Robi Butler: Multimodal Remote Interaction with a Household Robot Assistant

Anxing Xiao, Nuwan Janaka, Tianrun Hu, Anshul Gupta, Kaixin Li, Cunjun Yu, David Hsu

***Abstract*— Imagine a future when we can Zoom-call a robot to manage household chores remotely. This work takes one step in this direction. *Robi Butler* is a new household robot assistant that enables seamless multimodal remote interaction. It allows the human user to monitor its environment from a first-person view, issue voice or text commands, and specify target objects through hand-pointing gestures. At its core, a high-level behavior module, powered by Large Language Models (LLMs), interprets multimodal instructions to generate multistep action plans. Each plan consists of open-vocabulary primitives supported by vision-language models, enabling the robot to process both textual and gestural inputs. Zoom provides a convenient interface to implement remote interactions between the human and the robot. The integration of these components allows Robi Butler to ground remote multimodal instructions in real-world home environments in a zero-shot manner. We evaluated the system on various household tasks, demonstrating its ability to execute complex user commands with multimodal inputs. We also conducted a user study to examine how multimodal interaction influences user experiences in remote human-robot interaction. These results suggest that with the advances in robot foundation models, we are moving closer to the reality of remote household robot assistants.**

I. Introduction

Imagine a future where distance no longer constrains our ability to manage household tasks. Picture a robot assistant capable of remotely interpreting spoken commands and gestures to check your refrigerator or reheat a meal before you get home. Such a robotic system would fundamentally change the way we interact with our homes, bringing a new level of convenience and efficiency to daily life. In this work, we propose *Robi Butler*, a multimodal interaction system that enables seamless communication between remote users and household robots to execute various household tasks. *Robi Butler* allows users to leverage both natural language and gestures to command the robot to perform tasks remotely, see Fig. 1. Remote users can point to the desired object in the MR device and instruct the robot to manipulate it, move toward it, or ask questions about it, just like a real butler.

The core issue behind building such a robot assistant is how to allow the robot to remotely *receive*, *understand*, and *ground* the multimodal instructions into the executable actions in the home environment. To address this, we first design the communication interfaces consisting of a Zoom chat website and a gesture website for hand-pointing, which allows human users to send multimodal instructions using language and pointing remotely. To ground the received multimodal instructions in the home, the robot needs to have the ability to interpret and execute the open multimodal instructions in real-world environments. Inspired by the advanced capabilities of foundation models to achieve open vocabulary mobile manipulation in domestic environments [1]–[4], we aim to incorporate the LLM-based robots with the ability to make use of the language-related gestures. To allow the robot to ground both open language instruction and gesture selection, we first implement a mobile manipulation system that supports open vocabulary action primitives with pointing selection, driven by the recent advances in vision language models (VLMs). Then, we introduce a high-level behavior module powered by large language models (LLMs), which organizes and aligns the received language and gesture instructions to generate the plan.

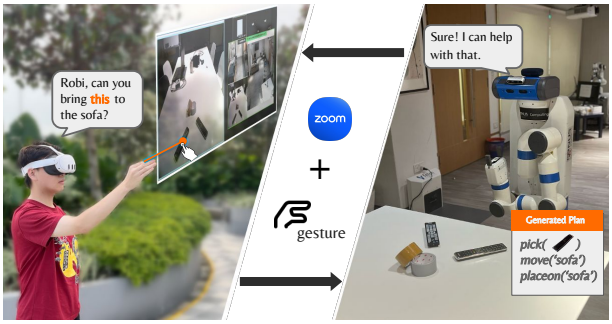


Fig. 1: The *Robi Butler* system enables the user to Zoom-call the butler robot remotely at home and interact with it naturally through both the *language* and *hand gestures*.

Overall, the proposed system, *Robi Butler*, is a multimodal interactive system for robotic home assistants that enables bi-directional remote human-robot interaction based on the real home environment through text, voice, video, and gesture. We evaluated the performance of *Robi Butler* on real-world daily household tasks and studied the benefits of such multimodal interaction in terms of efficiency and user experience in the remote human-robot interaction.

II. Related Work

A. Language and Gestures in Human-Robot Interaction

Effective communication interfaces are essential for Human-Robot Interaction (HRI). Natural language instruction for robots has been widely explored in prior research, employing both traditional methods [5]–[13] and large language models [1], [14]–[20]. However, language can be ambiguous and imprecise. Humans typically use nonverbal interaction, such as pointing, to supplement their verbal instructions [21]. Previous work explores the use of tools such as laser pointers [22] and point-and-click interfaces [23]

All authors are with the School of Computing, National University of Singapore, Singapore. Correspond to anxingx@comp.nus.edu.sg.

Nuwan Janaka, Tianrun Hu, and David Hsu are also with the Smart Systems Institute, National University of Singapore, Singapore.

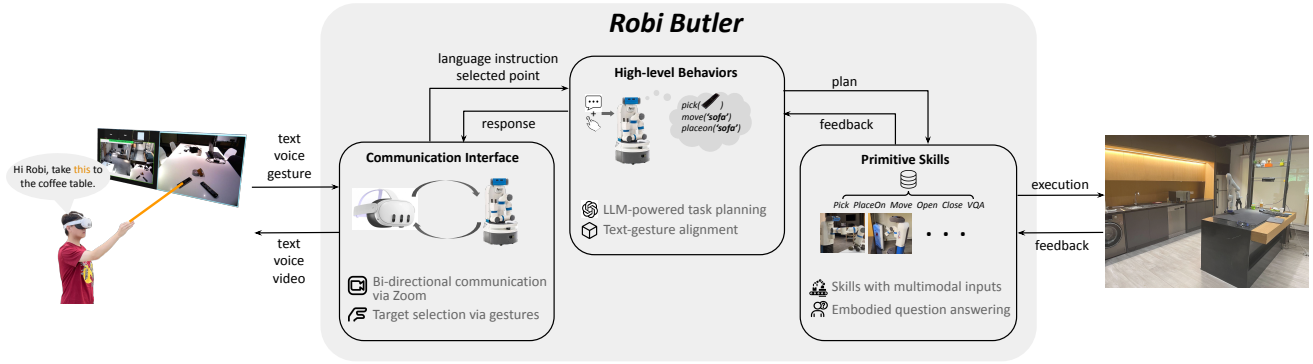


Fig. 2: An overview of *Robi Butler*. The system consists of three components: Communication Interface, High-level Behavior Module, and Primitive Skills. The Communication Interfaces transmit the inputs received from the remote user to the High-level Behavior Module, which composes the Primitive Skill to interact with the environment to fulfill the instructions or answer questions.

to improve instruction delivery and further integrate both speech and relevant gestures together [24]–[27] to specify the command more precisely. However, these systems typically rely on predefined instruction templates or task-specific in-domain model training, which limits generalization to open-ended multi-modal language instruction. Recent work uses LLMs to interpret gestures and commands [28], but only handles short speech inputs and requires the user to be within the third-person camera view. Our system is built on top of a multimodal communication interface to construct a *virtual clickable world* that allows the remote user to select the target by pointing while speaking, and the robot could interpret and execute the multimodal instructions in the home environment with a mobile manipulator.

B. Household Robot Assistant

Intelligent home robots with mobile manipulation capabilities can greatly expand functionality and integrate more seamlessly into daily routines. While past household mobile manipulation systems have been developed both in simulation [29]–[31] and real-world settings [32]–[36], they generally struggle with human-robot interaction due to their reliance on predefined tasks and limited language input. They would require users to select from fixed options or explicitly re-programme the robot. More recent approaches leverage vision-language-based models (VLMs) to enable open-vocabulary mobile manipulation in domestic environments [1]–[4], but they rely solely on language instructions and lack closed-loop human-robot interaction. Another area of research explores treating robot assistants as "physical avatars", which allows remote users to teleoperate the robots using VR controllers [37], haptic devices [38], haptic gloves [39], and hand tracking [40]. However, these approaches can result in a high cognitive workload [41], making them impractical for everyday use. In this paper, we present a human-robot interaction system for remote user to naturally instruct open-vocabulary mobile manipulation with multi-round interaction using both language and gestures.

III. Overview

This work addresses the problem of remote human-robot interaction for household robot assistants. We present a multimodal system, *Robi Butler*, that combines speech commands and gesture inputs, allowing remote users to naturally guide a robotic assistant to perform household tasks.

A. System Overview

The developed *Robi Butler* system is illustrated in Fig. 2. It enables seamless interaction between a user wearing a Mixed Reality (MR) Head-Mounted Display (HMD) and a robot. Users can send text/voice instructions L and gesture selections G to the robot while receiving video streams and text/voice feedback F in return. The robotic system comprises three key components. The communication interfaces C facilitate bidirectional communication, receiving user inputs and transmitting robot feedback. The high-level behavior module H, interprets user instructions L and gesture selections (G) to understand the intent, generating an action sequence $P = \{a_0, a_1, ..., a_N\}$ for the robot to execute, along with a response R to the user. This response can be low-level execution feedback or general information. The primitive skills A, provide core functionality that allows the robot to perceive and interact with the environment. These include basic mobile manipulation and Visual Question Answering (VQA) capabilities: *move(), pick(), placeon(), open(), close()* and *vqa()*. Note that all skills except *open()* and *close()* support both text and pointing queries.

B. Hardware Setup

Our system integrates multiple hardware components to facilitate effective human-robot interaction. The primary user interface is an Oculus Quest 3 MR headset, while the robotic platform consists of a Fetch mobile manipulator [42] with a differential-drive base and a 7-dof arm. Tasks that require heavy computation are distributed between a local workstation powered by an NVIDIA RTX 4090 GPU and a remote cloud server. To enhance user visual feedback, we incorporate two additional cameras that provide third-person views of the robot's operational environment.

IV. System Implementation

The system has a multimodal communication interface, a high-level behavior module, and low-level action modules.

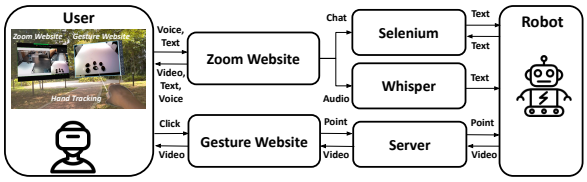


Fig. 3: The framework of communication interfaces.

A. Communication Interfaces

As shown in Fig. 3, the communication interfaces enable multimodal remote interaction between humans and robots, utilizing voice, text, and gestures. These interfaces consist of two main components: a Zoom platform and a gesture selection website. The Zoom platform supports voice, text, and video communication, while the Selenium library on the robot's server extracts specific text elements from the chat box during live sessions. For speech recognition, we employ the Whisper model [43]. For gesture-based interactions, we developed a website using Flask that allows users to select target objects by pointing. The site streams the robot's first-person video frames at 5 Hz, and the selected points are transmitted to the robot server in real-time, enabling immediate planning and execution. Since our design doesn't rely on a high-frequency control loop like in teleoperation, this interface is not sensitive to latency, allowing users to instruct the robot from anywhere in the world.

B. High-level Behavior Module

The high-level behavior module interprets and decomposes user multimodal instructions, comprising language inputs (L) and gesture inputs (G), into executable action sequences $P = \{a_0, a_1, ..., a_N | a_i \in A\}$, along with corresponding responses (R). This module processes both inputs, leveraging an LLM to generate structured responses and action plans. These are then passed to the execution module, which integrates the gesture inputs to ensure precise alignment between user gestures and robot actions, as depicted in Fig. 4.

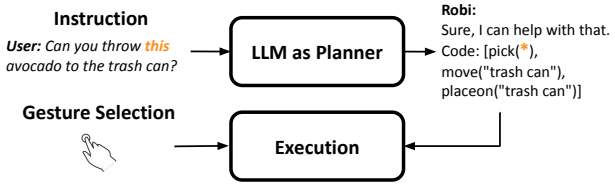
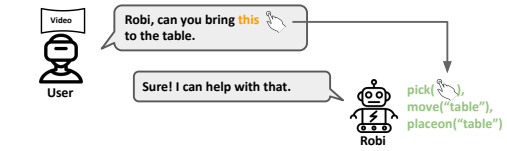
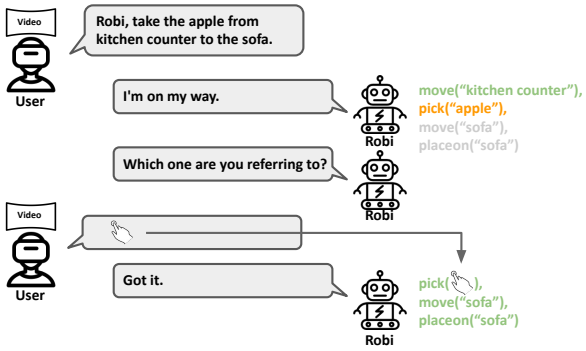


Fig. 4: The framework of high-level behavior module.

The task planner in the high-level behavior module, illustrated in Fig. 4, is powered by an LLM (OpenAI GPT-4o-2024-05-13) prompted to function as a household robot assistant. The prompt defines the robot's role, a list of known locations, primitive skills it can perform, and few-shot examples to demonstrate how these skills should be used. Full prompts for the LLM can be found at `https://robibutler.github.io`. To align instructions with gesture selections, we implement a rule: when inputs contain the keywords *"this"* or *"here"*, the planner generates *"*"* as an action parameter to resolve ambiguities, particularly



(a) The user directly gives multimodal instructions to the robot.



(b) The user gives language and gesture instructions separately.

Fig. 5: Human-Robot Remote interactions via language and gestures.

demonstrative pronouns [21]. For example, the instruction "*Robi, please pick this and put it on the plate*" results in the plan *[pick(*), placeon("plate")]*. During execution, the *"*"* is resolved using the latest gesture selection. We store the five most recent gesture selections and match them with the *"*"* parameters during execution. Additionally, the system supports gesture-only input for disambiguation when the detection model identifies multiple objects in response to a single query. In such cases, the robot prompts, "*Which one are you referring to?*", pausing for the user to select the target object. Fig. 5 illustrates the alignment between gesture selections and the LLM-generated plan.

C. Primitive Skills

1) Manipulation: For the robot to physically interact with the environment, it is equipped with manipulation skills such as picking/placing items, and opening/closing appliances.

Pick and Place Policy Fig. 6 illustrates the modular framework for the pick policy. The *pick()* function accepts either a text query *pick(text)* or a pointing query *pick(point)*. We employ the pre-trained open-vocabulary detection model OWLv2 [44] and the Segment Anything model [45] to generate the target object mask. This mask is then combined with the pre-trained grasping model Contact-GraspNet [46] to determine grasping poses. Grasping poses are filtered based on orientation and ranked by the score.

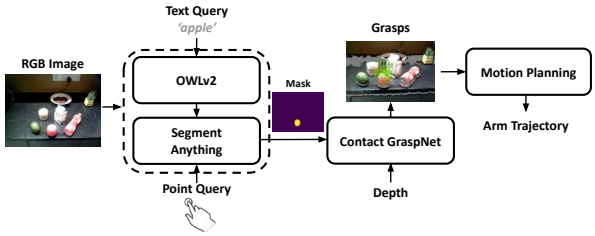


Fig. 6: The open-vocabulary pick pipeline.

Given the highest-scoring grasp, a straightforward pre-grasp and grasp strategy is applied, with arm trajectories generated using the motion planning tools from MoveIt [47].

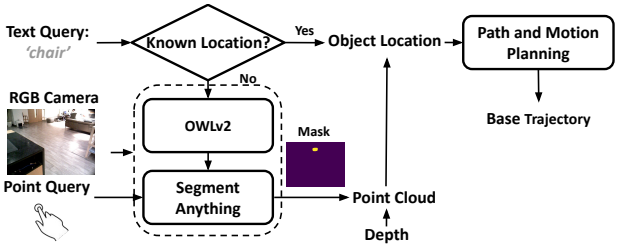


Fig. 7: The navigation pipeline.

The place policy, similar to the pick policy, utilizes the same perception modules and can handle both text and pointing queries. After obtaining the segmented point clouds, the center of the place position is calculated in the X-Y plane, while the height is determined by adding 0.2 meters to the highest point of the segmented point clouds. For larger fixed objects or locations, such as tables, counters, and trash cans, a pre-defined location is used to simplify the setting.

Open and Close Policy Similar to Chen et al. [48], the open/close policies rely on imitation learning to handle complex actions such as opening a refrigerator and a cabinet. We collected an average of 50 demonstrations per action using teleoperation. These demonstrations were used to train the policies using Action Chunking with Transformers (ACT) [49]. Demonstrations of the learned skills can be viewed at `https://youtu.be/ajfPVjjlBcI`.

2) Navigation: As shown in Fig. 7, our system integrates both predefined navigation places and open-world navigation to locate and move to the target object. First, we create an occupancy map using Gmapping [50] and define the navigation waypoint for the known locations in the map manually. In addition to predefined locations, the system also supports navigating to non-predefined locations via voice/text and gesture/point queries, similar to the perception pipeline in the pick policy (Sec IV-C.1). We utilize the off-the-shelf path and motion planning algorithm provided by the ROS Navigation Stack to generate the path and motion trajectory.

3) Visual Question Answering: Our system can also answer the user's open-ended questions about the status of the environment. Specifically, combined the actions *vqa()*, our system supports:

Question answering via mobile manipulation. To answer the question "*Do we have any beer left in the fridge?*", the robot should first navigate to the fridge, open it, and then query the VLM model. Our solution treats the VQA as a single action and uses the reasoning capabilities of LLMs to determine the necessary high-level steps before performing VQA. Given the question, the high-level behavior module decomposes the question into a series of actions to be executed before querying GPT-4o for the final answer.

Question answering via point referring. Text-only input may lack precision, so we enable the robot to answer verbal/textual questions combined with pointing selection (*vqa(text, pointing)*), as shown in Fig. 8. We use a simple visual prompting method for GPT-4o to answer specific questions by annotating the image with a mark.

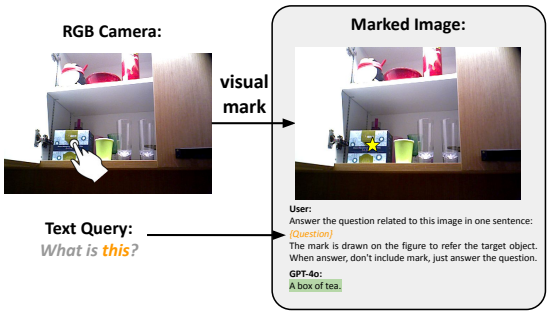


Fig. 8: Example of the question answering via point referring.

V. Experiments and Results

To understand the usage and impact of multimodal remote interaction in remote HRI, we evaluate the performance of the *Robi Butler* guided by the following research questions:

RQ1: *How effectively and robustly does the Robi Butler enable remote users to complete household tasks?*

RQ2: *How do the user interaction modalities (voice, gestures) affect the performance and usability of Robi Butler?*

A. Experiment I: Robi Butler Performance Evaluation

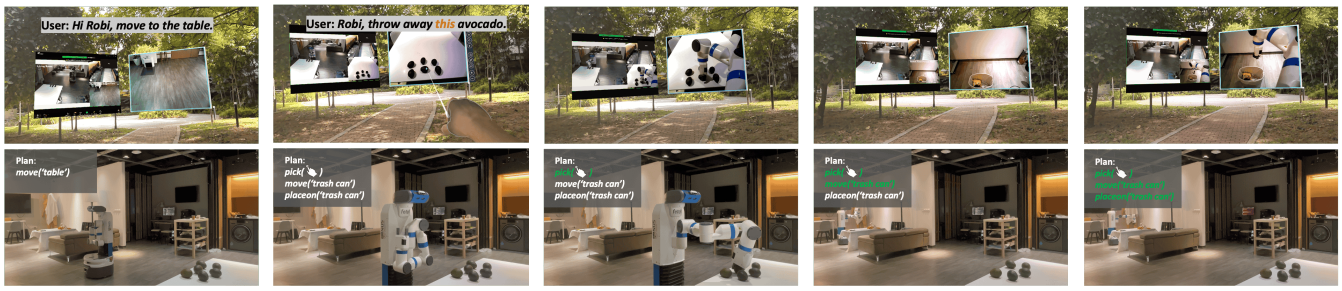
In this experiment, we evaluate the *Robi Butler* system on a set of daily household tasks to understand its effectiveness and answer ***RQ1***.

1) Experimental Design: The tasks were designed based on the American Time Use Survey [51]. These tasks fall under the common daily household activities, including *Food and drink preparation* (0.50 hr/day), *Interior cleaning* (0.35 hr/day), *Household & personal organization and planning* (0.11 hr/day), and *Medical and care services* (0.06 hr/day). The ten selected tasks (**T1-T10**) required the robot to interpret remote users' language and pointing gestures, then perform the corresponding actions (e.g., rearranging objects, answering questions). The object that requires disambiguation is highlighted in **bold**. To focus on remote human-robot interaction, we use objects compatible with the hardware, excluding those that are hard to manipulate, e.g., deformable or transparent.

T1. *Throw **avocado** into the trash can.*
T2. *Check the beer inside the fridge.*
T3. *Check medicine on the coffee table and bring **one** to the sofa.*
T4. *Describe the **object** in the cabinet.*
T5. *Bring the **drink** to the coffee table.*
T6. *Move the **cup** to the kitchen counter.*
T7. *Fetch the **remote** and place it on the sofa.*
T8. *Navigate to a **chair** and check if it's clean.*
T9. *Check if the laptop is open.*
T10. *Bring the **tool** to the table.*

To evaluate the effectiveness of *Robi Butler*, the following metrics were used: ***Task Success Rate (Task SR)***: defined as the percentage of tasks completed. ***Planning Success Rate (Planning SR)***: defined as the percentage of tasks completed when execution errors are ignored. ***Task Completion Time***, measuring the average time required to complete each task. ***Average Interactions***: calculating the average number of voice and gesture interactions required per task. A task is considered successful/completed if the goal is achieved or if correct answers are provided to the remote user within 5 minutes. After obtaining informed consent, the expert user evaluated *Robi Butler* on 10 tasks using free language in a fixed order, each repeated three times.

T1: *Throw* ***avocado*** *into the trash can.*



(a.1) *(a.2)* *(a.3)* *(a.4)* *(a.5)*

T4: *Describe the* ***object*** *in the cabinet.*

(b.1) *(b.2)* *(b.3)* *(b.4)* *(b.5)*

Fig. 9: Snapshots of completing tasks T1 and T4. *(a.1)*: User asks Robi to go to the table. *(a.2)*: User asks Robi to throw away the avocado. *(a.3)*: Robi picks up the avocado. *(a.4)*: Robi brings the avocado to the trash can. *(a.5)*: Robi throws away the avocado. *(b.1)*: User asks Robi to open the cabinet. *(b.2)*: Robi reaches the cabinet. *(b.3)*: Robi opens the cabinet. *(b.4)*: User asks Robi to identify an object. *(b.5)*: Robi identifies it as "A box of tea."

TABLE I: Real-world Experiments Result for Experiment I. Tasks that require the user's selection are indicated using *. Interactions include both Voice (V) and Gestures (G).

Task	Task SR	Planning SR	Time	Interactions (V + G)
T1*	3/3	3/3	119.7s	3 (2+1)
T2	3/3	3/3	153.0s	1 (1+0)
T3*	3/3	3/3	128.3s	3 (1+0)
T4*	2/3	3/3	147.0s	3 (2+1)
T5*	3/3	3/3	86.0s	2 (2+1)
T6*	3/3	3/3	95.3s	2 (1+1)
T7*	3/3	3/3	117.0s	3 (2+1)
T8*	3/3	3/3	64.0s	2 (1+1)
T9	3/3	3/3	57.3s	2 (2+0)
T10*	3/3	3/3	82.3s	2 (1+1)
Mean	96.7%	100%	105.0s	2.3 (1.5 + 0.8)

2) Analysis and Results: Table I presents the task performance results. Overall, *Robi Butler* achieved a high average task success rate of 96.7%, reflecting its strong ability to perform a variety of household tasks in real-world environments. However, the task success rate lags slightly behind the perfect planning success rate of 100%, indicating challenges related to low-level action execution rather than planning processes. For instance, in task T4, an error occurred when the system misidentified a green tea box as a tissue bag. On average, the system completed tasks in approximately 105 seconds, demonstrating its efficiency in performing household tasks in a complex environment. The system required an average of 2.3 interactions per task, with 1.5 voice commands and 0.8 gesture inputs. This low number of interactions demonstrates the system's efficiency in human-robot communication, requiring minimal user input to effectively guide the robot. While the overall performance of the system is generally satisfactory, answering ***RQ1***, further improvements in low-level action execution could help increase the overall performance and efficiency. Fig. 9 shows the process of two example tasks. The system has up to 0.2s delay in Singapore and stays within 0.5s for long distances like Singapore to Abu Dhabi. More videos of the tasks are available at `https://robibutler.github.io`.

B. Experiment II: The Effect of Modality on User Experience

To investigate user experience, the impact of multimodal communication, and challenges, we conducted this experiment with novice users to address ***RQ2***.

1) Experimental Design: We recruited twelve volunteers (P1–P12; 7 males, 5 females) from the university community with IRB approval. None had prior experience with AR/MR smart glasses. We compared the performance of *Robi Butler* with two baseline systems by removing user interaction modalities, similar to an ablation study, resulting in three systems: *Gesture-only*, *Voice-only*, and *Robi Butler* (*Gesture+Voice*). In the *Gesture-only* system, buttons were added for participants to select the action to be executed. For the *Voice-only* system, we adapted the interactive visual grounding model from [52]. Videos of two baseline systems are available on the website. Three representative tasks, *T1* (object rearrangement), *T2* (monitoring), and *T3* (object rearrangement + monitoring), were selected from the previous experiment (Sec V-A.1). As shown in Fig. 11, these tasks engaged the main areas of the home. Participants were instructed to use their preferred verbal expressions.

The study used a within-subject design with three system conditions as the independent variable, counterbalanced via a Latin Square, to minimize ordering effects. Tasks increased in difficulty and were presented in a fixed order. Participants completed all three tasks with each system (nine tasks total)

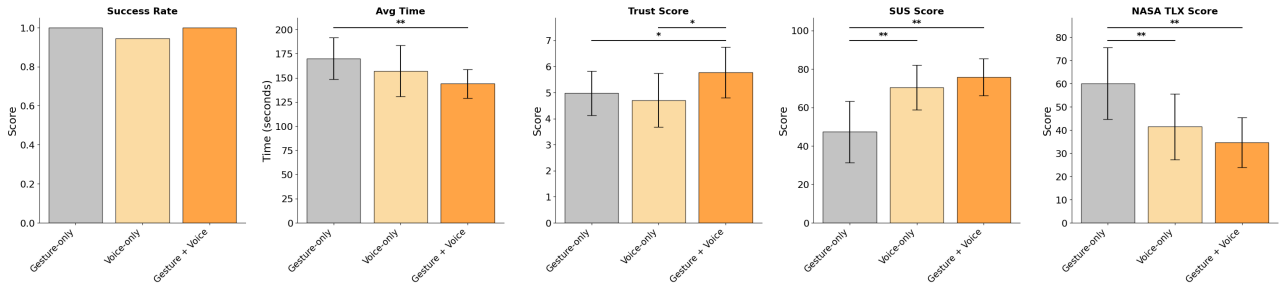


Fig. 10: Measures related to efficiency and user experiences of different systems with 12 participants. For Success Rate, Trust, and SUS, the higher, the better; for Avg Time and NASA TLX, the lower, the better. For statistical significance, one asterisk (*) is $p < 0.05$; two asterisks (**) is $p < 0.01$.

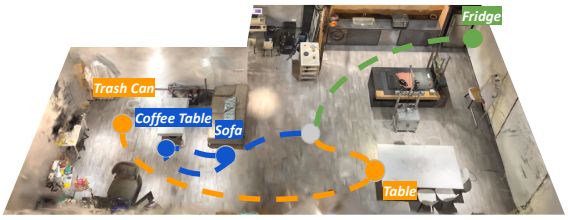


Fig. 11: Visualization of the experimental environment: Orange, green, and blue trajectories represent *T1*, *T2*, and *T3*, respectively.

and filled out a questionnaire after each system to assess their subjective experience. In addition to the *Task SR* and *Task Completion Time* measure from V-A.1, the following additional measures were used to assess user experience: ***NASA-TLX*** [53], assessing the perceived workload experienced by participants with each system. ***System Usability Scale (SUS)*** [54], evaluating perceived system usability. ***Trust*** [55], measuring the participants' trust. We used the reliable subscale under Capacity Trust.

2) Analysis and Results: Fig. 10 shows the task performance of the three systems. A one-way repeated measures ANOVA was conducted to analyze the quantitative data after confirming normality assumptions. Both the *Gesture-only* and *Gesture+Voice* (i.e., *Robi Butler*) systems achieved a perfect task success rate of 100%, while the voice system had a slightly lower, though non-significant, success rate of 94.4%. This difference was attributed to errors in target referencing with voice commands only. For example, the voice recognition system misinterpreted the word 'right' as 'red', leading to the grounding error. Additionally, *Robi Butler* ($M = 143.8$, $SD = 14.8$) had a significantly lower task completion time than the *Gesture-only* system ($M = 170.00$, $SD = 21.4$) ($p < 0.05$), but was not significantly lower than the *Voice-only* system ($M = 157.1$, $SD = 26.6$). The reduced task completion time for voice-supported systems primarily resulted from the ability to use voice commands to express combined queries, whereas with the *Gesture-only* system, participants had to perform multiple manual clicks, increasing task completion time.

Regarding the trust, the *Robi Butler* ($M = 5.77$, $SD = 0.97$) was perceived as significantly more trustworthy compared to both the *Gesture-only* system ($M = 4.98$, $SD = 0.85$, $p < 0.05$) and the *Voice-only* system ($M = 4.71$, $SD = 1.03$, $p < 0.05$). This suggests that combining gestures with voice enhances the system's reliability and consistency, outperforming single-modality systems. P2 reasoned that "*I trusted the gesture plus voice system the most because I found it easier to avoid making mistakes with it. For language only, sometimes it may misunderstand me. For gestures, I have to do the interaction multiple times.*" For the SUS, participants gave the *Gesture-only* the lowest usability score ($M = 47.29$, $SD = 15.90$), which significantly lower than both *Voice-only* ($M = 70.42$, $SD = 11.62$, $p < 0.01$) and *Robi Butler* ($M = 75.83$, $SD = 9.61$, $p < 0.01$). This also indicates that *Robi Butler* achieved 'Good' usability (i.e., $SUS > 75$ [56]) compared to the other systems.

Overall, the *Robi Butler* achieves the best performance, the highest usability, and the minimum perceived cognitive load among the baselines, answering ***RQ2***. This was primarily due to the complementary nature of voice and gestures—voice enabled natural queries, while gestures provided precise spatial annotations. Although multimodal interaction generally outperformed unimodal interaction, P10 expressed a negative sentiment, stating, "*Using both voice and gesture is [sometimes] hard, as I need to switch between two modalities. I prefer voice-only as I don't need to move my arm physically.*" Incorporating eye gaze tracking could reduce physical workload by minimizing hand interactions.

VI. Conclusion

This work introduces an interactive robotic assistant for household tasks using multimodal interactions with remote users. We outline three core components of the robot butler system and demonstrate its effectiveness in assistive question-answering and object rearrangement tasks. Experiments show *Robi Butler* grounds remote multimodal instructions with a high task success rate, reasonable time, and minimal interactions. Follow-up tests confirm that combining voice and gestures enhances usability and trust, and reduces cognitive load compared to unimodal systems. In future work, we aim to enhance *Robi Butler* with more adaptive skills, personalized interactions, and tactile feedback [57].

Acknowledgment

This research is supported in part by the National Research Foundation (NRF), Singapore and DSO National Laboratories under the AI Singapore Program (AISG Award No: AISG2-RP-2020-016) and by the Agency for Science, Technology & Research (A*STAR), Singapore under its National Robotics Program (No. M23NBK0091).

REFERENCES

[1] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*, 2022.

[2] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized robot assistance with large language models," *Autonomous Robots*, vol. 47, no. 8, pp. 1087–1102, 2023.

[3] S. Yenamandra, A. Ramachandran, K. Yadav, A. S. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. Clegg, J. M. Turner *et al.*, "Homerobot: Open-vocabulary mobile manipulation," in *Conference on Robot Learning*, 2023.

[4] P. Liu, Y. Orru, J. Vakil, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Demonstrating ok-robot: What really matters in integrating open-knowledge models for robotics," in *Robotics: Science and Systems (RSS)*, 2024.

[5] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 259–266.

[6] R. Cantrell, M. Scheutz, P. Schermerhorn, and X. Wu, "Robust spoken instruction understanding for hri," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 275–282.

[7] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011.

[8] S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell, "Open-vocabulary object retrieval." in *Robotics: science and systems*, vol. 2, no. 5, 2014, p. 6.

[9] D. K. Misra, J. Sung, K. Lee, and A. Saxena, "Tell me dave: Context-sensitive grounding of natural language to manipulation instructions," *The International Journal of Robotics Research*, 2016.

[10] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, "Interactively picking real-world objects with unconstrained spoken language instructions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[11] C. Paxton, Y. Bisk, J. Thomason, A. Byravan, and D. Foxl, "Prospection: Interpretable plans from language by predicting the future," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[12] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, 2020.

[13] H. Zhang, Y. Lu, C. Yu, D. Hsu, X. La, and N. Zheng, "Invigorate: Interactive visual grounding and grasping in clutter," in *Robotics: Science and Systems (RSS)*, 2021.

[14] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on robot learning*, 2021.

[15] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *International Conference on Machine Learning*, 2022.

[16] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," in *Conference on Robot Learning*, 2022.

[17] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[18] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: an embodied multimodal language model," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 8469–8488.

[19] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," in *Conference on Robot Learning*, 2023.

[20] L. X. Shi, Z. Hu, T. Z. Zhao, A. Sharma, K. Pertsch, J. Luo, S. Levine, and C. Finn, "Yell at your robot: Improving on-the-fly from language corrections," in *Robotics: Science and Systems (RSS)*, 2024.

[21] H. Diessel and K. R. Coventry, "Demonstratives in spatial language and social interaction: An interdisciplinary review," *Frontiers in Psychology*, vol. 11, 2020.

[22] H. Nguyen, A. Jain, C. Anderson, and C. C. Kemp, "A clickable world: Behavior selection through pointing and context for mobile manipulation," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.

[23] D. Kent, C. Saldanha, and S. Chernova, "A comparison of remote robot teleoperation interfaces for general object manipulation," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017.

[24] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox, "Learning from unscripted deictic gesture and language for human-robot interactions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014.

[25] D. Whitney, M. Eldon, J. Oberlin, and S. Tellex, "Interpreting multimodal referring expressions in real time," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[26] Y. Chen, Q. Li, D. Kong, Y. L. Kei, S.-C. Zhu, T. Gao, Y. Zhu, and S. Huang, "Yourefit: Embodied reference understanding with language and gesture," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[27] D. Weerakoon, V. Subbaraju, T. Tran, and A. Misra, "Cosm2ic: optimizing real-time multi-modal instruction comprehension," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 697–10 704, 2022.

[28] L.-H. Lin, Y. Cui, Y. Hao, F. Xia, and D. Sadigh, "Gesture-informed robot assistance via foundation models," in *7th Annual Conference on Robot Learning*, 2023.

[29] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets *et al.*, "Habitat 2.0: Training home assistants to rearrange their habitat," *Advances in Neural Information Processing Systems*, 2021.

[30] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. E. Vainio, C. Gokmen, G. Dharan, T. Jain *et al.*, "igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," in *Conference on Robot Learning*, 2021.

[31] S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, K. Liu *et al.*, "Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments," in *Conference on Robot Learning*, 2021.

[32] O. Khatib, "Mobile manipulation: The robotic assistant," *Robotics and Autonomous Systems*, 1999.

[33] U. Reiser, C. Connette, J. Fischer, J. Kubacki, A. Bubeck, F. Weisshardt, T. Jacobs, C. Parlitz, M. Hägele, and A. Verl, "Care-o-bot® 3-creating a product vision for service robot applications by integrating design and technology," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

[34] M. Ciocarlie, K. Hsiao, A. Leeper, and D. Gossow, "Mobile manipulation through an assistive home robot," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[35] G. Kazhoyan, S. Stelter, F. K. Kenfack, S. Koralewski, and M. Beetz, "The robot household marathon experiment," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[36] M. Bajracharya, J. Borders, R. Cheng, D. Helmick, L. Kaul, D. Kruse, J. Leichty, J. Ma, C. Matl, F. Michel *et al.*, "Demonstrating mobile manipulation in the wild: A metrics-driven approach," in *Robotics: Science and Systems (RSS)*, 2023.

[37] F. De Pace, G. Gorjup, H. Bai, A. Sanna, M. Liarokapis, and M. Billinghurst, "Leveraging enhanced virtual reality methods and environments for efficient, intuitive, and immersive teleoperation of robots," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 967–12 973.

[38] K. A. Wyrobek, E. H. Berger, H. M. Van der Loos, and J. K. Salisbury, "Towards a personal robotics development platform: Rationale and design of an intrinsically safe personal robot," in *2008 IEEE International Conference on Robotics and Automation*. IEEE, 2008, pp. 2165–2170.

[39] S. Dafarra, U. Pattacini, G. Romualdi, L. Rapetti, R. Grieco, K. Darvish, G. Milani, E. Valli, I. Sorrentino, P. M. Viceconte *et al.*, "icub3 avatar system: Enabling remote fully immersive embodiment of humanoid robots," *Science Robotics*, vol. 9, no. 86, p. eadh3834, 2024.

[40] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," in *Conference on Robot Learning*, 2024.

[41] R. Hetrick, N. Amerson, B. Kim, E. Rosen, E. J. de Visser, and E. Phillips, "Comparing virtual reality interfaces for the teleoperation of robots," in *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 2020, pp. 1–7.

[42] M. Wise, M. Ferguson, D. King, E. Diehr, and D. Dymesich, "Fetch and freight: Standard platforms for service robot applications," in *Workshop on autonomous mobile service robots*, 2016, pp. 1–6.

[43] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, 2023.

[44] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[45] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[46] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[47] S. Chitta, I. Sucan, and S. Cousins, "Moveit![ros topics]," *IEEE Robotics & Automation Magazine*, 2012.

[48] S. Chen, A. Xiao, and D. Hsu, "Llm-state: Open world state representation for long-horizon task planning with large language model," *arXiv preprint arXiv:2311.17406*, 2023.

[49] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023.

[50] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with rao-blackwellized particle filters," *IEEE transactions on Robotics*, 2007.

[51] U.S. Bureau of Labor Statistics, "American time use survey," 2022.

[52] J. Xu, H. Zhang, Q. Si, Y. Li, X. Lan, and T. Kong, "Towards unified interactive visual grounding in the wild," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 3288–3295.

[53] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, 2006.

[54] J. Brooke, "SUS - A quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, 1996.

[55] D. Ullman and B. F. Malle, "Mdmt: multi-dimensional measure of trust," 2019.

[56] A. Bangor, P. T. Kortum, and J. T. Miller, "An Empirical Evaluation of the System Usability Scale," *International Journal of Human–Computer Interaction*, vol. 24, Jul. 2008.

[57] S. Yu, K. Lin, A. Xiao, J. Duan, and H. Soh, "Octopi: Object property reasoning with large tactile-language models," in *Robotics: Science and Systems (RSS)*, 2024.