InteractGen: Enhancing Human-Involved Embodied Task Reasoning through LLM-Based Multi-Agent Collaboration

Nan Sun
Department of Computer Science and Technology
Tsinghua University
sunn24@mails.tsinghua.edu.cn

Chengming Shi
Institute for Interdisciplinary Information Sciences
Tsinghua University
scm23@mails.tsinghua.edu.cn

Yuwen Dong
Department of Automation
Tsinghua University
dyw23@mails.tsinghua.edu.cn

Abstract

This paper introduces InteractGen, a novel multi-agent reasoning framework that integrates humans, embodied robots, and LLM-powered agents for seamless collaboration in dynamic, real-world environments. InteractGen enhances task execution efficiency and adaptability through advanced reasoning, dynamic context-awareness, and interactive capabilities. A key contribution is EmboInteract, a new dataset incorporating real-time human interaction and evolving task challenges, addressing limitations of existing static datasets. Together, these innovations establish a robust foundation for advancing embodied AI, enabling agents to operate effectively in complex, unpredictable settings.

1 Introduction

The increasing demand for intelligent robotic systems capable of assisting humans in dynamic, real-world environments has driven significant advancements in artificial intelligence. Modern service robots are expected not only to interpret human instructions but also to execute complex tasks autonomously while navigating uncertainties. However, these systems still face significant limitations in adaptability, interactive collaboration, and reasoning, particularly in human-populated environments.

The use of mobile robots in such environments has emerged as a key area of research within robotics and embodied AI. Initially, studies concentrated on robots operating in structured settings with limited human interaction. As demand for robots in more dynamic and unpredictable contexts has grown, research has increasingly focused on improving adaptability and enhancing human-robot collaboration. For example, Chung et al. [1] explored how mobile robots can autonomously collect and transmit environmental data to support human activities. Various researchers, such as Zhang et al. [2], Trautman and Krause [3], Truong and Ngo [4], Trautman et al. [5], examined robust navigation strategies for mobile robots functioning in complex, human-centered environments. Additionally, Liang et al. [6] introduced a method enabling service robots to determine humans' dynamic locations through dialogue processing. Systems enabling robots to sense, learn, and model human social behaviors to make appropriate real-time decisions were developed by Triebel et al. [7]. Despite these advancements, achieving human-level adaptability and interactivity in diverse real-world tasks remains a significant challenge.

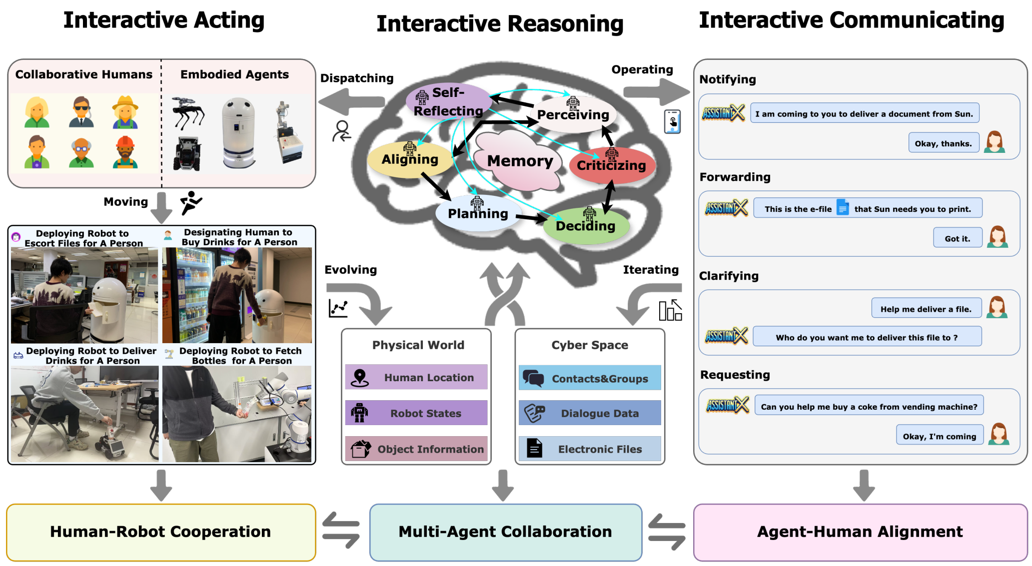


Figure 1: InteractGen exhibits powerful reasoning capabilities, enabling it to process rich, multi-faceted information with a strong awareness of interaction and coordination. It seamlessly integrates humans, embodied robots, and LLM-powered agents, bridging the virtual and physical worlds to support dynamic collaboration and adaptive decision-making. By unifying these components, InteractGen creates a cohesive system that enhances task execution efficiency and enables flexible, real-time operation in complex, evolving real-world environments.

A key driver of recent progress in multi-agent systems has been the rise of large language models, which have transformed how agents interact and collaborate. For instance, multi-agent frameworks have been employed to manage tasks such as GUI operations on smart devices [8, 9, 10, 11, 12, 13]. LLMs have also been used to autonomously assess and discuss the quality of generated responses [14]. Moreover, Abdelnabi et al. [15, 16] focused on evaluating LLMs within multi-agent systems, emphasizing their ability to deliberate and collaborate in environments requiring both cooperation and competition. These systems have also been applied in communication scenarios to gather detailed information through interaction [17, 18, 19], while Chen et al. [20] explored how cyber agents from different networks could collaborate and share intelligence to enhance overall performance.

Despite these advancements, significant challenges remain in applying multi-agent systems and embodied AI in real-world settings. Current embodied tasks are often highly specific and static, failing to account for the dynamic and unpredictable nature of real-world environments. Furthermore, existing systems typically neglect the complexities introduced by human factors, such as interactive behaviors, collaboration, and real-time decision-making under uncertainty. These limitations hinder the ability of embodied agents to perform robustly in human-populated, ever-changing environments. Moreover, agents in such tasks rarely integrate with physical robots to interact effectively with the real world, limiting their applicability in practical scenarios.

To address these challenges, we propose InteractGen, a unified multi-agent reasoning framework designed to enable seamless interaction and collaboration between humans, embodied robots, and LLM-powered agents across virtual and physical worlds (see Fig 1). InteractGen combines powerful reasoning capabilities with strong awareness of interaction and coordination, enabling it to process multi-faceted information, dynamically adjust to evolving conditions, and ensure context-aware decision-making. By bridging embodied intelligence, human collaboration, and virtual operations, InteractGen significantly improves task execution efficiency and adaptability in complex, real-world environments.

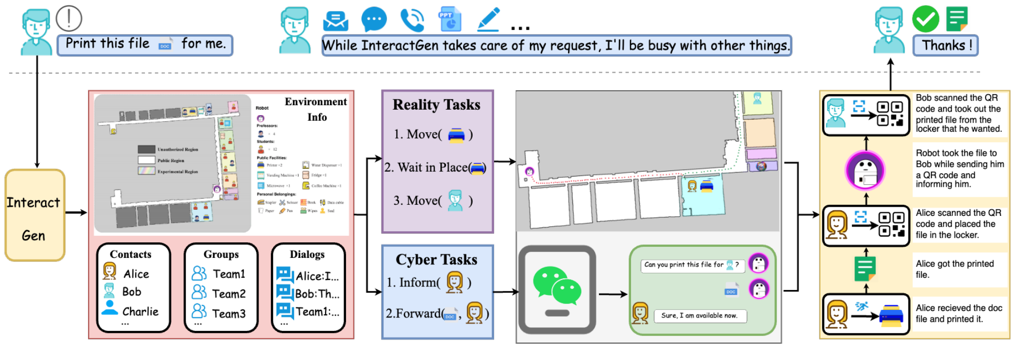


Figure 2: InteractGen ensures the task is executed dynamically and efficiently, without requiring constant human intervention. This flexibility and multi-agent cooperation overcome previous limitations of static systems, enabling real-time adaptability in complex, human-centered environments.

Another critical contribution of this work is the introduction of EmboInteract, a novel dataset designed to address the limitations of existing embodied task datasets. Unlike prior datasets such as ALFRED [21] and TEACh [22], which focus primarily on static task execution or predefined instructions, EmboInteract incorporates dynamic task execution that necessitates real-time human interaction. Inspired by previous works on proactive agent datasets [18], EmboInteract introduces exceptional circumstances during instruction execution through dialogue interactions and simulated robot perception states. This dataset construction follows the EBD (Elements Extraction, Base Instruction Generation, Dynamic Interaction Formation) methodology to enhance dataset quality and diversity by utilizing structured data representation, conditional prompting, and in-context demonstrations. This methodology generates comprehensive task instructions while simulating ambiguous directives, thereby promoting the development of agents capable of proactive clarification and dynamic adjustment. EmboInteract represents the first dataset of its kind to integrate multi-person interactions, dynamic challenges, and embodied task execution in evolving environments.

In summary, this work makes three key contributions:

- **InteractGen**: A multi-agent reasoning framework that unifies LLM-powered agents, embodied robots, and human collaborators to enable dynamic, context-aware task execution across physical and virtual environments.
- **EBD**: A novel dataset construction methodology for creating dynamic, interactive task datasets, addressing the shortcomings of existing static embodied task datasets.

These contributions provide a foundation for the advancement of embodied AI, enabling agents to navigate the complexities of real-world environments with adaptability, interactivity, and collaboration.

2 Related Work

2.1 Human-Centric Robotic Systems

Human-centric robotic systems aim to seamlessly integrate robots into human environments by emphasizing adaptability, collaboration, and user-centric design. Early research [23, 24, 25] laid the groundwork by improving communication between humans and robots, while multimodal interaction techniques, combining modalities such as speech, gestures, and vision, were introduced to make these interactions more natural [26, 27, 28]. Collaborative robotics, often referred to as cobots, expanded on these efforts by focusing on joint human-robot tasks in industrial and service contexts [29, 30].

With the advent of artificial intelligence, robots have gained the ability to learn from human behaviors and adapt to increasingly complex tasks [31, 32, 33]. Emotional intelligence, as explored in [24], further enhanced the human-robot connection by enabling robots to recognize and respond to human emotions, fostering more engaging interactions. Despite these advances, current systems are predominantly designed for structured and predictable environments. This restricts their capacity to handle the ambiguity and dynamic changes inherent in real-world scenarios. Moreover, while robotic intelligence has seen substantial progress, these systems still lack the deep reasoning capabilities required for proactive collaboration and decision-making under uncertainty, limiting their effectiveness in unstructured settings.

2.2 LLMs for Embodied Tasks

Building upon the limitations of traditional human-centric robotic systems, Large Language Models have emerged as a transformative technology for embodied tasks. These tasks demand agents navigate physical or simulated environments while executing complex instructions. Modular reasoning frameworks have been proposed to dissect the capabilities of LLM-centric agents, breaking tasks into manageable components for more effective execution [34]. Similarly, continuous learning paradigms refine agents' performance through iterative feedback loops [35], and interactive learning approaches have enhanced agents' ability to adapt to socially dynamic contexts [36].

Advancements in multimodal systems that integrate vision and language have further improved robotic control and task execution [37, 38]. Additionally, embedding language models in physical contexts through embodied experiences has been shown to enhance their reasoning and action capabilities [39]. Frameworks like Think-on-Graph [40] leverage structured knowledge representations to refine decision-making in complex, multi-step tasks. Nevertheless, these systems often fall short when confronted with real-world dynamics. Many remain constrained by predefined tasks and lack the adaptability to operate effectively in environments with frequent and unpredictable changes. Furthermore, their limited integration with human collaborators hampers their ability to refine decision-making processes collaboratively, which is crucial for tackling tasks in evolving and uncertain scenarios.

2.3 LLM-Based Multi-Agent Collaboration

As the demands of real-world applications grow, LLM-based multi-agent collaboration has become a pivotal research direction, addressing challenges in scalability, adaptability, and coordination. These systems enable agents to engage in structured cooperation, negotiation, and role adaptation. Frameworks such as AutoGen [41] and AgentVerse [42] facilitate dynamic role adjustment and collaboration on complex, multi-agent tasks. Meanwhile, policy optimization techniques, like those explored in Agent-Pro [43], focus on enabling agents to iteratively improve performance over time.

Contributions to this field also emphasize the importance of adaptability and strategic coordination. LLMArena [44] highlights real-time decision-making as a critical component for effective collaboration, while AgentCoord [45] demonstrates the value of visual exploration strategies in multi-agent scenarios. Theory of Mind approaches [46] provide insights into how agents can infer and reason about the intentions of others, a vital skill for teamwork. Beyond technical frameworks, studies from a social psychology perspective explore how group dynamics influence agent behavior and decision-making outcomes [47]. Despite these advances, many of these systems focus primarily on intra-agent communication, often at the expense of robust integration with physical agents and real-time human collaboration. This gap limits their potential to address the unpredictability and complexity of dynamic, human-populated environments. Additionally, while scalability has been explored in multi-agent systems [48, 49], further research is needed to manage the increased complexity associated with diverse, large-scale collaborations. Unlike existing approaches that focus on isolated reasoning or predefined collaboration strategies, InteractGen integrates interactive reasoning, human intention alignment, and human-robot cooperation into a unified framework. By leveraging seamless interaction across virtual and physical domains, InteractGen facilitates dynamic, adaptive, and context-aware task execution, significantly advancing the capabilities of embodied AI in complex, human-centered real-world scenarios (see Fig 2).

3 Methodology

3.1 Dataset Construction

In real-world, evolving environments, even the most meticulously designed plans are prone to deviations, leading to task execution failures. Inspired by previous research on proactive agent datasets

[18], we address these challenges by introducing EmboInteract, a dataset that incorporates dynamic task execution requiring human interaction within office contexts. EmboInteract is constructed to simulate exceptional scenarios during task execution, facilitated through dialogue interactions or simulated robot perception states. This approach is designed to encourage dynamic behaviors in agents and humans alike, fostering proactive interactions and adaptive strategies.

The dataset construction pipeline consists of three core phases: (1) elements extraction, (2) base instruction generation, and (3) dynamic interaction formation (see Fig 5). Initially, we establish a seed set comprising 30 character roles, 12 private items, and 6 public facilities that correspond directly to elements within the physical office environment. Each element is carefully annotated with attributes such as ownership, functionality, and interdependencies. During the elements extraction phase, random initialization of parameters determines task specifications, including the scope of participants, objects, and locations involved.

To generate structured and contextually grounded task instructions, we adopt a template-driven approach, which provides clear structure while mitigating uncertainty in content generation [50]. Task templates embed domain-specific constraints—such as resource limitations, temporal relationships, and element interdependencies—to enhance logical coherence and ensure task diversity [51, 52]. The resulting instructions are represented as JSON files, defining key task attributes and forming the basis for base instruction generation. Using conditional prompting, we transform the structured JSON data into coherent, context-aware task instructions, explicitly encoding desired behaviors and constraints. This process generates high-quality, diverse instructions that adapt to dynamic office scenarios while surpassing conventional direct prompting methods in both precision and instruction quality.

To simulate practical ambiguities inherent in real-world tasks, we intentionally obscure one field in the JSON file, except for the origin. This enables the LLM to generate vague instructions while prompting clarifying questions, mirroring realistic agent behaviors when handling incomplete information. To further improve task diversity, dynamic interaction formation integrates embodied perceptual variations and personnel availability changes into the generated instructions. For example, scenarios include absent personnel, unresponsive assistance requests, or incomplete robotic perception.

Existing datasets, such as ALFRED [21], TEACh [22], and TouchDown [53], focus primarily on static, step-by-step navigation and planning tasks, lacking real-time iterative adjustments and direct human-agent collaboration. Similarly, frameworks like PaLM-E [37] and OPEx [34] address embodied reasoning but overlook unexpected variations in task execution and the role of human involvement in refining strategies. In contrast, EmboInteract introduces dynamic interactions and multi-person collaboration, addressing two critical gaps: adaptability to evolving scenarios and proactive agent-human coordination.

Building on the PPDR4X framework [54], we developed an annotator capable of chain-of-thought-style decomposition reasoning. This annotator extracts task-relevant details, annotating individuals, objects, and actions while enriching instructions with dynamic elements. Such enriched scenarios require agents to exhibit interactive behaviors, such as adapting to absent individuals, addressing unfulfilled requests, or responding to incomplete sensory information.

EmboInteract, to the best of our knowledge, is the first dynamic, interactive embodied task dataset designed to simulate multi-participant collaboration with real-time adaptability. It overcomes the limitations of existing datasets that focus on simplistic tasks, static instructions, and minimal human-agent interaction. By introducing unexpected variations and multi-agent dynamics, EmboInteract provides a robust benchmark for evaluating the resilience, adaptability, and proactive engagement of embodied agents operating in complex, evolving environments.

3.2 Multi-Agent Framework

To address the inadequacy of inference capabilities in current service robot systems, we present a multi-agent framework for InteractGen (see Fig.4). In a given office scenario, InteractGen is capable of accurately perceiving the surroundings and human intentions, thereby formulating comprehensive plans based on user instructions. It can also autonomously execute tasks and engage in self-reflection, even when the instructions are complex and lacking in detail. Multi-agent collaboration equips InteractGen with a problem-solving mindset similar to that of a human assistant, facilitating seamless integration into authentic work environments for autonomous effective interaction with other individuals.

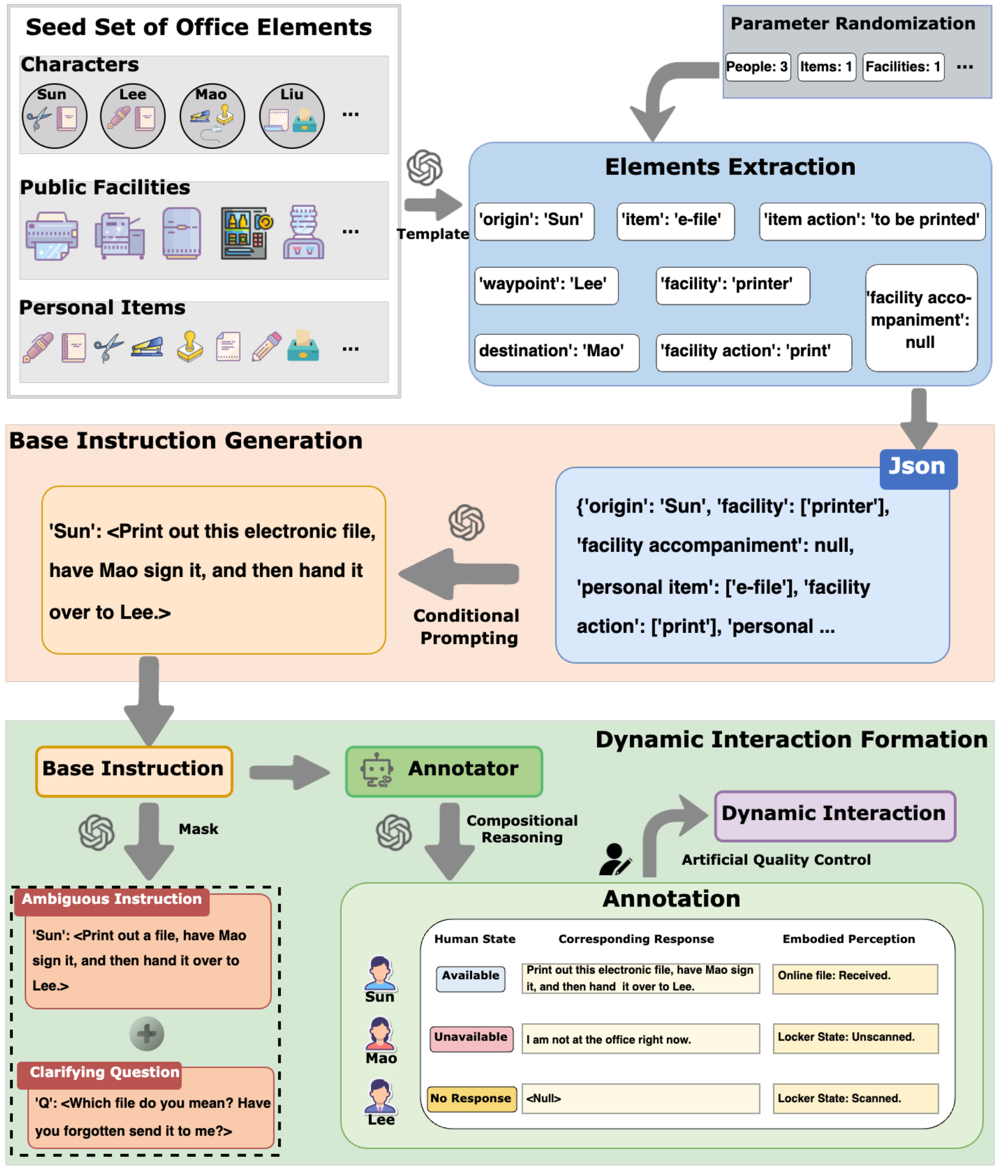


Figure 3: The EBA method generates dynamic, interactive task scenarios by combining structured office elements, conditional prompting, and compositional reasoning in natural language. It simulates realistic, evolving environments where agents handle ambiguities, adapt to changing human states, and interactively refine task execution through proactive reasoning and human-agent communication.

Memory Unit serves as the fundamental cornerstone of the entire framework, storing the initial dynamic map data provided by human operators. As InteractGen carries out commands, it updates the relevant virtual and physical world information. Concurrently, the agent's cognitive processes and actions throughout this procedure are recorded. To enhance efficiency in planning and executing consecutive operations, Memory Unit stores both individual and group chat records generated during instruction execution. It encapsulates, processes, and organizes both long-term memory (dynamic map information in $\mathcal{E}$) and short-term memory (comprising dialogue data $\mathcal{D}$, thoughts generated by agents, and executed cyber tasks $\mathcal{TC}$ and real-world tasks $\mathcal{TR}$). We use $\mathcal{M}_t$ to denote the memory package that encompasses all stored memory data at t time for effective utilization by the Perception Agent.

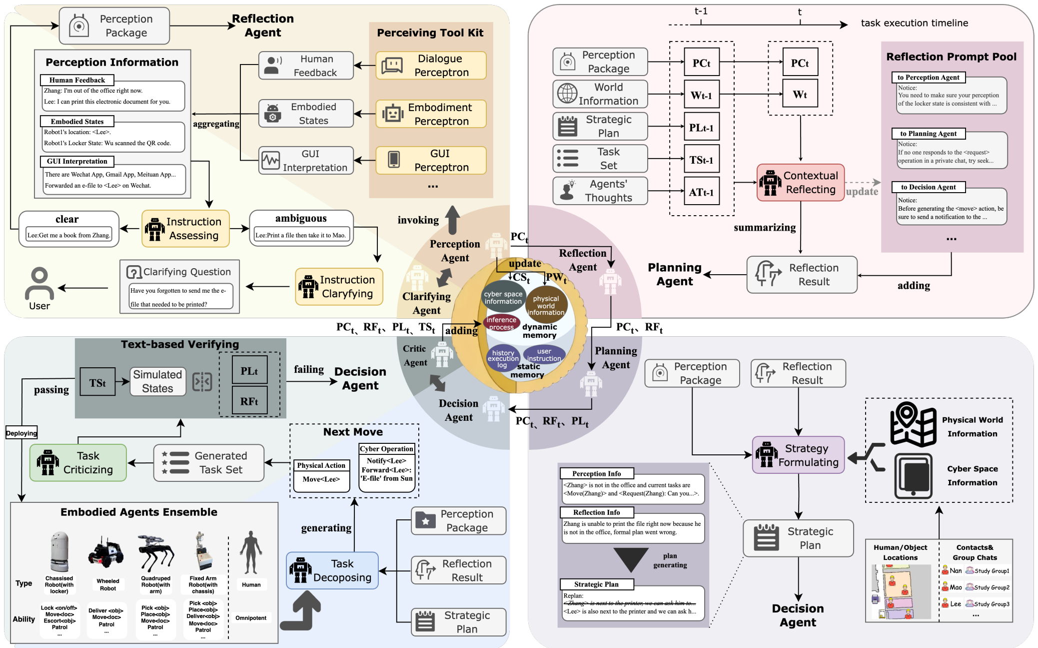


Figure 4: Our multi-agent collaboration framework integrates perception, planning, decision-making, and reflection agents to dynamically adapt, clarify instructions, and execute tasks efficiently across cyber-physical environments.

InteractGen is designed to possess human-like capabilities in perceiving user instructions, virtual environment information, and real-world states. Perception Agent serves as the starting phase, processing diverse input data to create a comprehensive perception package encapsulating the user's intentions, current surroundings, and physical states. The perceptual process can be articulated as follows:

$$\mathcal{PC}_t = perceive(\mathcal{I}, \mathcal{M}_t, \mathcal{SO}_{t-1}) \tag{1}$$

where $perceive(\cdot)$ represents the perceiving process of LLM and $\mathcal{PC}_t$ denotes the current perception package at t time, $\mathcal{M}_t$ is the memory package derived from the Memory Unit, and $\mathcal{SO}_{t-1}$ represents summarized history operations at $t-1$ time.

The goal of planning is to ensure that the generated plan aligns with user intentions while optimizing efficiency. This involves continuous evaluation and refinement as new information emerges or partial tasks are completed. Planning Agent analyzes the t time perception package $\mathcal{PC}_t$ while integrating historical information from the Memory Unit. To maintain computational efficiency, the agent summarizes accumulated historical operations, denoted as $\mathcal{SO}$, before formulating a detailed plan:

$$\mathcal{PL}_t = plan(\mathcal{M}_t, \mathcal{PC}_t, \mathcal{SO}_t, \mathcal{R}_{t-1}) \tag{2}$$

where $plan(\cdot)$ represents the planning process of LLM, $\mathcal{PL}_t$ is the newly generated plan at t time, and $\mathcal{R}_{t-1}$ is the reflection result from the previous step.

Decision Agent determines the specific actions InteractGen must execute to fulfill the user's instructions. Acting as the executor of strategic plans, it translates high-level objectives into operational steps. To facilitate smooth execution and ensure task alignment, we define an Action Space for each type of embodied agents, which guides the decision-making process. The agent also evaluates reflective outcomes from the previous step to prevent overlooked tasks or misalignments. The decision process is formalized as:

$$\mathcal{TC}_t, \mathcal{TR}_t = decide(\mathcal{M}_t, \mathcal{PC}_t, \mathcal{PL}_t, \mathcal{TC}_{t-1}, \mathcal{TR}_{t-1}, \mathcal{R}_{t-1}) \tag{3}$$

where $decide(\cdot)$ represents the decision process of LLM.

After $\mathcal{TC}_t$ and $\mathcal{TR}_t$ are executed, corresponding alterations occur in the virtual environment, real-world context, and robot state, reflected in $\mathcal{M}_{t+1}$ and $\mathcal{PC}_{t+1}$. Reflection Agent assesses these outcomes and renders binary judgments—'Y' for success or 'N' for deviation—while providing reflective reasons. This cohesive reflection result informs future planning and decision-making processes:

$$\mathcal{R}_t = reflect(\mathcal{M}_t, \mathcal{PC}_t, \mathcal{PL}_t, \mathcal{TC}_t, \mathcal{TR}_t, \mathcal{M}_{t+1}, \mathcal{PC}_{t+1}) \tag{4}$$

where $reflect(\cdot)$ represents the reflective process of LLM.

Clarifying Agent and Critic Agent serve as essential supplements to the perception and decision-making phases, respectively. The Clarifying Agent operates during the perception phase, ensuring that the received instructions are clear and aligned with the user's intent. It achieves this by proactively asking clarifying questions, for which we fine-tuned a LLaMA-7B model to handle this process effectively. On the other hand, the Critic Agent acts as a reflective mechanism prior to executing an action. By leveraging natural language processing to simulate changes in the real-world environment, embodied states, and task outcomes, it evaluates whether the intended execution aligns with the expected results. This pre-execution reflection ensures task consistency and reduces the likelihood of failures during operation.

4 Experiments

To assess the effectiveness of our architecture, we set six evaluation metrics in Table 1 (in Appendix). We select GPT-4o as the base model for our framework. The comprehensive test results of our architecture can be found in Table 2 (in Appendix), where it indicates that InteractGen offers strong effectiveness and stability. We annotated a total of 23 locations on the semantic map, including 16 individual workstations and 7 public facilities (see Fig 5). To enrich the contextual information, we incorporated details about each individual's personal belongings, ensuring that every person has at least three personal items. A service robot equipped with a mobile chassis, a mounted robot shell, and a smart locker is introduced to assist with tool deployment in the framework of InteractGen. Users can interact with InteractGen by issuing commands through a one-on-one messaging interface or by tagging '@InteractGen' in group chats to initiate the instruction $\mathcal{I}$.

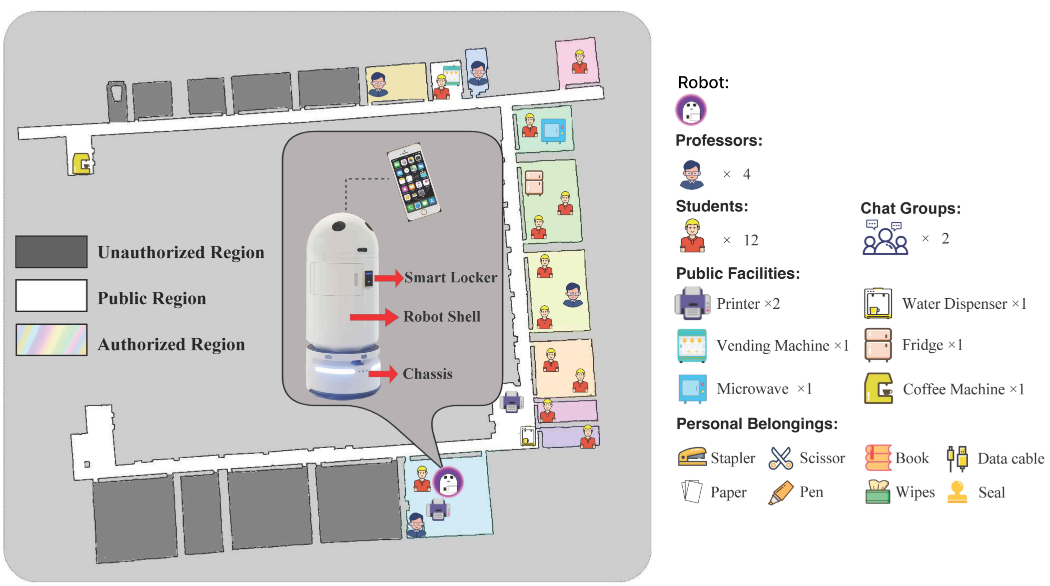


Figure 5: Our experiment consists of a semantic map and a customized service robot equipped with a smartphone.

We utilize a red-black tree structure to represent the branching relationships between the base instructions and their dynamic versions (see Fig.6). The black height of the instruction red-black tree represents the number of individuals that InteractGen must sequentially engage with, from top to bottom, to fully accomplish a task. This metric is also referred to as the task hop count associated with the instruction. In the tree, black nodes represent individuals capable of assisting the robot in completing certain tasks, while red nodes indicate that the person is unable to provide support. An illegal status occurs when a red node has a child node that is also red, which corresponds to a real-world scenario where, after person A is unable to assist with the task, the others are also unable

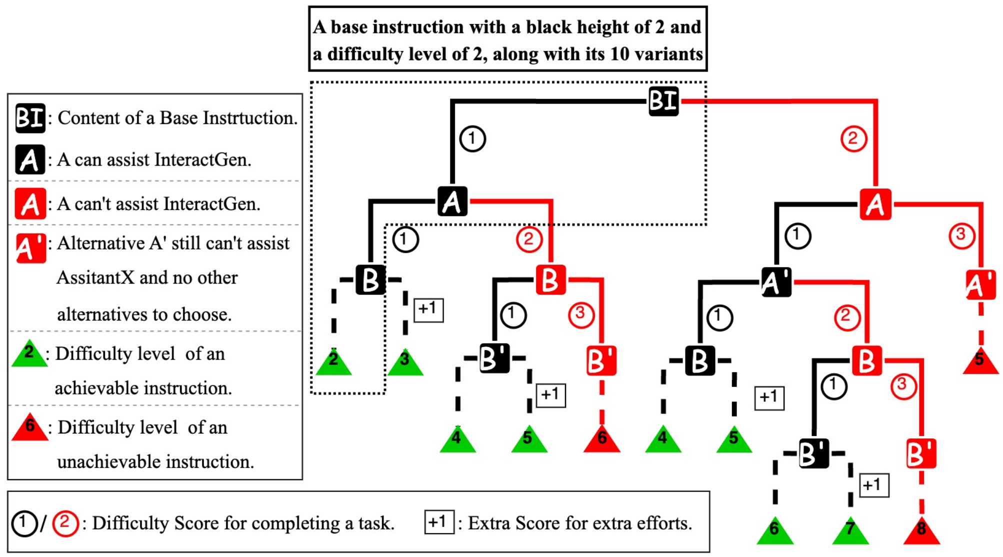


Figure 6: The red-black tree structure illustrates the branching process of the base instructions with their variants and the strategy to evaluate their difficulty level.

to provide support. This scenario results in the corresponding instructions being unachievable. By leveraging this structure, we can assess the difficulty level of any user instruction and its dynamic versions (see Fig.7 in Appendix).

To further validate the effectiveness of each agent, we conducted ablation experiments on EmboInteract, with detailed results also shown in Table 2. Our findings indicate that Planning Agent are crucial for effective instruction execution, as its removal in ablation experiments led to a significant decrease in both the success and completion rates of instructions. Meanwhile, Reflection Agent plays a key role in improving the redundant rates. Perception Agent further enhance performance, even when the framework is already functioning optimally, demonstrating the significant impact on overall robustness(see Fig.8 in Appendix).

5 Conclusion

In this study, we introduce InteractGen, a unified multi-agent reasoning framework powered by large language models that seamlessly integrates humans, embodied robots, and virtual agents for collaborative task execution in dynamic, real-world office environments. InteractGen processes complex, multi-modal inputs by autonomously perceiving user intentions, understanding its environment, and managing task flows across cyber and physical domains. By dynamically combining reasoning, task planning, and reflective mechanisms, InteractGen can adapt to uncertainties, clarify ambiguous instructions, and maintain efficient operation under evolving conditions. This integrated approach ensures that tasks requiring both physical actions—like navigating to a location—and cyber operations—such as managing files or communicating with users—are executed cohesively and contextually.

The experimental results validate the effectiveness and robustness of InteractGen in handling multi-agent coordination, ambiguous inputs, and real-time uncertainties. Through dynamic reasoning and iterative adjustments, InteractGen achieves precise task execution while incorporating feedback and resolving operational failures. The framework demonstrates significant improvements in handling collaborative tasks, where real-world scenarios often involve varying human states, unavailable resources, or incomplete perceptions. Our findings show that InteractGen achieves up to a 30 % improvement in operational efficiency compared to static systems, while reducing reliance on constant human intervention through its adaptive task-decomposition and clarification strategies.

Future work will focus on enhancing InteractGen's contextual reasoning capabilities to further improve its understanding of complex, evolving tasks. We aim to expand its physical action repertoire, allowing it to interact with more diverse objects and environments, while strengthening its ability to collaborate with larger groups of agents and humans. Additionally, we will explore the scalability of InteractGen in more intricate, multi-agent environments with higher task variability and uncertainty. This study establishes a solid foundation for creating embodied AI systems that bridge virtual and physical worlds, enabling more natural and productive integration into everyday human-centered settings, ultimately revolutionizing collaborative task execution.

References

[1] Michael Jae-Yoon Chung, Andrzej Pronobis, Maya Cakmak, Dieter Fox, and Rajesh P. N. Rao. Autonomous question answering with mobile robots in human-populated environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 823–830, 2016. doi: 10.1109/IROS.2016.7759146.

[2] Xuebo Zhang, Jiarui Wang, Yongchun Fang, and Jing Yuan. Multilevel humanlike motion planning for mobile robots in complex indoor environments. *IEEE Transactions on Automation Science and Engineering*, 16(3):1244–1258, 2019. doi: 10.1109/TASE.2018.2880245.

[3] Peter Trautman and Andreas Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 797–803, 2010. doi: 10.1109/IROS.2010.5654369.

[4] Xuan-Tung Truong and Trung Dung Ngo. Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model. *IEEE Transactions on Automation Science and Engineering*, 14(4):1743–1760, 2017. doi: 10.1109/TASE.2017.2731371.

[5] Peter Trautman, Jeremy Ma, Richard M. Murray, and Andreas Krause. Robot navigation in dense human crowds: the case for cooperation. In *2013 IEEE International Conference on Robotics and Automation*, pages 2153–2160, 2013. doi: 10.1109/ICRA.2013.6630866.

[6] Lanjun Liang, Ganghui Bian, Huailin Zhao, Yanzhi Dong, and Huaping Liu. Extracting dynamic navigation goal from natural language dialogue. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3539–3545, 2023. doi: 10.1109/IROS55552.2023.10342509.

[7] Rudolph Triebel, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, Daniel Cremers, Vanessa Evers, Michelangelo Fiore, Hayley Hung, Omar A. Islas Ramírez, Michiel Joosse, Harmish Khambhaita, Tomasz Kucner, Bastian Leibe, Achim J. Lilienthal, Timm Linder, Manja Lohse, Martin Magnusson, Billy Okal, Luigi Palmieri, Umer Rafi, Marieke van Rooij, and Lu Zhang. *SPENCER: A Socially Aware Service Robot for Passenger Guidance and Help in Busy Airports*, pages 607–622. Springer International Publishing, Cham, 2016. ISBN 978-3-319-27702-8. doi: 10.1007/978-3-319-27702-8_40. URL `https://doi.org/10.1007/978-3-319-27702-8_40`.

[8] Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception, 2024. URL `https://arxiv.org/abs/2401.16158`.

[9] Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration, 2024. URL `https://arxiv.org/abs/2406.01014`.

[10] Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. Coco-agent: A comprehensive cognitive mllm agent for smartphone gui automation, 2024. URL `https://arxiv.org/abs/2402.11941`.

[11] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents, 2024. URL `https://arxiv.org/abs/2401.10935`.

[12] Weihao Tan, Wentao Zhang, Xinrun Xu, Haochong Xia, Ziluo Ding, Boyu Li, Bohan Zhou, Junpeng Yue, Jiechuan Jiang, Yewen Li, Ruyi An, Molei Qin, Chuqiao Zong, Longtao Zheng, Yujie Wu, Xiaoqiang Chai, Yifei Bi, Tianbao Xie, Pengjie Gu, Xiyun Li, Ceyao Zhang, Long Tian, Chaojie Wang, Xinrun Wang, Börje F. Karlsson, Bo An, Shuicheng Yan, and Zongqing Lu. Cradle: Empowering foundation agents towards general computer control, 2024. URL `https://arxiv.org/abs/2403.03186`.

[13] Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users, 2023. URL `https://arxiv.org/abs/2312.13771`.

[14] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023. URL `https://arxiv.org/abs/2308.07201`.

[15] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games. *ArXiv*, abs/2309.17234, 2023. URL `https://api.semanticscholar.org/CorpusID:271270974`.

[16] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation, 2024. URL `https://arxiv.org/abs/2309.17234`.

[17] Anonymous. LLM-powered multi-agent proactive communication system for embodied intelligence. In *Submitted to ACL Rolling Review - June 2024*, 2024. URL `https://openreview.net/forum?id=n9dV9E7RVj`. under review.

[18] Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, and Tat-Seng Chua. Ask-before-plan: Proactive language agents for real-world planning, 2024. URL `https://arxiv.org/abs/2406.12639`.

[19] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners, 2023. URL `https://arxiv.org/abs/2307.01928`.

[20] Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence, 2024. URL `https://arxiv.org/abs/2407.07061`.

[21] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks, 2020. URL `https://arxiv.org/abs/1912.01734`.

[22] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat, 2021. URL `https://arxiv.org/abs/2110.00534`.

[23] M. A. Goodrich and A. C. Schultz. Human-robot interaction: A survey. *arXiv preprint cs/0507008*, 2005. URL `https://arxiv.org/abs/cs/0507008`.

[24] Cynthia Breazeal. Emotion and sociable humanoid robots. *arXiv preprint cs/0308031*, 2003. URL `https://arxiv.org/abs/cs/0308031`.

[25] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT Press, 2005. URL `https://arxiv.org/abs/cs/0508013`.

[26] Rainer Stiefelhagen, Hazim Kemal Ekenel, Christian Fugen, et al. Enabling multimodal human-robot interaction for a receptionist robot. *arXiv preprint cs/0703101*, 2007. URL `https://arxiv.org/abs/cs/0703101`.

[27] Kerstin Dautenhahn, Michael L. Walters, Sarah N. Woods, et al. How may i serve you? a robot companion approaching a seated person in a helping context. *arXiv preprint cs/0603023*, 2006. URL https://arxiv.org/abs/cs/0603023.

[28] Peng Liang, Arjun Ganesan, and Anand Panangadan. Multimodal interaction for human-robot collaboration. *arXiv preprint arXiv:2103.12345*, 2021. URL https://arxiv.org/abs/2103.12345.

[29] Valeria Villani, Francesco Pini, Francesco Leali, and Cristian Secchi. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *arXiv preprint arXiv:1803.12345*, 2018. URL https://arxiv.org/abs/1803.12345.

[30] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, et al. Progress and prospects of human-robot collaboration. *arXiv preprint arXiv:1709.00001*, 2017. URL https://arxiv.org/abs/1709.00001.

[31] A. D. Dragan, S. Bauman, J. Forlizzi, and S. Srinivasa. Effects of robot motion on human-robot collaboration. *arXiv preprint arXiv:1504.00001*, 2015. URL https://arxiv.org/abs/1504.00001.

[32] Karthik Talamadupula, J. Benton, Paul Schermerhorn, et al. Open-world robots: Planning under uncertainty. *arXiv preprint arXiv:1205.00001*, 2014. URL https://arxiv.org/abs/1205.00001.

[33] Michael Laskey, Jane Lee, Sile Zhou, et al. Dart: Noise injection for robust imitation learning. *arXiv preprint arXiv:1703.00001*, 2017. URL https://arxiv.org/abs/1703.00001.

[34] Haochen Shi, Zhiyuan Sun, Xingdi Yuan, Marc-Alexandre Côté, and Bang Liu. Opex: A component-wise analysis of llm-centric agents in embodied instruction following, 2024. URL https://arxiv.org/abs/2403.03017.

[35] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

[36] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents, 2024. URL https://arxiv.org/abs/2310.11667.

[37] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. URL https://arxiv.org/abs/2303.03378.

[38] Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. Embodied task planning with large language models, 2023. URL https://arxiv.org/abs/2307.01848.

[39] Yu-Ren Liu, Biwei Huang, Zhengmao Zhu, Honglong Tian, Mingming Gong, Yang Yu, and Kun Zhang. Learning world models with identifiable factorization, 2023. URL https://arxiv.org/abs/2306.06561.

[40] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph, 2024. URL https://arxiv.org/abs/2307.07697.

[41] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023. URL https://arxiv.org/abs/2308.08155.

[42] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors, 2023. URL https://arxiv.org/abs/2308.10848.

[43] Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. Agent-pro: Learning to evolve via policy-level reflection and optimization, 2024. URL https://arxiv.org/abs/2402.17574.

[44] Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. *arXiv preprint arXiv:2402.16499*, 2024.

[45] Bo Pan, Jiaying Lu, Ke Wang, Li Zheng, Zhen Wen, Yingchaojie Feng, Minfeng Zhu, and Wei Chen. Agentcoord: Visually exploring coordination strategy for llm-based multi-agent collaboration. *arXiv preprint arXiv:2404.11943*, 2024.

[46] Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*, 2023.

[47] Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.782. URL https://aclanthology.org/2024.acl-long.782.

[48] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic llm-powered agent network for task-oriented agent collaboration, 2024. URL https://arxiv.org/abs/2310.02170.

[49] Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large-language-model-based multi-agent collaboration, 2024. URL https://arxiv.org/abs/2406.07155.

[50] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners, 2021. URL https://arxiv.org/abs/2009.07118.

[51] Canyu Chen and Kai Shu. Promptda: Label-guided data augmentation for prompt-based few-shot learners, 2023. URL https://arxiv.org/abs/2205.09229.

[52] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey, 2024. URL https://arxiv.org/abs/2406.15126.

[53] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments, 2020. URL https://arxiv.org/abs/1811.12354.

[54] Nan Sun, Bo Mao, Yongchang Li, Lumeng Ma, Di Guo, and Huaping Liu. Assistantx: An llm-powered proactive assistant in collaborative human-populated environment, 2024. URL https://arxiv.org/abs/2409.17655.

Appendix

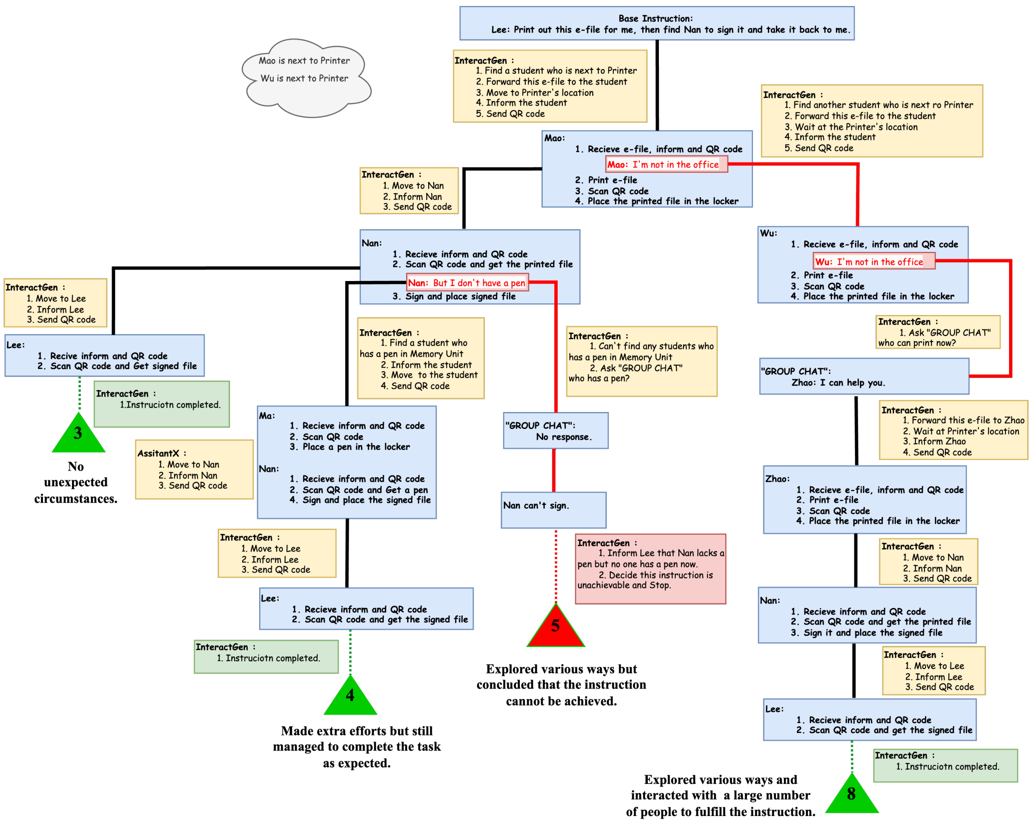
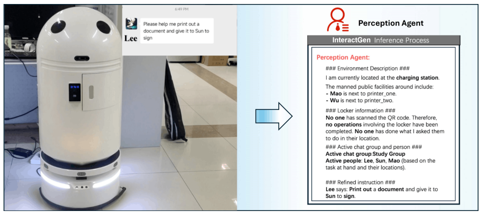
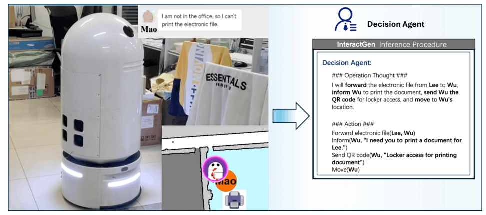


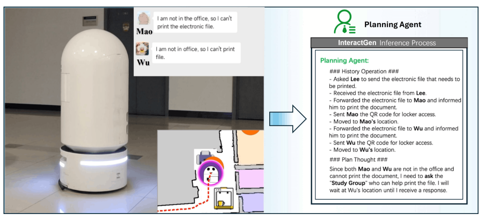
Figure 7: An illustration of how instructions are evaluated for difficulty levels.



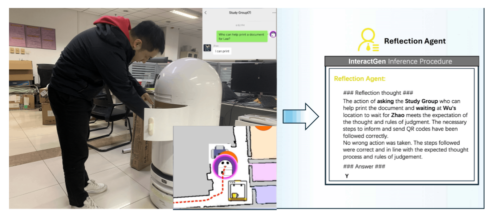
(a) Perception Agent is capable of perceiving online and real-world environments, while refining the user's initial instructions to capture key details of the task.



(b) When the corresponding personnel is absent and InteractGen needs to find another person, Decision agent can generate both cyber and real-world tasks, performing them synchronously.



(c) Planning Agent can retrieve relevant information from its memory to formulate alternative plans. If still failing, it will direct Decision Agent to engage with others in the group chat for new insights and replan.



(d) Reflection Agent reflects on the actions generated by Decision Agent and evaluates the outcomes, ensuring that each task is executed accurately.

Figure 8: Our multi-agent framework showcased impressive reasoning abilities during the experiments.

Figure 9: Our framework can sustain a remarkably low task error rate, even in scenarios where the task depth is significantly increased.

Table 1: The metrics that we used in evaluation.

Evaluation Metric	Description
SR: Success Rate	Success Rate is measured as the percentage of instructions that InteractGen successfully completed across various scenarios.
CR: Completion Rate	Completion Rate is calculated by dividing the depth value of the deepest successfully executed node by the total height of the instruction branch. This metric indicates how far InteractGen is able to progress in fulfilling the given instruction.
RR: Redundant Rate	Redundancy Rate is calculated by dividing the redundancy hop count by the instruction's black height. The redundant hop count is the value obtained by subtracting the fixed hop count from the actual hop count, where the actual hop count is represented as the black height in the corresponding red-black tree structure.
CTA: Cyber Task Accuracy	The proportion of correct cyber tasks out of the total number of cyber tasks generated by Decision Agent while executing user instructions.
RTA: Real-World Task Accuracy	The proportion of correct real-world tasks out of the total number of real-world tasks generated by Decision Agent while executing user instructions.
RA: Reflection Accuracy	The proportion of correctly generated reflection results by Reflection Agent out of the total number of reflection results produced during the execution of user instructions.

Table 2: Ablation Evaluation

Perception	✔				✗				✔				✔			
Reflection	✔				✔				✗				✔			
Planning	✔				✔				✔				✗			
	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4
SR	0.98	0.87	0.73	0.67	0.96	0.81	0.55	0.37	0.94	0.79	0.59	0.34	0.85	0.57	0.18	0.01
CR	0.99	0.92	0.80	0.74	0.98	0.87	0.65	0.49	0.95	0.83	0.68	0.45	0.91	0.64	0.31	0.18
RR	0.06	0.04	0.02	0.06	0.01	0.03	0.01	0.01	0.06	0.02	0.04	0.01	0.01	0.03	0.04	0
CTA	0.99	0.89	0.78	0.73	0.97	0.85	0.63	0.49	0.95	0.82	0.67	0.44	0.92	0.63	0.32	0.17
RTA	0.99	0.89	0.79	0.71	0.97	0.86	0.65	0.50	0.95	0.82	0.69	0.45	0.93	0.63	0.31	0.19

* L1 represents Difficulty Level 1-3, L2 represents Difficulty Level 4-6, L3 represents Difficulty Level 7-8, L4 represents Difficulty Level 9+