Supplementary Document for REVERIE: Remote Embodied Visural Referring Expressions in Real indoor Environments

Yuankai Qi[1,2] Qi Wu[1*] Peter Anderson[3†] Xin Wang[4] William Yang Wang[4] Chunhua Shen[1] Anton van den Hengel[1]
[1]Australia Centre for Robotic Vision, The University of Adelaide [2]Harbin Institute of Technology, Weihai
[3]Georgia Institute of Technology [4]University of California, Santa Barbara
qykshr@gmail.com {qi.wu01, chunhua.shen, anton.vandenhengel}@adelaide.edu.au
peter.anderson@gatech.edu {xwang, william}@cs.ucsb.edu

In this supplementary material, we provide detailed explanation of evaluation metrics, examples of collected data, data collecting tools, how human test are performed and visualisation of several REVERIE results.

1. Evaluation Metrics

- Navigation Success: a navigation is considered successful only if the target object can be observed at the stop viewpoint. Please note that to encourage the agent to approach closer to the target object, we set the objects visible if they are within 3 meters away from the current location.

- Navigation Oracle Success: a navigation is considered oracle successful if the target object can be observed at one of its passed viewpoints.

- Navigation SPL: it is the navigation success weighted by the length of navigation path, which is

$$\frac{1}{N}\sum_{i=1}^{N} S_i \frac{\ell_i}{\max(\ell_i, p_i)} \tag{1}$$

  where N is the number of tasks, $S_i \in \{0,1\}$ is a binary indicator of success of task i, ℓ_i is the shortest length between the starting viewpoint and the goal viewpoint of task i, and p_i is the path length of an agent for task i.

- Navigation Length: trajectory length in meters.

- REVERIE Success: a task is considered REVERIE successful if the output bounding box has an IoU (intersection over union) ≥ 0.5 with the ground truth.

2. Typical Samples of The REVERIE Task

In Figure 1, we present several typical samples of the proposed REVERIE task. It shows the diversity in object category, goal region, path instruction, and target object referring expression.

3. Data Collecting Tools

To collect data for the REVERIE task, we develop a WebGL based data collecting tool as shown in Figure 2 and Figure 3. To facilitate the workers, we provide real-time updated reference information in the web page according to the location of the agent, including the current level/total level, the current region, and the number of regions in the build having the same function. At the goal location, in addition to highlighting the target object with a red 3D rectangle, we also provide the label of the target object and the number of objects falling in the same category with the target object. Text and video instructions are provided for workers to understand how to make high quality annotations as shown in Figure 2.

[*] Corresponding author

[†] Now at Google

Methods	Val Seen					Val UnSeen					Test (Unseen)				
	Navigation Acc.				REVERIE	Navigation Acc.				REVERIE	Navigation Acc.				REVERIE
	Succ.	OSucc.	SPL	Length	Succ.	Succ.	OSucc.	SPL	Length	Succ.	Succ.	OSucc.	SPL	Length	Succ.
Random	2.74	8.92	1.91	11.99	0.14	1.76	11.93	1.01	10.76	0.00	2.30	8.88	1.44	10.34	0.08
Shortest	100	100	100	10.46	5.83	100	100	100	9.47	3.61	100	100	100	9.39	4.64
R2R-TF	7.38	10.75	6.40	11.19	0.42	3.21	4.94	2.80	11.22	0.09	3.94	6.40	3.30	10.07	0.27
R2R-SF	29.59	35.70	24.01	12.88	0.49	4.20	8.07	2.84	11.07	0.23	3.99	6.88	3.09	10.89	0.17
RCM	23.33	29.44	21.82	10.70	0.28	9.29	14.23	6.97	11.98	0.11	7.84	11.68	6.67	10.60	0.22
SelfMonitor	41.25	43.29	39.61	7.54	2.52	8.15	11.28	6.44	9.07	0.37	5.80	8.39	4.53	9.23	0.59
FAST-Short	45.12	49.68	40.18	13.22	2.74	10.08	20.48	6.17	29.70	0.57	14.18	23.36	8.74	30.69	1.08
FAST-Lan-Only	8.36	23.61	3.67	49.43	0.14	9.37	29.76	3.65	45.03	0.40	8.15	28.45	2.88	46.19	0.40
Ours	**28.74**	**41.81**	**18.81**	37.96	**1.62**	**11.87**	**28.97**	**5.58**	50.89	**0.60**	**9.92**	**25.14**	**3.79**	48.58	**0.48**
Human	–	–	–	–	–	–	–	–	–	–	81.51	86.83	53.66	21.18	77.84

Table 1. REVERIE success rate achieved by combining state-of-the-art navigation methods with the RefExp method MAttNet [?] using detected bounding boxes.

	Val Seen	Val UnSeen	Test
Baseline	1.69	0.85	0.97
MAttNet	5.83	3.61	4.64
CM-Erase	6.11	4.89	4.26
Human	–	–	90.76

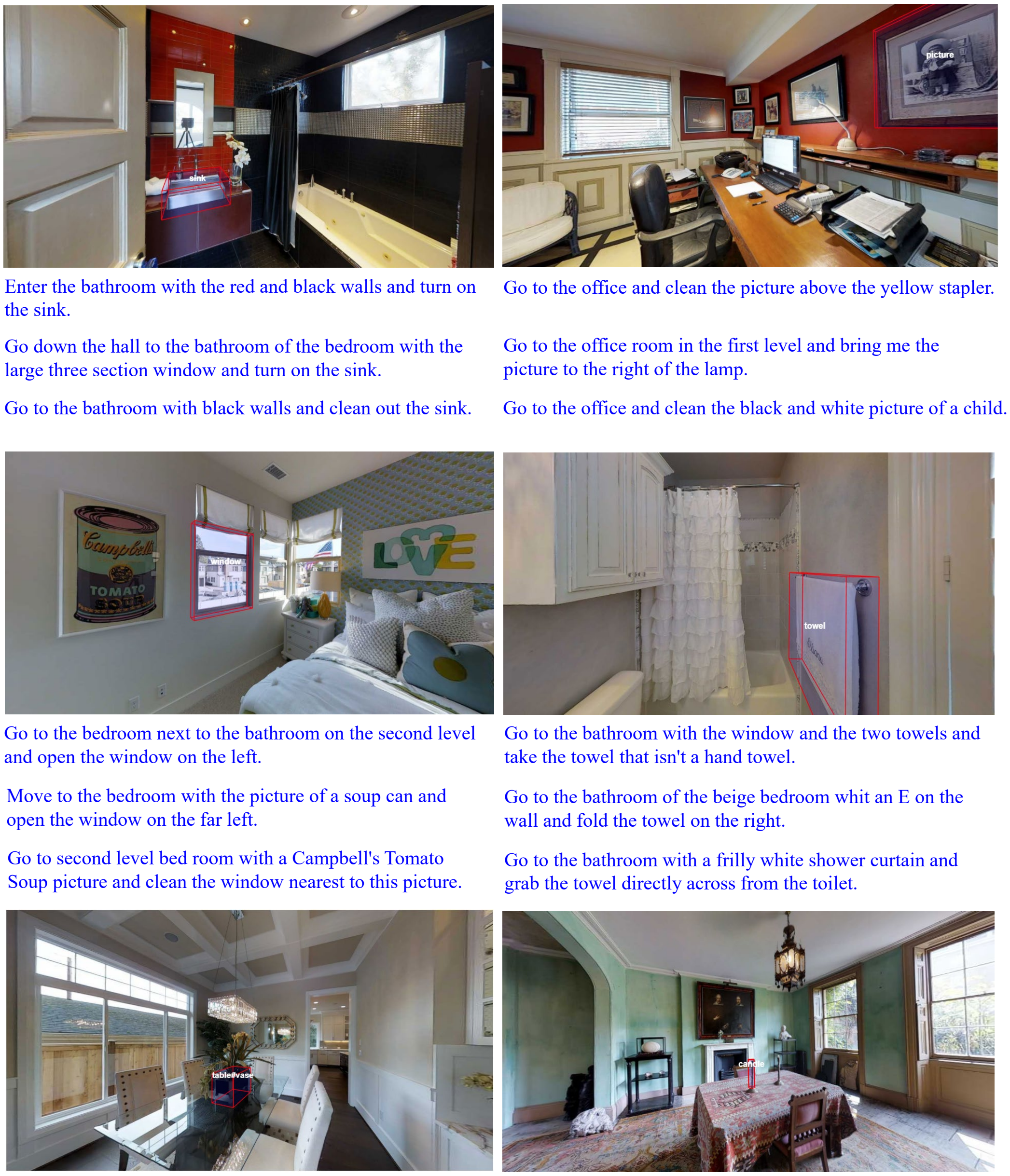
Table 2. Referring expression comprehension success rate (%) at the ground truth goal viewpoint of our REVERIE dataset using detected bounding boxes.

4. Human Performance Test

To obtain the machine-human performance gap, we develop a WebGL based tool as shown in Figure 4 to test human performance. In the tool, we show an instruction about a remote object to the worker. Then the worker needs to navigate to the goal location and select one object as the target object from a range of object candidates. The worker can look around and go forward/backword by dragging or clicking.

5. Visualisation of REVERIE Results

In Figure 5, we provide the visualisation of several REVERIE results obtained by the typical state-of-the-art method, FAST-short, and the typical baseline method, R2R-SF.



Enter the bathroom with the red and black walls and turn on the sink.

Go down the hall to the bathroom of the bedroom with the large three section window and turn on the sink.

Go to the bathroom with black walls and clean out the sink.

Go to the office and clean the picture above the yellow stapler.

Go to the office room in the first level and bring me the picture to the right of the lamp.

Go to the office and clean the black and white picture of a child.

Go to the bedroom next to the bathroom on the second level and open the window on the left.

Move to the bedroom with the picture of a soup can and open the window on the far left.

Go to second level bed room with a Campbell's Tomato Soup picture and clean the window nearest to this picture.

Go to the bathroom with the window and the two towels and take the towel that isn't a hand towel.

Go to the bathroom of the beige bedroom whit an E on the wall and fold the towel on the right.

Go to the bathroom with a frilly white shower curtain and grab the towel directly across from the toilet.

Bring me the vase on the dining room table.

Go into the dining room and take the vase off the table.

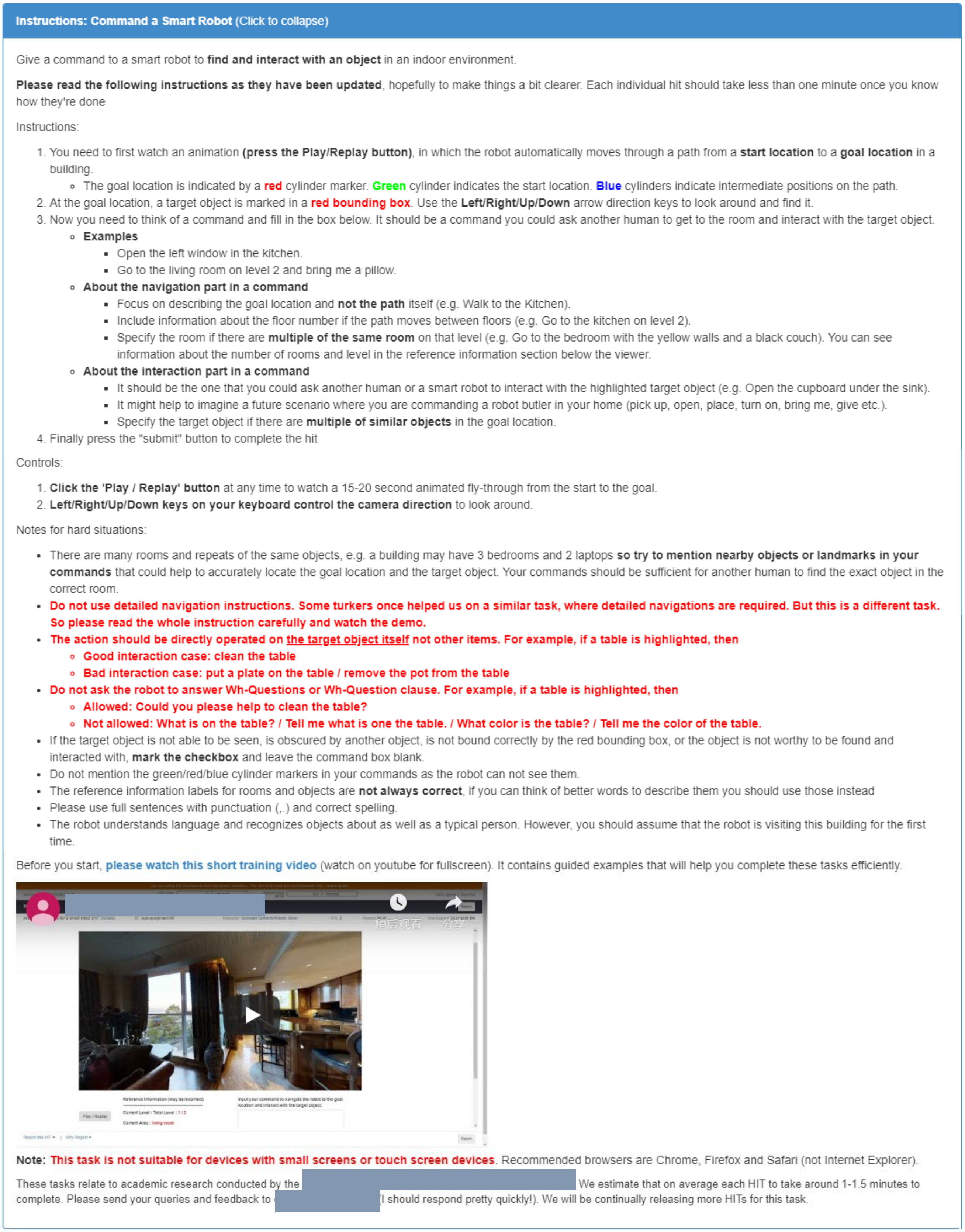
Go to the dining room and take the vase.

Up the stairs in the room with the black paining on the wall take the left candle off the table.

Light the candle furthest from the windows in the lounge with green walls

Move to the lounge on level 2 with the egg sculpture and light the candle furthest from the windows

Figure 1. Several typical samples of the collected dataset, which involves various object category, goal region, path instruction, and object referring expression.



Instructions: Command a Smart Robot (Click to collapse)

Give a command to a smart robot to **find and interact with an object** in an indoor environment.

Please read the following instructions as they have been updated, hopefully to make things a bit clearer. Each individual hit should take less than one minute once you know how they're done

Instructions:

1. You need to first watch an animation **(press the Play/Replay button)**, in which the robot automatically moves through a path from a **start location** to a **goal location** in a building.
   - The goal location is indicated by a **red** cylinder marker. **Green** cylinder indicates the start location. **Blue** cylinders indicate intermediate positions on the path.
2. At the goal location, a target object is marked in a **red bounding box**. Use the **Left/Right/Up/Down** arrow direction keys to look around and find it.
3. Now you need to think of a command and fill in the box below. It should be a command you could ask another human to get to the room and interact with the target object.
   - **Examples**
     - Open the left window in the kitchen.
     - Go to the living room on level 2 and bring me a pillow.
   - **About the navigation part in a command**
     - Focus on describing the goal location and **not the path** itself (e.g. Walk to the Kitchen).
     - Include information about the floor number if the path moves between floors (e.g. Go to the kitchen on level 2).
     - Specify the room if there are **multiple of the same room** on that level (e.g. Go to the bedroom with the yellow walls and a black couch). You can see information about the number of rooms and level in the reference information section below the viewer.
   - **About the interaction part in a command**
     - It should be the one that you could ask another human or a smart robot to interact with the highlighted target object (e.g. Open the cupboard under the sink).
     - It might help to imagine a future scenario where you are commanding a robot butler in your home (pick up, open, place, turn on, bring me, give etc.).
     - Specify the target object if there are **multiple of similar objects** in the goal location.
4. Finally press the "submit" button to complete the hit

Controls:

1. **Click the 'Play / Replay' button** at any time to watch a 15-20 second animated fly-through from the start to the goal.
2. **Left/Right/Up/Down keys on your keyboard control the camera direction** to look around.

Notes for hard situations:

- There are many rooms and repeats of the same objects, e.g. a building may have 3 bedrooms and 2 laptops **so try to mention nearby objects or landmarks in your commands** that could help to accurately locate the goal location and the target object. Your commands should be sufficient for another human to find the exact object in the correct room.
- **Do not use detailed navigation instructions. Some turkers once helped us on a similar task, where detailed navigations are required. But this is a different task. So please read the whole instruction carefully and watch the demo.**
- **The action should be directly operated on the target object itself not other items. For example, if a table is highlighted, then**
  - **Good interaction case: clean the table**
  - **Bad interaction case: put a plate on the table / remove the pot from the table**
- **Do not ask the robot to answer Wh-Questions or Wh-Question clause. For example, if a table is highlighted, then**
  - **Allowed: Could you please help to clean the table?**
  - **Not allowed: What is on the table? / Tell me what is one the table. / What color is the table? / Tell me the color of the table.**
- If the target object is not able to be seen, is obscured by another object, is not bound correctly by the red bounding box, or the object is not worthy to be found and interacted with, **mark the checkbox** and leave the command box blank.
- Do not mention the green/red/blue cylinder markers in your commands as the robot can not see them.
- The reference information labels for rooms and objects are **not always correct**, if you can think of better words to describe them you should use those instead
- Please use full sentences with punctuation (,.) and correct spelling.
- The robot understands language and recognizes objects about as well as a typical person. However, you should assume that the robot is visiting this building for the first time.

Before you start, **please watch this short training video** (watch on youtube for fullscreen). It contains guided examples that will help you complete these tasks efficiently.

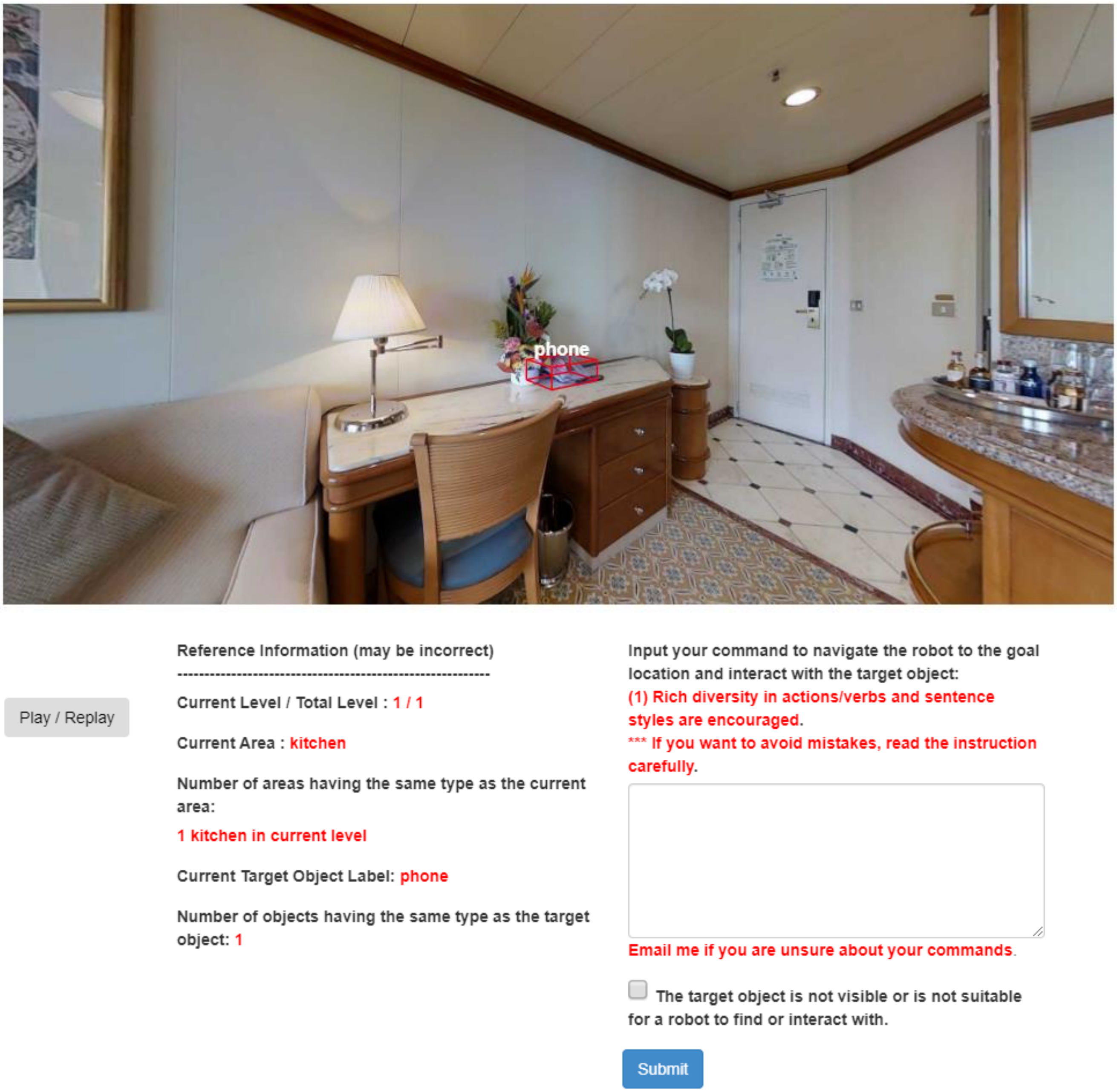
[redacted]

稍后观看
分享
Play / Replay
Reference information (may be incorrect)
Current Level / Total Level : 1 / 2
Current Area : living room
Input your command to navigate the robot to the goal location and interact with the target object:

Note: **This task is not suitable for devices with small screens or touch screen devices**. Recommended browsers are Chrome, Firefox and Safari (not Internet Explorer).

These tasks relate to academic research conducted by the [redacted] We estimate that on average each HIT to take around 1-1.5 minutes to complete. Please send your queries and feedback to [redacted] (I should respond pretty quickly!). We will be continually releasing more HITs for this task.

Figure 2. Data collecting interface part I: instructions for AMT workers.



Reference Information (may be incorrect)

Current Level / Total Level : 1 / 1

Current Area : kitchen

Number of areas having the same type as the current area:

1 kitchen in current level

Current Target Object Label: phone

Number of objects having the same type as the target object: 1

Input your command to navigate the robot to the goal location and interact with the target object:

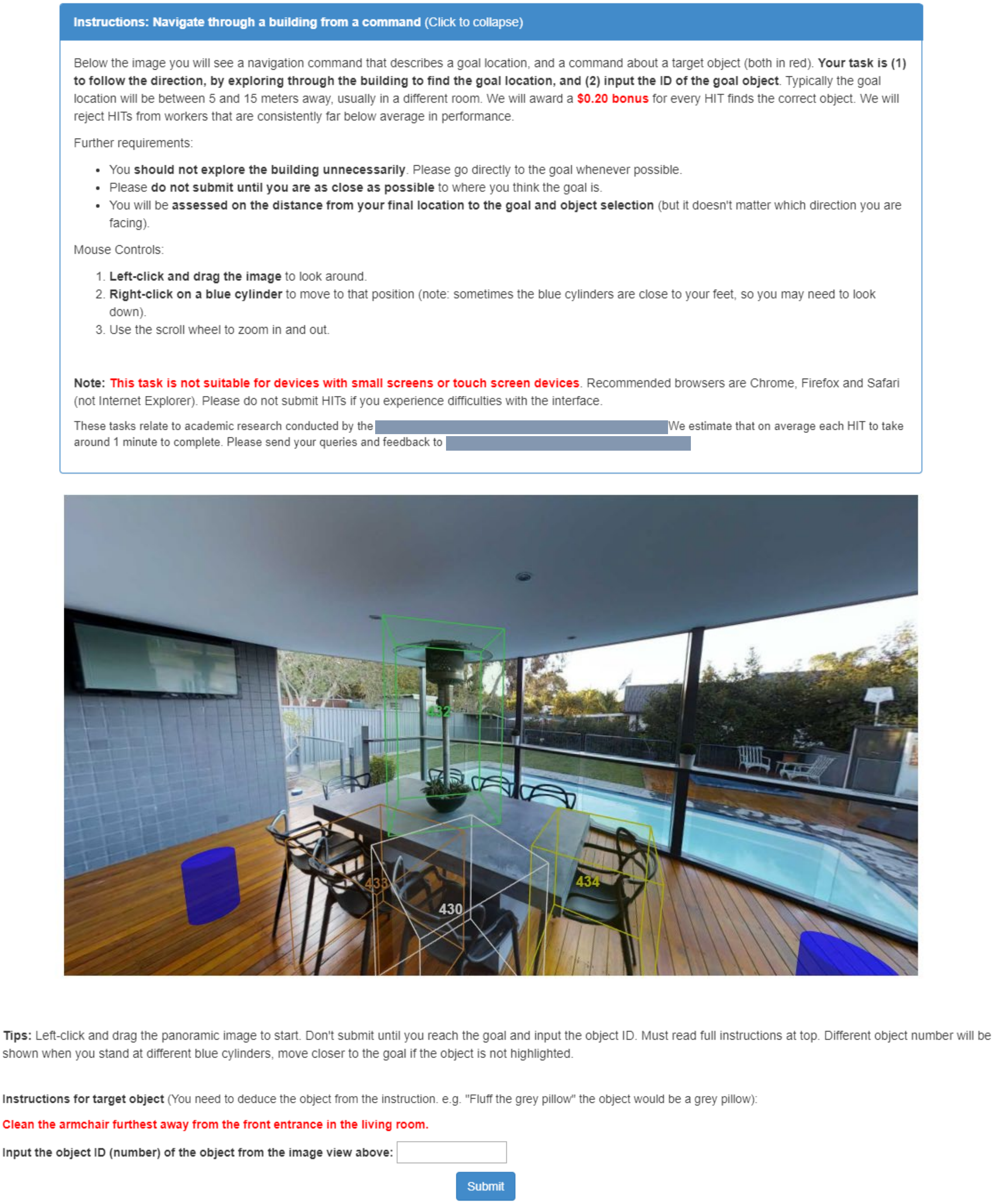
(1) Rich diversity in actions/verbs and sentence styles are encouraged.

*** If you want to avoid mistakes, read the instruction carefully.

Email me if you are unsure about your commands.

□ The target object is not visible or is not suitable for a robot to find or interact with.

Figure 3. Data collecting interface Part II: assistant information and user input field.



Instructions: Navigate through a building from a command (Click to collapse)

Below the image you will see a navigation command that describes a goal location, and a command about a target object (both in red). **Your task is (1) to follow the direction, by exploring through the building to find the goal location, and (2) input the ID of the goal object**. Typically the goal location will be between 5 and 15 meters away, usually in a different room. We will award a **$0.20 bonus** for every HIT finds the correct object. We will reject HITs from workers that are consistently far below average in performance.

Further requirements:

- You **should not explore the building unnecessarily**. Please go directly to the goal whenever possible.
- Please **do not submit until you are as close as possible** to where you think the goal is.
- You will be **assessed on the distance from your final location to the goal and object selection** (but it doesn't matter which direction you are facing).

Mouse Controls:

1. **Left-click and drag the image** to look around.
2. **Right-click on a blue cylinder** to move to that position (note: sometimes the blue cylinders are close to your feet, so you may need to look down).
3. Use the scroll wheel to zoom in and out.

Note: This task is not suitable for devices with small screens or touch screen devices. Recommended browsers are Chrome, Firefox and Safari (not Internet Explorer). Please do not submit HITs if you experience difficulties with the interface.

These tasks relate to academic research conducted by the [redacted] We estimate that on average each HIT to take around 1 minute to complete. Please send your queries and feedback to [redacted]

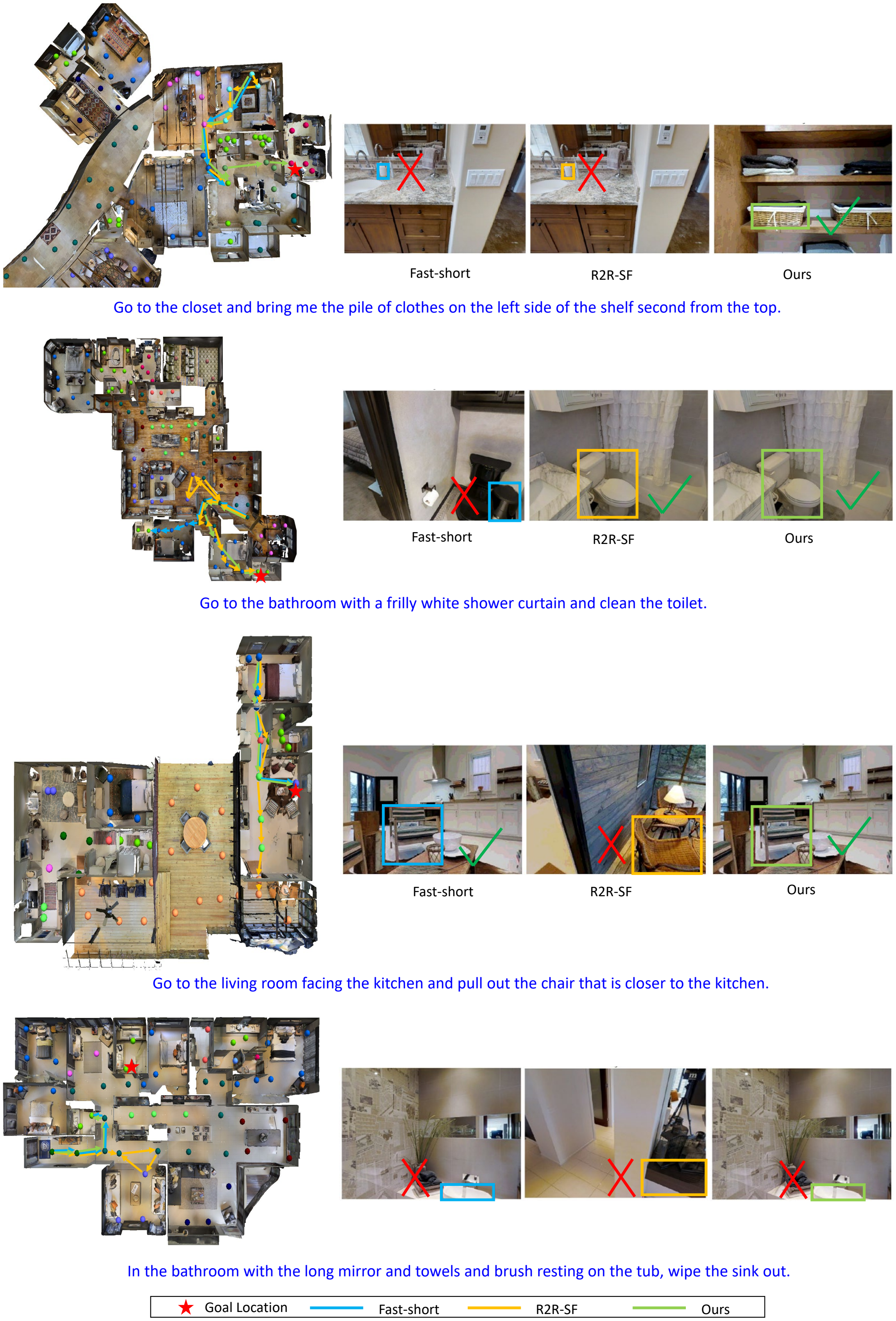
Tips: Left-click and drag the panoramic image to start. Don't submit until you reach the goal and input the object ID. Must read full instructions at top. Different object number will be shown when you stand at different blue cylinders, move closer to the goal if the object is not highlighted.

Instructions for target object (You need to deduce the object from the instruction. e.g. "Fluff the grey pillow" the object would be a grey pillow):

Clean the armchair furthest away from the front entrance in the living room.

Input the object ID (number) of the object from the image view above:

Figure 4. Human test interface. Workers need first to navigate to the goal location by clicking or dragging mouse, and then identify the target object. Only objects within 3 meters from the current location are highlighted. Different colors are used to facilitate workers to distinguish one object from others.



Go to the closet and bring me the pile of clothes on the left side of the shelf second from the top.

Go to the bathroom with a frilly white shower curtain and clean the toilet.

Go to the living room facing the kitchen and pull out the chair that is closer to the kitchen.

In the bathroom with the long mirror and towels and brush resting on the tub, wipe the sink out.

Figure 5. Visualisation of several REVERIE results, including trajectories and referring expression grounding of three typical methods. Colorized dots denote reachable locations, and different colors mark locations belonging to different regions according to the Matterport dataset.