Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI Supplementary Materials

Santhosh K. Ramakrishnan[1,2], **Aaron Gokaslan**[1,5], **Erik Wijmans**[1,3], **Oleksandr Maksymets**[1], **Alex Clegg**[1], **John Turner**[1], **Eric Undersander**[1], **Wojciech Galuba**[1], **Andrew Westbury**[1], **Angel X. Chang**[4], **Manolis Savva**[4], **Yili Zhao**[1], **Dhruv Batra**[1,3]
[1]Facebook AI Research [2]UT Austin [3]Georgia Tech [4]Simon Fraser University [5] Cornell University

This document provides additional information about the HM3D dataset as well as qualitative and quantitative results to support the analysis from the main paper. Below is a summary of the sections in the supplementary file:

S1 Limitations of Habitat-Matterport 3D dataset

Data acquisition: The dataset is currently limited to scans from 38 countries. The dataset is restricted to contain data from building-owners who can afford to purchase the Matterport Pro2 sensor (which costs $\sim$ 3,000$) and have internet access to upload data to the cloud. The dataset also excludes regions where the Matterport Pro2 is not available to purchase. Due to these factors, we are limited in the types of regions and neighborhoods which can be included in the dataset. This can introduce an unintended bias into the algorithms developed based on our dataset, where the algorithms work only in a subset of buildings that we encounter in the real world. Nevertheless, this dataset is a significant leap from past building-scale datasets [1–3] that were restricted to labs, residences, and offices. We hope to expand our data set in the future to include scans from many more diverse backgrounds and countries.

Task support: The dataset only supports geometric tasks in static (i.e., unchanging) environments and does not include semantic annotations. We plan to investigate augmenting the dataset with semantic annotations to tackle high-level understanding tasks like object retrieval. We also plan to study dynamic and changing environments so that our simulations will be fluid rather than static. This would bring simulated training environments closer to the real world, where people and pets freely move around and where everyday objects such as mobile phones, wallets, and shoes are not always in the same spot throughout the day.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

S2 Hyperparameters for PointNav experiments

We use the publicly available implementation of DD-PPO [4] from Habitat Lab. We use the same hyperparameters as Wijmans et al. [4] for our experiments. We use a ResNet-50 [5] backbone and an LSTM with 512-D hidden states and 2 layers. Following Wijmans et al. [4], we replace BatchNorm with GroupNorm layers in the ResNet-50 backbone. We use a PPO clip parameter of 0.2, 2 PPO epochs, 2 mini-batches, a value loss coefficient of 0.5, entropy coefficient of 0.01 and a learning rate of 0.00025. Please see the default configuration here for more details. We train each model for 1.5 billion steps (sufficient for convergence) with 256 parallel environments divided between 8 nodes, 4 workers (i.e., 4 GPUs) per node and 8 environments per worker.

S3 Computational requirements for experiments

The PointNav experiments were the most computationally expensive of all our experiments. Each experiment is run in our internal cluster in a distributed fashion over 8 nodes, with 4 GPUs per node. Each GPU (32 in total) is a Volta 16/32 GB. Training an agent takes 2-3 days with depth inputs, and 4-5 days with RGB inputs.

S4 Accessing Habitat-Matterport 3D dataset

HM3D is free and available for academic, non-commercial research here:
`https://matterport.com/habitat-matterport-3d-research-dataset`
The terms of use are available here:
`https://matterport.com/matterport-end-user-license-agreement-academic-use-model-data`

S5 Habitat-Matterport 3D dataset collection process

The 1000 scans in HM3D were collected by Matterport Inc. in collaboration with the Habitat team at Facebook AI Research. Matterport directly contacted its users explicitly requesting them to contribute their scans for open-sourced Embodied AI research (see mailer snippet below).

> Imagine if firefighters could ask a robot to detect where smoke is coming from within your house, then command it to find people who need help. Or, if you could ask an AI assistant to locate your car keys. To realize innovations like these, robots and AI assistants need to be trained in how to act in multiple environments. They must learn to recognize and navigate through 3D spaces. That's where you come in. We have identified your Matterport 3D model as an ideal space for an open-source AI project focused on furthering such causes.

Each user agreed to the following terms while contributing their scans for the dataset.

> I agree to allow Matterport to use the Space(s) (including all related imagery) that I have designated in this form for academic and/or non-commercial purposes as further provided in the Matterport Terms and Conditions for Academic and Non-Commercial Use of Spaces, without payment by Matterport for such use. I affirm that I have all necessary rights, consents and permissions relating to my Space(s) necessary to grant the foregoing permission. By checking this box, I specifically agree to all of the provisions of the Matterport Terms and Conditions for Academic and Non-Commercial Use of Spaces & Matterport Privacy Policy.

After obtaining scans from users, Matterport used commercially reasonable efforts to try and obscure personally identifiable information such as pictures of people or faces, names, documents with personal information, diplomas, driver's licenses, email addresses, phone numbers, street addresses, personal notes / letters / envelopes, employer information, license plates, and street names. Human reviewers were asked to preview images from every scanned location to check for the above information and annotated each instance of personal information using a label. Any personal information identified in the previous step was blurred using a pixel-wise blur mask. The blurred data was used to recreate the 3D scans. A helpful FAQ regarding this process can be found here:

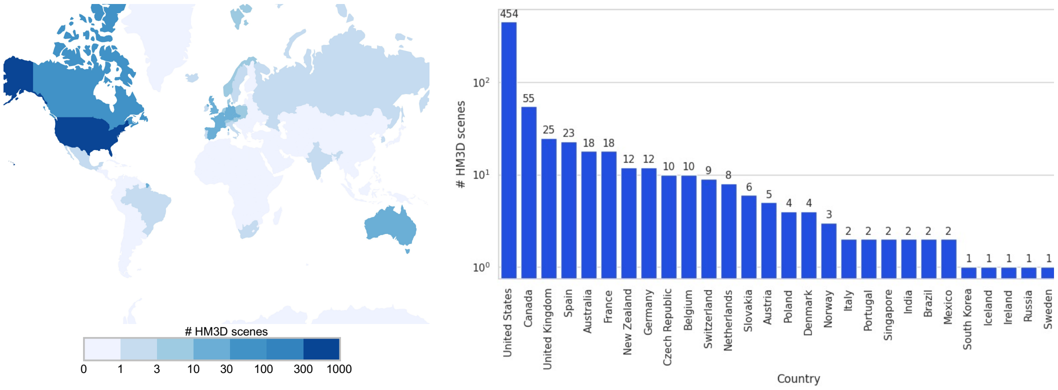

Figure S1: Visualizing the distribution of countries included in HM3D scenes. In total, we obtain scans from users in 38 countries. The left plot shows the distribution over different countries in the world map with the hue representing the number of scenes. The right plot shows a histogram over the different countries. Note: Due to missing data from certain scans, we only visualize the statistics over 30/38 countries in this figure.

https://go.matterport.com/ProjectHabitat.html. Overall, the dataset contains scans from users in 38 countries (see Figure S1).

S6 Rejection criteria for HM3D dataset collection

In the previous section, we provided details about how HM3D was collected and the steps taken to de-identify personal information. We now provide the criteria used to reject and exclude scenes from the dataset. We rejected scenes that were incomplete, had significant reconstruction artifacts, or included people. Specifically, we rejected scenes which had:

- Large holes/cracks in the floors, walls, or staircases which lead to separated "islands" within a single scene. These impact navigability between different parts of the scene for embodied agents in simulation.
- Large holes/cracks on object surfaces which can impact perception as well as potentially hinder future semantic annotations.
- Reconstruction artifacts/errors (i.e. bogus meshes) due to reflections and lighting. The scanning process occasionally misinterpreted unusual lighting/shadows as objects.
- Large outdoor areas in which many parts were out of range of the depth sensors.
- Uninteresting content with limited to no interactive potential (e.g., large empty parking lots, basement storage areas, abandoned warehouses, empty halls, hazardous waste store room).
- Objects that changed positions in the middle of the scanning process. For example, the door to a room may be open during a part of the scanning process, and closed for the latter part. This gave rise to artificial debris in the room entrance.
- Many doors that were scanned partially or fully closed (affects navigability between rooms).
- Brand names which were clearly visible (usually commercial buildings).
- Humans that were partially or fully visible in the scan.
- Other scenes which were naturally hard to scan and had poor reconstruction quality (e.g., submarine and ship exhibits, open-pit mines with equipment like excavators).

It is possible that these choices may impact the distribution over real scenes since some artifacts are correlated with either challenging surface materials or lighting conditions in the real world. However, we chose to focus on interiors, and tried to minimize reconstruction issues that affect navigability in simulation. Particularly, we prioritized surface completeness and the absence of spurious geometric artifacts as these have the biggest impact on RGB-D rendering in simulation (as we note in Sec. 3.2).

		Gibson (val)		MP3D (val)		HM3D (val)	
	Dataset	Success ↑	SPL ↑	Success ↑	SPL ↑	Success ↑	SPL ↑
Depth	MP3D	0.98 ± 0.00	0.92 ± 0.00	0.93 ± 0.01	0.84 ± 0.01	0.96 ± 0.00	0.87 ± 0.00
	Gibson 4+	0.96 ± 0.00	0.91 ± 0.00	0.88 ± 0.00	0.76 ± 0.00	0.93 ± 0.00	0.84 ± 0.00
	Gibson	0.99 ± 0.00	0.94 ± 0.00	0.95 ± 0.00	0.86 ± 0.01	0.98 ± 0.00	0.90 ± 0.00
	HM3D	**1.00** ± 0.00	**0.95** ± 0.00	**0.96** ± 0.00	**0.87** ± 0.00	**0.99** ± 0.00	**0.91** ± 0.00
RGB	MP3D	0.93 ± 0.01	0.78 ± 0.00	0.78 ± 0.01	0.59 ± 0.01	0.84 ± 0.00	0.67 ± 0.00
	Gibson 4+	0.88 ± 0.00	0.78 ± 0.01	0.49 ± 0.00	0.36 ± 0.00	0.77 ± 0.01	0.62 ± 0.00
	Gibson	**0.99** ± 0.00	**0.91** ± 0.00	0.86 ± 0.01	0.69 ± 0.01	0.94 ± 0.00	0.82 ± 0.00
	HM3D	**0.99** ± 0.00	**0.91** ± 0.00	**0.92** ± 0.01	**0.74** ± 0.01	**0.97** ± 0.00	**0.87** ± 0.00

Table S1: **PointNav val performance** on multiple navigation metrics. We report the mean and standard deviation by training on 1 random seed, and evaluating on 3 random seeds. The 1^{st} column indicates whether the agent uses depth or RGB inputs. The HM3D agents reach 100% navigation success for both sensors on Gibson (val). In the majority of cases, HM3D agents significantly outperform the other agents on both metrics. Thus, training on HM3D greatly benefits embodied agents.

S7 PointNav validation results

In Figure 6 from the main paper, we presented the validation performance as a function of the training steps. Now, we present the final validation performance of the best checkpoint (analogous to Table 2 in the main paper). See Table S1. We observe trends similar to the ones observed in the main paper. The HM3D agent matches the Gibson agent on Gibson (val) with RGB inputs. On all other cases, the HM3D agents outperform the other agents by a good margin, particularly in the RGB case.

S8 Dataset characteristics that impact PointNav performance

In the main paper, we compared datasets along different characteristics such as navigable area, visual fidelity, 3D reconstruction quality, and the utility for training agents for the PointNav task. In Section 4 and Figure 6, we analyzed the impact of dataset size on the PointNav performance, and noted that total navigable area in the training scans are highly correlated with the PointNav results. Here, we perform a complete analysis of how the following factors affect the PointNav performance (on the val splits).

1. EMD (train, val) measures the dissimilarity between episodes in the train and val splits. We calculate the normalized histogram of geodesic distances between the start and goal locations for each episode in the train and val splits (independently). We then measure the distribution shift between the train and val episodes. This is done by computing the Earth Mover's Distance (EMD) a.k.a Wasserstein Distance between the normalized histograms of geodesic distances for the train and val splits.
2. KID (mean) is a measure of visual fidelity of images rendered from each dataset. This is calculated as the mean of KID (Gibson real) and KID (MP3D real) from Table 5(b) in the main paper.
3. % defects is a measure of reconstruction completeness for the 3D scans. For each dataset, this is calculated as the mean of "% defects" values from Figure 4 in the main paper.
4. Navigable area (m^2) measures the dataset size. It is computed as the overall navigable area in the training scans for each dataset.

We compute the above metrics for all the train datasets[1]. For a given PointNav val set, we measure the Pearson's correlation between each of the above metrics for a train dataset and the navigation SPL achieved by agents trained on the same dataset (see Table S2). As noted in the main paper, we observe that the navigable area is highly correlated with the PointNav performance (0.82 to 0.97) indicating that large-scale datasets are critical for achieving high-quality navigation. For both MP3D (val) and HM3D (val), there is a strong negative correlation of -0.4 to -0.7 between EMD (train, val) and SPL. This indicates that large distribution shifts between the train and val episodes leads to worse performance. Next, we observe that KID (mean) is weakly correlated with the SPL (-0.30 to 0.1), indicating that visual fidelity may not strongly impact PointNav performance in simulation. In most cases, $\%$ defects is generally uncorrelated with SPL. However, we find a slightly *positive*

[1] We compute EMD(train, val) for all pairs of train and val sets.

	PointNav (depth)			PointNav (RGB)		
Factor	Gibson (val)	MP3D (val)	HM3D (val)	Gibson (val)	MP3D (val)	HM3D (val)
EMD (train, val)	$+0.118$	-0.693	-0.412	-0.129	-0.589	-0.584
KID (mean)	-0.100	$+0.016$	-0.145	-0.167	-0.062	-0.299
% defects	-0.102	$+0.302$	$+0.019$	-0.290	$+0.168$	-0.208
Navigable area (m^2)	$+0.968$	$+0.822$	$+0.900$	$+0.886$	$+0.864$	$+0.884$

Table S2: We measure the correlation between different dataset characteristics (row) on the PointNav performance (column). EMD (train, val) measures the distribution mismatch between the difficulties of episodes in the train and val PointNav datasets. KID (mean) measures the visual fidelity of images from the train scans. % defects measures the reconstruction completeness of scans in the train scans. Navigable area measures the total area that is navigable across all the train scans from that dataset.

correlation (0.17-0.32) with SPL on MP3D (val). This may be due to the fact that MP3D val scenes have significantly more mesh reconstruction artifacts than Gibson val scenes[2] (see Figure 4 in main paper). Agents trained on scenes with more mesh reconstruction artifacts adapt better to such testing conditions. Note that these results are not very indicative of the transfer performance to a real robot. It is possible that higher visual fidelity and lower % defects may be necessary for real-world transfer.

S9 Example scenes from Habitat-Matterport 3D dataset

We provide more examples of scenes from Habitat-Matterport 3D in the same style as Figure 2 in the main paper. In Figure S2, we visualize 5 residences, and in Figure S3, we visualize 5 diverse scenes such as offices, gyms, restaurants, and nightclubs. All 900 scenes from the train and val splits of HM3D can be visualized on the dataset website: https://aihabitat.org/datasets/hm3d/

[2] Only Gibson 4+ scenes are used for Gibson (val) and Gibson (test).

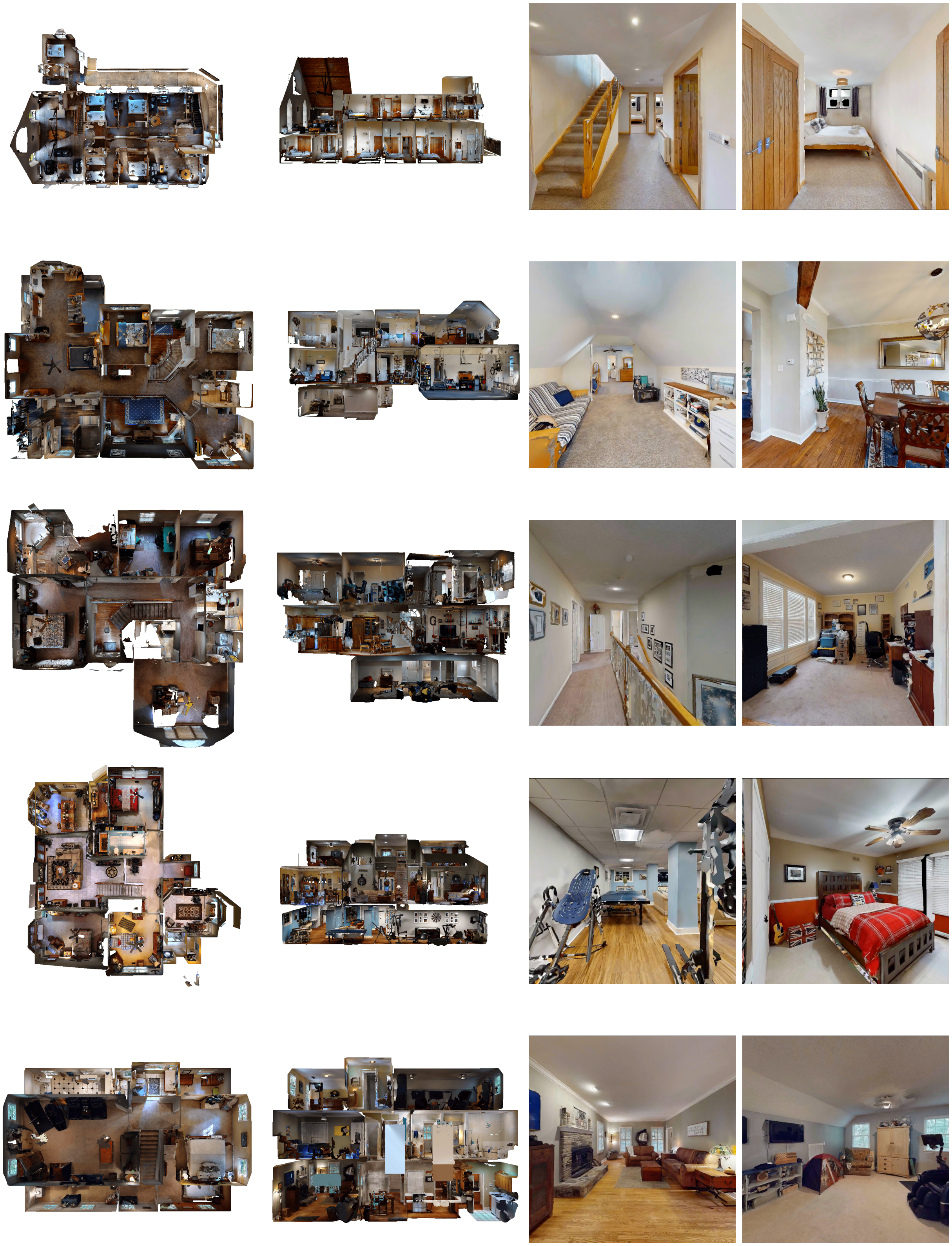

Figure S2: Five example residences from the HM3D dataset. From left to right in each row: top-down view, cross section view, and two egocentric views from navigable positions in the scene. The dataset contains multi-floor buildings spanning a wide range of sizes.

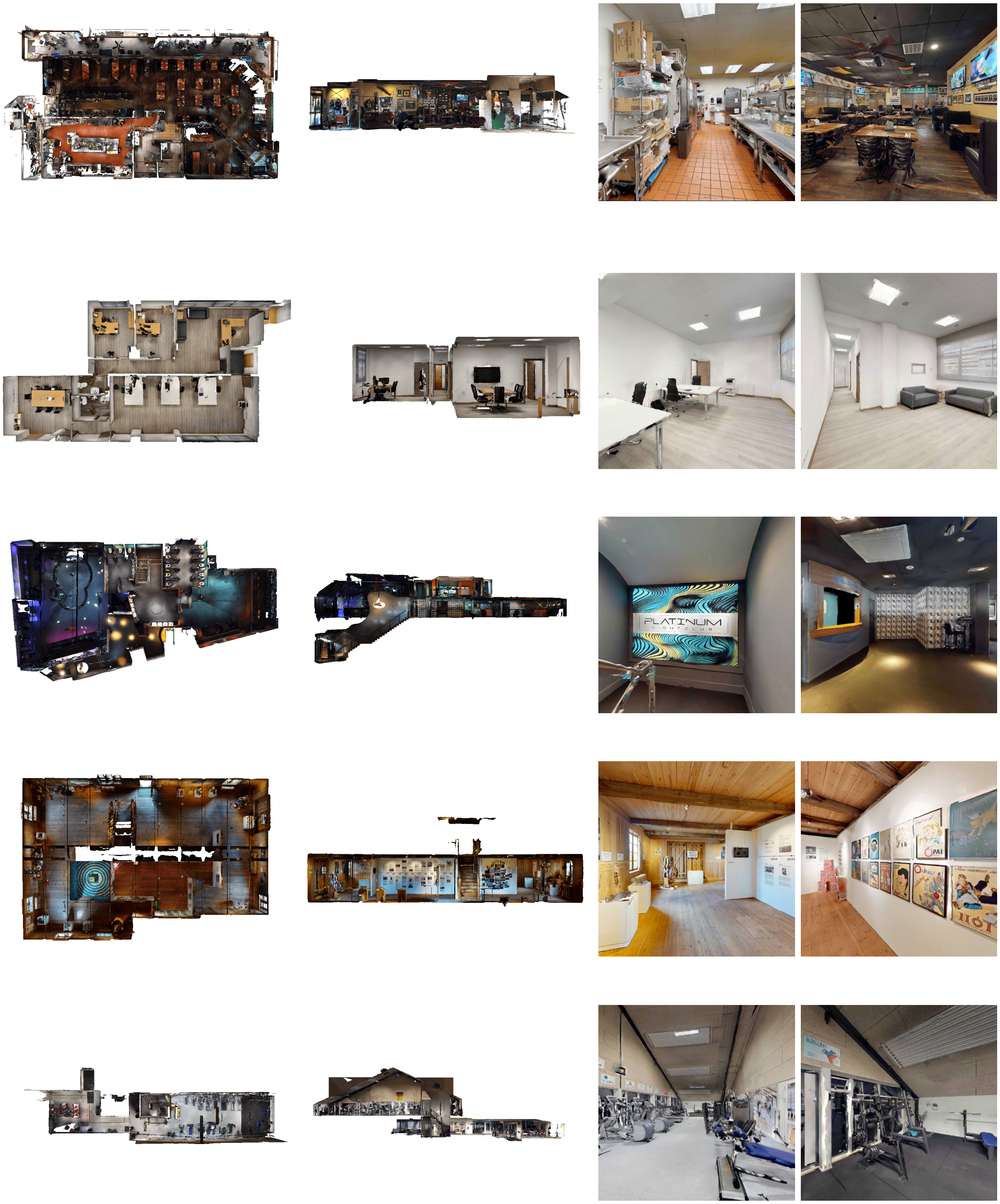

Figure S3: Five examples of diverse scenes from the HM3D dataset. From left to right in each row: top-down view, cross section view, and two egocentric views from navigable positions in the scene. The dataset contains diverse scans such as restaurants (row 1), office buildings (row 2), a night club (row 3), art studio (row 4), and gyms (row 5).

S10 PointNav qualitative results

We show sample PointNav episodes of the HM3D agents in Figure S4 and Figure S5. We present the qualitative results in a format similar to Wijmans et al. [4]. The episodes are categorized based on the difficulty (i.e., the geodesic distance b/w start and goal), and the agent performance (in SPL).

References

[1] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 1

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *Fifth International Conference on 3D Vision (3DV)*, 2017.

[3] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016. 1

[4] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 8

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

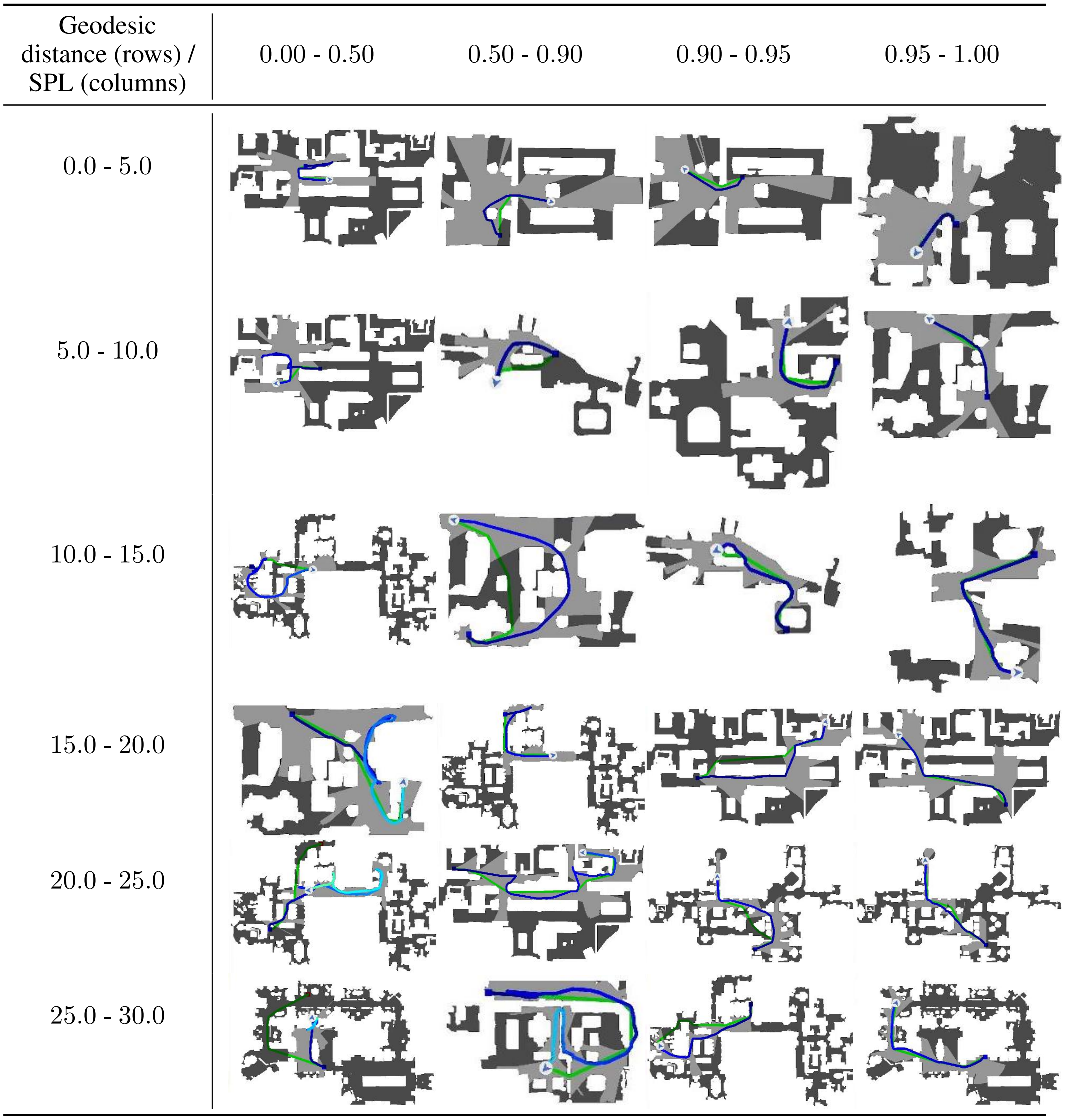

Figure S4: **PointNav with depth sensor:** Example episodes for the HM3D-agent with depth inputs broken down by geodesic distance between agent's start and goal locations (on rows) vs SPL achieved by the agent (on columns). Gray represents navigable regions on the map while white is non-navigable. The agent starts at the blue square and navigates to the red square. The green line shows the shortest path on the map (or oracle navigation). The blue line shows the agent's trajectory. The SPL score is higher if the blue trajectory closely matches the green trajectory.

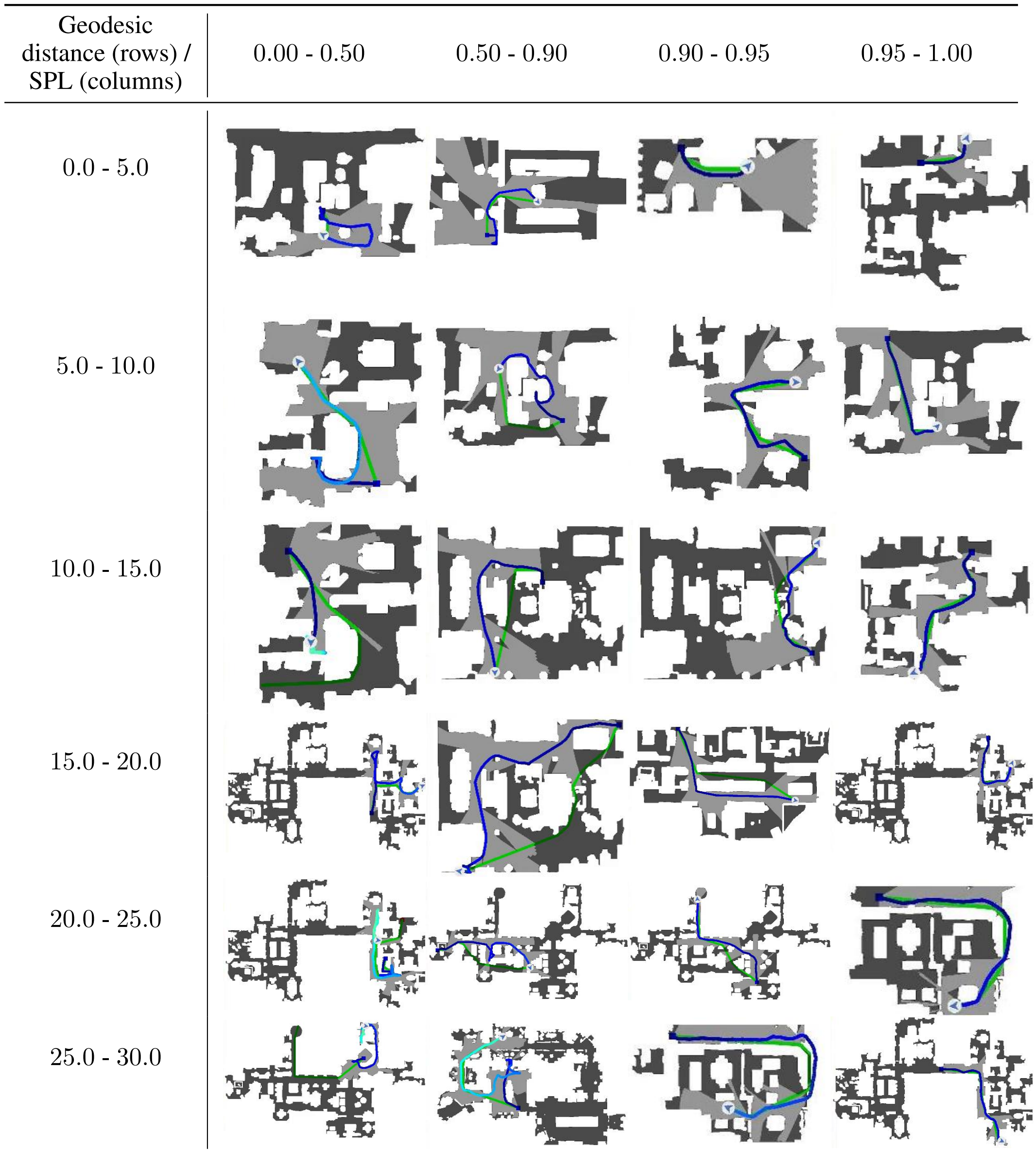

Figure S5: **PointNav with RGB sensor:** Example episodes for the HM3D-agent with RGB inputs broken down by geodesic distance between agent's start and goal locations (on rows) vs SPL achieved by the agent (on columns). Gray represents navigable regions on the map while white is non-navigable. The agent starts at the blue square and navigates to the red square. The green line shows the shortest path on the map (or oracle navigation). The blue line shows the agent's trajectory. The SPL score is higher if the blue trajectory closely matches the green trajectory.