9. Supplementary

9.1. VO Prediction Error

We report Mean Absolute Error (MAE) Eq. (3) between ground-truth and estimated egomotion in Tab. 4.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left(|x-\hat{x}|+|y-\hat{y}|+|z-\hat{z}|\right)+\frac{1}{N}\sum_{i=1}^{N}|\theta-\hat{\theta}| \tag{3}$$

	Dataset	VO		Embedding		Train time		Epoch	Translation MAE (cm)				Rotation MAE (centi-radians)			
	size(M)	Encoder	Size(M)	1FC	2FC	Flip	Swap		Total	Forward	Left	Right	Total	Forward	Left	Right
1	0.5	ResNet18	4.2					50	2.65	2.21	3.14	3.30	1.00	0.66	1.41	1.45
2	0.5	ResNet18	4.2	✓				43	2.45	1.82	3.17	3.37	0.90	0.56	1.30	1.39
3	0.5	ResNet18	4.2	✓	✓			44	2.38	1.78	3.08	3.22	0.86	0.55	1.23	1.31
4	0.5	ResNet18	4.2	✓	✓		✓	48	2.60	2.22	3.06	3.12	0.90	0.66	1.21	1.2
5	0.5	ResNet18	4.2	✓	✓	✓		50	2.26	1.77	2.86	2.92	0.75	0.49	1.09	1.1
6	0.5	ResNet18	4.2	✓	✓	✓	✓	50	2.26	2.02	2.56	2.56	0.75	0.55	0.98	1.03
7	1.5	ResNet18	4.2	✓	✓			48	1.94	1.33	2.72	2.71	0.69	0.43	1.03	1.02
8	1.5	ResNet18	4.2	✓	✓	✓	✓	50	1.7	1.48	1.94	2.02	0.61	0.48	0.75	0.81
9	1.5	ResNet50	7.6	✓	✓	✓	✓	48	1.48	1.34	1.63	1.68	0.49	0.38	0.57	0.67

Table 4. Mean Absolute Error (MAE) between ground-truth and estimated egomotion for all types of actions in total and for each type of actions separately (corresponding to Tab. 1).

9.2. Qualitative Results

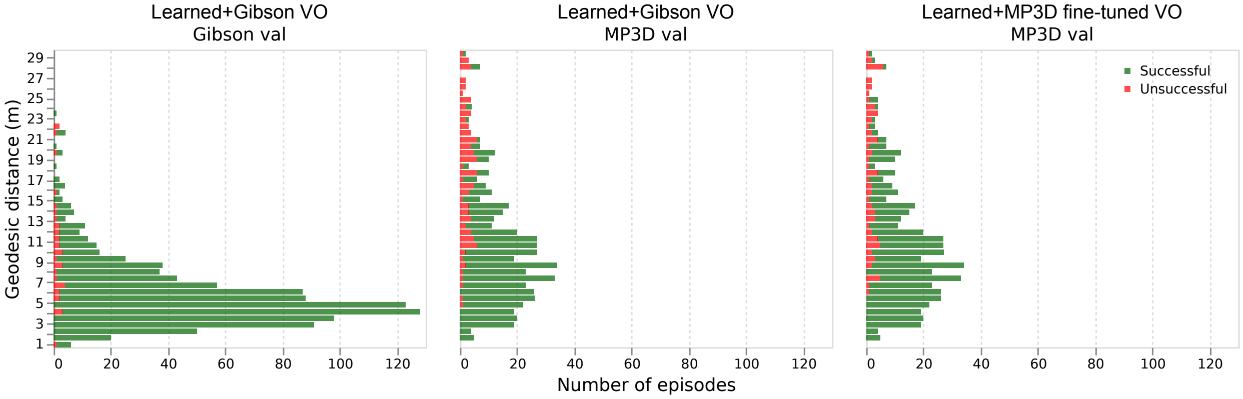
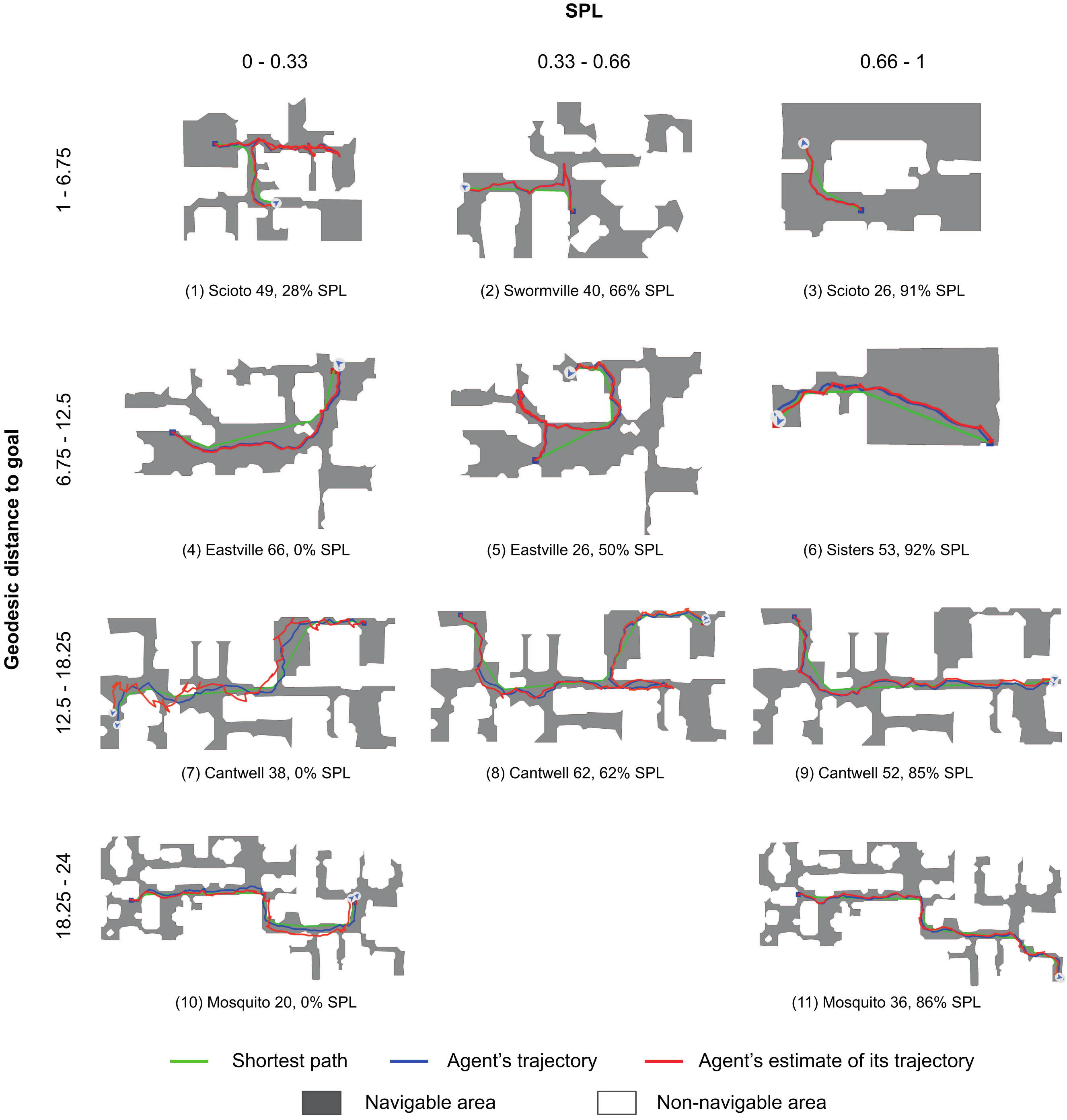


Figure 6. Success vs. path length. Our method performs worse on longer episodes. Fine-tuning on MP3D improves performance.

Complementary to Tab. 2 we provide additional qualitative results when integrating the navigation policy with our VO model. On Gibson 4+ val scenes dataset (navigation trajectories illustrated in Fig. 7) our navigation agent follows a near-perfect path for both: episodes with a relatively small geodesic distance to the target (row 1 – row 2) and episodes with a large geodesic distance to the target (row 3 – row 4) and can do backtracking when the wrong way was chosen (top-down maps (1), (2), (5) and (8)). We found that on Matterport3D (MP3D) (navigation trajectories illustrated in Fig. 8) the navigation performance suffers (see Tab. 2). First of all, the Matterport3D navigation episodes are 'longer': $10.92m$ average geodesic distance to goal on MP3D vs $5.89m$ average geodesic distance to goal on Gibson 4+. Larger scenes usually have more than one way to the target place and if agent chooses longer way it affects navigation metrics (top-down maps (1), (2), (5) and (8)). We also noticed that it is harder for agent to backtrack in larger scenes (top-down map (10)).



(1) Scioto 49, 28% SPL

(2) Swormville 40, 66% SPL

(3) Scioto 26, 91% SPL

(4) Eastville 66, 0% SPL

(5) Eastville 26, 50% SPL

(6) Sisters 53, 92% SPL

(7) Cantwell 38, 0% SPL

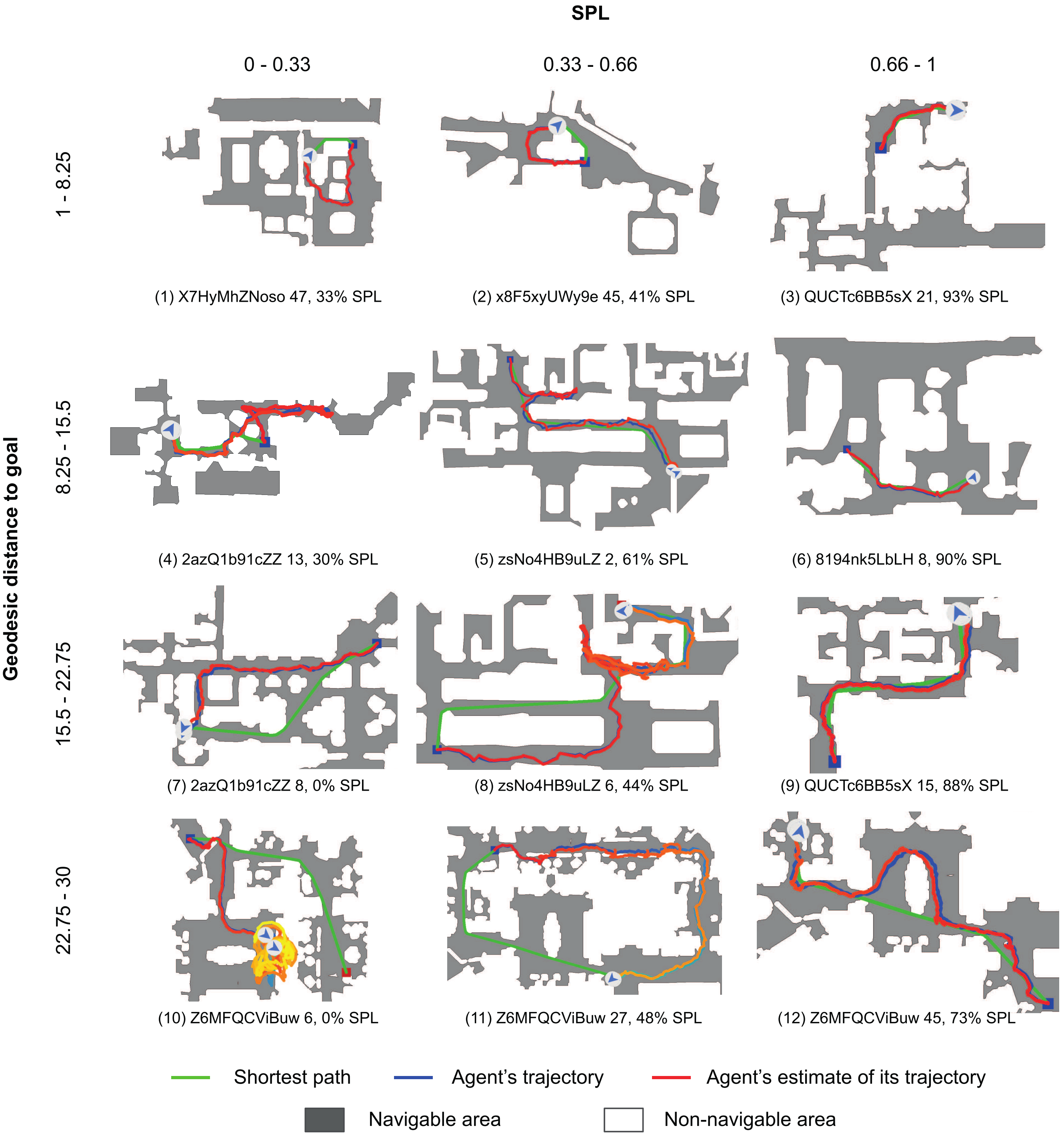
(8) Cantwell 62, 62% SPL

(9) Cantwell 52, 85% SPL

(10) Mosquito 20, 0% SPL

(11) Mosquito 36, 86% SPL

Figure 7. Our best HC 2021 PointNav agent's (row 16 in Tab. 1) navigation trajectories on Gibson 4+ val scenes (and Gibson-v2 PointGoal navigation episodes) broken down by geodesic distance between agent's spawn location and target (on rows) vs SPL achieved by the agent (on cols). The color of the trajectory changes from dark to light over time (`cv2.COLORMAP_WINTER` for agent's trajectory, `cv2.COLORMAP_AUTUMN` for agent's estimate of its trajectory).



(1) X7HyMhZNoso 47, 33% SPL

(2) x8F5xyUWy9e 45, 41% SPL

(3) QUCTc6BB5sX 21, 93% SPL

(4) 2azQ1b91cZZ 13, 30% SPL

(5) zsNo4HB9uLZ 2, 61% SPL

(6) 8194nk5LbLH 8, 90% SPL

(7) 2azQ1b91cZZ 8, 0% SPL

(8) zsNo4HB9uLZ 6, 44% SPL

(9) QUCTc6BB5sX 15, 88% SPL

(10) Z6MFQCViBuw 6, 0% SPL

(11) Z6MFQCViBuw 27, 48% SPL

(12) Z6MFQCViBuw 45, 73% SPL

Figure 8. Our best HC 2021 PointNav agent's (row 16 in Tab. 1) navigation trajectories on MP3D val scenes (and MP3D-v2 PointGoal navigation episodes) broken down by geodesic distance between agent's spawn location and target (on rows) vs SPL achieved by the agent (on cols). The color of the trajectory changes from dark to light over time (`cv2.COLORMAP_WINTER` for agent's trajectory, `cv2.COLORMAP_AUTUMN` for agent's estimate of its trajectory).

10. Simulation-to-reality Transfer

Episode name	Run	Path length (m)		Navigation metrics			Success			
		Shortest	Agent's	d_G	SoftSuccess	SoftSPL	$d_G < 0.36$	$d_G < 0.395$	$d_G < 0.45$	$d_G < 0.70$
kitchen2couch	1	4.94	7.04	0.39	0.92	0.70	0	1	1	1
kitchen2couch	2	4.94	9.06	0.44	0.91	0.54	0	0	1	1
kitchen2couch	3	4.94	7.69	0.25	0.95	0.64	1	1	1	1
desk2bathroom	1	8.37	11.26	0.38	0.96	0.74	0	1	1	1
desk2bathroom	2	8.37	12.32	0.69	0.92	0.68	0	0	0	1
desk2bathroom	3	8.37	9.01	0.76	0.91	0.93	0	0	0	0
bed2desk	1	8.23	11.82	0.65	0.92	0.70	0	0	0	1
bed2desk	2	8.23	10.77	1.00	0.88	0.76	0	0	0	0
bed2desk	3	8.23	11.98	0.60	0.93	0.69	0	0	0	1
			Average:	0.57	0.92	0.71	0.11	0.33	0.44	0.78

Table 5. **Sim2real transfer.** The shortest path length was calculated using RRT* [38] for 5000 iterations.

Policy Training. We use a similar training recipe as in the main paper to train our policy with the follow modifications. We alter the simulated camera config (e.g., FOV, mounting height) to match the Intel RealSense D435 depth camera attached on our robot. Additionally, we use all of Gibson [36], MP3D [5], and HM3D [25] for training scenes to increase diversity.

Visual Odometery Training. For visual odometry training, we again modify the camera config to match both the RGB and depth cameras of the Intel RealSense D435. Note that we use the unaligned RGB and depth output from the camera (i.e., we do not do a re-centered crop on the depth images) to leverage the fact that the depth camera's field-of-view is larger than that of the RGB camera and increase the amount of overlap between observations after a rotation action.

Deployment details. We use a LoCoBot [11] with the Kobuki base. We mount an Intel RealSense D435 camera 0.61 meters from the ground with a tilt angle of 0 degrees. The policy only uses the RGB and depth images from the RealSense camera as input. The depth images were de-noised using a median filter. Mapping and localization using the robot's LiDAR are used solely for visualization and calculating the lengths of the shortest path and the agent's path.