Robi Butler: Multimodal Remote Interaction with Household Robotic Assistants

Anxing Xiao, Anshul Gupta, Yuhong Deng, Kaixin Li, David Hsu

***Abstract*— In this paper, we introduce Robi Butler, a novel household robotic system that enables multimodal interaction with the user. Leveraging advanced communication interfaces, Robi Butler enables users to monitor the robot's status, give text/voice instruction, and select target objects with hand pointing. At the core of our robotic system are the high-level behavior module powered by Large Language Models (LLMs) that interpret received multimodal instructions to generate plans, and open-vocabulary primitives supported by the Vision-Language Models (VLMs) for executing the planned actions with text and pointing queries. The integration of above components allows Robi Butler to ground remote multimodal instruction in the real-world home environment in a zero-shot manner. We demonstrate the efficacy and efficiency of this system with a variety of daily household tasks involving remote users, such as question answering via interactive mobile manipulation, and object disambiguation for manipulation through gesture. Link: https://robibutler.github.io/**

I. Introduction

A robotic assistant capable of assisting remote users with household tasks could greatly improve the convenience and efficiency of our daily lives. In this work, we aim to develop a multimodal remote interactive system for household robot assistants to enable bidirectional remote human-robot communication and interaction. Imagine you are out and want to *check the ingredients in the refrigerator and prepare the heated food*, the intelligent robots should have the ability to *receive*, *interpret*, and *execute* the instructions given by your natural expressions such as language and gesture.

There are several issues behind building such a robot butler. The first is human-robot communication: allowing remote users to give instructions using natural expressions and receive feedback from the robot. We humans usually use both language and gestures to express our needs and preferences. Relying only on voice, text, and video streaming limits the instructions users can send, resulting in a less natural experience. To address this issue, we designed a communication interface consisting of a zoom chat and a hand pointing website that allows human users to send multimodal instructions using language and pointing. Moreover, grounding the received multimodal instructions is also a challenge, the robot needs to have the ability to interpret and execute the open multimodal instructions in the real-world environments. While some recent work has exploited the advanced capabilities of foundation models to achieve open vocabulary mobile manipulation in domestic environments [1]–[4], the action executed only supports pure language instructions as parameters without additional gesture modality. Previous work focusing on nonverbal interaction typically ignores the need to interpret language-related gestures [5]–[7], rely on a hand-designed closed set of instructions and in-domain training [8]–[10], or use short fixed language and limit pointing selection in third-person camera [11]. To allow the robot to ground both open language instruction and open pointing selection, we first implement a mobile manipulation system that supports open vocabulary action primitives with pointing selection in real-world household environments, driven by the recent advances in vision language models (VLMs). Then, we introduce a high-level behavior manager, powered by large language models (LLMs), which organizes and aligns the received speech and gesture instructions with the human-in-the-loop to generate compositions of action primitives to solve the task.

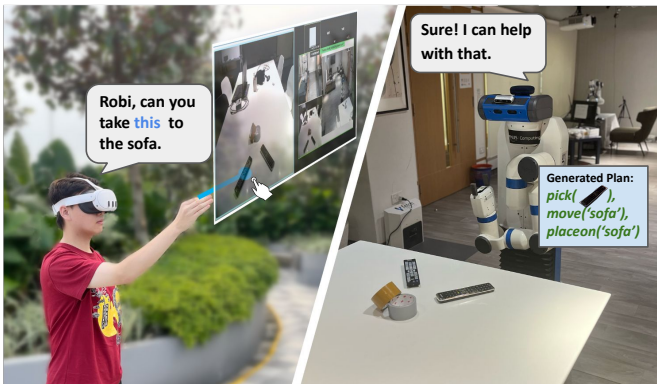

Fig. 1: The illustration of the proposed system. Our system allows remote human users to efficiently and naturally select the target and instruct the robot to perform tasks using mixed language and gestures.

We call the integrated system **Robi Butler**. It is a multimodal remote interactive system for robotic home assistants with mobile manipulators that enables bi-directional remote human-robot interaction grounded on the real home environment through text, voice, video and gesture.

A. Related Work

Language and Gesture in Human-Robot Interaction Human-Robot Interaction (HRI) is strongly influenced by the communication interface. Language is the most powerful natural interface for human communication. Instructing robots with natural language has been widely explored in previous work with traditional method [12]–[17] and Foundation Model powered method [1], [18]–[23]. However, language can be ambiguous and inaccurate. Humans typically use nonverbal interaction, such as pointing, to support their

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

David Hsu is also with the Smart Systems Institute, National University of Singapore, Singapore.

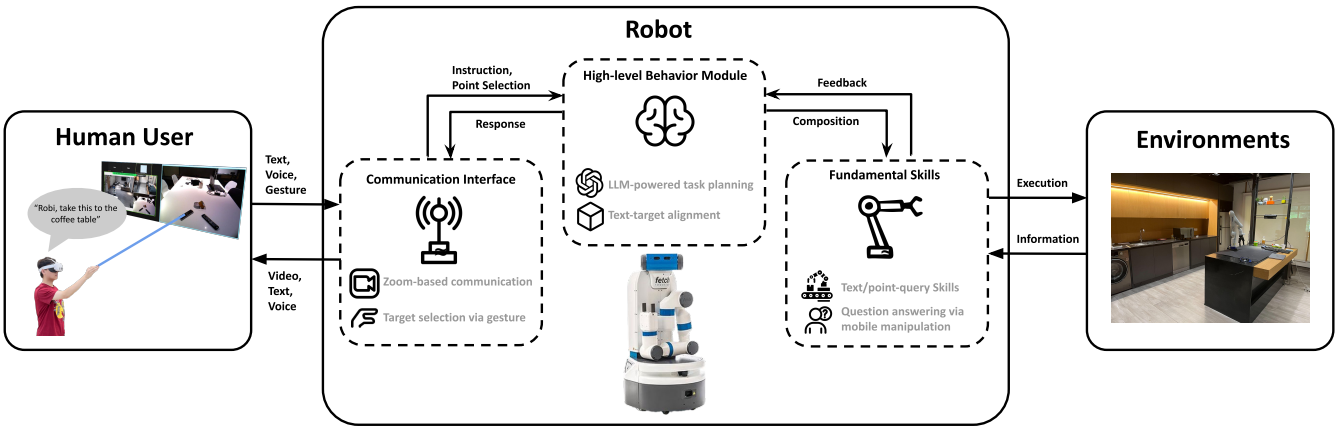

Fig. 2: The conceptual framework of the proposed system, as discussed in II-A. The robot system consists of three components: Communication Interfaces, High-level Behavior Manager, and Fundamental Skills. The Communication Interfaces transmit the inputs received from the remote user to the High-level Behavior Module, which composes the Fundamental Skill to interact with the environment to fulfill the instructions or answer questions.

references. [5] introduced the concept of the *clickable world* which allow users to use the laser pointer to trigger robots' actions. The remote point-and-click interface for grasping has also been explored in [6]. These works focus solely on non-verbal interaction itself, ignoring the need to interpret speech-related gestures [7]. Prior works have investigated how to interpret language and speech-related gestures together [8]–[10]. However, these methods require hand-designed word sets or in-domain training, thus fail to deal with open language instructions. While recent work leverages the power of LLM to interpret gesture and speech instructions [11], the speech instructions are short and rarely varied, and the interface relies on the third-person camera, limiting the remote users to specify the target. 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 the received open multimodal instructions and generate and execute the actions sequence in the real world.

Household Robot Assistant The pursuit of home robot assistants has been a long-standing dream within the robotics community and has evolved significantly over the past few decades. Intelligent home robot assistants with mobile manipulation capabilities would enable a wider range of functionalities remotely and deeper integration into daily routines. Household mobile manipulation systems have been explored in the past both in simulation platforms [24]–[26] and real-world systems [27]–[31], these classical systems have poor generalization in terms of human-robot interaction because they do not incorporate open language, one can only specify one of a limited number of goals or options using fixed language or modifying code. More recent work has exploited the advanced capabilities of vision and language base models to achieve open vocabulary mobile manipulation in domestic environments [1]–[4]. However, the input is pure language instruction without additional modality, and there is no closed-loop interaction between the human and the robot. In this paper, we explored to method and system to build an open-vocabulary mobile manipulation with the support of multi-round language and gesture interaction.

II. Methods and Implementation

In this work, we address the problem of remote human-robot interaction for household robotic assistants. Specifically, our goal is to build a multimodal interactive household robotic system that allows human users to communicate and interact with the robot with language instructions and gesture selections to perform household tasks remotely and naturally.

A. Overall System

The proposed system developed in this paper is illustrated in Fig. 2. When wearing the AR devices, the user can send text/voice instructions L and pointing selections G to the robot and receive video stream, text/voice feedback F from it. For the robotic system, there are three crucial components: communication interfaces C, high-level behavior module H, and fundamental skills A. The communication interfaces enable bidirectional communication between the user and the robot, allowing receiving the text, voice commands, and pointing selections from the remote user device and sending the video stream, text, and voice feedback back. Given the received language instructions L and pointing selections G, the high-level behavior module interprets and corresponds the language instructions L and pointing selections G to understand the user's intent, and then generates the action sequence $P = \{a_0, a_1, ..., a_N\}, a_t \in A$ for the robot to interact with the environment, together with the response R to the user. The response can be low-level execution feedback or a general response to the user.

The fundamental capabilities A include the core functionalities and skills that the robot needs to perceive and interact with the environment, consisting of basic mobile manipulation and question-answering skills, including *move(), pick(), placeon(), open(), vqa()*. Noticed that our skills support both open vocabulary and pointing queries except action *open()*.

B. Communication Interfaces

The communication interfaces are designed to enable multimodal remote interaction between humans and robots. These interfaces support a variety of bi-directional communication channels, including voice, text, and gesture-based interactions. The communication interfaces include the Zoom

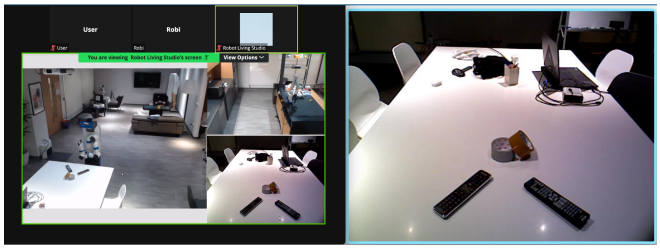

Fig. 3: The visualization of communication interfaces.

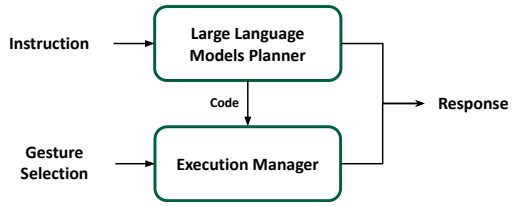

Fig. 4: The frameork of high-level behaviour module

website and a gesture selection website (Fig. 3). The Zoom website handles voice, text and video communication. On the robot server side, we use the Selenium library to extract specific text communication elements from the Zoom website chat box during live sessions. The Zoom platform's live transcription feature is used to convert the user's voice commands into text. To support the pointing selection of target objects, we developed a gesture website that allows users to make point selections that are then sent back to the robot server. The website server is developed using Flask, and video frames are sent to the user's web browser at a frequency of 5 Hz. The results of the point selections are immediately sent to the robot server.

C. High-level Behavior Module

The primary purpose of the high-level behavior module is to composes the Fundamental Skill that interacts with the environment to fulfil the instructions and answer the questions given by the user. As illustrated in Fig. 4, this module takes the received instruction and gesture selection as input, the response is send back to the user through our communication interface before or during the execution process of actions. The module uses large language model (LLM) to reason and formulate the response as well as a structured plan through code. The plan generated by the LLM will be send to the execution manager, which is responsible for the alignment of gesture with language.

The task planner is built around a large language model with a prompt instructing it to behave as a household robotic assistant. We define the robot's role, known location list, a list of fundamental skills that the robot can perform and few-shot example to demonstrate how to use these skills. We also write the rule to help the alignment of instruction and gesture selection. When the language input involves the keyword *'this'*, the planner will generate *'*'* as the action parameters. For example, if the instruction is *"Robi, please pick this and put it on the plate"*, the generated plan will be *"pick(*), placeon('plate')"*. In the execution manager, the *'*'* will be aligned with the gesture selection. We store the latest five gesture selections and match all the '*' in the plan with the gesture selections. The action feedback will also be sent

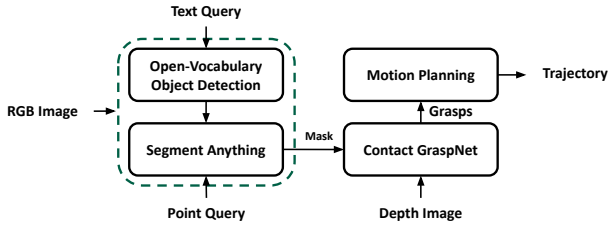

Fig. 5: The open-vocabulary pick pipeline

to the user for further information. We also deploy a simple process to monitor the feedback from the execution. When the detection system returns two objects, the robot will send *"Which one are you refering to?"* and wait for the user to choose the target object. Full prompts for the LLM can be found at the website. We choose GPT-4 [32] from OpenAI as the language model used in our system. We maintain a history chat message history that includes the feedback from the execution manager.

D. Fundamental Skills

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

Open-Vocabulary Pick Policy The modular framework for pick policy is visualized in Fig. 5. We use a pre-trained open-vocabulary detection [33] and segmentation model [34] to generate the target object mask and combine it with a pre-trained grasping model Contact-GraspNet [35] to get the grasping pose. Once the grasping pose is obtained, we can use the motion planning tools [36] to generate a trajectory for the robot arm to perform the grasping.

Open-Vocabulary Place Policy Similar to the pick policy, the place policy also uses the open vocabulary object segmentation. After obtaining the segmented point clouds, we calculate the center of the point clouds in the X-Y plane and the height is calculated by adding 0.2 meters to the highest point of the segmented point clouds. For the larger fixed receptacle and location such as table, counter, and trash can, we use a fixed place location to simply the setting.

Learning-based Open Policy Similar to [37], we used imitation learning [38] to enable complex actions such as opening a microwave and fridge. A human demonstrator used the VR controller to teleoperate the pose and state of the robot's gripper, and the joint angles were computed by solving inverse kinematics (IK). An average of 30 trajectory demonstrations per primitive action are collected with the real robot.

2) Navigation: The navigation within our system, integrates both the predifined navigation pose and the open-vocabulary navigation to identify and move to the target location. First, we create an occupancy map using Gmapping [39] and define the navigation waypoint for the known locations in the map. We choose to manually define the navigation waypoint for the known location because it can simplify the subsequent manipulation process. The navigation also supports navigation to the location that is not in the known

TABLE I: Real-world Experiments Result. Objects that require user's selection by pointing are highlighted in **bold**.

Task Name	Task SR	Planning SR	Average Time	Average Interactions
*Put **overripe avocado** in the trash can.*	3/3	3/3	156s	1
*Move the **cup** from the coffee table to the kitchen counter.*	1/3	3/3	182s	2
Check the item in the fridge.	3/3	3/3	145s	1
*Take the **drink** to the coffee table.*	2/3	3/3	117s	1
*Navigate to the **location** and check if the surface is clean.*	3/3	3/3	77s	2
Mean	80%	100%	135s	1.4

location lists by text query and point query, similiar to the open vocabulary object segmentation module discussed in the pick policy II-D.1. We use the off-the-shelf path and motion planning algorithm in ROS Navigation Stack to generate the navigation path and motion trajectory.

3) Visual Question Answering: Our system is capable of answering users' open-ended questions about the objects in the robot's environment. Specifically, for the action *vqa(text)* and *vqa(text, pointing)*, our system applies GPT-4V [40] and supports two kinds of tasks:

Question answering via mobile manipulation For real-world environments, answering the question directly from the visual sensor input ignores the robot's ability to perform navigation and manipulation. Previous works [41], [42] to address these problems rely heavily on in-distribution training. Our solution leverages the reasoning ability of LLM. Given the question q, we use LLM to generate the sequence of actions: $\{a_i\}$ before querying the GPT-4V.

Question answering via point referring While text-only input allows users to ask questions, the single modality may fall short in terms of precise specification in the question. We hope to allow the robot to answer the question together with a pointing selection given by the user, i.e. *vqa(text, pointing)*. To achieve this, we apply a visual prompting method for GPT-4V similar to [43]. We use SAM to get the segmentation and draw a mask and a point in the image. The processed image is sent to the GPT-4V to answer the question.

III. Preliminary Experiments

A. Setting

Task Description In this section, we evaluate our system on a set of daily household tasks that require remote language and gesture instruction from humans. The tasks were designed with reference to the American Time Use Survey [44], which records how people spend their time. These specific tasks are under the common daily household class of *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 tasks required the robot to be able to answer questions and rearrange objects with remote language and pointing instructions. The names of the tasks can be found in Table. I. The robot is in a home environment and the user sits in a different room from the robot to send the voice and gesture command while viewing the shared zoom screen. More details can be found on the website.

We evaluate the system based on the following metrics: ***Task Success Rate:*** We define a successful task as one in which the goal was achieved or the correct answers were returned to the remote user within 5 minutes. ***Planning Success Rate:*** A plan is successful if the generated plan can lead to a success task assuming the low-level skills are executed perfectly. ***Average Time per Completed Task:*** The average time spend to successfully complete the task. ***Average Interactions:*** The average number of interactions per task.

B. Results and Analysis

For each task, we record and compute the efficacy and efficiency measures described in III-A. The main results of the experiments are shown in Table. I. Some recorded videos of the experiments can be found on the website for demonstration. The average task success rate, an indicator of efficacy, stands at 80%, highlighting the system's ability to complete the diverse tasks in real-world environment as intended. The 100% planning success rate suggests that the gap is mainly due to the imperfect low-level action execution. Our results also show that the system takes about 135 seconds on average and requires at most two interactions to complete a task, demonstrating the efficiency of the proposed system. Note that most of the time is spent on the navigation and manipulation process.

Failure Modes Analysis Despite the overall success, there is still gap to stop the task from performing in real-world robustly. Our analysis identified several key areas where the system's performance could falter:

- Grasp execution: An error occurs while gripping the target cup. Although the pose is generated, the gripper accidentally touches the cup during the approach process.
- Accidental pointing: The user accidentally pointed to the wrong target when giving the instruction, leading to failure.
- Motion Planning: While the arm is moving, the drink in the robot's gripper hits the body, causing the drink to fall to the floor.

IV. Conclusion

This work presents an interactive robotic assistant designed to improve household tasks by facilitating multimodal interactions with remote users. We have presented the three key components for realizing such a robot butler system. We also demonstrate the practical assistive question-answering and object rearranging application implemented on a mobile manipulator. The experimental evidence shows the efficacy and efficiency of our system. In future work, we hope to test the system in more everyday tasks with more users and to extend the system with more diverse manipulation skills.

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*. PMLR, 2023, pp. 287–318.

[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," *arXiv preprint arXiv:2305.05658*, 2023.

[3] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. W. Clegg, J. Turner *et al.*, "Homerobot: Open-vocabulary mobile manipulation," *arXiv preprint arXiv:2306.11565*, 2023.

[4] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Ok-robot: What really matters in integrating open-knowledge models for robotics," *arXiv preprint arXiv:2401.12202*, 2024.

[5] 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, pp. 787–793.

[6] 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, pp. 371–379.

[7] T. Vuletic, A. Duffy, L. Hay, C. McTeague, G. Campbell, and M. Grealy, "Systematic literature review of hand gestures used in human computer interaction interfaces," *International Journal of Human-Computer Studies*, vol. 129, pp. 74–94, 2019.

[8] 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*, vol. 28, no. 1, 2014.

[9] 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)*. IEEE, 2016, pp. 3331–3338.

[10] 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, pp. 1385–1395.

[11] 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. [Online]. Available: https://openreview.net/forum?id=Ffn8Z4Q-zU

[12] 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*, vol. 25, no. 1, 2011, pp. 1507–1514.

[13] 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*, vol. 35, no. 1-3, pp. 281–300, 2016.

[14] 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)*. IEEE, 2018, pp. 3774–3781.

[15] 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)*. IEEE, 2019, pp. 6942–6948.

[16] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 217–232, 2020.

[17] 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.

[18] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 894–906.

[19] 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*. PMLR, 2022, pp. 9118–9147.

[20] 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*. PMLR, 2023, pp. 1769–1782.

[21] 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, pp. 2998–3009.

[22] 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," *arXiv preprint arXiv:2303.03378*, 2023.

[23] 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 *7th Annual Conference on Robot Learning*, 2023.

[24] 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*, vol. 34, pp. 251–266, 2021.

[25] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain *et al.*, "igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," *arXiv preprint arXiv:2108.03272*, 2021.

[26] 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*. PMLR, 2022, pp. 477–490.

[27] O. Khatib, "Mobile manipulation: The robotic assistant," *Robotics and Autonomous Systems*, vol. 26, no. 2-3, pp. 175–183, 1999.

[28] 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*. IEEE, 2009, pp. 1992–1998.

[29] 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*. IEEE, 2012, pp. 5313–5320.

[30] 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)*. IEEE, 2021, pp. 9382–9388.

[31] 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," *arXiv preprint arXiv:2401.01474*, 2024.

[32] OpenAI, "Gpt-4 technical report," 2023.

[33] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," *arXiv preprint arXiv:2306.09683*, 2023.

[34] 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," *arXiv preprint arXiv:2304.02643*, 2023.

[35] 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)*. IEEE, 2021, pp. 13 438–13 444.

[36] S. Chitta, I. Sucan, and S. Cousins, "Moveit![ros topics]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 1, pp. 18–19, 2012.

[37] S. Chen, A. Xiao, and D. Hsu, "Llm-state: Expandable state representation for long-horizon task planning in the open world," *arXiv preprint arXiv:2311.17406*, 2023.

[38] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[39] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with rao-blackwellized particle filters," *IEEE transactions on Robotics*, vol. 23, no. 1, pp. 34–46, 2007.

[40] OpenAI, "Gpt-4v(ision) system card," https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023, accessed: 2024-02-03.

[41] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4089–4098.

[42] Y. Deng, D. Guo, X. Guo, N. Zhang, H. Liu, and F. Sun, "Mqa: Answering the question via robotic manipulation," in *Robotics: Science and Systems (RSS)*, 2020.
[43] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023.
[44] U.S. Bureau of Labor Statistics, "American time use survey," https://www.bls.gov/tus/, 2022.