Supplementary Material: Sim-to-Real Transfer for Vision-and-Language Navigation

Peter Anderson[1]* **Ayush Shrivastava**[1] **Joanne Truong**[1] **Arjun Majumdar**[1]
Devi Parikh[1,2] **Dhruv Batra**[1,2] **Stefan Lee**[3]

[1]Georgia Institute of Technology [2]Facebook AI Research [3]Oregon State University

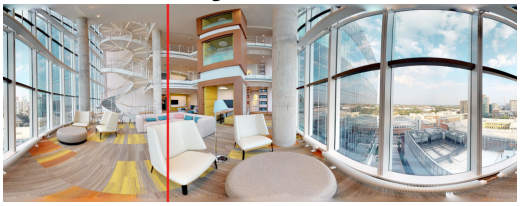
		Matterport3D / R2R Dataset					
		Train ($n = 61$)			Val-Unseen ($n = 11$)		
	Coda	Min	Avg	Max	Min	Avg	Max
Num Viewpoints	59	8	125	345	20	87	215
Navigation Graph Degree	3.6	2.2	4.0	5.4	3.1	3.8	4.9
Avg Edge Distance (m)	2.8	1.3	2.2	3.1	1.8	2.2	2.8
Num Instructions	111	6	230	300	18	214	300
Avg Instruction Length (words)	25	20	29	35	22	28	32
Avg Trajectory Length (m)	12.0	5.3	9.7	15.0	6.1	9.2	11.1
Avg Trajectory Edges	4.8	3.0	4.9	5.4	3.9	4.7	5.2

Table 1: Comparison of per-environment average statistics between Coda and R2R, suggesting that Coda is fairly typical of environments found in the Matterport3D / R2R dataset.

*Now at Google.

Turn right, move through the open open double doors. Turn right, move to the end of the hall. Turn right, wait in front of the drinking fountain.

Walk straight towards the exit, but stop and make a right when you see the wall with art and stop.



Make a right in front of the couch, veer straight, then right to proceed past the windows and stop between the first two bookshelves in front of the first hanging piece of art.

Go between the first set of bookshelves. Turn left and go straight until you are at the end of the hallway.

Turn around and walk to towards the end of the hall. At the intersection, turn and head into the men's room.

Continue forward with the whiteboards on your left. Keep walking and you will be in a new room. Stop before you reach the peach-colored couch on your right.

Figure 1: Additional examples of navigation instructions in the Coda environment. Each instruction is shown with the panoramic view from the starting pose, with the initial heading indicated in red.

Figure 2: Panoramic captures in Coda from the Matterport3D camera (row 1) and the smaller and cheaper Ricoh Theta V camera (mounted on the robot) collected on three different days (rows 2-4). The robot camera's limited dynamic range and loss of detail as compared to the Matterport3D camera (used for training the VLN agent) is clearly evident. Images collected on different days (with the robot camera) illustrate the variations in shadows, lighting, and precise object placement that confront the robot in the real physical environment.

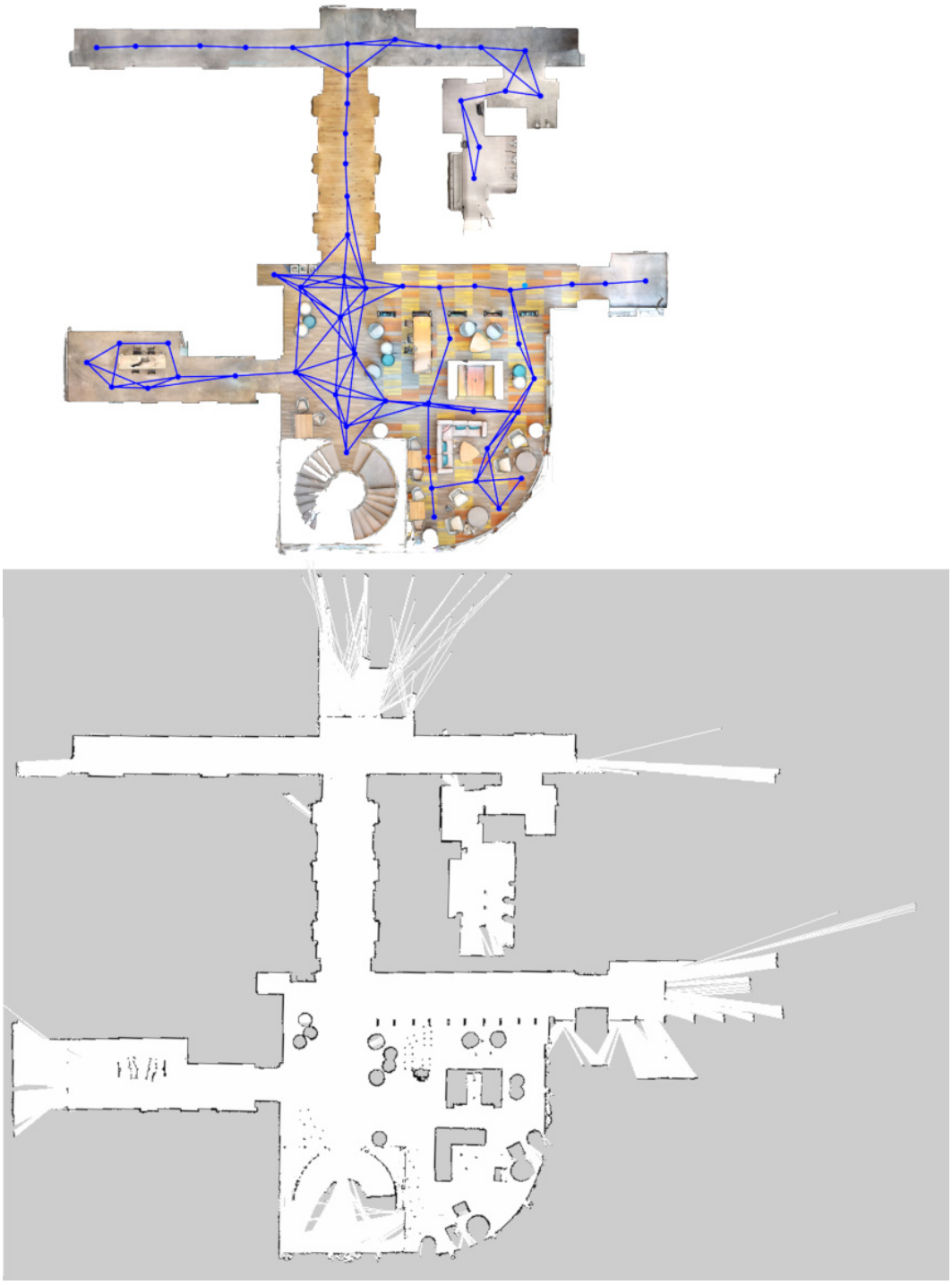


Figure 3: Floorplan view of Coda, showing the Matterport reconstruction and simulator navigation graph (top), and its close alignment to the 2D laser scan used for robot pose tracking (bottom).

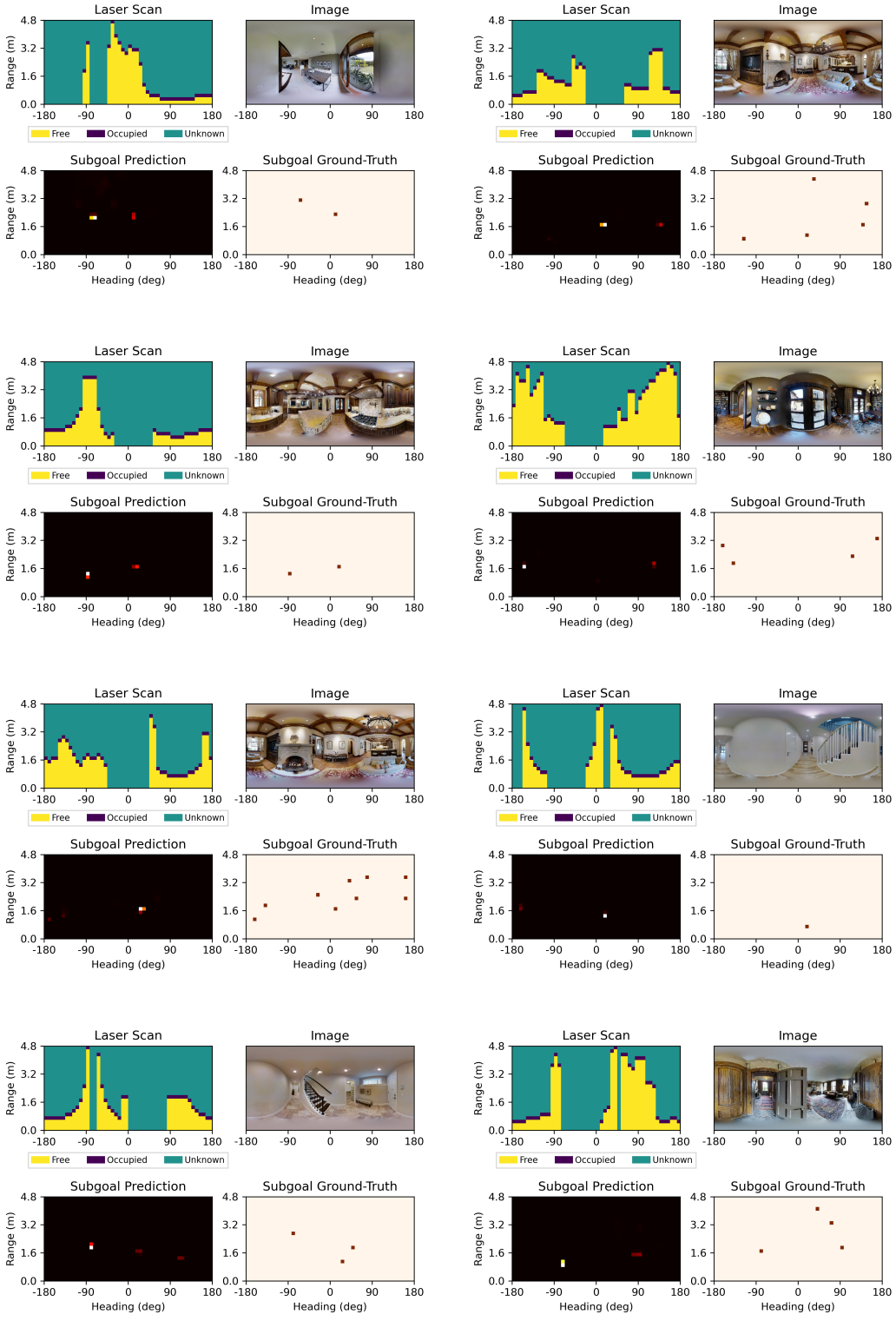
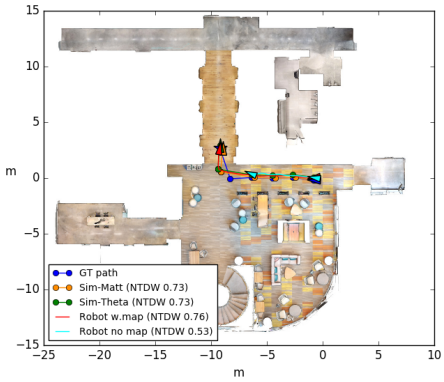
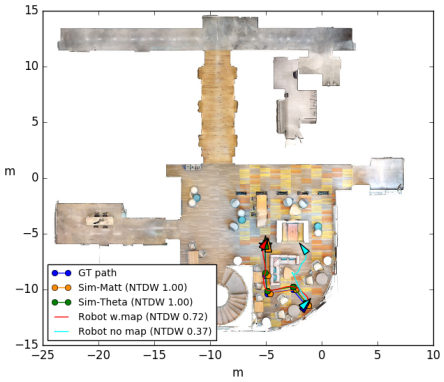


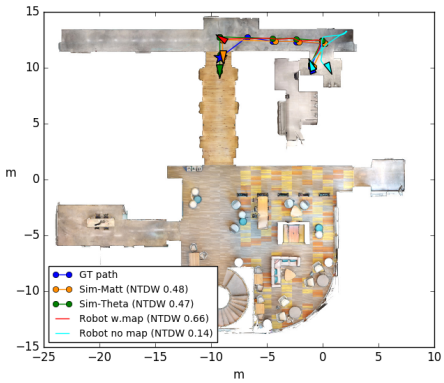
Figure 4: Waypoint predictions from the subgoal model on 8 randomly selected viewpoints from the Matterport validation set.



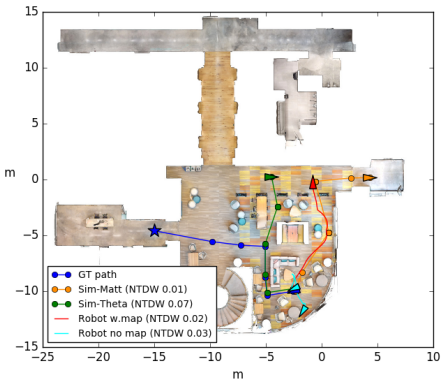
Instruction: Walk down the hall with the brown shelving units on your left. Turn right into the hallway with the elevators and then stop.



Instruction: Go between the table bookcase and the sectional sofa on this floor.



Instruction: Walk into hallway. Make a left at closed brown door. Walk down hall and make a left and stop by open white doors.



Instruction: Walk around the back of the large couch. Turn left towards the open green doorway. Walk through the green doors and stop.

Figure 5: Examples of Coda trajectories in sim and real for various instructions. While the robot's trajectory often resembles the simulator (top-left), subgoal prediction errors can lead to divergences between the 'with map' and 'no map' settings (top-right), particularly in areas of the building with floor-to-ceiling glass walls that are not easily detected (bottom-left). In the last example (bottom-right) the agent fails in both sim and real, highlighting the challenging nature of the VLN task.

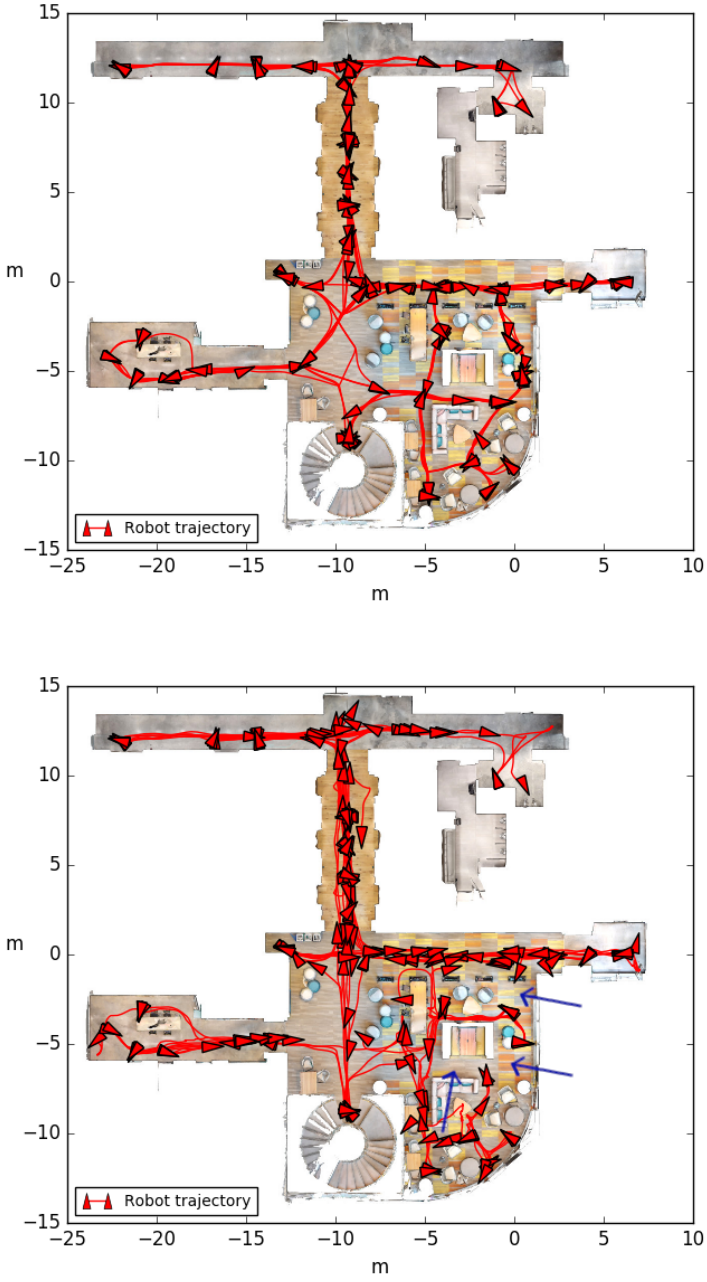
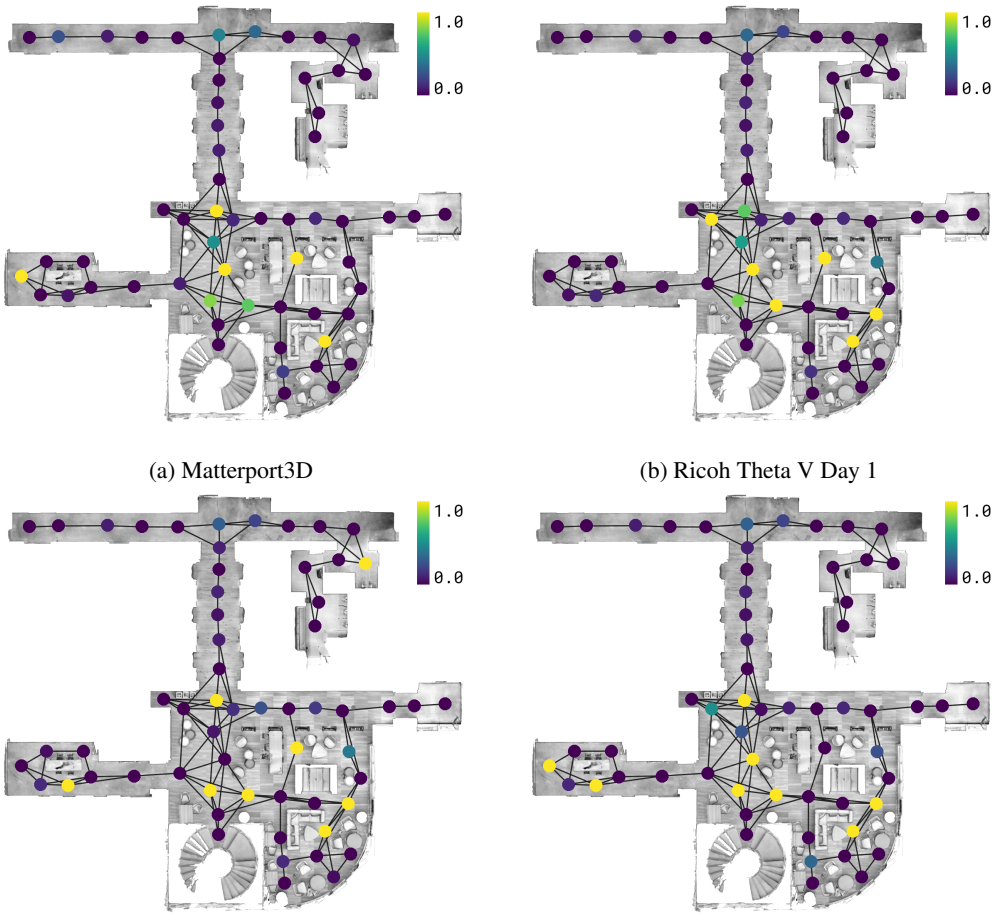


Figure 6: Illustration of all 111 of trajectories traversed by the robot under the 'with map' setting (top) and the 'no map' setting (bottom). With a map, the robot traversed the entire space without any collisions or navigation failures. Without a map, certain trajectory segments (highlighted) with blue arrows are never traversed, indicating that the subgoal model failed to predict these waypoints.



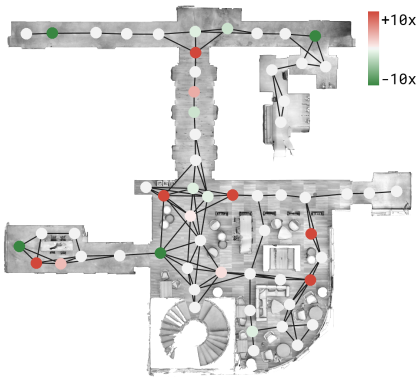
(a) Matterport3D

(b) Ricoh Theta V Day 1

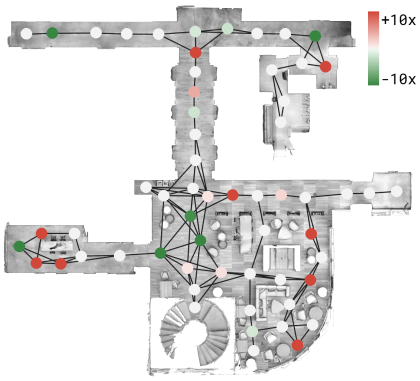
(c) Ricoh Theta V Day 2

(d) Ricoh Theta V Day 3

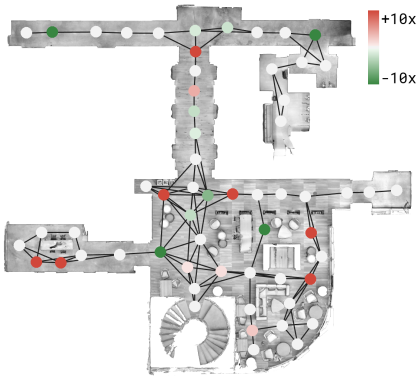
Figure 7: Illustration of the VLN agent's failure rates (in simulation) at each node in the navigation graph. The failure rates are consistently higher (yellow) in the bottom right of the map across all four data collections with the Matterport3d and Ricoh Theta V cameras.



(a) Failure Rate Ratio - Ricoh Theta V Day 1 to Matterport3D



(b) Failure Rate Ratio - Ricoh Theta V Day 2 to Matterport3D



(c) Failure Rate Ratio - Ricoh Theta V Day 3 to Matterport3D

Figure 8: Illustration of the log of the ratio between the failure rates with the Ricoh Theta V camera and the Matterport3D camera. A positive ratio, illustrated in red, indicates that the VLN agent was more likely to fail when processing data from the Ricoh Theta V camera. A negative ratio (in green) indicates the opposite. Across all three days, the agent was more likely to fail at nodes to the top and bottom of the elevator area and at nodes near the glass windows in the open space in the bottom right. Surprisingly, there are also some nodes (in green) at which the agent consistently performs better using the Ricoh Theta V panoramas.