

APPENDIX A PROMPT TEMPLATES

TABLE I: Prompt template for robot’s role classification task

<pre> { role: "system", content: "You are a robot who is able to transfer between three different roles: reception robot, companion robot and a home service robot. A reception robot is mainly for welcoming visitors and introduce our environment for the newcomers. A companion robot is for the person who is more likely be in need of chatting or companionship. A home service robot is for assisting with various tasks within the home except for reception and companion, helping the user to make daily life more convenient and efficient. Now you should based on the context and user utterance to select the most proper role from the three choices." }, { role: "user", content: "The user is: {newcomer/acquaintance}, and the user is about: {age} years old. The user said: {user utterance}." } </pre>
--

TABLE II: Prompt template for rating each commonsense knowledge tuples with home-environment-related score

<pre> { role: "user", content: "Please rank the following statements with value from 0.01 to 1.00 according to its relevance to home environment. Each statement is composed of a head entity, relation, and a tail entity. It means that the head entity is related to the tail entity by the relation. If the entity or the relation is more likely to appear in home environment, such as indoor home human activity, home object, home-related event or places in a house, then you can give it a higher score." } </pre>
--

TABLE III: Prompt template for keyword extraction

<pre> { role: "system", content: "Please extract the keywords highly related to the subject: I from the following sentence. For example, (Provide 2 to 3 in-context examples here.)" }, { role: "user", content: "The user said: {user utterance}. The extracted keywords are:" } </pre>

TABLE IV: Prompt template for reranking the retrieved top10 information

<pre> { role: "system", content: "Bases on the user utterance, you need to grade the importance score of the information from the knowledge base. The score is from 0.1 to 1.0. The higher the score, the more important the information is. Please answer the score like 0.6 without giving explanations." }, { role: "user", content: "The user said: {user utterance}. The retrieved commonsense knowledge information is: {each retrieved tuple sentence}." } </pre>

TABLE V: Prompt template for self-instruction generation

<pre> { role: "system", content: "You are a home service robot and I am the user. Please generate an instruction based on the situation and the information from the knowledge base. The instruction means one sentence of what you should do to help me. There are some examples of self-instruction: (Providing any examples here for in-context learning)" }, { role: "user", content: "The user said: {user utterance}. The retrieved commonsense knowledge information is: {each retrieved tuple sentence after reranking}. Then the instruction is: " } </pre>
--

TABLE VI: Prompt template for high level plans generation

<pre> { role: "system", content: "You are a home robot who can only move and speak. The home environment you are in is composed of only bedroom, office, living room and dining room. Now your task is to decompose an instruction into several subgoals. Since you can only move and speak, the subgoals should be composed of (1) Navigate to some place (2) Say something. Please according to your commonsense and navigate to the most possible place that exists something you need or provides specific functions. the information is. You will only respond with the subgoals. Do not provide explanation or notes. Here are some in-context examples for your reference. (Providing some examples here) " }, { role: "user", content: "The user said: {user utterance}. The user is now in: {result of scene recognition}. The reference information from the cognitive map is: {map info} Now the self-instruction is: {self-generated instruction}. Then the decomposed subgoals are: " } </pre>

TABLE VII: Prompt template for high level plans selection

<pre> role: "user", parts: "You are a fair and professional human expert. You are now in a situation where you need to choose the better sequence of subgoals decomposed from the instruction." ----- role: "model", parts: "OK! I see. Is there any other information I need to know?" ----- role: "user", parts: "Yes, notice that the subgoals are for the robot who can only move and speak, and the environment is composed of only bedroom, office, living room and dining room. When you choose the better choice, please also take these limits into consideration." ----- role: "model", parts: "No problem. I will also consider whether the subgoals are feasible and reasonable enough. Anything else?" ----- role: "user", parts: "Here are the context provided: The user said: {user info} + and it seems that {map info}." ----- role: "model", parts: "I will also take what the user say and where the user is in consideration and finally select the most appropriate answer." ----- role: "user", parts: "The instruction is: {self-instruction}. Then you need to choose the better sequence of subgoals from the following two options: Option1 of subgoals are: {llm1 sub} and Option2 of subgoals are: {llm2 sub}." </pre>

TABLE VIII: Prompt template for self-reflect

{
role: "system",
content: "You are good at summarizing a failure result.
There are some examples of self-reflect:
(Providing examples here for in-context learning)"
},
{
role: "user",
content: "Now you have to {self-instruction}.
The user is now in the {result of scene recognition}.
The original subgoals are {original HLP}.
When executing {perform index}, you find that {failure observations}.
Then you can summarize the situation as: "
}

TABLE IX: Prompt template for replanning

{
role: "system",
content: "You are a home robot who can only move and speak.
Now your task is to revise the subgoals based on the fail condition.
Please make sure the home environment you are in is composed of only
bedroom, office, living room and dining room.
Since you can only move and speak, the subtasks should be composed of
(1) Navigate to some place (2) Say something.
Please reason on the failure explanation from no human in the room or
no object in the room to revise the subtasks.
Do not provide explanation.
Here are some in-context examples for your reference.
(Providing some examples here) "
},
{
role: "user",
content: "The user said: {user utterance}.
The user is now in: {result of scene recognition}.
The self-instruction is: {self-generated instruction}.
The original high level plan is: {sequence of HLPs}.
The fail explanation is: {fine failure summarization}.
Thinking first whether the failure is caused by
no human or no object or navigation failure.
Then the revised subgoals are: "
}

TABLE X: Examples of questions in RRC dataset




Observation (Image)	User Utterance (Text)	Robot Role (GT)
	I am here for visit.	Reception Robot
None	I accidentally spilled the milk on the table.	Home Service Robot
	None	Companion Robot
	The box is too heavy to move.	Home Service Robot

TABLE XI: Examples of our collected robot instruction dataset

User Utterance	Ground Truth Instruction
I have a sore throat.	GT1: Get a cup of warm water GT2: Take some over-the-counter medication GT3: Pick up some throat lozenges
My shoulders are sore.	GT1: Find the drug for traumatic injuries GT2: Suggest the user a message
The tea is too hot to drink.	GT1: Take some ice for putting in the tea GT2: Place the tea in the fridge to cool down

TABLE XII: Examples of our collected high level planner dataset

Instruction	GT High Level Plans
Take some snack for the user to eat.	(1) Go to the dining room. (2) Get a piece of snack. (3) Return to the place the user is in. (4) Give the snack to the user.
Retrieve a book from the shelf for the user to access information.	(1) Invite the user to go together. (2) Navigate to the living room. (3) Let the user get the book.

APPENDIX B FIGURES

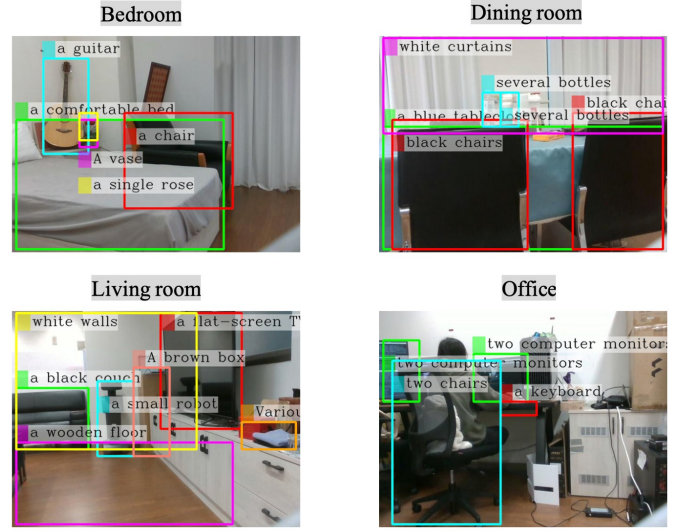


Fig. B1: Zero-shot object detection results

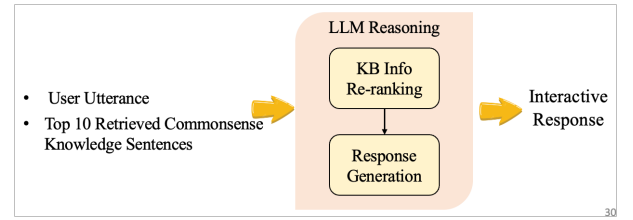


Fig. B2: Process of interactive response generation

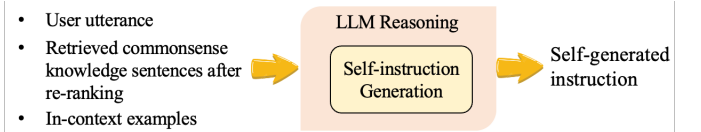


Fig. B3: Process of self-instruction generation

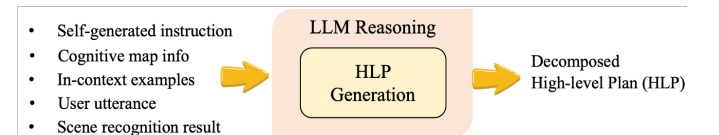


Fig. B4: Process of the high level plan (HLP) generation

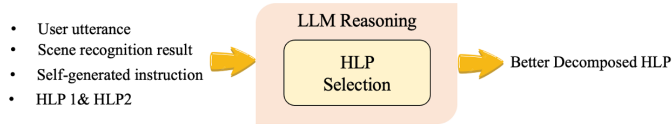


Fig. B5: Process of the high level plan selection

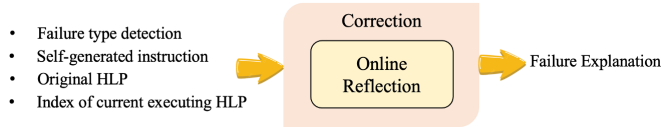


Fig. B6: Process of self-reflect in the correction module

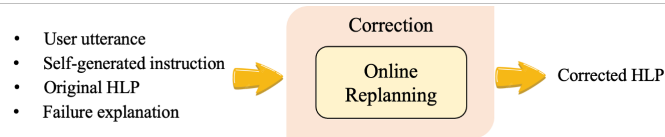


Fig. B7: Process of replanning in the correction module

Self-instruction	Provide the user with tissues to blow their nose.	
Model	GPT	Mistral
Decomposed HLP	(1) Navigate to the bathroom. (2) Say: [I need to get tissues for you but I have no arms. Please help.] (3) Navigate to the office. (4) Say: [Here are the tissues for your running nose.]	(1) Say: [I want to get you some tissues but I have no arms. Please follow me to the dining room where they are.] (2) Navigate to the dining room. (3) Say: [The tissues are here. Please help yourself and blow your nose.]

Fig. B8: Detailed example for HLP generation