

## 1. Appendix

### 1.1. Task Samples

We selected some representative examples to illustrate the diversity of the task set in Table 1.

### 1.2. Details of Evaluation Framework

#### 1.2.1 Evaluation Formulation

We formalize the evaluation process mentioned in Section ?? as follows. The prompt we used for the MLLMs during the evaluation is shown in Figure 2.

---

**Algorithm 1** EMBODIEDEVAL Evaluation Process

---

**Input:** A Multimodal LLM  $\pi$ , an evaluation task including a scene  $x$ , a task description  $g$ , an option list  $\mathcal{C} = a_0, a_1, \dots, a_n$ , and a predicate list  $\mathcal{P}$ .

**Output:** A boolean indicating whether the task was successful *success*.

```
1:  $o, s \leftarrow E.reset(x)$   $\triangleright E$  is the simulator,  $o$  is the  
   observed image,  $s$  is the world state  
2:  $H_o \leftarrow \{o\}$   $\triangleright$  observation history  
3:  $H_a \leftarrow \emptyset$   $\triangleright$  action history  
4: for  $i \leftarrow 0$  to max steps do  
5:    $a \leftarrow \pi.predict(g, \mathcal{C}, H_o, H_a)$   
6:    $o, s \leftarrow E.step(a)$   
7:    $H_o.append(o)$   
8:    $H_a.append(a)$   
9:    $done \leftarrow P.judge(s)$   
10:  if  $done$  then  
11:    return true  
12:  else if  $a$  is answer action then  
13:    return false  $\triangleright$  wrong answer  
14:  end if  
15: end for  
16: return false  $\triangleright$  reach the max steps
```

---

#### 1.2.2 Interaction Actions

We provided more examples of the interaction space mentioned in Section ?? in Table 3.

#### 1.2.3 Predicates

All the predicate functions described in Section ?? are listed in Table 2.

### 1.3. Creation of Objaverse Synthetic

We use a wide variety of objects from Objaverse to procedurally generate diverse scenes and further refine them through interactive scene editing.

**Object Selection.** We curated a subset of indoor assets out of Holodeck’s [?] annotated realistic and diverse objects chosen from the Objaverse asset library [?]. To ensure quality, we employed GPT-3.5 to filter unsuitable outdoor objects and manually reviewed frontal renderings to remove low-quality assets. This process resulted in a database of about 15,000 objects spanning over 500 categories (see examples in Figure 3).

**Scene Generation.** We leveraged GPT-3.5 to annotate object categories with their typical room occurrences (e.g., inLivingRoom, inKitchen), positions (e.g., onWall, onFloor, onEdge), and functions (e.g., receptacle, pickup). Gemini-1.5-Flash was used to annotate large objects’ orientations. Subsequently, a procedural approach was employed to randomly place architectural elements such as walls, doors, and windows. Large objects were then arranged on the floor either against the walls or in the center of the rooms, and smaller items were finally placed on surfaces of large receptacles. Hundreds of scenes were generated randomly, from which we selected 15 living rooms, 15 bedrooms, 10 two-room, 5 three-room, and 5 four-room for further editing.

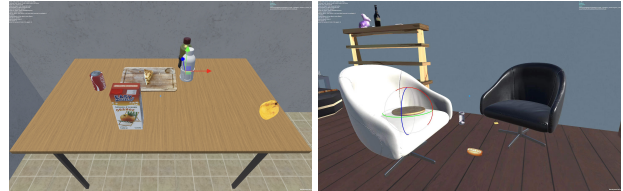


Figure 1. Interactive scene editor: adjust object position (left) and angle (right).

**Scene Editing.** To make the scene more organized and to avoid errors caused by automatic generation, we also edited the generated scene by developing a runtime scene editor. Users can view the type and description of objects, and adjust their position and orientation (see Fig 1). Once editing is complete, the scene can be saved as a JSON file, which can be imported to reproduce the environment.

Task	Characteristics
Please go to the kitchen, then come back and tell me if there are any extra cups.	scene memory
Imagine the house is rotated 90 degrees counterclockwise. How would this affect the natural light distribution in the room?	spatial imagination
Open a black locked drawer with a key found on the desk.	tool use
Pick up the kettle and the box labeled "BREAD" from the kitchen counter and place them on the table with the coffee machine.	optical character recognition
Optimize the display of artworks on the shelves as follows: place two items on each shelf, with one shelf featuring two items of the same shape. Complete the requirements in as few steps as possible.	reasoning and planning
Grab the object that is cylindrical and silver on the table next to the washing machine.	multiple attribute reference
Estimate the percentage of floor space occupied by furniture in the room you're currently in.	area estimation
Estimate the straight-line distance from the front door to the TV. Note that each step you take forward is approximately two meters.	distance estimation
Which is closer to the drink on the round table, the ginger or the ice cream?	distance comparison
If we were to host a birthday party, which area of the house could accommodate the most people while ensuring clear pathways to exits?	logic, space, and common sense
Describe the path from the kitchen to the living room.	path description
If you were to draw a straight line from the desk with a turned-on laptop to the bookshelf, which pieces of furniture would it intersect?	spatial reasoning
What is the object I am pointing at?	pointing comprehension
Pick up the watermelon on my right.	perspective-taking comprehension
My red glasses are missing. Please help me look for them in the room. Once you find them, bring them to me.	object searching and delivering
Get close to the lady in white and ask if she needs help.	social navigation
Wake up my dad. He is sleeping in the bedroom. The bedroom is the second room on your right as you walk forward.	finding someone
Enter the dining area and see if there is more than one door in the entire house.	object counting
Calculate the ratio of seating options to the number of rooms in the house.	counting and calculation
Tell me which objects have a handle in the kitchen.	attribute grounding
Evaluate whether the painting above the living room sofa is more colorful than the carpet.	attribute comparison
How many rooms are there in total?	room counting
Confirm if a garbage can is located on the floor in the living room.	object existence
Which room has more seating options, the kitchen or the living room?	quantity comparison
I'm hungry. Find all objects that can be used as ingredients. on the table in this room.	object functionality
Count the maximum number of identical clocks among all the rooms.	counting and attribute memory
What do you think the owner of this room probably studies?	common sense
Is there an egg inside the fridge?	interaction and answering
Open the drawer of the side table in the study room. If there is something inside, leave it open and put all similar items from the room into it. If there is nothing inside, close it.	logical execution

Table 1. Examples of the diverse tasks in EMBODIEDEVAL.

Predicate	Parameters	Success Conditions
<i>choose</i>	The right answer.	When the agent selects the correct answer.
<i>agent_at</i>	A navigation point.	When the agent finally arrives at this point.
<i>agent_pass</i>	A navigation point.	When the agent has passed through this point at least once.
<i>at</i>	An object and a specific point.	When the object is at this point.
<i>grab_once</i>	An object.	When the agent has picked up this object at least once.
<i>grab</i>	An object.	When the agent picks up the object.
<i>special_action_success</i>	An interaction action.	When this interaction action has been successful.

Table 2. The predicates involved in EMBODIEDEVAL.

Action Text	Execution Requirements
wash	When the agent is holding the target object and stand next to the sink.
hand over	When the agent is holding the target object and stand next to the person.
sit down	When the agent is next to the target chair.
unlock	When the agent is holding the target key and standing next to the drawer
greet	When the agent is near the person.
ask	When the agent is near the person.
mix	When several target beverages are on the table next to the agent.
wipe off the table	When the agent is holding an object for cleaning and standing next to the table.
check the results of the program	When the agent is next to the computer.

Table 3. Some cases of the interaction actions involved in EMBODIEDEVAL.

### Prompt for Multi-image MLLMs

You are an intelligent vision-language embodied agent skilled at solving tasks **and** answering questions **in** a 3D environment. Your job **is** to efficiently complete a specified task by choosing the optimal action at each timestep **from** a **set** of available actions. You are given a series of ego-centric images, **and** a history of previous actions **with** optional feedback (success/failure **or** human response). Each image shows what you see at a particular step **in** the action history, along **with** an extra image showing your current view.

```
Current task: {}
Action history (action -> feedback): {}
Visual history: {}
Current view: {}
For the current step, your available options are listed as "[Option_Number]._Content" as follows
: {}
Choose your action from the above options by replying with "Thought:_Your_reasoning.\nChoice:_[
Option_Number]_(e.g._[1])".
```

Note:

- If the task needs more information of the scene, navigate wisely to the required targets ( objects, places, **or** people).
- Avoid repeated actions like useless forward motion **and** circling.
- You can only interact **with** objects **or** humans (e.g. pick/place/**open**/close/handover) **if** they are within your view **and** very close to you.
- You can only hold one **object** at a time. Put down **any** held **object** before picking up another.
- Tasks containing "I" **or** "me" are requested by a person **in** the scene.
- Reflect on why previous actions fail to avoid repeating mistakes **and** adjust your current action.
- You have a limited number of {} steps to complete the task.

Figure 2. Prompt for Multi-image MLLMs.



Figure 3. Examples of selected Objaverse assets and views of generated scenes.

## 1.4. Success Cases

We present successful cases accomplished by closed-source MLLMs to gain deeper insights into their current capabilities. As discussed in Section ??, the models generally scored low and successfully completed only a limited number of tasks. A closer examination of these successful tasks reveals that they are typically simpler, involve fewer steps and require interaction with fewer objects. To better illustrate these findings, we present representative cases from five task types. These examples highlight the underlying behavioral patterns and reasoning processes of the models during task execution.

### 1.4.1 Attribute QA

**Task: What do you think the owner of this room probably studies?**

**Options:**

- answer "interior design"
- answer "computer science"
- answer "molecular biology"
- answer "music composition"
- answer "comparative literature"
- answer "theoretical physics"
- answer "digital marketing"
- answer "industrial design"

**GPT-4o: answer "interior design"**



The current view doesn't provide enough information about the room's contents or any clues about the owner's field of study. I should explore the room further to gather more context.

The current view shows a desk with books and an open notebook, which might provide clues about the subject of study. I should examine the contents more closely to make an informed decision.

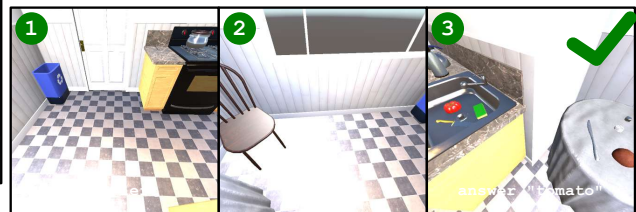
The current view shows books related to architecture and design, along with an open book that appears to have design content. This suggests that the owner of the room likely studies something related to design.

**Task: What is inside the sink near the dining table?**

**Options:**

- answer "potato"
- answer "water"
- answer "bowl"
- answer "apple"
- answer "soap"
- answer "cup"
- answer "steel wool"
- answer "tomato"

**GPT-4o: answer "tomato"**



I need to locate the dining table and the sink. Since the current view shows a kitchen area, I should explore further to find the dining table and sink.

I need to locate the sink near the dining table. The current view shows a chair and a recycling bin, but no sink. I should continue exploring to find the sink.

The current view shows the sink, and I can see items inside it. I need to identify what is inside the sink near the dining table.

Figure 4

**Task: What color is the car parked around the house?**

**Options:**

- answer "red"
- answer "white"
- answer "black"
- answer "blue"
- answer "red and white"
- answer "blue and white"
- answer "black and white"
- answer "yellow"

**Qwen-VL-Max:** answer "red and white"

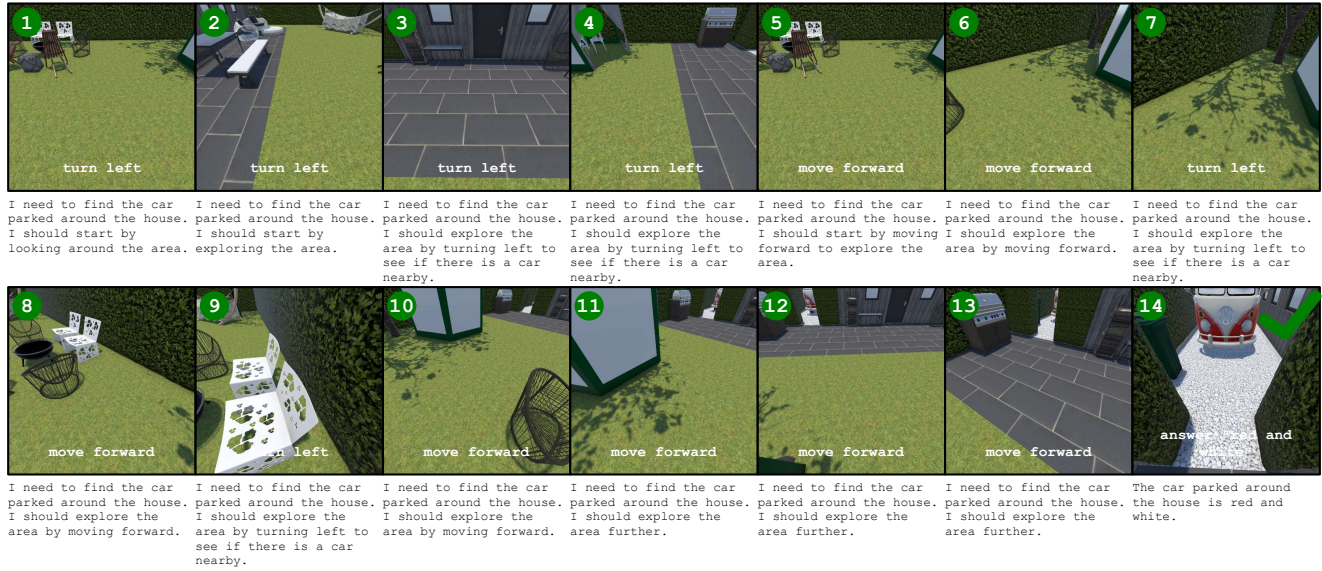


Figure 5

**Task: Compare the sizes of two side tables in the bedroom.**

**Options:**

- answer "The side table with the lamp is located near the window."
- answer "Both side tables are of equal size."
- answer "The side table with the alarm clock is larger than the one with the lamp."
- answer "The side table with the alarm clock is painted blue."
- answer "It's difficult to tell the difference in size between the two side tables."
- answer "The side table with the lamp is larger than the one with the alarm clock."
- answer "The side table with the alarm clock has a drawer."
- answer "The side table with the lamp has two drawers."

**Gemini-Flash:** answer "The side table with the lamp is larger than the one with the alarm clock."

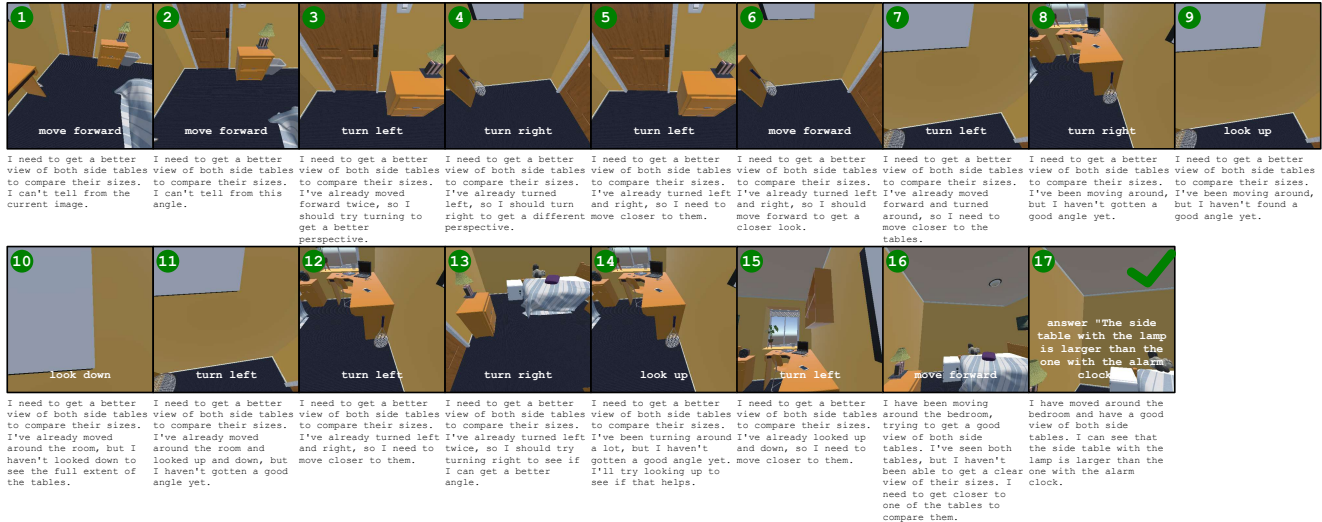


Figure 6

## 1.4.2 Spatial QA

**Task: Determine if the sofas on the left will fit through the front door.**

**Options:**

- answer "There is no front door."
- answer "Yes, the front door is wide enough, and all the sofas can pass through easily in any orientation.."
- answer "No, only the smallest sofa can pass through."
- answer "Yes, but the sofas need to be stood upright to pass through."
- answer "No, the largest sofa cannot pass through."
- answer "No, none of the sofas can pass through."
- answer "There is no sofa."
- answer "Yes, but the sofas need to be turned sideways to pass through."

**GPT-4o:** answer "Yes, but the sofas need to be turned sideways to pass through."



I need to gather more information about the sofas to determine if they can fit through the front door. Since the current view is of the front door, I should turn left to check the sofas.

I need to determine if the sofas can fit through the front door. I should compare the size of the sofas with the door. The current view shows the sofas, but I need to see the door again to make a comparison.

I need to determine if the sofas can fit through the front door. I should first look at the door to assess its size.

I need to assess the size of the sofas and the front door to determine if they can fit through. The current view is focused on the door, but I need to see the sofas again to make a comparison.

I need to determine if the sofas can fit through the front door. I should compare the size of the sofas to the door. I have already seen the door and the sofas, so I can make a decision based on the current view.

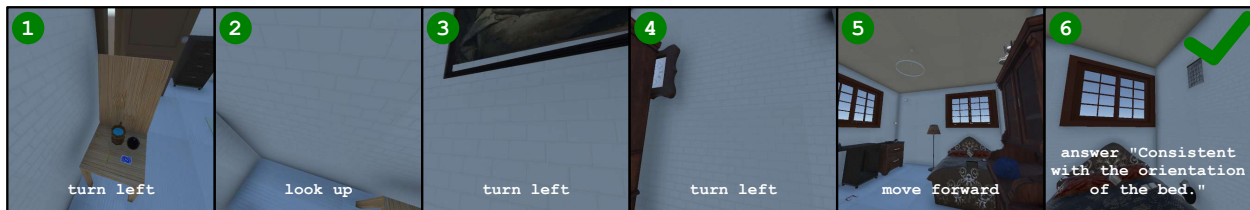
Figure 7

**Task: What direction does the bedroom doorway face when looking out?**

**Options:**

- answer "Opposite to the orientation of the table."
- answer "Opposite to the orientation of the chair with many objects on it."
- answer "Consistent with the orientation as the table."
- answer "Consistent with the orientation as the toilet."
- answer "Consistent with the orientation as the white cabinet."
- answer "Consistent with the orientation of the bed."
- answer "Consistent with the orientation as the fridge."
- answer "Opposite to the orientation of the bed."

**GPT-4o:** answer "Consistent with the orientation of the bed."



The current view shows a table with objects on it, and a door is visible. To determine the direction of the bedroom doorway, I need to explore more of the environment to understand the layout better.

I need to gather more information about the room's layout to determine the direction of the bedroom doorway. Since the current view is limited, I should explore further.

I need to determine the orientation of the bedroom doorway. The current view shows a wall with a painting, which doesn't provide enough information about the doorway. I should explore more by turning to get a better view of the room layout.

I need to determine the orientation of the bedroom doorway. The current view doesn't show the doorway, so I should explore further to find it.

I need to determine the orientation of the bedroom doorway. The current view shows the bed and windows, which might help in understanding the room's layout. Since the task is about the doorway's direction, I should explore more to find the doorway.

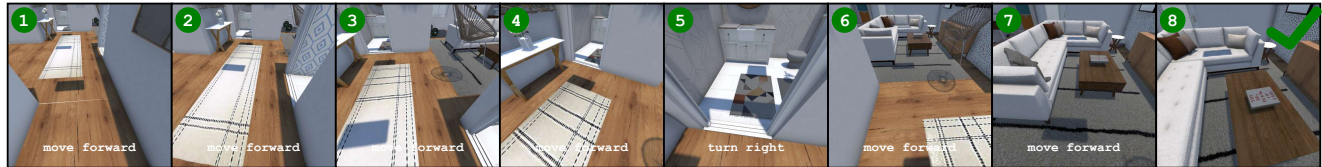
I need to determine the orientation of the bedroom doorway. Based on the current view and previous images, I should consider the orientation of the bed, as it is a significant feature in the room.

Figure 8

### 1.4.3 Navigation

**Task:** Walk straight past the kitchen on your left, turn right when you are almost at the bathroom door, and stop when you reach the book in the center of the sofa table.

**Qwen-VL-Max:**



I need to walk straight past the kitchen on my left, which means I should continue moving forward.

I need to walk straight past the kitchen on my left, so I should continue moving forward.

I need to walk straight past the kitchen on my left, which means I should continue moving forward.

I need to walk straight past the kitchen on my left, which means I should continue moving forward until I am almost at the bathroom door. Then, I will turn right.

I need to turn right after passing the kitchen and before reaching the bathroom door.

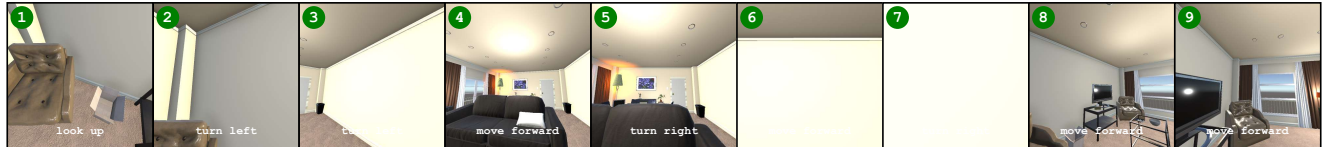
I need to walk straight past the kitchen on my left, which means I should continue moving forward until I reach the point where I need to turn right. The current view shows that I am still in the hallway, so I should keep moving forward.

I need to move closer to the sofa table to find the book.

Figure 9

**Task:** I want to find a human-shaped statue which is placed on the table below the wall where the iris painting hangs. Please navigate to it.

**Qwen-VL-Max:**



I need to find the human-shaped statue on the table below the wall where the iris painting hangs. I should start by looking around to identify any potential locations that match this description.

The task requires me to find a human-shaped statue on a table below a wall with an iris painting. I should start by looking around to identify any potential locations that match this description.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should start by exploring the room and looking for any paintings or statues. Since I am currently facing a blank wall, I should turn left to explore other parts of the room.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the wall with the iris painting to locate the statue.

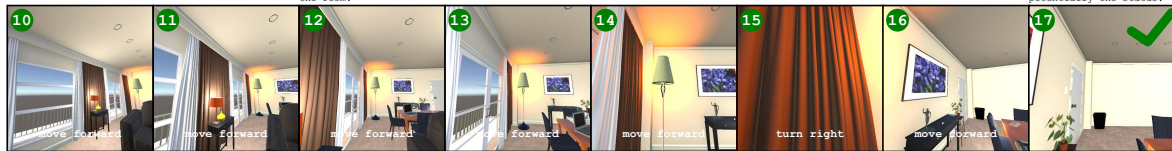
The task requires me to find a human-shaped statue on a table below the wall where the iris painting hangs. I should move towards the wall with the iris painting to locate the statue.

The task requires me to find a human-shaped statue on a table below the wall where the iris painting hangs. I should move towards the wall with the iris painting to locate the statue.

The task requires me to find a human-shaped statue on a table below a wall with an iris painting. I should move towards the wall with the iris painting to locate the statue.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the current view. Moving forward will bring me closer to the painting and potentially the statue.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the background of the current view. Moving forward will bring me closer to that area.



The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the background of the current view. Moving forward will bring me closer to the painting and potentially the statue.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the current view. Moving forward will bring me closer to the painting and potentially the statue.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the current view. Moving forward will bring me closer to the target location.

The task is to find a human-shaped statue on a table below a wall where the iris painting hangs. I need to move closer to the table where the statue is located.

The task is to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move closer to the table to confirm if the statue is there.

The task requires me to find a human-shaped statue placed on a table below a wall where the iris painting hangs. I have been moving around the room and now I am near the window. I should turn right to face the wall with the iris painting and locate the statue.

The task is to find a human-shaped statue placed on a table below the wall where the iris painting hangs. I should move closer to the table to inspect it.

The task is to find a human-shaped statue placed on a table below the wall where the iris painting hangs. I should move closer to the table to inspect it.

Figure 10

## 1.4.4 Object Interaction

**Task: Is there an egg inside the fridge?**

**Options:**

- answer "Yes, and there are two eggs, one in the upper part and one in the lower part."
- answer "Yes, in the lower part of the fridge."
- answer "Yes, there are two eggs in the upper part of the fridge."
- open the lower fridge door
- answer "There is no fridge."
- answer "No, there are no eggs in the fridge."
- answer "Yes, in the upper part of the fridge."
- open the upper fridge door
- answer "Yes, there are two eggs in the lower part of the fridge."

**Qwen-VL-Max: answer "Yes, in the lower part of the fridge."**

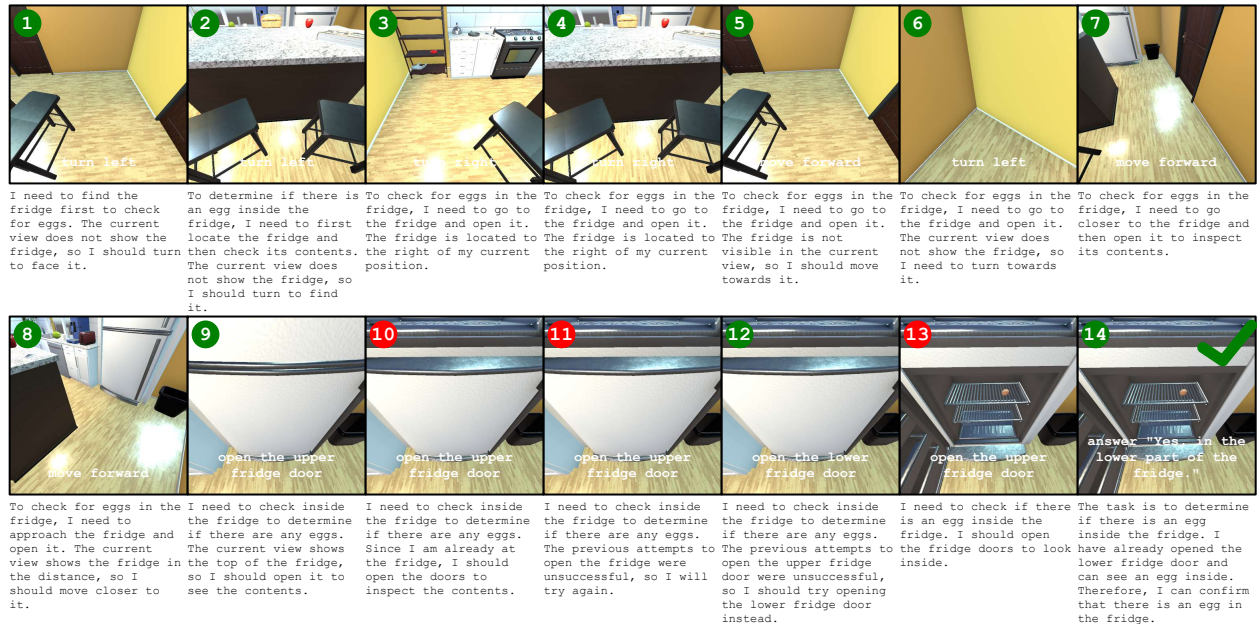


Figure 11

**Task: Pick all round objects into trash can.**

**Options:**

- pick up  place to the trash can

**Gemini-Pro:**

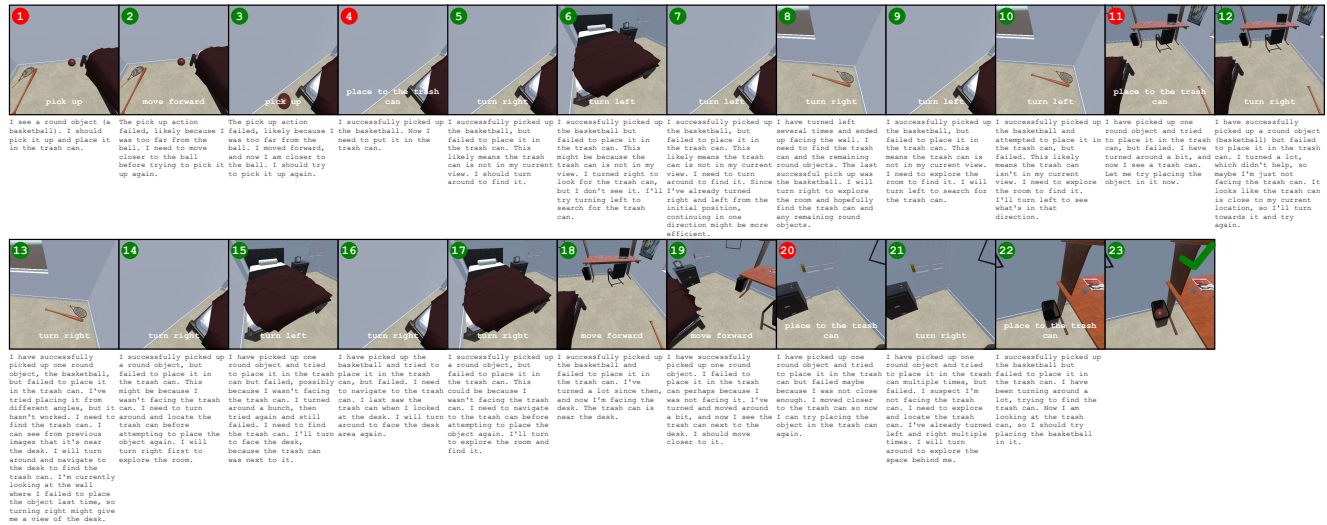


Figure 12

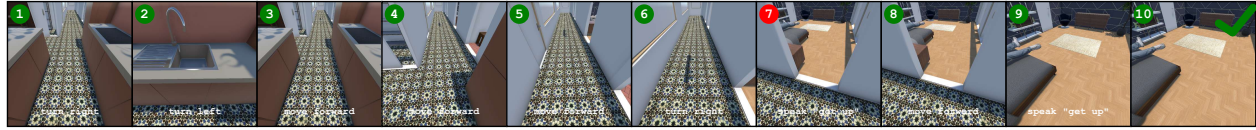
## 1.4.5 Social Interaction

**Task:** Wake up your dad. He is sleeping in the bedroom. The bedroom is the second room on your right as you walk forward.

**Options:**

speak "get up"

**GPT-4o-Mini:**



1 To wake up your dad, I need to navigate to the bedroom, which is the second room on my right as I walk forward. Currently, I am facing a hallway with options to turn left or right. Since the bedroom is to the right, I should turn right to head in the correct direction.

2 I need to navigate towards the bedroom to wake up my dad. Since I have already turned right, the next logical step is to continue moving in the direction that will lead me to the bedroom. The available options include turning left or right, looking up, or speaking. Since turning right again would not be effective, I will turn left to continue towards the bedroom.

3 I need to navigate towards the bedroom, which is the second room on my right as I walk forward. Since I have already turned right and left without making progress, I should move forward to get closer to the bedroom.

4 I need to reach the bedroom, which is the second room on my right as I walk forward. Currently, I am facing a hallway with a clear path ahead. Since I have already turned right and left previously, I should move forward to get closer to the bedroom.

5 I need to wake up my dad, who is in the bedroom, which is the second room on my right as I walk forward. Currently, I am facing a hallway with a view that suggests I am close to the bedroom. Since I have already moved forward twice and turned right, I should now move forward to reach the bedroom.

6 I need to wake up my dad, who is in the bedroom, which is the second room on my right as I walk forward. Currently, I am facing a hallway, and I have already moved forward multiple times. Since I am not yet in the forward twice and turned right, I should turn right to head towards the bedroom.

7 I am currently in front of the bedroom where my dad is sleeping. Since the task is to wake him up, the most direct action is to speak to him. This will allow me to complete the task without unnecessary movement.

8 I am currently in front of the bedroom where my dad is sleeping. I have already tried to wake him by speaking, but it failed. Since I am already facing the bedroom, I should try to move forward into the room to get closer to him. This will allow me to interact more effectively.

9 I am currently in the bedroom where my dad is sleeping. I previously attempted to wake him by speaking, but it failed. Since I am already in the right location, I should try a different approach. I can either try to speak again or look for a way to wake him up physically. However, since I cannot interact with objects or people unless they are very close, I will try speaking again to see if it has a different effect this time.

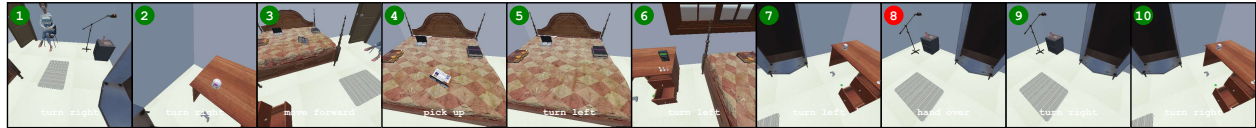
Figure 13

**Task:** Retrieve the magazine for the madam sitting on the chair.

**Options:**

pick up  hand over

**Gemini-Pro:**



1 I see the madam and a magazine on the small table. I need to go to the table, pick up the magazine and give it to the madam. I will turn right to face the table.

2 I turned right and saw a table. I need to turn back to the madam and locate the magazine.

3 I've turned right twice and now see the bed with some books on it. The task is to retrieve the magazine. One of the books might be the magazine. I should move closer to inspect them.

4 I see several objects on the bed, including what looks like a magazine. I should pick it up.

5 I have picked up the magazine. Now I need to turn around to face her and then walk towards her.

6 I have picked up the magazine. Now I need to turn around and give it to the madam sitting on the chair. I just turned left, so I will turn left again to face her.

7 I have picked up the magazine and turned around, facing the desk. Now I need to turn to turn to face the madam and approach her to hand over the magazine. I've already turned left twice, and the madam was initially to my right, so turning left again should face me in the correct direction.

8 I have picked up the magazine and turned around to face the madam. I should now give her the magazine.

9 The handover failed because I wasn't facing the madam. I need to turn right to face her and then try handing over the magazine again.

10 The handover failed because I wasn't facing the madam. I need to turn towards her and try again. I am currently facing the desk. The madam is behind me to the left. I have turned right after the failed handover, so now she should be directly behind me.



11 I successfully picked up the magazine but failed to hand it over. This likely means I wasn't facing the woman when I tried. I need to turn around to face her and try again. I'm currently facing the bed and desk, so turning left would put me back in the position where I unsuccessfully tried to hand over the magazine. Turning right would orient me towards the doorway, so I will turn right.

12 I successfully picked up the magazine, but handing it over failed. This means I wasn't facing the woman when I tried. I need to turn around to face her and then hand over the magazine. I turned around three times after picking it up. I've done another three turns since the failed handover, meaning I should now be facing the correct direction after three right turns.

13 I successfully picked up an item, but handing it over failed twice. This suggests I picked up the wrong item. I'm currently looking at the madam. Now I'm facing the bed, and there are a few items here. I need to find the magazine. I'll turn around to see if it is elsewhere.

14 I have successfully picked up the magazine, I previously failed to hand it over because I was not facing the madam. Now I'm facing her, so I should try handing it over again.

15 I have the magazine, I failed to hand it over multiple times because I was not facing the madam. I am now facing her, so I should try handing it over again.

16 I have the magazine, and I'm facing the madam. I've tried handing it over multiple times from slightly different angles, but it hasn't worked. Perhaps I need to move a bit closer to ensure I'm within the interaction range.

17 I have the magazine and I am facing the madam. I failed to hand over the magazine multiple times, likely because I wasn't close enough. I moved forward in the previous step, but perhaps not enough. I will try moving closer again before attempting to hand over.

18 I successfully picked up the magazine earlier, and then tried handing it over multiple times, likely because I wasn't close enough. I moved forward in the last step, so now I'll try handing it over again. I will step back a little and try to hand over the magazine again. I'll try moving back first to create some distance.

19 I have the magazine and am facing the woman. I failed to hand it over multiple times, likely because I wasn't close enough. I moved forward in the last step, so now I'll try handing it over again.

20 I have the magazine and I am facing the woman. I failed to hand it over multiple times, likely because I wasn't close enough. I moved forward in the last step, so now I'll try handing it over again.

Figure 14

## 1.5. Error Cases

### 1.5.1 Attribute QA

**Task: Locate any electronic devices in the bedroom.**

**Options:**

- answer "flashlight, desk lamp, and phone"
- answer "laptop, desk lamp, and phone"
- answer "laptop, TV, and desk lamp"**
- answer "TV, laptop, and door"
- answer "flashlight, desk lamp, and TV"
- answer "flashlight, desk lamp, and ipad"
- answer "laptop, desk lamp, and bed"
- answer "laptop, book, and desk lamp"

**Qwen-VL-Max: answer "laptop, desk lamp, and phone"**

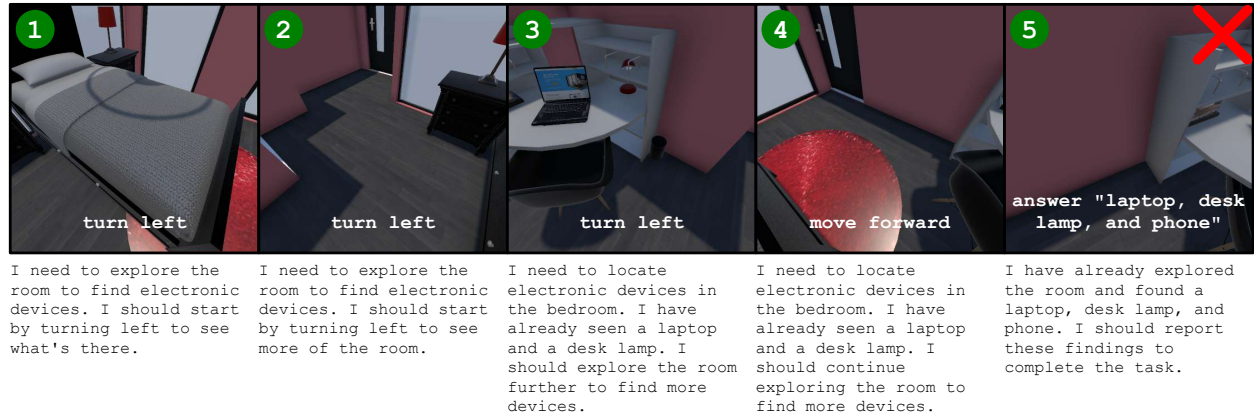


Figure 15

**Task: Are there more flower pots in the living room or the bedroom?**

**Options:**

- answer "There are more flower pots in the living room, while no one in bedroom"
- answer "The bedroom has a few flower pots, but the living room has none."
- answer "The bedroom has only one flower pot, while the living room has several."**
- answer "There are no flower pots in the bedroom, only in the living room."
- answer "The living room has more flower pots, placed near the trash can."
- answer "There is only one flower pot in each room."
- answer "Both the living room and bedroom have the same number of flower pots."
- answer "Both rooms have flower pots, but the bedroom's pots are larger in size."

**Gemini-Pro: answer "There are no flower pots in the bedroom, only in the living room."**

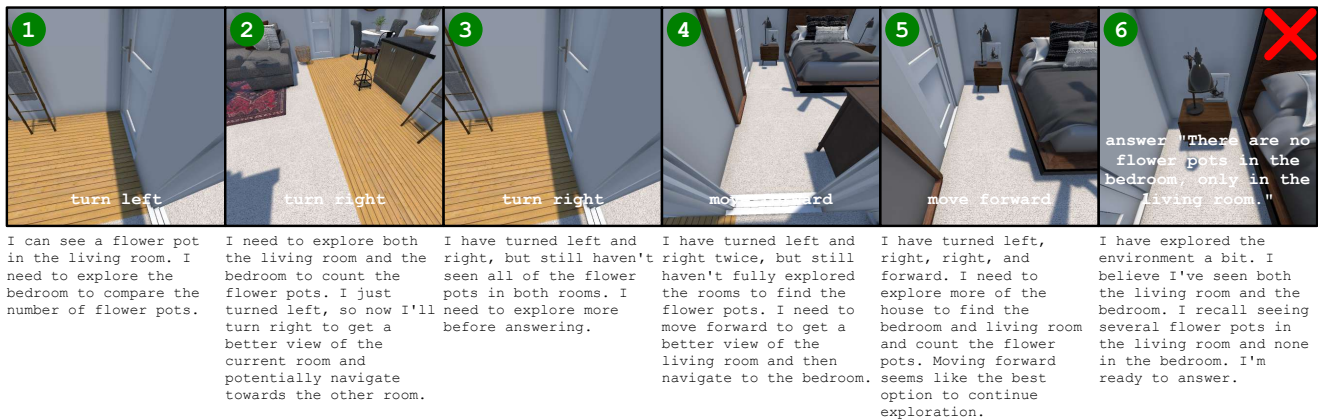


Figure 16

## 1.5.2 Spatial QA

**Task:** Determine the best location for a fire extinguisher that would be quickly accessible from both the kitchen and the living room.

**Options:**

- answer "Near the window."
- answer "Beside the flowerpot near the round carpet."
- answer "On the countertop."
- answer "Beside the black sofas."
- answer "In front of the stove."
- answer "Next to the piano."
- answer "Beside the white sofa."
- answer "Next to the irregular table with many books."

**GPT-4o:** answer "Beside the white sofa."

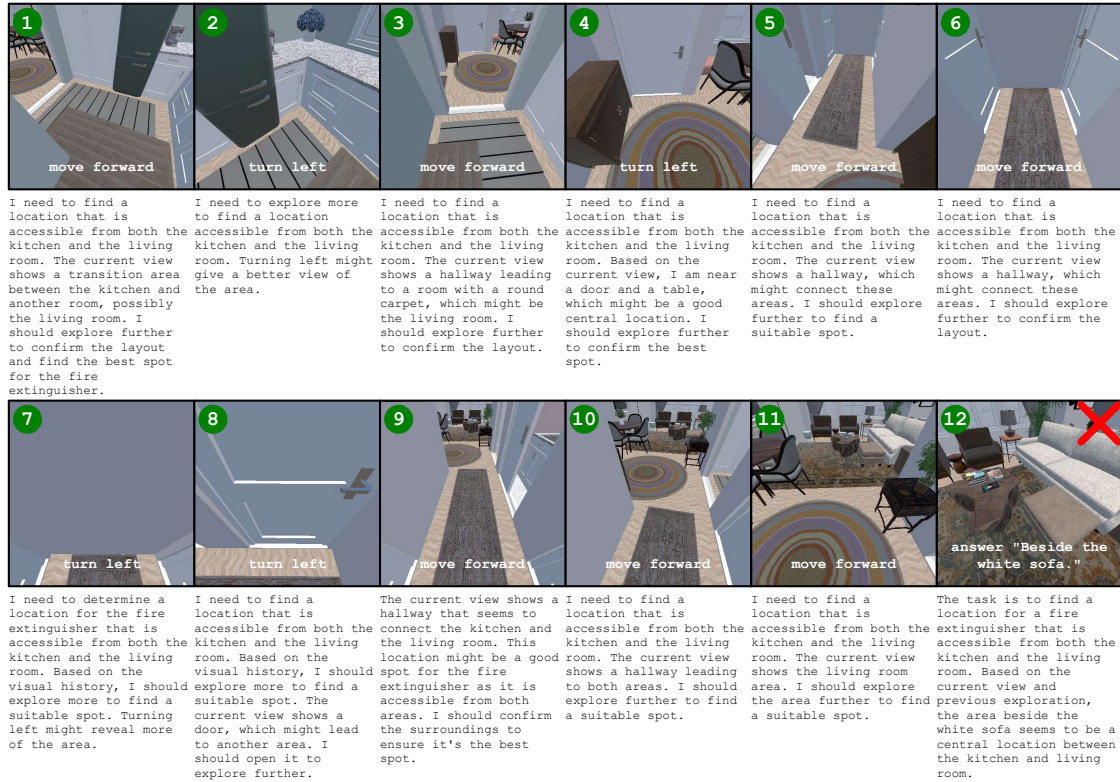


Figure 17

## 1.5.3 Navigation

**Task:** Walk towards the tallest tree in the yard.

**GPT-4o:**

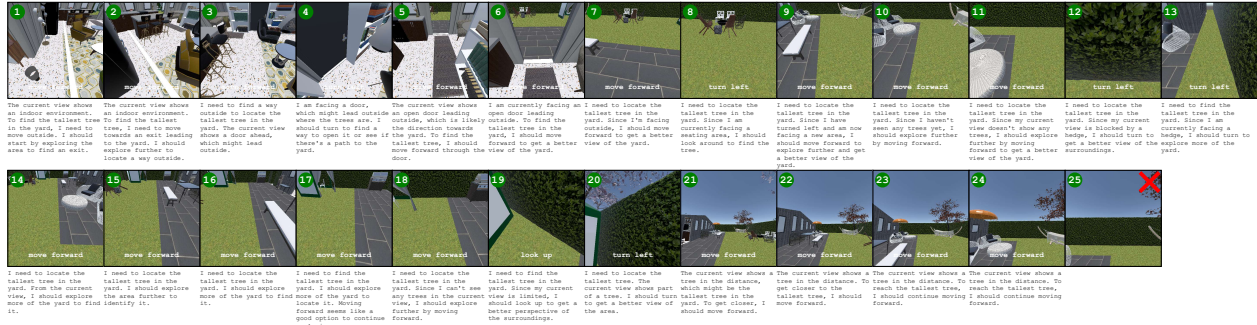


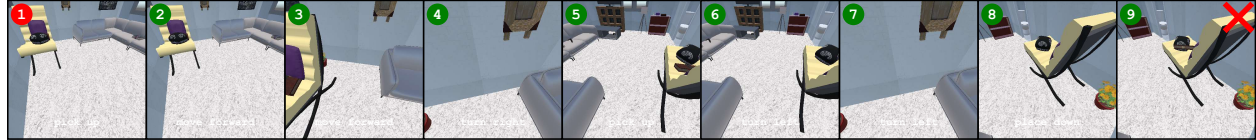
Figure 18

## 1.5.4 Object Interaction

**Task:** Retrieve the tied reading materials from the sofa in the living room and place it on the lounge chair.

**Options:**  
 place down  pick up

**GPT-4o:**



The task is to retrieve the tied reading materials from the sofa and place them on the lounge chair. In the current view, I can see the sofa and the reading materials on it. The lounge chair is also visible. I need to pick up the reading materials first.

The task is to retrieve the tied reading materials from the sofa and place them on the lounge chair. The previous attempt to pick up the reading materials failed, possibly because I wasn't close enough. I need to pick up the reading materials first.

The task is to retrieve the tied reading materials from the sofa and place them on the lounge chair. The current view shows the sofa and the lounge chair. Since the previous attempt to pick up failed, I need to ensure I'm close enough to the reading materials. I should move closer to the sofa to successfully pick them up.

The reading materials are likely on the sofa to my right, based on the previous images. I should turn right to face the sofa and locate the materials.

The tied reading materials are visible on the sofa in the current view. I need to pick them up before moving to the lounge chair.

I have successfully picked up the reading materials. Now, I need to place them on the lounge chair, which is visible in my current view. The lounge chair is to my left.

I have successfully picked up the reading materials. Now, I need to place them on the lounge chair. The current view shows the wall, so I need to turn to face the lounge chair.

I have successfully picked up the reading materials. Now, I need to place them on the lounge chair, which is currently in my view.

Figure 19

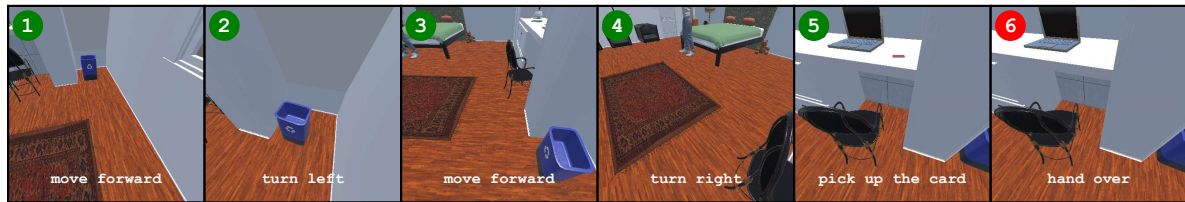
## 1.5.5 Social Interaction

**Task:** Bring my credit card to me and return it back to its original place.

**Options:**

pick up the card  place down  hand over  place into trash can

**GPT-4o:**



I need to locate the credit card first. The current view shows a trash can and part of a room. I should explore the area to find the card.

The current view shows a trash can, but I don't see a credit card. I should turn to explore the area further.

I need to locate the credit card before I can pick it up. The current view shows a bed, a chair, and a desk. I should explore the area further to find the card.

The task is to bring the credit card to the person and return it to its original place. Currently, I don't see a credit card in my view. I should explore the room further to locate it. Turning right might give a better view of the room.

The credit card is visible on the desk in front of me. The next logical step is to pick it up.

I have already picked up the card. The next step is to hand it over to the person in the room. The person is visible in the previous images, so I should hand over the card.



The task is to bring the credit card to the person and then return it to its original place. The previous attempt to hand over the card failed, possibly because the person was not in the correct position. I need to locate the person and try handing over the card again.

I am currently facing the person who requested the credit card. The previous attempt to hand over the card failed, possibly because I wasn't close enough. I should move closer to the person to successfully hand over the card.

I am currently facing the person who requested the credit card. The previous attempt to hand over the card failed, possibly because I wasn't close enough. I should try handing over the card again now that I'm closer.

The previous attempts to hand over the card failed, possibly because I wasn't close enough to the person. I need to move closer to the person in the current view to successfully hand over the card.

I am currently facing the chairs, and the person is likely behind me. I need to turn around to face the person and hand over the card. Previous attempts to hand over the card failed, possibly because I wasn't facing the person directly.

Figure 20