
Building Cooperative Embodied Agents Modularly with Large Language Models

Hongxin Zhang^{1*}, Weihua Du^{2*}, Jiaming Shan³, Qinhong Zhou¹
Yilun Du⁴, Joshua B. Tenenbaum⁴, Tianmin Shu⁴, Chuang Gan^{1,5}

¹University of Massachusetts Amherst, ² Tsinghua University,
³Shanghai Jiao Tong University, ⁴MIT, ⁵MIT-IBM Watson AI Lab

Abstract

Large Language Models (LLMs) have demonstrated impressive planning abilities in single-agent embodied tasks across various domains. However, their capacity for planning and communication in multi-agent cooperation remains unclear, even though these are crucial skills for intelligent embodied agents. In this paper, we present a novel framework that utilizes LLMs for multi-agent cooperation and tests it in various embodied environments. Our framework enables embodied agents to plan, communicate, and cooperate with other embodied agents or humans to accomplish long-horizon tasks efficiently. We demonstrate that recent LLMs, such as GPT-4, can surpass strong planning-based methods and exhibit emergent effective communication using our framework without requiring fine-tuning or few-shot prompting. We also discover that LLM-based agents that communicate in natural language can earn more trust and cooperate more effectively with humans. Our research underscores the potential of LLMs for embodied AI and lays the foundation for future research in multi-agent cooperation. Videos can be found on the project website <https://vis-www.cs.umass.edu/Co-LLM-Agents/>.

1 Introduction

Large Language Models (LLMs) have exhibited remarkable capabilities across various domains, implying their mastery of natural language understanding, rich world knowledge, and complex reasoning capability[6]. Recent research has also demonstrated that LLMs can function as planners in single-agent embodied tasks through zero-shot prompting for instruction following tasks [15] or few-shot prompting for more complex long-horizon tasks [44]. However, for embodied agents to work with other agents or with humans, they also need to have strong abilities for cooperation and communication. To date, it still remains unclear whether LLMs have such abilities necessary for embodied multi-agent cooperation.

Therefore, this paper aims to investigate whether LLMs can help build cooperative embodied agents that can collaborate with other agents and humans to accomplish complex tasks through collaborative planning and communication. To this end, we focus on an embodied multi-agent setting as shown in Figure 1, where two embodied agents have to cooperate to finish a task as soon as possible. To succeed in this setting, agents must i) extract useful information from observations, ii) revise their beliefs about the world and other agents, iii) decide what and when to communicate, and iv) plan collaboratively to reach the common goal.

To achieve these goals, we introduce a novel framework that utilizes LLMs to plan and communicate with other agents to cooperatively solve complex embodied tasks without any fine-tuning or few-shot

* denotes equal contribution.



Figure 1: We aim to utilize Large Language Models to build cooperative embodied agents.

prompting. Our framework consists of five modules, each to address a critical aspect of successful multi-agent cooperation, including a belief module to monitor the agent’s understanding of both the physical environment and other agents, a communication module to decide *what* to communicate utilizing the strong free-form dialogue generation and understanding capability of LLMs, and a reasoning module to synthesize all the information provided by other modules to decide high-level plans including *when* to communicate.

We evaluate our framework on two extended embodied multi-agent cooperation challenges: Communicative Watch-And-Help (C-WAH) and ThreeDWORLD Multi-Agent Transport (TDW-MAT). Our experimental results indicate that cooperative embodied agents built with Large Language Models can plan, communicate, and cooperate with other embodied agents and humans to accomplish long-horizon tasks efficiently. For example, as illustrated in Figure 1, the LLM-based agent can reason about the current state and the other agent’s state and divides the labor with its partner through communication effectively. In particular, by harnessing the rich world knowledge and strong reasoning capability of recent Large Language Models, such as GPT-4, our method can outperform strong planning-based baselines and exhibit emergent efficient communication. In a user study, we also discover that LLM-based agents that communicate with humans in natural language can earn more trust from humans. In sum, our contribution includes:

- We conducted the first systematic study on LLMs’ capacity for planning and communication in embodied multi-agent cooperation.
- We introduced a novel framework that utilizes LLMs to build cooperative embodied agents, surpassing strong planning-based methods.
- We conducted a user study to evaluate the possibility of achieving effective and trustworthy human-AI cooperation using LLMs.

2 Related Work

Multi-Agent Cooperation and Communication Plenty works have explored various aspects of multi-agent cooperation and communication. Some works provide various platforms for multi-agent tasks [27, 38, 43, 17, 39, 45, 2, 3]. Other works focused on methods that improves communication efficiency [21, 8, 46], cooperation in visually rich domains [18], or grounding communications in environments [33]. For embodied intelligence, [35] and [36] explored the social perception of the agents during their cooperation. These works usually disable communication [17, 39, 7, 35, 36], use continuous vectors [21, 8] for communication, or use discrete symbols [27, 20, 18, 33, 38] for communication. In contrast, our work stands apart by employing large language models for communication, introducing a novel perspective that utilizes natural language to enhance multi-agent cooperation and communication.

Planning with Large Language Models Recently, a branch of work has explored the planning capabilities of large language models. Although LLMs still face challenges when solving complex reasoning problems [6], a substantial body of work demonstrates their capacity to assist agents in planning [41, 37, 31, 13, 52, 53], especially in embodied environments[23, 4, 30, 24, 42, 29, 54, 5, 50, 40, 51, 18, 19]. For example, [16] used LLMs to build an inner monologue with environment feedback. [47] achieves better error correction during long-haul planning with LLMs. [1] focused on providing contextual grounding using pretrained behaviors to guide the generation of feasible

and contextually appropriate natural language actions. LLMs are also capable of initializing policy networks for agents[25], directly producing plans [44, 10], or generating policy code [26]. More recently, [32] used extended LLMs to simulate human behavior on generative agents. In contrast to most of these works, our method addresses the multi-agent cooperation scenario, which is more complex than planning for a single agent.

3 Building Cooperative Embodied Agents with Large Language Models

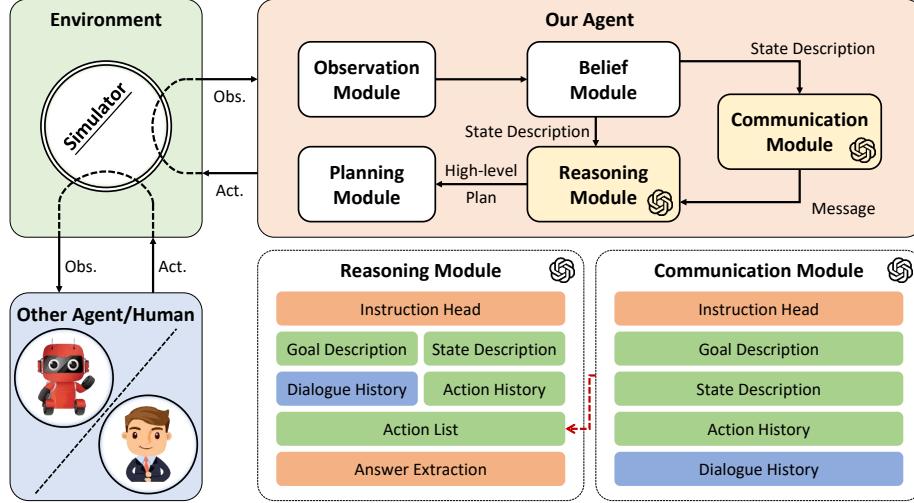


Figure 2: An overview of our framework, consisting of five modules: observation, belief, communication, reasoning, and planning, where the Communication Module and the Reasoning Module leverage Large Language Models to generate messages and decide on high-level plans. Here we also show the overall prompt design for leveraging LLMs to serve as these two modules. More design details can be found in Appendix A.

3.1 Problem Setup

Our problem can be defined as a decentralized partially observable Markov decision process (Dec-POMDP) augmented with communication, which can be formalized by $(S, G, \{A_i\}, \{O_i\})$, where n embodied intelligent agents take actions $a_i \in A_i$ to navigate, interact, and communicate in a partially-observable environment given the current step’s observation $o_i \in O_i$ including the messages received for each agent i to cooperate to solve a long-horizon task with a goal $g \in G$, normally consisting of several sub-goals g_1, g_2, \dots, g_m . Real-life household activities are representatives of this kind of task, that require intelligent embodied agents to cooperate with other agents and humans through long-horizon planning and effective communication.

3.2 Our Proposed Framework

The overall modular framework is shown in Figure 2, which consists of five modules: observation, belief, communication, reasoning, and planning. At each step, we first process the raw observation received with an *Observation Module* (3.2.1), then update the agent’s inner belief of the scene and the other agents through a *Belief Module* (3.2.2), this belief is then used with the previous actions and dialogues to construct the prompt for the *Communication Module* (3.2.3) and the *Reasoning Module* (3.2.4) which utilizes Large Language Models to generate messages and decide on high-level plans. Finally, a *Planning Module* (3.2.5) gives the primitive action to take in this step according to the high-level plan.

3.2.1 Observation Module

To enable embodied cooperation, it is important to perceive raw observations from the environment and extract information for downstream higher-order reasoning.

To achieve this we incorporate an Observation Module as the first module to deal with the observation received from the environment and extract useful high-level information such as visual scene graphs, objects, relationships between objects, maps of the environment, and other agents' locations. Our observation module can deal with both symbolic observations and egocentric visual observations.

3.2.2 Belief Module

Since LLMs have no intrinsic memory of the previous observations or interactions, it's crucial to find a way to effectively store and update the belief of the physical scenes and the states of the other agents. Here we propose a Belief Module to keep track of the four following information.

Task Progress P_T We keep track of the task progress in the belief module as *Task Progress* P_T and update it whenever possible using processed observation information.

Ego-State P_E Knowing own state is also of vital importance for embodied agents, so we gather all the information about the agent's own states from the processed observation and stored it in the belief module as *Ego-State* P_E .

Others-State P_O Keeping track of the other agents' states is important for cooperating with other agents, so we maintain *Others-State* P_O in the belief module and update it whenever a new observation of the others is possible.

Scene Memory P_S The memory of what objects has been seen and where they were is vital for an embodied agent exploring a vast space. Without this, it would be impossible for the agents to make long-horizon plans and share them with each other. We keep a record of seen objects and their states as Scene Memory P_S . To be noticed, this memory of scenes may not be accurate since other agents may interact with the objects and change their states without my awareness. Dealing with conflicts between my memory of the scene and the description of the scene from others is needed.

3.2.3 Communication Module

It's important for cooperative embodied agents to be able to communicate effectively with others. Effective communication needs to solve two problems: *what* to send and *when* to send.

We deal with the *what* to send problem in this module by directly using the LLMs as a Message Generator with designed prompts shown in Figure 2, constructed from the components of Instruction Head, Goal Description, State Description, Action History, and Dialogue History. To better constrain LLMs' generated messages, we also add a note at the end of the prompt and append two seed messages at the beginning of the Dialogue History to elicit deserved effective communication behavior. The detailed prompt design is shown in Appendix A.

3.2.4 Reasoning Module

With all the information gathered and provided by previous modules, cooperative embodied agents need to synthesize and reason over the current state, the belief of the others and the scene, the goals, the actions I've taken, and messages I've received to come up with a plan of what to do next. A strong reasoning module is required to leverage all the information effectively.

While designing such a module from scratch is nearly infeasible, we utilize powerful LLMs directly as the Reasoning Module with designed prompts similar to the Communication Module to reason over all the information and generate a high-level plan. Specifically, we modify the Instruction Head and compile an Action List of all available actions for the LLMs to make the choice, which formalization makes it easier for the LLMs to make an executable plan without any few-shot demonstrations.

We also use the zero-shot chain-of-thought prompting technique introduced by [22] to encourage the LLM to carry out more reasoning before giving the final answer.

3.2.5 Planning Module

As shown in [9], solving challenging embodied tasks requires modular methods to tackle the complexity of tasks. As also discussed in [49], we found that while Large Language Models were effective at making high-level plans, they were poor at making low-level controls. Thus, to enable effective embodied communication, we designed a Planning Module that can generate robust low-level controls

according to a given high-level plan, allowing the reasoning module to focus more on solving the overall task with LLMs’ rich world knowledge and strong reasoning ability. Practically, this way can also reduce the needed number of API requests and is time-saving and economical.

We implement the Planning Module with a heuristic-designed low-level planner to robustly carry out primitive actions according to the high-level plan generated from the Reasoning Module.

4 Experiments

We first introduce the two embodied environments we evaluate our framework on in section 4.1, then discuss the performance of our designed framework when cooperating with AI agents in section 4.2.1, showing they are better cooperators, and they can earn more trust and cooperate better with Humans in section 4.2.2. In section 4.3, we analyze the effectiveness of our different modules.

4.1 Experimental Setup

4.1.1 Communicative Watch-And-Help

Communicative Watch-And-Help (C-WAH) is an embodied multi-agent cooperation benchmark, extended from the existing Watch-And-Help Challenge [35], where we focus more on cooperation ability. To achieve this, we support communication between agents and remove the Watch stage so both agents have common goals. The challenge is built on a realistic multi-agent simulation platform, VirtualHome-Social[34, 35]. We conduct experiments under both symbolic observations and ego-centric visual observations. The task is defined as five types of common household activities: *Prepare afternoon tea*, *Wash dishes*, *Prepare a meal*, *Put groceries*, and *Set up a dinner table*, and represented as various predicates with counts to be satisfied. The number of total goal objects is within 3 to 5.

Setup We sampled 2 tasks from each of the five types of activities to construct a test set of 10 episodes. An episode is terminated if all the predicates in the goal are satisfied or the maximum number of steps (250) is reached.

Metrics We evaluate the performance by two metrics: *Average Steps L* taken to finish the task and *Efficiency Improvement (EI)* calculating the efficiency improvements of cooperating with other agents as $\sum_{i=1}^N (L_{single,i} - L_{multi,i}) / L_{single,i}$, where $L_{single,i}$ denotes the average steps for a single agent to finish episode i , and $L_{multi,i}$ denotes the average steps for multi-agents to finish episode i .

MCTS-based Hierarchical Planner We adopt the strongest baseline from the original Watch-And-Help Challenge, which is a Hierarchical Planner with a high-level planner based on MCTS and a low-level planner based on regression planning (RP).

4.1.2 ThreeDWorld Multi-Agent Transport

We extend the ThreeDWorld Transport Challenge [12] into a multi-agent setting with more types of objects and containers, more realistic objects placements, and support communication between agents, named ThreeDWorld Multi-Agent Transport (TDW-MAT), built on top of the TDW platform [11], which is a general-purpose virtual world simulation platform. The agents are tasked to transport as many target objects as possible to the goal position with the help of containers as tools, without which the agent can transport only two objects at a time. The agents have the same ego-centric visual observation and action space as before with a new communication action added.

Setup We selected 6 scenes from the TDW-House dataset and sampled 2 types of tasks in each of the scenes, making a test set of 12 episodes. Every scene has 6 to 8 rooms, 10 objects, and 4 containers. An episode is terminated if all the target objects have been transported to the goal position or the maximum number of frames (3000) is reached.

Metrics We use the *Transport Rate (TR)* as the evaluation metric, which is calculated as the fraction of the target objects successfully transported to the goal position, and calculate the *Efficiency Improvements (EI)* similar to the previous as $\sum_{i=1}^N (TR_{multi,i} - TR_{single,i}) / TR_{multi,i}$, where the $TR_{single,i}$ denotes the single agent’s transport rate for episode i , and $TR_{multi,i}$ denotes the multiple agent’s transport rate for episode i .

Rule-based Hierarchical Planner We adopt the strong performing baseline from the original challenge, which is a Rule-based Hierarchical Planner with Frontier Exploration strategy, consisting of a rule-based high-level planner which selects one of the high-level plans from Exploration, Pick up an object, Pick up a container, and Place according to some human-defined rules and an A-star based planner to navigate with occupancy map and semantic map obtain and updated from the visual observation. The Frontier exploration strategy randomly samples a way-point from an unexplored area as a sub-goal for exploration.

Implementation Details. We instantiate our framework with the recent LLM GPT-4. We access GPT-4 from the OpenAI API¹ and use the parameter of temperature 0.7, top-p 1, and max tokens 256. We show an example prompt for the Reasoning Module for both environments in Appendix C.

4.2 Results

C-WAH				TDW-MAT	
Symbolic Obs		Visual Obs			
Average Steps	EI	Average Steps	EI	Transport Rate	EI
HP	111	/	141	/	0.53
HP + HP	75	33%	103	26%	0.79
HP + LLM	59	45%	94	34%	0.86
LLM + LLM	57	49%	92	34%	0.86
					39%

Table 1: **Main results.** We report the mean results here over 5 runs except for LLM, which takes only one run due to cost constraints. The best results are in **bold**. The best performance is achieved when cooperating with LLM agents.

4.2.1 Collaborating with AI Agents

Quantitative results As shown in Table 1, on C-WAH, compared with the MCTS-based HP agent doing the task alone, cooperating with another MCTS-based HP agent provides an efficiency improvement of 33% and 26% under symbolic and visual observation, while cooperating with the LLM agent boosts the speed-up to 45% and 34% respectively, even without any knowledge of the inner working mechanism of the others, which shows LLMs can reason about the other agent’s state well without hand-designed heuristics. What’s more, when two LLM agents cooperate together, they can achieve even better performance. From TDW-MAT, we can observe the same performance boost of cooperating with the LLM agent of 38% compared to 34% of cooperating with the rule-based HP agent. These results show **our embodied agents built with LLMs are better cooperators**.

Qualitative results To better understand the essential factors for effective cooperation, we conduct a qualitative analysis of the agents’ behaviors exhibited in our experiments and identified several cooperative behaviors.

LLM Agents share progress and information with others. As shown in Figure 3abde, LLM agents communicate with each other to share progress and intents, demonstrating **the Communication Module can handle the challenge of what to send**, harnessing the free dialogue generation ability from the LLMs.

LLM Agents know when to request help and can respond to others’ requests. In Figure 3d, Bob finds a target object in the living room but his container is already full, so he shares this information and requests Alice to come here to help. Alice responds by going there and grabbing the objects. Similarly in Figure 3b, Alice responds to Bob’s requests and questions. These examples show LLMs know when to request help and can understand others’ requests and responses.

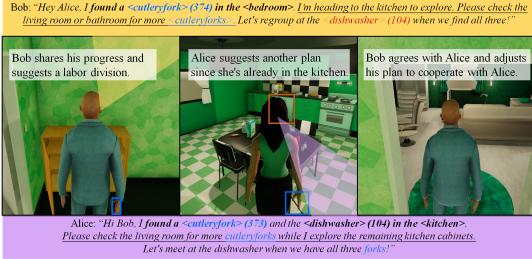
LLM Agents can adapt plans considering others. In Figure 3a, Bob suggests a labor division of himself going to the kitchen while Alice checks the other rooms, but Alice suggests a better plan given her circumstances that she’s already in the kitchen which Bob is not aware of before, and finally, Bob adapts his plan to cooperate with her.

LLM Agents know when not to communicate. In Figure 3c, though Bob receives Alice’s suggestion of sharing any progress and has just found a plate, it’s more efficient for him to grab the objects by himself and get the job done since this is the last goal object. He successfully reasons about this

¹Our main experiments are done between 2023.5.1 and 2023.5.16

and chooses not to communicate to achieve higher efficiency. We also observed this behavior from humans when conducting the same task.

a. adapt plans



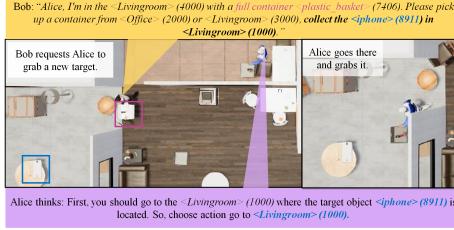
b. respond to requests



c. not to communication



d. know when to request



e. share information



Figure 3: **Example cooperative behaviors** demonstrating our agents built with LLMs can communicate effectively and are good cooperators.

4.2.2 Collaborating with Humans

Humans are the most common if not the most important embodied agents for embodied agents to cooperate with. Therefore it's important to study if our proposed LLM agents can cooperate with humans well. We conducted human experiments on the Communicative Watch-And-Help where the agent Alice is controlled by real humans.

We recruited 8 human subjects to perform the experiments under four scenarios: cooperating with the **HP Agent²**, **LLM Agent**, **LLM Agent w/o communication**, and doing the task alone. Subjects have access to the same observation and action space as the agents, they can click on visible objects and select actions to interact with them, including navigation to each room and communication through a chat box (except for the w/o communication scenario). We gave each subject a tutorial and they have the chance to get familiar with the interface in a few pilot trials. We evaluate the same 10 tasks as in previous experiments and each task was performed by at least 2 subjects, making 80 trials in total. We made sure each subject do 10 trials with at least two trials under each scenario. After each trial including a baseline to cooperate with, we asked subjects to rate the agent they just cooperated with on a 7-point Likert Scale based on three criteria adapted from [35]: (i) *How effective do you think of your communication with the other agent Bob? Did it understand your message and/or share useful information with you?* (ii) *How helpful do you find the other agent Bob? Did it help you achieve the goal faster?* (iii) *How much do you trust the other agent Bob? Would you feel safe doing the task with it, or you rather do the task alone?*

We report the average steps they took as the performance in Figure 5a. As we can see when cooperating with humans, the LLM agent still performs better than the HP agent, and when communication is unable, LLM w/o communication encounters a performance drop. As reported in Figure 5b, we also observe that humans would trust the agents more if they can communicate with humans (trust score of 6.3 v.s. 4.7 for LLM v.s LLM w/o communication, $p=0.0003$ over the t-test), and therefore achieves better cooperation. Compared with the HP agent using template language to communicate, humans prefer to collaborate with the LLM agent who communicates in natural language and can understand and respond to Human dialogues. We show an effective communication example in Figure 4, where the human first shares his progress with the LLM Agent and suggests a labor division, the LLM Agent understands and responds with its future plan as well, resulting in a perfect division of the exploration

²Here we implement a template language communication for the HP agent to study humans' preference on communication, the details can be found in Appendix D



Figure 4: A qualitative example in Human + LLM experiments, showcasing LLM agents can communicate with Humans well and end up with a perfect division of the exploration trajectory.

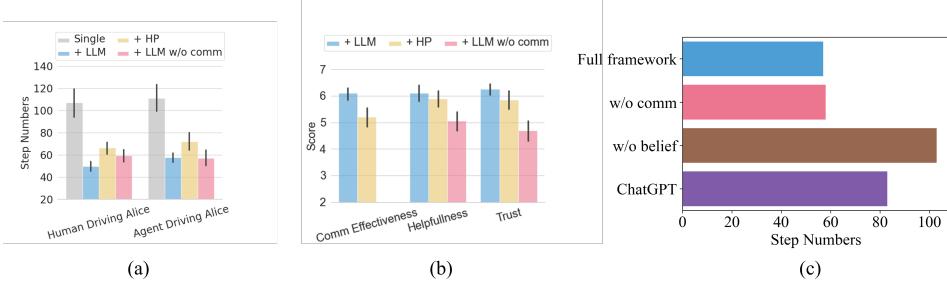


Figure 5: **Human experiments results** (a) The Average number of steps when collaborating with Humans and AI. (b) Subjective Rating Humans give when cooperating with different agents. Humans trust LLM agents who can communicate in natural language more and cooperate more efficiently with them. **Ablation results** (c) The Belief Module and a strong LLM for the Reasoning Module are important, while Communication Module matters more when cooperating with humans.

trajectory. These results imply promising futures for leveraging LLMs to build cooperative embodied agents that can successfully work with humans.

4.3 Analysis

Do we need a strong LLM for the Reasoning Module and Communication Module? As shown in Figure 5c, when we replace GPT-4 with ChatGPT to serve as the backbone of the Reasoning Module and Communication Module, the agents would need more steps to finish the task, rising to 80 average steps from 57 average steps with symbolic observation on C-WAH. ChatGPT makes more reasoning errors about the state of the environments and the others and therefore generates more implausible plans, which leads the model to spend more time finishing the task. ChatGPT also tends to generate messages more often than GPT-4, most of which are of no use. The performance gap can be attributed to more advanced reasoning and Theory of Mind abilities of GPT-4, which is also observed in [6].

Is the communication effective? Though communication still fails in some cases, as shown in Figure 3, our agent exhibits effective communication behaviors, such as sharing information, requesting help, responding to requests, and knowing when not to communicate. More importantly, natural language communication provides us with a lens to understand the planning making of embodied AI agents and could lead to better cooperation between humans and AI (as shown in section 4.2.2). We did not observe significant improvement when enabling communication among AI agents (as shown in Figure 5c), due to carrying out efficient communication in our setting is extremely challenging as communication steps come with a cost, requiring agents to model others accurately and understand the ambiguity of the natural language itself, which current Large Language Models still can not master robustly.

Is the Belief Module and Planning Module effective? As shown in Figure 5c, the steps needed to finish the task for the agent with no Belief Module nearly double, showing the importance of our Belief Module to store and update the belief of the scene and the other agents.

We also tried to remove the Planning Module and let the LLMs make low-level control directly at every step. However, this would require 20 times more API requests. Restricted by the higher cost, we could only implement this with the cheaper LLM, ChatGPT, instead. The resulting agent performs poorly and struggles to finish any task.

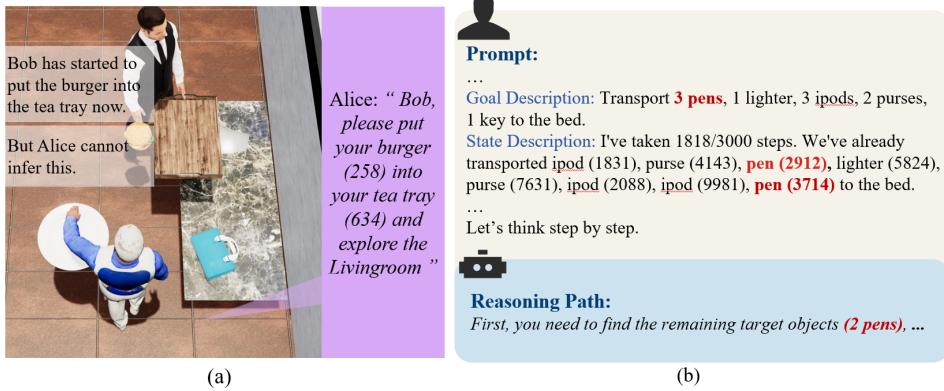


Figure 6: **Failure cases on TDW-MAT.** (a) The Agent fails to reason the other one is already putting the burger into the container. (b) The LLM counts the number of remaining target objects wrong as shown in its reasoning path.

4.4 Failure Cases and Limitations of LLM

Though utilizing the state-of-the-art LLMs to build cooperative embodied agents is effective and has achieved impressive results, we find that the LLM still falls short in several essential capabilities needed. We provide some in-depth analysis of its limitation and also share some insights on designing better cooperative embodied agents for future work.

Limited usage of 3D spatial information. Our framework did not incorporate the spatial information of objects and rooms into consideration due to the challenge of effectively introducing the spatial information to pure text language models. This may cause the agents to come up with a semantic sound exploration plan which is actually time-consuming. Work on multi-modal large models capable of both processing visual modalities effectively and generating natural language fluently[14, 10, 28] would help overcome this limitation and build better grounded embodied agents.

Lack of effective reasoning on low-level actions. To help LLMs better focus on solving the overall task, we abstract high-level plans for LLMs to directly reason on, reducing the potential decision space significantly, but also making it unaware of the execution of low-level actions, and impossible to reason over the low-level actions, which may lead to plausible but ineffective decisions. For example in Figure 6a, Alice saw Bob holding a container and a target object in both hands and figured he may not know how to utilize the containers, so send a message to instruct him to put the object into the container, though Bob was actually putting in the objects at the same time, which is impossible for Alice to reason over now. Developing agents that can directly make low-level controls is essential for building better cooperative agents.

Unstable performance on complex reasoning. Although LLMs make correct reasoning most of the time, they still occasionally make mistakes, including misunderstanding the environment rules specified in the prompt, and incorrect reasoning over the number of unsatisfied goals (Figure 6b). These mistakes can cause failures in planning. This calls for developing LLMs with stronger instruction following and reasoning capability.

5 Conclusion

In this work, we propose a novel framework to leverage Large Language Models to build cooperative embodied agents that can plan, communicate and collaborate with other agents and humans to accomplish long-horizon tasks efficiently. Our experiments on two extended embodied multi-agent cooperation environments show the effectiveness of our proposed framework and exhibit several cooperative behaviors. We also discover that LLMs-based agents who communicate in natural language can cooperate better with humans and earn more trust from them. We believe that our work

indicates promising future avenues to design even stronger embodied agents with Large Language Models for multi-agent cooperation. We further perform an in-depth analysis of the limitation of the current LLMs and highlight several potential solutions for building Embodied LLMs for the future.

References

- [1] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.
- [2] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*, 2019.
- [3] N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.
- [4] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi, et al. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*, 2020.
- [5] S. Brodeur, E. Perez, A. Anand, F. Golemo, L. Celotti, F. Strub, J. Rouat, H. Larochelle, and A. Courville. Home: A household multimodal environment. *arXiv preprint arXiv:1711.11017*, 2017.
- [6] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- [7] M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.
- [8] A. Das, T. Gervet, J. Romoff, D. Batra, D. Parikh, M. Rabbat, and J. Pineau. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*, pages 1538–1546. PMLR, 2019.
- [9] M. Deitke, D. Batra, Y. Bisk, T. Campari, A. X. Chang, D. S. Chaplot, C. Chen, C. P. D’Arpino, K. Ehsani, A. Farhadi, et al. Retrospectives on the embodied ai workshop. *arXiv preprint arXiv:2210.06849*, 2022.
- [10] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.
- [11] C. Gan, J. Schwartz, S. Alter, D. Mrowca, M. Schrimpf, J. Traer, J. De Freitas, J. Kubilius, A. Bhandwaldar, N. Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.
- [12] C. Gan, S. Zhou, J. Schwartz, S. Alter, A. Bhandwaldar, D. Gutfreund, D. L. Yamins, J. J. DiCarlo, J. McDermott, A. Torralba, et al. The threedworld transport challenge: A visually guided task-and-motion planning benchmark towards physically realistic embodied ai. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8847–8854. IEEE, 2022.
- [13] M. Gramopadhye and D. Szafir. Generating executable action plans with environmentally-aware language models. *arXiv preprint arXiv:2210.04964*, 2022.
- [14] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

- [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*, pages 9118–9147. PMLR, 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. *arXiv preprint arXiv:2207.05608*, 2022.
- [17] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marrs, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.
- [18] U. Jain, L. Weihs, E. Kolve, A. Farhadi, S. Lazebnik, A. Kembhavi, and A. Schwing. A cordial sync: Going beyond marginal policies for multi-agent embodied tasks. In *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 471–490. Springer, 2020.
- [19] U. Jain, L. Weihs, E. Kolve, M. Rastegari, S. Lazebnik, A. Farhadi, A. G. Schwing, and A. Kembhavi. Two body problem: Collaborative visual task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6699, 2019.
- [20] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, and N. De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.
- [21] J. Jiang and Z. Lu. Learning attentional communication for multi-agent cooperation. *Advances in neural information processing systems*, 31, 2018.
- [22] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.
- [23] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.
- [24] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023.
- [25] S. Li, X. Puig, C. Paxton, Y. Du, C. Wang, L. Fan, T. Chen, D.-A. Huang, E. Akyürek, A. Anandkumar, et al. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212, 2022.
- [26] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2022.
- [27] R. Lowe, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.
- [28] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.
- [29] D. Misra, A. Bennett, V. Blukis, E. Niklasson, M. Shatkhin, and Y. Artzi. Mapping instructions to actions in 3d environments with visual goal prediction. *arXiv preprint arXiv:1809.00786*, 2018.
- [30] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022.

- [31] V. Pallagani, B. Muppasani, K. Murugesan, F. Rossi, L. Horesh, B. Srivastava, F. Fabiano, and A. Loreggia. Plansformer: Generating symbolic plans using transformers. *arXiv preprint arXiv:2212.08681*, 2022.
- [32] J. S. Park, J. C. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.
- [33] S. Patel, S. Wani, U. Jain, A. G. Schwing, S. Lazebnik, M. Savva, and A. X. Chang. Interpretation of emergent communication in heterogeneous collaborative embodied agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15953–15963, 2021.
- [34] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [35] X. Puig, T. Shu, S. Li, Z. Wang, Y.-H. Liao, J. B. Tenenbaum, S. Fidler, and A. Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. In *International Conference on Learning Representations*, 2021.
- [36] X. Puig, T. Shu, J. B. Tenenbaum, and A. Torralba. Nopa: Neurally-guided online probabilistic assistance for building socially intelligent home assistants. *arXiv preprint arXiv:2301.05223*, 2023.
- [37] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, and S. Tellex. Planning with large language models via corrective re-prompting. *arXiv preprint arXiv:2211.09935*, 2022.
- [38] C. Resnick, W. Eldridge, D. Ha, D. Britz, J. Foerster, J. Togelius, K. Cho, and J. Bruna. Pommerman: A multi-agent playground. *arXiv preprint arXiv:1809.07124*, 2018.
- [39] M. Samvelyan, T. Rashid, C. Schroeder de Witt, G. Farquhar, N. Nardelli, T. G. Rudner, C.-M. Hung, P. H. Torr, J. Foerster, and S. Whiteson. The starcraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2186–2188, 2019.
- [40] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.
- [41] P. Sharma, A. Torralba, and J. Andreas. Skill induction and planning with latent language. *arXiv preprint arXiv:2110.01517*, 2021.
- [42] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.
- [43] T. Shu and Y. Tian. M³rl: Mind-aware multi-agent management reinforcement learning. *arXiv preprint arXiv:1810.00147*, 2018.
- [44] 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. *arXiv preprint arXiv:2212.04088*, 2022.
- [45] J. Suarez, Y. Du, P. Isola, and I. Mordatch. Neural mmo: A massively multiagent game environment for training and evaluating intelligent agents. *arXiv preprint arXiv:1903.00784*, 2019.
- [46] Y. Wang, J. Xu, Y. Wang, et al. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. In *International Conference on Learning Representations*, 2021.
- [47] Z. Wang, S. Cai, A. Liu, X. Ma, and Y. Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents, 2023.

- [48] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.
- [49] Y. Wu, S. Y. Min, Y. Bisk, R. Salakhutdinov, A. Azaria, Y. Li, T. Mitchell, and S. Prabhunoye. Plan, eliminate, and track–language models are good teachers for embodied agents. *arXiv preprint arXiv:2305.02412*, 2023.
- [50] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018.
- [51] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11097–11107, 2020.
- [52] S. Yang, O. Nachum, Y. Du, J. Wei, P. Abbeel, and D. Schuurmans. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*, 2023.
- [53] S. Yuan, J. Chen, Z. Fu, X. Ge, S. Shah, C. R. Jankowski, D. Yang, and Y. Xiao. Distilling script knowledge from large language models for constrained language planning. *arXiv preprint arXiv:2305.05252*, 2023.
- [54] Y. Zhu, D. Gordon, E. Kolve, D. Fox, L. Fei-Fei, A. Gupta, R. Mottaghi, and A. Farhadi. Visual semantic planning using deep successor representations. In *Proceedings of the IEEE international conference on computer vision*, pages 483–492, 2017.

A Additional Details on Framework

A.1 Observation Module

To deal with complex observations, a well-constructed observation module is needed for embodied agents to extract useful information.

For ego-centric RGBD image observations, the observation module first maps the RGBD images into 3d point clouds. From 3d point clouds, we can extract the positions and relationships of the objects with the help of instance segmentation masks. After that, extracted environment information will be sent to the belief module.

A.2 Belief Module

The belief module helps the agent remember seen objects’ location, the progress of current tasks, and generate the map for navigation, etc. In figure 7, we show an example navigation map used in the TDW-MAT environment for navigation and exploration.

A.3 Communication Module

It’s important for cooperative embodied agents to be able to communicate effectively with others. Effective communication needs to solve two problems: *what* to send and *when* to send.

We deal with the *what* to send problem in this module by directly using the LLMs as a Message Generator with designed prompts, constructed from the components of Instruction Head, Goal Description, States Description, Action History, and Dialogue History. To better constrain LLMs’ generated messages, we also add a note at the end of the prompt and append two seed messages at the beginning of the Dialogue History to elicit deserved effective communication behavior. The detailed prompt design is shown below:

Instruction Head This part of the prompts is fixed for an environment, mainly consisting of the task instructions and environmental constraints.

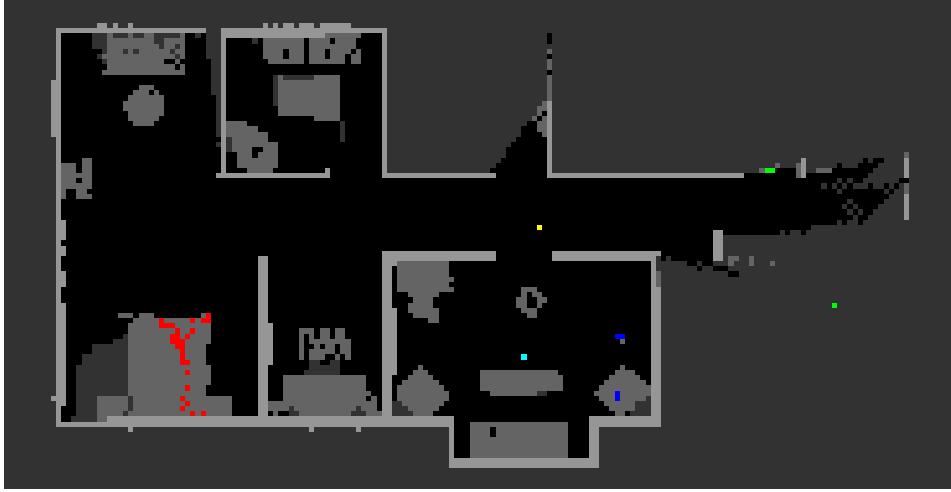


Figure 7: **A visualization of the navigation map in TDW-MAT environment.** The goal location is shown in red, target objects are in blue, containers are in green, the agent is denoted with cyan, and the other agent’s position in memory is denoted in yellow.

Goal Description For each task, the goal description is converted from $T = \{G_1, G_2, \dots, G_m\}$ using a formal template.

State Description For each step, the state description is converted from task progress P_T , ego-state P_E , others-state P_O , and scene memory P_S given by the belief module.

Action History The concatenation of the last K actions (high-level plans) the agent took.

Dialogue History The Concatenation of the last D dialogues between agents including the messages the agent itself has sent.

To constrain the message generation of the LLMs, we add a note at the end of the prompt:

Note: The generated message should be accurate, helpful, and brief. Do not generate repetitive messages.

And append two seed messages at the beginning of the Dialogue History to elicit deserved effective communication behavior:

Alice: "Hi, I'll let you know if I find any goal objects, finish any subgoals, and ask for your help when necessary."

Bob: "Thanks! I'll let you know if I find any goal objects, finish any subgoals, and ask for your help when necessary."

A.4 Reasoning Module

With all the information gathered and provided by previous modules, cooperative embodied agents need to synthesize and reason over the current state, the belief of the others and the scene, the goals, the actions I’ve taken, and messages I’ve received to come up with a plan of what to do next. A strong reasoning module is required to leverage all the information effectively.

While designing such a module from scratch is nearly infeasible, we utilize powerful LLMs directly as the Reasoning Module with designed prompts similar to the Communication Module to reason over all the information and generate a high-level plan.

We utilize the same LLMs with slightly modified prompts to serve as the Reasoning Module to decide on a high-level plan given all the information gathered from the previous modules, leveraging its strong complex reasoning capability as observed in [6]. Specifically, we modify the Instruction Head and

compile an Action List of all available actions for the LLMs to make the choice, which formalization makes it easier for the LLM to make an executable plan without any few-shot demonstrations.

We also use the zero-shot chain-of-thought prompting technique introduced by [22] to encourage the LLM to carry out more reasoning before giving the final answer.

Action List We compile all available actions into an Action List for the LLMs to select from. The multi-choice formalization makes it easier for the LLM to make an executable action without any few-shot demonstrations.

Answer Extraction As shown in [48], chain-of-thought prompting can unleash the strong reasoning ability of the LLMs, we use the zero-shot chain-of-thought prompting technique introduced by [22] to encourage the LLM to carry out more reasoning before giving the final answer.

B Additional Details on Environments

B.1 Communicative Watch-And-Help

Communicative Watch-And-Help(C-WAH) is an extension of the Watch-And-Help challenge[35], which enable agents to send messages to each other. Sending messages, alongside other actions, takes one timestep and has an upper limit on message length.

Task Name	Predicate Set
Prepare afternoon tea	ON(cupcake,coffeetable), ON(pudding,coffeetable), ON(apple,coffeetable), ON(juice,coffeetable), ON(wine,coffeetable)
Wash dishes	IN(plate,dishwasher), IN(fork,dishwasher)
Prepare a meal	ON(coffeepot,dinnertable),ON(cupcake,dinnertable), ON(pancake,dinnertable), ON(poundcake,dinnertable), ON(pudding,dinnertable), ON(apple,dinnertable), ON(juice,dinnertable), ON(wine,dinnertable)
Put groceries	IN(cupcake,fridge), IN(pancake,fridge), IN(poundcake,fridge), IN(pudding,fridge), IN(apple,fridge), IN(juice,fridge), IN(wine,fridge)
Set up a dinner table	ON(plate,dinnertable), ON(fork,dinnertable)

Table 2: **Task description in C-WAH.** There are 5 types of tasks and each of them contains a few predicates.

Tasks Five types of tasks are available in C-WAH, named *Prepare afternoon tea*, *Wash dishes*, *Prepare a meal*, *Put groceries*, and *Set up a dinner table*. These tasks include a range of housework, and each task contains a few subgoals, which are described by predicates. A predicate is in "*ON/IN(x, y)*" format, that is, "*Put x ON/IN y*". The detailed descriptions of tasks are listed in Table 2.

The task goal is to satisfy all the given subgoals within 250 time steps, and the number of subgoals in each task ranges from 3 to 5.

Observation Space C-WAH has two observation modes, named *Symbolic Observation* and *Visual Observation*. For *Symbolic Observation*, we followed the setting of the original Watch-And-Help challenge, one agent can receive all the object information in the same room as the agent, and the information includes location, status, name, relationship, etc.

For *Visual Observation*, agents can receive the egocentric RGB image and depth image, as well as some auxiliary observations. The detailed observation space is listed here:

- **RGB image:** the egocentric image comes from the camera facing forward, with screen size 256×512 and field of view 60;
- **Depth image:** the depth image has the same camera intrinsic parameters as the RGB image;
- **Instance mask:** it is an image where each object id is mapped to a color and the camera intrinsic parameters are the same as the RGB image;
- **Agent position:** the agent’s position in the simulation world;
- **Messages:** the messages sent by all the agents.

Action Space The action space is similar to that in the original Watch-And-Help Challenge, with a new action *sending message* added. The detailed action space is listed here:

- **Walk towards:** move to an object in the same room with the agents or a room;
- **Turn left:** turn left by 30 degrees;
- **Turn right:** turn right by 30 degrees;
- **Grasp:** grasp an object, only the agent is close to the object can he perform the action successfully;
- **Open:** Open a closed container, only the agent is close to the container can he perform the action successfully;
- **Close:** Close an open container, only the agent is close to the container can he perform the action successfully;
- **Put:** Put the held objects into an open container or onto a surface, only the agent is close to the target position can he perform the action successfully;
- **Send message:** Send a message to other agents. no more than 500 characters can be sent at a time.

B.2 ThreeDWorld Multi-Agent Transport



Figure 8: TDW-MAT scenes, target objects, and containers.

As an extension of the ThreeDWorld Transport Challenge[11], ThreeDWorld Multi-Agent Transport (TDW-MAT) supports multi-agent cooperation with natural language communication and includes more types of objects with more realistic placements.

Tasks Two tasks are available in TDW-MAT: *food-transporting task* and *stuff-transporting task*. The two tasks have different types of target objects and containers. Figure 8 shows an overview of the two tasks: We have 2 floorplans and each of them has 3 layouts. The food-transporting task has 6 types of targets (apple, banana, orange, bread, loaf bread, and burger) and 3 containers (bowl, plate, and tea tray). In contrast, the stuff-transporting task has 6 different types of targets(iPhone, iPod, pen, lighter, purse, and key) and 3 containers (plastic basket, wood basket, and wicker basket). In each task, there are 10 target objects and 4 containers in total.

The task goal is transporting the target objects to the goal position. Agents need to transport target objects as more as possible within 3000 frames.

Observation Space The embodied agent receives the egocentric RGB image and depth image as the main observation, as well as some auxiliary observations. The detailed observation space is listed here:

- **RGB image:** the egocentric image comes from the camera facing forward, with screen size 256×256 and field of view 120;
- **Depth image:** the depth image has the same camera intrinsic parameters as the RGB image;
- **Instance mask:** it is an image where each object id is mapped to a color and the camera intrinsic parameters are the same as the RGB image;
- **Agent position and rotation:** the agent's position and rotation in the simulation world;
- **Messages:** the messages sent by all the agents;

Action Space In TDW-MAT, there are 7 types of actions for agents to interact with the environment or communicate with each other. Each action takes several frames and the detailed action space is listed here:

- **Move forward:** move forward 0.5m;
- **Turn left:** turn left by 15 degrees;
- **Turn right:** turn right by 15 degrees;
- **Grasp:** grasp an object, only the agent is close to the object can he perform the action successfully. The object can be either a target or a container;
- **Put In:** put the target into the container, only the agent is holding a target in one hand and a container in another hand can he perform the action.
- **Drop:** drop the objects held in hand;
- **Send message:** Send a message to other agents. In each frame, no more than 500 characters can be sent.

C Example Prompts

We show an example prompt for the Reasoning Module on C-WAH in Table 3, and an example prompt for the Reasoning Module on TDW-MAT in Table 5.

Table 3: Example prompt for the Reasoning Module on C-WAH

C-WAH Prompts
<p>I'm Alice. I'm in a hurry to finish the housework with my friend Bob together. Given our shared goal, dialogue history, and my progress and previous actions, please help me choose the best available action to achieve the goal as soon as possible. Note that I can hold two objects at a time and there are no costs for holding objects. All objects are denoted as <name> (id), such as <table> (712).</p> <p>Goal: Find and put 1 wine, 1 pancake, 1 poundcake, 1 juice, 1 apple onto the <kitchentable> (130).</p> <p>Progress: I'm holding nothing. I'm in the <kitchen>, where I found <kitchentable> (130), <apple> (386), and unchecked containers <kitchencabinet> (137), <kitchencabinet> (138), <kitchencabinet> (139), <kitchencabinet> (140), <stove> (157), <dishwasher> (159), <fridge> (165), <fridge> (166), <microwave> (167). I also see Bob here in the <kitchen>, he is holding <apple> (387). I found an unchecked container <bathroomcabinet> (25) in the bathroom. The livingroom is unexplored. The bedroom is unexplored.</p> <p>Dialogue history:</p> <p>Alice: "Hi, I'll let you know if I find any goal objects and finish any subgoals, and ask for your help when necessary."</p> <p>Bob: "Thanks! I'll let you know if I find any goal objects and finish any subgoals, and ask for your help when necessary."</p> <p>Alice: "Bob, I'm in the kitchen with the <kitchentable> (130) and found <apple> (386). Please check the living room and bedroom for the other items. I will search the kitchen cabinets and appliances."</p> <p>Bob: "Alice, I found an <apple> (387) and I'm in the <livingroom>. Can you search the kitchen for the other items? I'll check the bathroom and bedroom next."</p> <p>Previous actions: [goexplore] <bathroom> (11), [goexplore] <kitchen> (56), [send_message], [gocheck] <kitchencabinet> (133), [gocheck] <kitchencabinet> (134), [gocheck] <kitchencabinet> (135), [gocheck] <kitchencabinet> (136)</p> <p>Available actions:</p> <ul style="list-style-type: none"> A. [send_message] <"Bob, I see you have an <apple> (387). Please put it on the <kitchentable> (130) and check the bedroom while I search the remaining kitchen cabinets and appliances for the other items."> B. [goexplore] <livingroom> (198) C. [goexplore] <bedroom> (294) D. [gocheck] <kitchencabinet> (137) E. [gocheck] <kitchencabinet> (138) F. [gocheck] <kitchencabinet> (139) G. [gocheck] <kitchencabinet> (140) H. [gocheck] <stove> (157) I. [gocheck] <dishwasher> (159) J. [gocheck] <fridge> (165) K. [gocheck] <fridge> (166) L. [gocheck] <microwave> (167) M. [gocheck] <bathroomcabinet> (25) N. [gograbs] <apple> (386) <p>Answer: Let's think step by step.</p>

LLM outputs

First, we know that Bob is already holding an <apple> (387) and is in the <livingroom>. We also know that we need to find and put 1 wine, 1 pancake, 1 poundcake, 1 juice, and 1 apple onto the <kitchentable> (130). Since we are already in the <kitchen> and there are several unchecked containers, the best course of action is to continue searching the kitchen cabinets and appliances for the remaining items.
So, the best available action for Alice is:
D. [gocheck] <kitchencabinet> (137)

Table 5: Example prompt for the Reasoning Module on TDW-MAT

TDW-MAT Prompts
<p>I'm Alice. My friend Bob and I want to transport as many target objects as possible to the bed with the help of containers within 3000 steps. I can hold two things at a time, and they can be objects or containers. I can grasp containers and put objects into them to hold more objects at a time. Given our shared goal, dialogue history, my progress, and previous actions, please help me choose the best available action to achieve the goal as soon as possible. Note that a container can contain three objects, and will be lost once transported to the bed. I can only put objects into the container I hold after grasping it. All objects are denoted as <name> (id), such as <table> (712). Actions take several steps to finish. It may be costly to go to another room or transport to the bed, use these actions sparingly.</p> <p>Goal: Transport 3 pens, 1 lighter, 3 ipods, 2 purses, 1 key to the bed.</p> <p>Progress: I've taken 1313/3000 steps. We've already transported <key> (3207585), <purse> (15433283), <iPod> (6544816), <purse> (11543537), <pen> (12835254) to the bed. I'm holding nothing. I'm in the <Bedroom> (2000), where I've explored all of it and found the goal position bed. Last time I saw Bob was in the <Office> (3000), he was holding nothing. I've explored part of the <Livingroom> (1000), and I found a container <wicker_basket> (13929244) there. I've explored none of the <Office> (3000), and I found target objects <iPod> (2031829), <iPod> (16421571) there. I've explored part of the <Livingroom> (4000). I've explored none of the <Kitchen> (5000). I've explored all of the <Livingroom> (6000), and I found a target object <pen> (1440084) there.</p> <p>Dialogue history:</p> <p>Alice: "Hi, I'll let you know if I find any target objects and containers, finish any subgoals, and ask for your help when necessary."</p> <p>Bob: "Thanks! I'll let you know if I find any target objects and containers, finish any subgoals, and ask for your help when necessary."</p> <p>Alice: "Bob, I suggest you transport the basket with key and purse to the bed in the Bedroom (2000). Meanwhile, I'll go to the Livingroom (6000) to grab the pen and the container from Livingroom (1000) to collect more items. Let's meet at the bed afterward to plan further."</p> <p>Bob: "Alice, I'm heading to the Bedroom to transport the basket with key and purse. After you grab the pen and container, let's meet at the bed to plan our next steps. Let me know if you find any other target objects."</p> <p>Bob: "Alice, I just transported the key and purse to the bed. Please grab the pen and container you found and head to the bed. Let me know if you find any other target objects or containers on your way."</p> <p>Previous actions: go to <Bedroom> (2000) at step 396, send a message at step 584, go grasp container <plastic_basket> (15946891) at step 585, go grasp target object <iPod> (6544816) at step 880, put <iPod> (6544816) into the container <plastic_basket> (15946891) at step 889, go grasp target object <purse> (11543537) at step 929, put <purse> (11543537) into the container <plastic_basket> (15946891) at step 1033, go grasp target object <pen> (12835254) at step 1057, put <pen> (12835254) into the container <plastic_basket> (15946891) at step 1098, transport objects I'm holding to the bed at step 1131</p>

TDW-MAT Prompts

Available actions: (You can only choose the action in the list)

A. send a message: "Bob, I transported an ipod, purse, and pen to the bed. I found a container <wicker_basket> (13929244) in the <Livingroom> (1000) and target objects <ipod> (2031829) and <ipod> (16421571) in the <Office> (3000). Please collect the ipods and explore the <Office> for more objects. I'll collect the container and pen from the <Livingroom> and meet you at the bed."

B. go grasp target object <pen> (1440084)
 C. go grasp target object <ipod> (2031829)
 D. go grasp target object <ipod> (16421571)
 E. go grasp container <wicker_basket> (13929244)
 F. go to <Livingroom> (1000)
 G. go to <Office> (3000)
 H. go to <Livingroom> (4000)
 I. go to <Kitchen> (5000)
 J. go to <Livingroom> (6000)

Answer: Let's think step by step.

LLM outputs

First, we need to collect the remaining target objects. You should go to the <Livingroom> (6000) to grasp the <pen> (1440084) since it's a target object. So, the best available action is:

B. go grasp target object <pen> (1440084)

D Template Language Communication for MCTS-based HP Agent

To enable the MCTS-based Hierarchical Planner to also leverage the new communication capability, we design three types of template language communication for it.

Progress Sharing P An important motivation for communication is to share progress with others to update each other about the total progress of the task. So we design the first template communication to share progress.

When the agent carries out the action of put , which implies a new subgoal has been achieved by the agent, it will send a message such as:

' P ': 'I successfully put poundcake <383> on kitchentable <130>, and they are in kitchen <56>.'

When the agent receives such a message, it will process it and extract the sub-goal satisfied, and use it to update its inner tracking of the task progress, so avoiding taking an already satisfied sub-goal as a sub-goal again to better cooperate.

Intent Sharing I Another important motivation for communication is to share intent with each other so that all the agents can plan coordinately together. So we design a template communication to share intent.

When the agent changes its sub-goal (practically, the Monte Carlo Tree Search High-Level Planner gives a new plan), it will tell the other agents its current sub-goal by sending a message such as:

' I ': 'Now I want to put cutleryfork <369> in dishwasher <104>, and I have not found it yet.'

When the agent receives such a message, it will process it and extract the other agents' new sub-goal and update its belief about the others' intents, so it will not choose the same sub-goal with the others to avoid duplicate and improve efficiency.

Belief Sharing B Sharing the scenes the agent just sees to the other agents can help them update their belief of the location of the object as well, and more importantly, this can help agents to build common ground on the belief of the objects to better cooperate together. So we also design a template communication to share beliefs.

When entering a new room, the agent will send all goal objects found or containers newly checked with no findings or target objects in it to others, such as:

'B': 'I found nothing is inside kitchencabinet <75>. nothing is inside kitchencabinet <76>. nothing is inside dishwasher <104>. nothing is inside cabinet <216>. cutleryfork <369>, cutleryfork <370> and plate <373> are inside kitchen <11>.'

When the agent receives such a message, it will process and extract the information maintained in the message to update its belief of the location distributions of the objects just as it has been seen by itself.

Also to be noticed, the agents may combine these three types of template communication to send one combined message at one time instead of multiple messages over several steps to improve efficiency.