promptly. The patrol route covers four specific locations within the office, and the task requires optimizing the route to minimize unnecessary repetition. Participants are required to construct a service that manages the robot's route planning and employee notification process, ensuring efficiency in both patrol and communication.

(2) **High-complexity Task H1: Visitor Guidance at Reception**
This task involves configuring the robot to assist with visitor guidance in a corporate environment. Each day, a significant number of scheduled visitors arrive at the company, and the robot is tasked with guiding them to their respective meeting locations. Employees will provide the robot with visitor information (e.g., names, destinations) during each service call, and the robot must accurately guide visitors accordingly. Participants are required to construct a service that supports the following functionality:
- The ability for employees to input visitor information, including names and destination details.
- A guidance process for leading visitors to the specified locations.
- A contingency plan for handling visitors who are not pre-registered in the system.
- Ensuring the robot successfully escorts visitors to their destinations.

(3) **Low-complexity Task L2 : Employee Guidance and Area Introduction**
In this task, the robot is responsible for guiding employees to two predefined office areas and delivering an introduction about the functionality of each area (with the content designed by the user). The robot must inquire whether the employee understood the explanation and, based on their response, either repeat the explanation or continue guiding them to the next area. Participants are required to design a service that ensures smooth and adaptive interaction during the employee's guided tour.

(4) **High-complexity Task H2:Employee Gathering and Preparation Work**
In this task, the robot is assigned to locate an employee at the request of a manager. The manager inputs the employee's details (including name and current location), and the robot must proceed to the designated office area to ask whether the employee is ready to meet with the manager. If the employee is not ready, the robot must inquire how much additional time is required and relay this information to the manager. Participants are tasked with constructing a service that satisfies the following conditions:
- The ability for the manager to input employee information, including name and location.
- The capability to handle multiple employee searches in an efficient sequence.
- Optimization of the task execution order to ensure timely completion of all assigned tasks.

C.2 Semi-structured Interview questions

- Could you compare the overall experience between System 1 and System 2? What are their respective strengths and weaknesses?
- When using both systems, were there any differences in your thought process or the steps you followed to complete the tasks? Which system met your expectations better?
- How do you assess the role of the graphical interface in System 2 for facilitating communication? What are its advantages and areas for improvement?
- Does the integration of the graphical interface and voice-based multimodal interaction feel seamless? How do you evaluate the balance of information presentation? Do you have any suggestions for improvement?
- To what extent does the system align with your personal preferences and communication style? In what scenarios do you think it would be most practically valuable?
- Which features or functionalities in the system impressed you or exceeded your expectations? Do you have any additional recommendations for features?