Supplementary Material
PEANUT: Predicting and Navigating to Unseen Targets

Abstract

In the following supplementary material, we provide additional experimental details (Sec. 1) and additional qualitative results (Sec. 2 to Sec. 4). We invite readers to watch the supplementary video (`supp.mp4`) for further visualization.

1. Additional Experimental Details

We first provide additional details about our prediction network architecture, our semantic mapping module, and our procedure for finetuning Mask-RCNN on HM3D.

1.1. Prediction Network Architecture Details

Our target prediction network has a PSPNet-based architecture and largely follows the base implementation provided in [3]. The backbone is a ResNet50 [6] with dilated convolutions as proposed in [2]. This outputs feature maps that are $1/8$ times the size of the input. It is followed by a pyramid pooling module with pooling scales of 1, 2, 3, and 6. The fused features are passed through a 3×3 convolution layer with 512 channels and then a 1×1 convolution for final prediction. Intermediate supervision is provided through an FCN [8] prediction head before the last stage of the ResNet50.

1.2. Semantic Mapping Details

Our semantic mapper is essentially the same as that of SemExp [1]. The map resolution is $5\,\mathrm{cm}$, and the global map size is $H \times W = 960 \times 960$. The 2D segmentation and depth images are downsampled by $4\times$ before being projected to the map in order to save computation. Points within the height range $[0.25\,\mathrm{m}, 0.88\,\mathrm{m}]$ are marked as obstacles. This is because $0.88\,\mathrm{m}$ is the agent's height, and ignoring points close to the ground allows for some robustness to minor elevation changes.

1.3. Mask-RCNN Finetuning Details

For HM3D, we finetune a COCO-pretrained Mask-RCNN [5] on images from Habitat. We collect images by letting an agent wander around according to the baseline exploration policy described in [9]. In total, we collect 80K images from the HM3D train split and 20K images from the HM3D val split – 1K from each scene. We use the associated semantic annotations to extract instance segmentation masks for each image, and discard masks that are either less than 1000 pixels in area or have a bounding box whose aspect ratio is either less than 0.1 or greater than 10.

We finetune the Mask-RCNN for $25\,000$ iterations using an SGD optimizer with a learning rate of 0.02, momentum of 0.9, and batch size of 16. The learning rate was decayed by a factor of 10 after $20\,000$ iterations. During training, we apply random resizing and horizontal flipping for data augmentation.

2. Qualitative Results on Target Prediction

We provide additional visualizations of predictions of target objects made by PEANUT's prediction model in Fig. 1. They demonstrate that the model uses spatial regularities and semantic cues to make highly informative predictions about unexplored areas. In many cases, the model seems to be more confident than a human would be. We also compare the predictions to those made using only single-frame egocentric context, as described in Sec. 4.4 of the main paper, and visualize the results in Fig. 2. The predictions made based on the egocentric crop are diffuse and significantly less informative than those that use the global context.

3. Qualitative Results on Navigation Episodes

We provide additional visualizations of navigation using PEANUT in Fig. 3. Overall, they suggest that PEANUT searches through rooms in an efficient manner. The prediction-based goal selection usually picks goals with a large amount of unexplored area nearby and does not exhibit excessive backtracking.

4. Failure Case Analysis

We visualize some failure cases of PEANUT in Fig. 4. They are representative of PEANUT's most common failure modes, which are segmentation errors and scenes with stairs. In the first case, a sofa is misclassified as a chair. In the second case, a bathtub is misclassified as a bed. These result in false positive detections of the target category, causing the agent to move to the misclassified object and stop, failing the episode. In the third failure case, the target category does not exist on the agent's starting floor, so the agent must traverse a staircase to search a different floor. However, the mapping module marks stairs as obstacles, preventing their traversal.

5. HM3D Test-Standard Leaderboard Evaluation

We note that our method (entry name: "Finding NIMO (PEANUT)") currently ranks 3rd in terms of SPL on the public leaderboard for the HM3D test-standard split (as of March 14th, 2023). Since the rank 1 and 2 methods are not published, PEANUT's ranking is higher than that of any published method. The next highest ranking published method is ProcTHOR [4], which relies on additional training scenes generated using the AI2-THOR simulator [7]. Our method is trained only on the default HM3D ObjectNav dataset, but achieves better performance.

6. MP3D Test-Standard Leaderboard Evaluation

We tried submitting our MP3D agent to the 2021 Habitat Challenge leaderboard on EvalAI, but our submissions repeatedly failed for unknown reasons. Thus, we report results on the MP3D val split in the main paper. We will continue to investigate the cause of this issue.

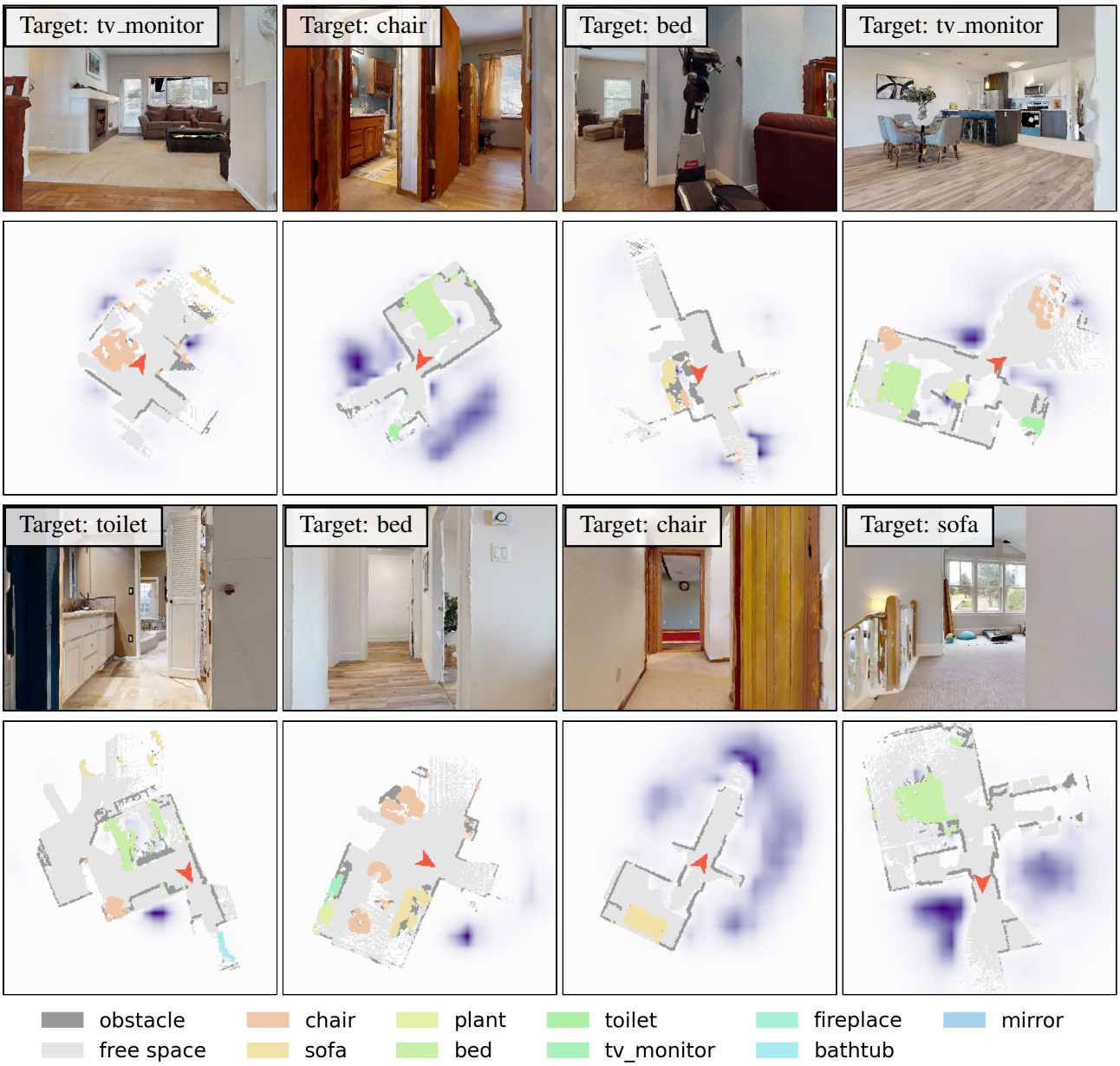

Figure 1. **Example target predictions**. We visualize predictions made by our model in scenes from HM3D (val). The top row shows the agent's RGB observation, and the bottom row shows the incomplete semantic map overlaid with the target probability prediction. Note that the prediction network only has access to the semantic map, not the RGB images.

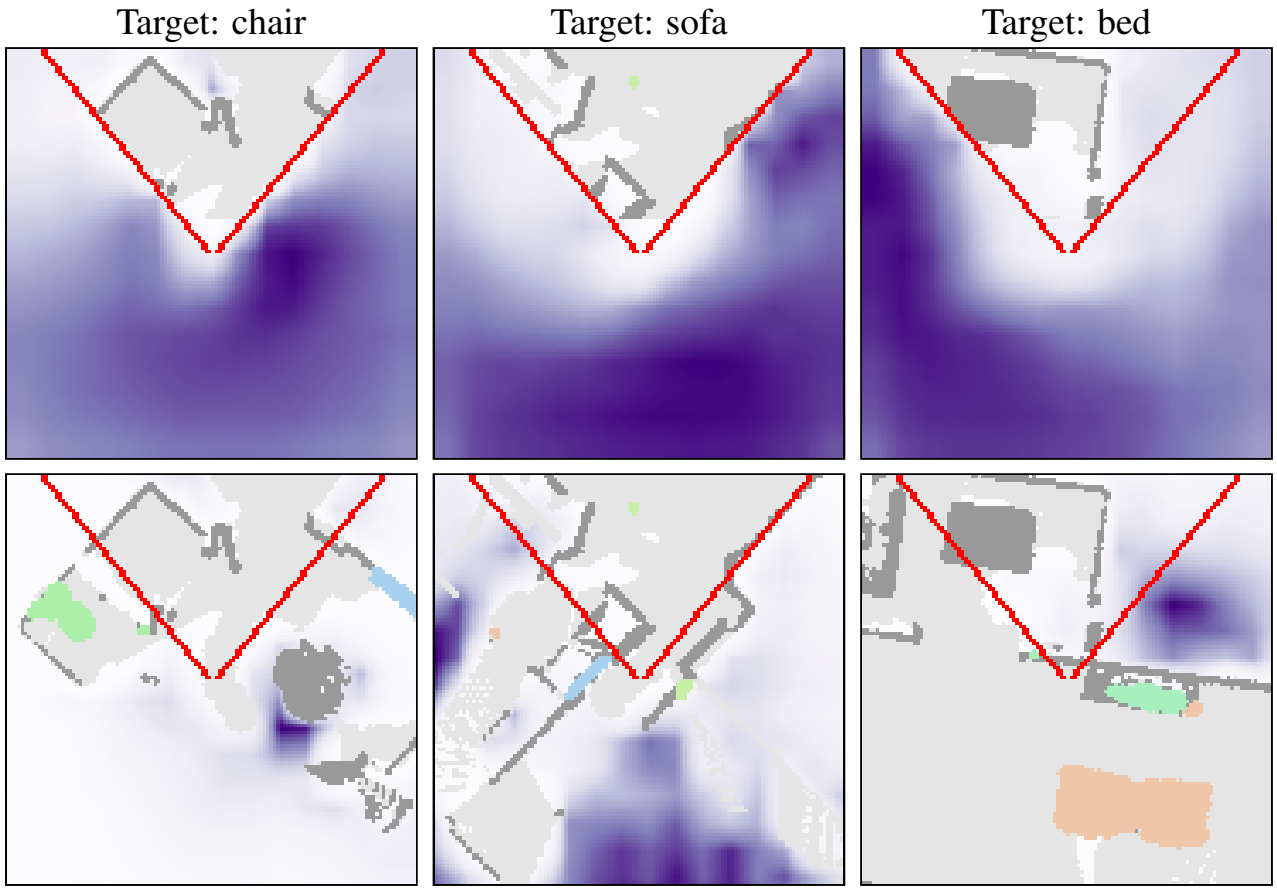

Figure 2. **Effect of context on target predictions.** We visualize predictions made with and without global context on maps from HM3D (val). The predictions made based on the egocentric crop are shown in the top row, and the predictions made using the global context are shown in the bottom row. The red lines mark the visible region for the egocentric crop.

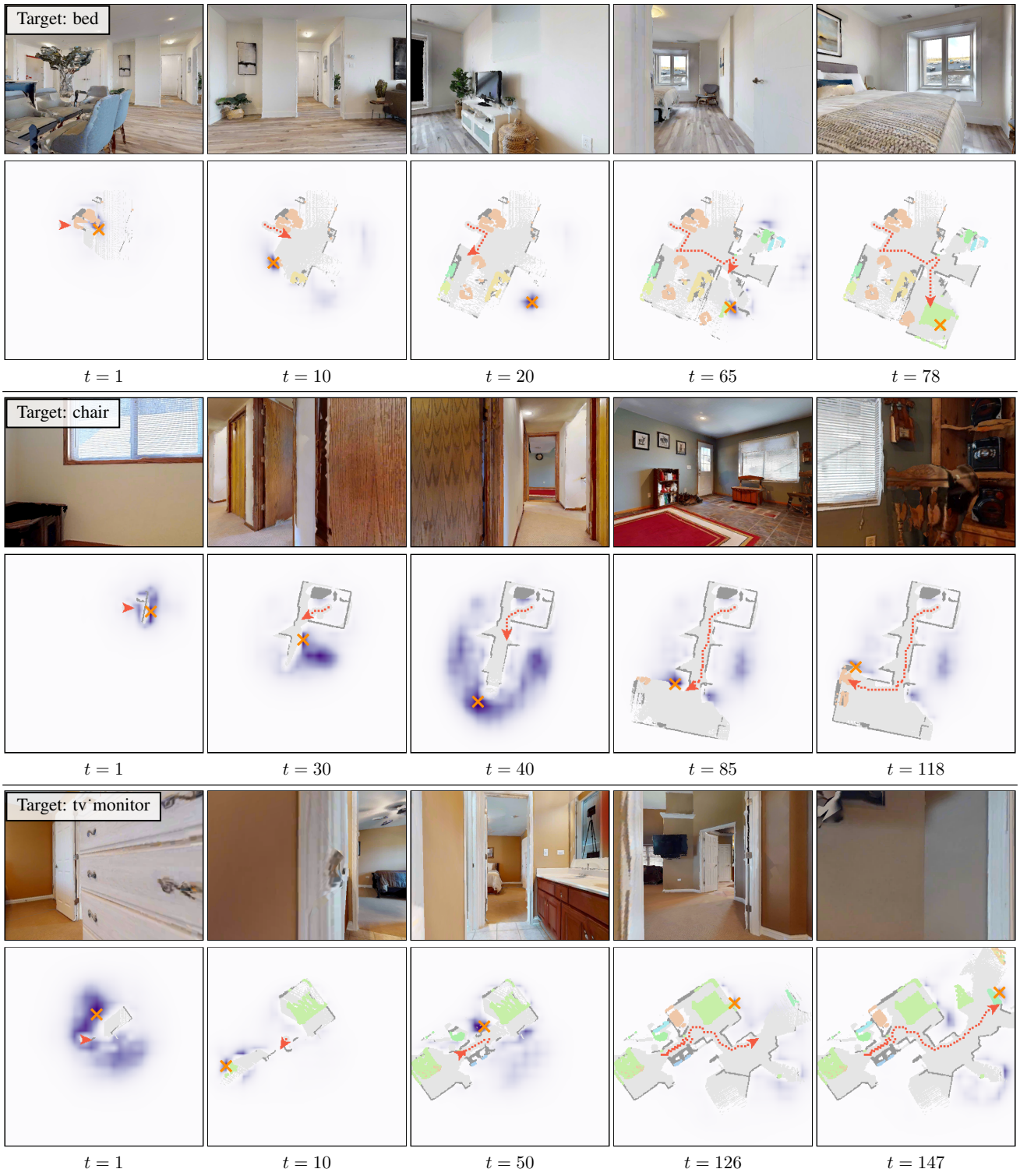

Figure 3. **Example navigation episodes.** We visualize episodes of navigation in scenes from HM3D (val). The top row shows the agent's RGB observation, and the bottom row shows the incomplete semantic map overlaid with the target probability prediction. The agent's selected long-term goal is marked by an orange cross.

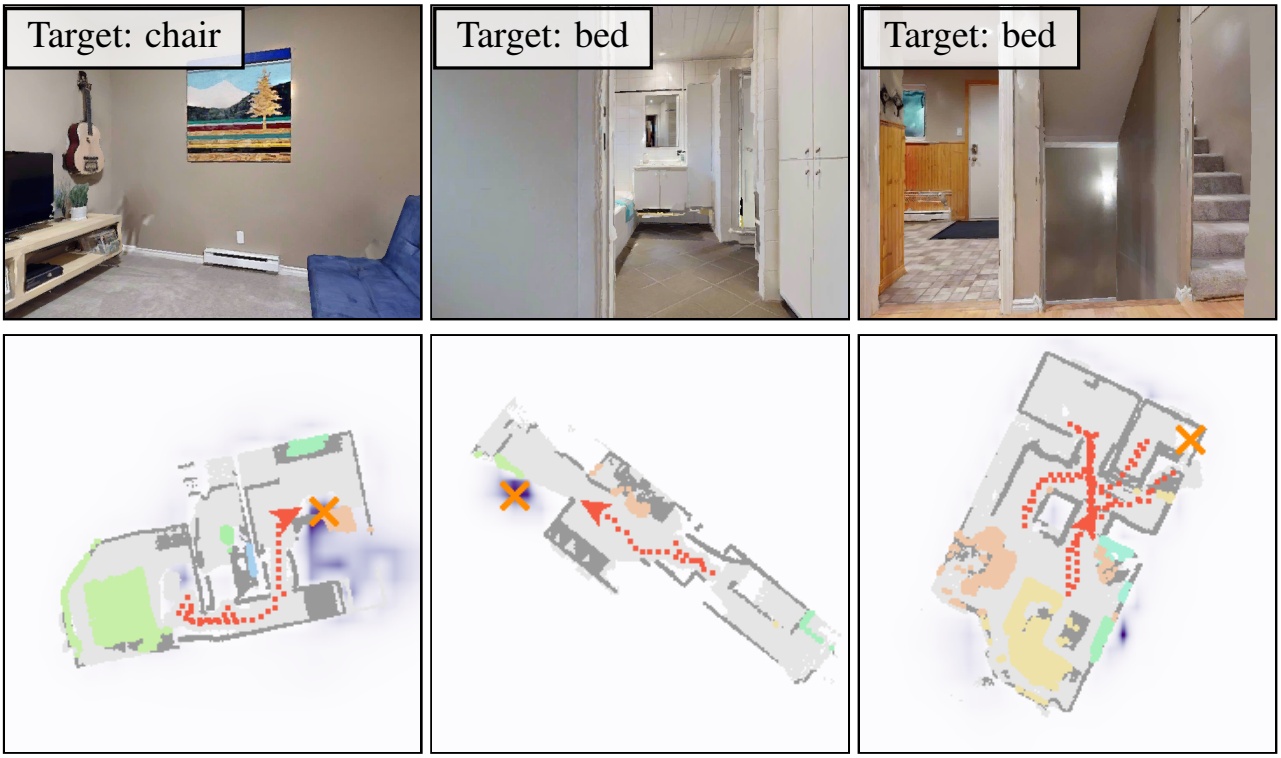

Figure 4. **Example failure cases**. We visualize some of PEANUT's failure cases on HM3D (val). The first two are caused by false positive detections of the target. The third results from the agent's inability to traverse stairs and search in another floor (due to the stairs appearing as obstacles on the agent's map).

References

[1] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *NeurIPS*, 2020. 1

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2017. 1

[3] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 1

[4] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, et al. Procthor: Large-scale embodied ai using procedural generation. *arXiv preprint arXiv:2206.06994*, 2022. 2

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[7] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 2

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[9] Haokuan Luo, Albert Yue, Zhang-Wei Hong, and Pulkit Agrawal. Stubborn: A strong baseline for indoor object navigation. *arXiv preprint arXiv:2203.07359*, 2022. 1