Translating Navigation Instructions in Natural Language to a High-Level Plan for Behavioral Robot Navigation – Supplementary Appendix –

Xiaoxue Zang[1]*, Ashwini Pokle[1]*, Marynel Vázquez[2], Kevin Chen[1], Juan Carlos Niebles[1], Alvaro Soto[3], Silvio Savarese[1]
[1] Stanford University, [2] Yale University, [3] P. Universidad Católica de Chile
[1]{xzang, ashwinipokle, kchen92, jniebles, ssilvio}@stanford.edu,
[2]marynel.vazquez@yale.edu, [3]asoto@ing.puc.cl

A Behavioral Navigation through Natural Language Instructions

We translate instructions in natural language to high-level navigation plans in the context of behavioral robot navigation (Sepúlveda et al., 2018). The next section details the landmarks that we consider for behavioral navigation graphs in this work. Afterwards, we describe the data collection process that we used to build a new dataset of navigation instructions and corresponding high-level navigation plans. We also summarize relevant dataset statistics.

A.1 Landmarks

The environmental landmarks that we consider in this work are: chair, table, vase, clock, lamp, printer, computer, fridge, window, sofa, dustbin, bed, shoes, TV, shelf, bike, bookshelf, sink, photo, and locker.

A.2 Data Collection

Our dataset for behavioral navigation through natural language instructions was collected in three steps:

1. Generation of Indoor Environments. We first generated 100 distinct layouts of indoor environments. For each of these layouts, we also created a corresponding behavioral navigation graph by positioning nodes on all relevant semantic locations of the environment. Edges were defined by finding the behavior from Table 1 (main paper) that best matched the navigation route between nearby nodes.

While creating graphs, we avoided defining edges that summarized other routes. For instance, if the triplets $\langle \text{node\_1}; \text{edge\_1}; \text{node\_2} \rangle$ and $\langle \text{node\_2}; \text{edge\_2}; \text{node\_3} \rangle$ were valid, we did not include in the graph another triplet that connected the first and third nodes, e.g., $\langle \text{node\_1}; \text{edge\_3}; \text{node\_3} \rangle$. This restriction was particularly relevant for situations where the environment included a long corridor, with multiple nodes connected by "follow the corridor" edges.

We used the Open Motion Planning Library (OMPL) (Şucan et al., 2012) to compute the shortest route between all pairs of rooms in all navigation graphs. We then randomly chose a subset of these routes for humans to provide a free-form natural language description.

2. Collection of Instructions. We collected natural language instructions that described the chosen routes from step 1 via crowdsourcing on Amazon Mechanical Turk (MTurk). Each MTurk Human Intelligence Task (HIT) showed the corresponding worker a 2D map of an environment with a navigation route to annotate. Fig. 1 shows the HIT instructions, and Fig. 2 shows the input user interface. The interface asked the worker to first identify the start location and the destination of the route in the map. Then, (s)he was asked to provide a free-form natural language description of the route. To ensure sufficient diversity in the language used to describe the instructions, each route in the training dataset was displayed to two different workers; except for the Test-new set, where we only had one HIT per route. A total of 822 workers provided route descriptions. We have provided a few examples of instructions of our dataset in section E.

3. Verification of Instructions. The natural language instructions provided in the previous step were verified through Mechanical Turk as well. Each instruction was shown to two different workers along with the corresponding 2D map of the environment (using a similar interface as for the previous step). Workers were asked to label the

*Both authors contributed equally to this work.

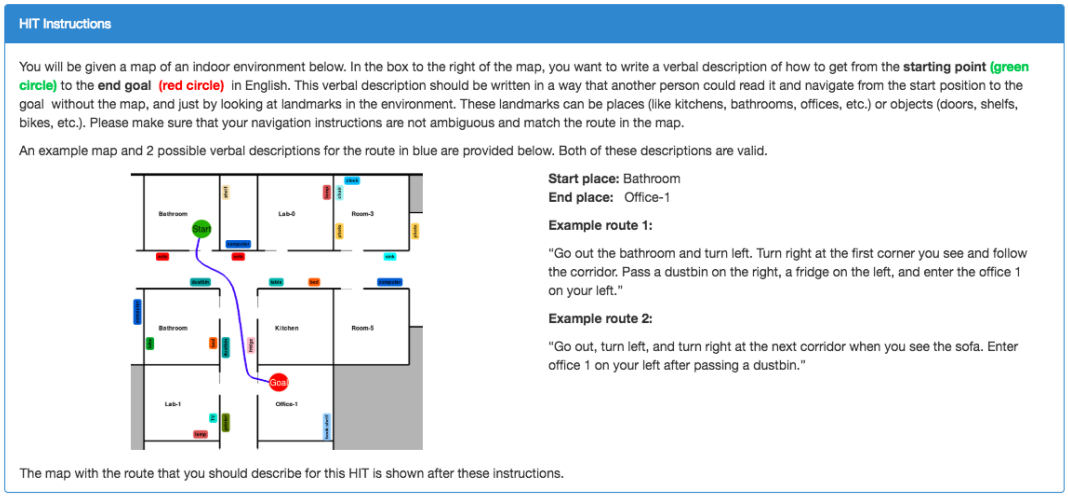


Figure 1: Instructions displayed to workers before accepting the HIT in Mechanical Turk.

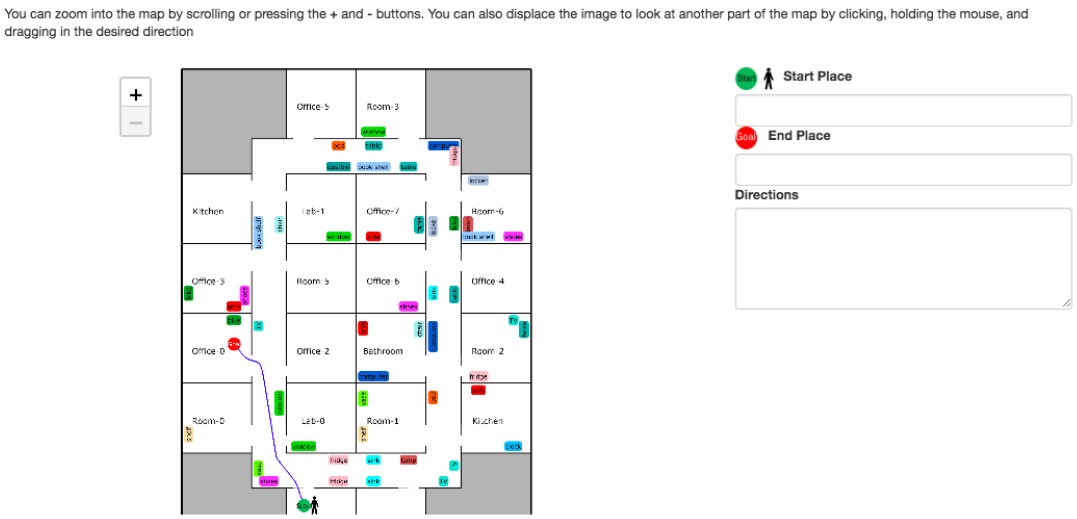


Figure 2: User interface used for data collection through Mechanical Turk.

starting point and destination of the route in the map. They were then asked to verify if the instruction provided from step 2 above was unambiguous and represented the route displayed in the map. We asked the workers to label a description as incorrect under the following conditions:

– The description referred to a route that started or ended at a location that was different from the actual starting location or destination of the route in the map.
– The description was ambiguous and did not have sufficient information about relative locations of landmarks to navigate the graph.
– The description mentioned landmarks that were not present near the route.
– The description provided incorrect directions to reach the destination.

Workers were also given an option to mark an instruction as incorrect for some other reason. They had to provide a justification in this case.

We included a navigation instruction in our dataset only if both of the workers that verified its description labeled it as correct.

A.3 Dataset Statistics

While selecting a subset of routes for annotation, we ensured that our training dataset and test set with new maps (Test-New Set) had similar distribution of routes in terms of their length, as illus-

Data Statistics	No. of Graph Triplets (Train)	No. of Graph Triplets (Test-new)	Instruction Length (Train)	Instruction Length (Test-new)	Behavior Freq (Train)	Behavior Freq (Test-new)
Min	27	51	2	7	1	1
Max	379	379	239	145	22	19
Mean	176.38	209.48	33.72	31.01	7.15	7.08
Std Dev	72.01	104.8	17.95	16.94	3.67	3.66

Table 1: Dataset statistics - summary of distribution of the number of triplets in graph, number of words in instructions, and number of behaviors in the predicted routes for training and testing (test-new) dataset.

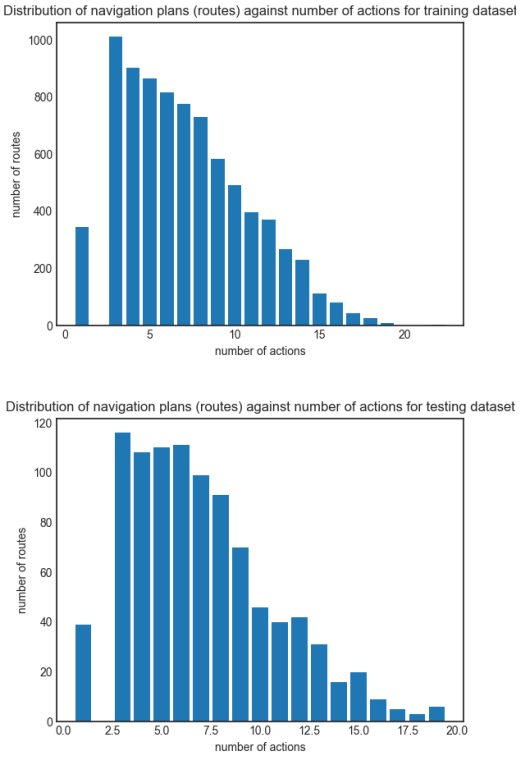


Figure 3: Distribution of routes by their number of behaviors for the training dataset (top) and for the Test-New set (bottom).

trated in Figure 3.

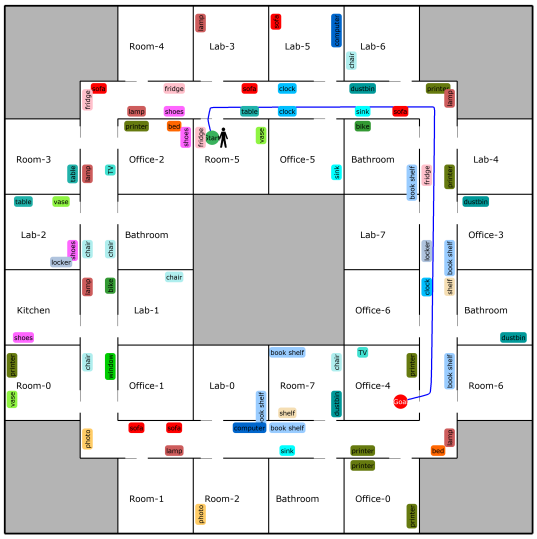
General statistics about the distribution of behaviors, routes and instructions in the training dataset and the Test-New dataset are summarized in Table 1. The table reports the minimum value, the maximum value, the mean and standard deviation of the number of triplets in the graph, and the number of words and behaviors in the instructions.

Note that we truncated the graph and instructions to a maximum of 300 tokens and 150 tokens, respectively, when we input them to translation models. The truncation was helpful to save memory at training time, and we expected it to have a minor effect on performance as only 6.4% (5.4%) of the unique graphs in the training (validation) set had more than 300 triplets. Additionally, less than 0.15% of the natural language instructions in these sets had more than 150 words.

A.4 Examples Environments in the Dataset

This section presents a example environments that we included in our dataset. Fig. 4 shows small environments, while Fig. 5 and 6 show environments of medium and big size, respectively.

Figure 4: Examples of small-sized maps ($\leq$ 20 rooms used for data collection. Best viewed in digital.



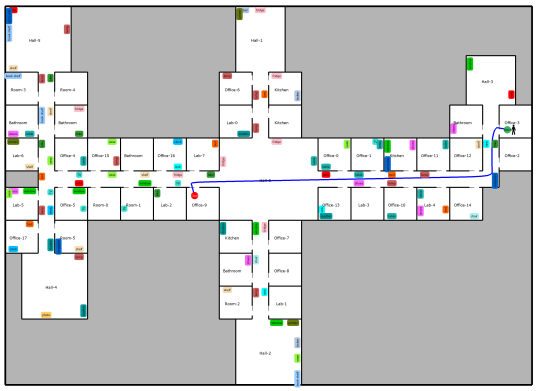
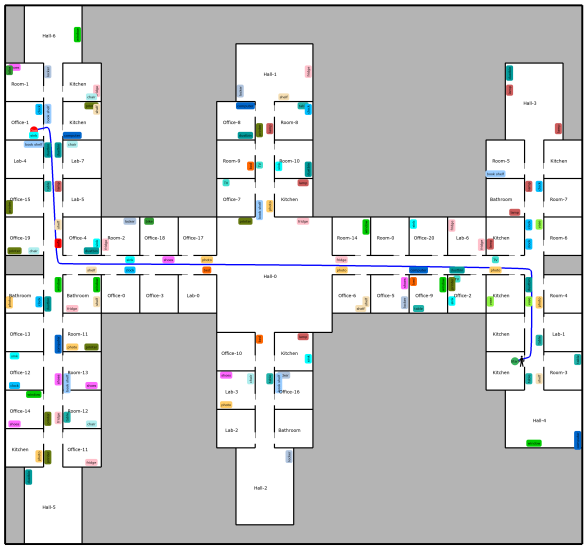
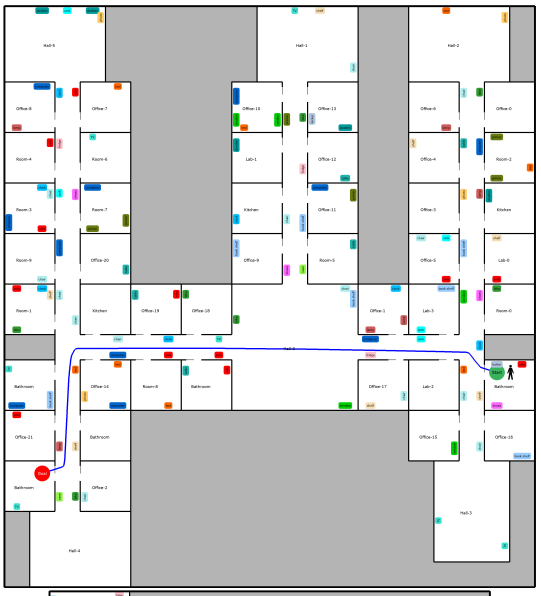


Figure 5: Examples of medium-sized maps (20 $<$ rooms $\leq$ 40) used for data collection.





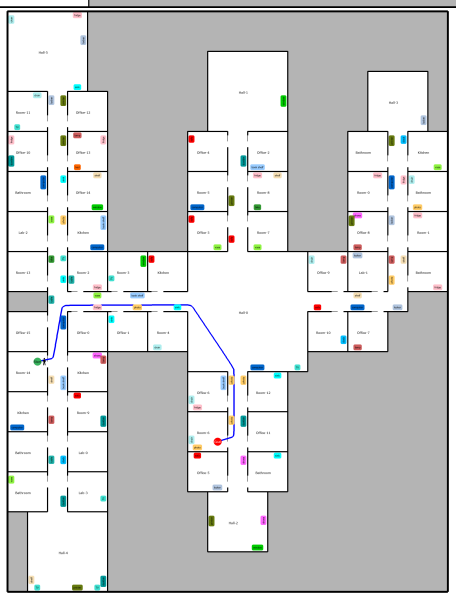


Figure 6: Examples of large-sized maps (rooms $>$ 40) used for data collection.

The green circles identified as "Start" in the Figures indicate the starting location for the navigation task. The red circles ("Goal") indicate the desired end location. The navigation route that we asked Turkers to describe during our data collection effort is drawn in blue.

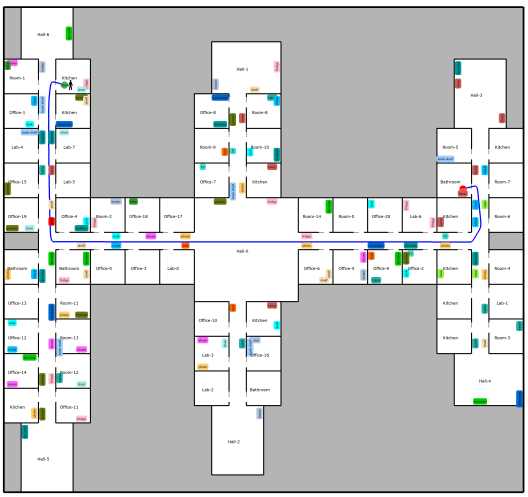
B Model Predictions

This section shows several examples of the predictions made by the proposed model. The predictions are encoded as a sequence of behaviors with the start location and the end location at both ends. The navigation behaviors that we considered in this work are defined in Table 1 of the main paper. Parameterized behaviors are associated with a turning direction <d> (either "l" or "r" for turning left or right, respectively).

In the examples presented in this section, we abbreviate room types by the type's first letter, e.g. the abbreviation for a hall is H and for a lab is L. Relevant locations for the navigation task are then encoded as the abbreviated room type followed by a room number. For example, K-8 means Kitchen 8, and R-4 corresponds to Room 4.

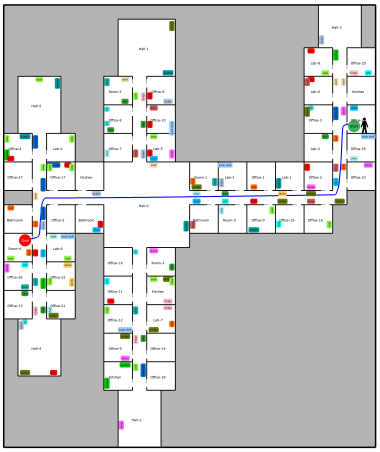
B.1 Examples

1. **Instruction:** *"Turn left as you exit the Kitchen and walk down the corridor until you see a sofa. Turn left at the sofa, walk through Hall-0, then turn left at the end of the corridor. Pass two clocks, and you will see the bathroom on your left."*



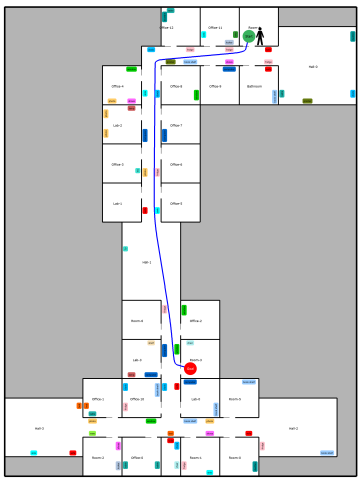
Prediction: [Correct] K-8 ool cf cf cf cf lt cf cf cf cf chs cf cf cf cf cf lt cf iol B-3

2. **Instruction:** *"Exit Office turning to the left, continue down until end of hall and make right at printer. Continue straight down this hall through Hall 0 going to the end, make a left turn at the computer. Go down this hall a short way and just past the bed turn right into Room 4."*



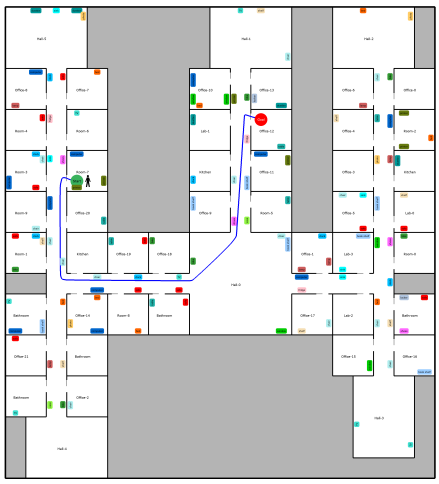
Prediction [Correct] O-24 ool cf cf rt cf cf cf cf cf chs cf cf lt cf ior R-4

3. **Instruction:** *"Exit Room-4 and turn right and continue to walk through the corridor. Then turn left at the corner with clock. Continue down the hallway, through hall-1, then enter the second door on the left, just past the window."*



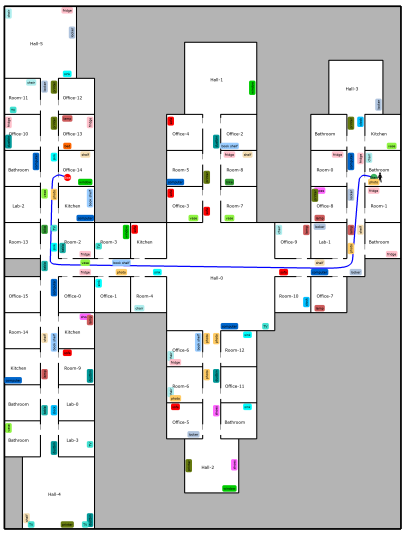
Prediction [Correct] R-4 oor cf cf lt cf cf cf cf chs cf cf iol R-3

4. **Instruction** *"Exit Room 7, turn to the left. Keep straight until you come to the chair on the left, turn left onto that hall keep straight until you get to Hall 0. Go to the left in Hall 0 and down the hall where there are shoes on the left and a vase on the right. Keep straight down this hall until you come to the fridge on the right. Turn right into Office 12 after the fridge."*



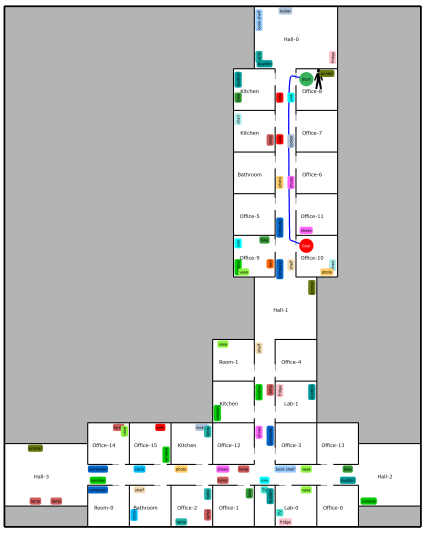
Prediction [Correct] R-7 ool cf cf lt cf cf cf chl cf cf cf ior O-12

5. **Instruction:** *"Go out the bathroom and turn left. Follow the corridor, which will turn to the right. Pass straight through the Hall-0 area. At the end of the corridor, turn right. Go between a vase and a photo and enter the next door on the right."*



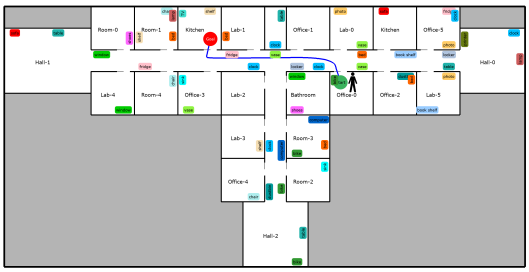
Prediction [Incorrect] B-1 ool cf cf rt cf cf chs cf cf cf rt cf cf cf cf ior O-12
Ground Truth B-1 ool cf cf rt cf cf chs cf cf cf rt cf cf ior O-14

6. **Instruction:** *"Exit office 8 and turn left. The fourth door on the left will be Office 10."*



Prediction [Incorrect] O-8 ool cf cf cf iol O-11
Ground Truth O-8 ool cf cf cf cf iol O-10

7. **Instruction:** *"Exit office 0 to the left, pass the locker, pass the vase, after the fridge enter the first doorway on the right, the kitchen."*



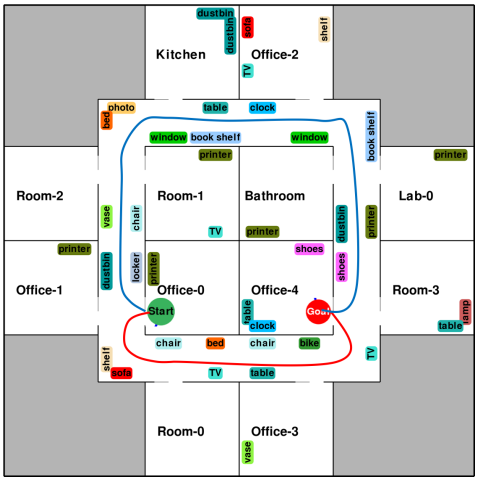
Prediction [Correct] O-0 ool cf sp cf ior K-0

C Examples of Sub-Optimal Paths

In this section, we show several examples of the model's predictions when the desired route is not necessarily the optimal path between the start location and the destination. These examples suggest that although our dataset only consisted of

the shortest routes between two nodes, our model is capable of following the desired navigation instruction.

1. Figure below describes two routes of the same start location and the destination. The red route is optimal and the blue route is sub-optimal. Our model made the correct predictions in both cases.



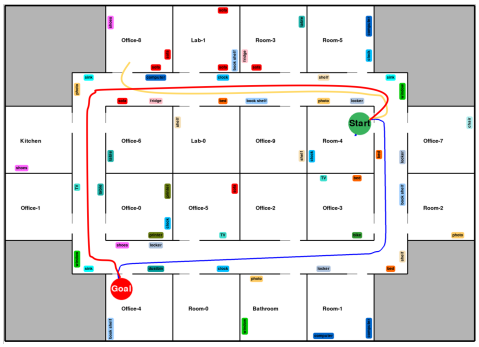
Red route [Correct]

- Instruction: *"Walk out of office and turn left. Turn left at the corner with sofa and shelf and continue to walk straight past a TV and a table on your right till you reach the corridor intersection with TV. Turn left at the intersection and walk into the first room on your left."*

Blue route [Correct]

- Instruction: *"Make a right after going out of the office. Turn right at the corner. Follow the corridor and turn right again at the next corner. Go straight and enter room on your right after passing the shoes."*

2. Both routes described in the figure below are from Room-4 to Office-4. Our model successfully followed the instruction describing the blue route but failed to follow the instruction describing the red route. Despite the error, the general direction of the the predicted routes match the desired routes. Thus, our model was capable of utilizing the input instructions. The cause of the error in this example could be the length of the red route. Our model seemed to have trouble processing the entire information correctly.



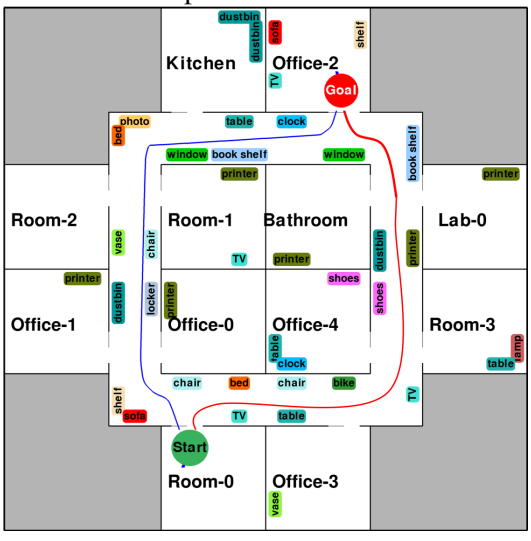
Red route [Incorrect]

- Instruction: *"Exit the room and turn left. Walk till you reach the corner and turn left. Continue to walk straight till you reach the opposite end of the corridor with a photo and a sink. Turn left and walk till the corridor intersection with a sink and a window. Turn left and enter the first room on your right."*
- Prediction: R-4 ool lt cf cf cf ior O-8
  A route which is from Room-4 to Office-8 as described in the yellow line in the figure above.

Blue route [Correct]

- Instruction: *"Take a right out of room 4 and walk down the corridor till you reach the intersection with a bed and shelf. Turn right at the corner. Walk down the corridor to enter the fourth door on your left opposite shoes."*

3. Figure below shows an example of two optimal routes. Both routes are from Room-0 to Office-2 on the map. The model was able to predict the desired routes.



Red route [Correct]

- Instruction: *"Exit the room 0 and turn right, go to the end of the corridor and turn left, go straight to the end of the corridor and turn left again. After passing bookshelf on your left and table on your right, enter the kitchen on your right."*

Blue route [Correct]

- Instruction: *"Exit the room and turn left. Turn right at the corner and go straight to the end. Make a right again and enter the first room on your left. "*

D Attention Visualization

Figure 7 displays examples of the attention distributions from the decoder layer of the proposed model. The color-coded and numbered regions on the 2D map (left) correspond to the triplets that are highlighted with the corresponding color in the attention plot (right). All three examples achieve correct predictions.

Figure 7a and Figure 7c are examples where attention map is interpretable from the graph. Figure 7b depicts an attention map that is rather hard to interpret, although the prediction is correct.

E Examples of Natural Language Instructions in the Dataset

To emphasize the diversity and complexity of our dataset, we provide a few examples of natural language instructions that describe the same route in a map.

- *– "Exit Office-7 into corridor. Turn right and advance through corridor past the shelf and bed at the corner. Turn left at the shelf and proceed to the second door on your right. Turn right at that door and advance into Office 7."*

  – "Go out, turn right, walk to the end of the corridor, turn left at the bed, keep walking straight and enter Office-7 to the right at the first door past the bed and locker."

- *– "Take a left out of Room-4 and another left at the corner. Walk down the hallway to enter the fourth door on your right."*

  – "Exit Room-4 into corridor. Turn left and advance to the corner with sink and window. Turn left and advance through corridor. Take a right turn to enter the room immediately after computer and fridge."

- *– "Go out Office-11 and turn right. Turn left at the first corner where there is a printer on your left. Enter the second door on your left, which is Office-7."*

  – "Turn right after leaving office 11. Pass a fridge on the right and a bookshelf and a printer on your left. Turn left at the corner and go past a clock on the left and a sink on the right. Turn left into the room immediately after clock to enter office 7."

- *– "Exit Room-3, turn right, and continue down the hall into hall-1. Continue down the hallway, turn right at the corner with the clock, then take the second right door, past the shoes, into the bathroom."*

  – "Exit Room-3 and make a right in the hallway. Continue all the way down the hallway, through Hall-1 and down the next corridor till you reach the corner with clock. After taking a right turn at the corner, enter the second door on your right just past the shoes."

- *– "Turn left out of Office-4 and walk down the corridor opposite of it. Walk into the first door on the right to enter Office-9."*

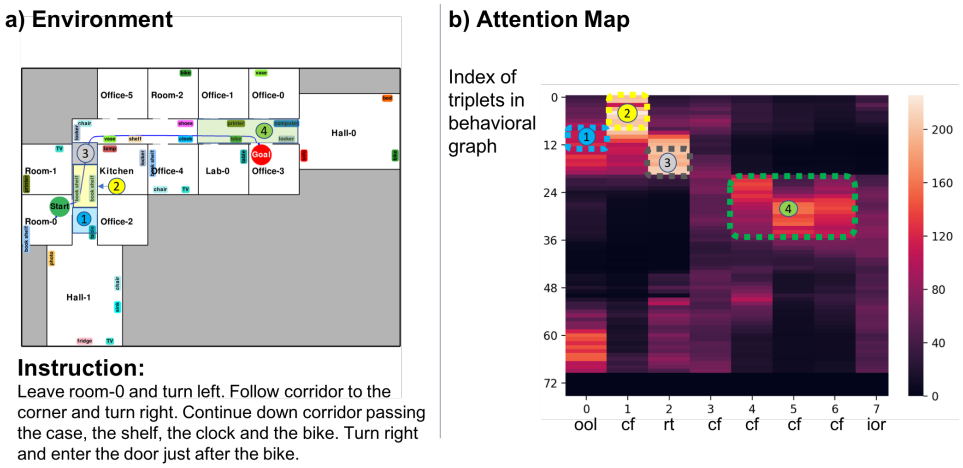
  – "Exit office-4, turn left then right at the clock. Enter the first door on the right just past the bookshelf."

- *– '"Get out of the office-6 and walk right, get into the office on the left which is the last office before the end of the corridor and immediately after the sink."*

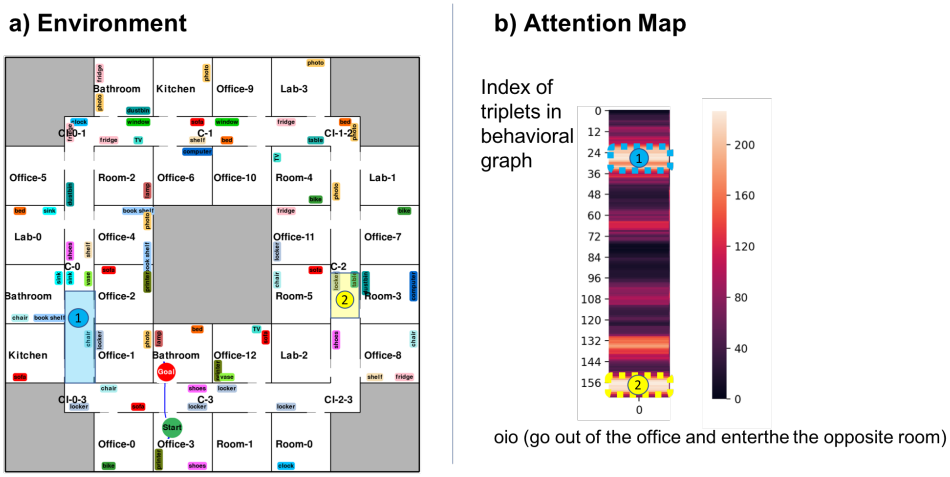
  – "Exit Office-6 and turn right, enter the second door on the left, just past the sink. "

References

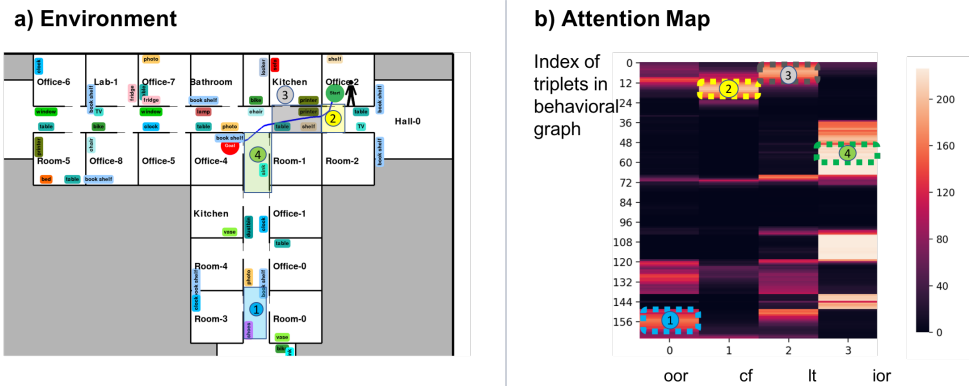
Ioan A. Şucan, Mark Moll, and Lydia E. Kavraki. 2012. The Open Motion Planning Library. *IEEE Robotics & Automation Magazine*, 19(4):72–82. http://ompl.kavrakilab.org.



(a) First attention example



(b) Second attention example



(c) Third attention example

Figure 7: Attention visualizations. The color-coded and numbered regions on the map (left) correspond to the triplets that are highlighted with the corresponding color in the attention map (right). In all the cases, model achieves the correct prediction.