Phone2Proc: Bringing Robust Robots Into Our Chaotic World Supplementary Material

Matt Deitke[*†ψ], **Rose Hendrix**[*†], **Ali Farhadi**[ψ],
Kiana Ehsani[†], **Aniruddha Kembhavi**[†ψ]
[†]PRIOR @ Allen Institute for AI, [ψ]University of Washington, Seattle
https://phone2proc.allenai.org/

A. Implementation Details

For all experiments, we use the same architectures and process from EmbCLIP [4] and adopt the same hyperparameters as ProcTHOR [2]. The 3x224x224 RGB images are processed with a frozen CLIP-ResNet-50 architecture [3, 5]. This visual embedding is compressed with a 2-layer CNN, concatenated with a goal object type embedding, and compressed with a 2-layer CNN. This is flattened and combined with an embedding of the previous action, then passed through a single-layer GRU [1] policy with a hidden belief state of size 512. An actor and critic are used to generate the next action probability distribution and current state value estimates, respectively. The agent's next action is sampled from the actor distribution.

The following updates from [2] are made for policy and goal encoder fine-tuning:

1. Learning rate is lowered to 3e-5 (10x lower than that of ProcTHOR [2]).
2. A small penalty of -0.05 is assessed if the agent runs into objects.
3. If the agent is about to run into an object, it will randomly move and rotate in small increments to coarsely emulate unmodeled and unintended physical environment interactions.
4. The neck actions are limited to looking 30° above and below the horizon, as on our physical platforms.
5. The horizontal field of view for a fixed aspect ratio is randomized by episode (uniformly sampled in 0.2° increments between 48° and 65°), and the vertical field of view/aspect ratio is modified to more closely resemble the Intel RealSense D435.

We use multi-node training on 3 or 4 (depending on the environment size) AWS g4dn.12xlarge machines with 16 processes per machine.

The Habitat baseline used in Table 1 is trained on the 80 HM3D [6] set training scenes used for the 2022 Habitat challenge [7] using 80 processes and 8 A100 GPUs. The model trained for 200M steps and we used the model which achieved the best performance on a validation set of 200 episodes. It was trained with the same updated field of view as every other model and baseline evaluated in this work.

B. Failure Cases

The most common failure case for PHONE2PROC models was semantic confusion; that is, not being able to recognize particular instances of objects or mistaking instances of other object categories for the target object. For example, a couch with a cover on it in the 6-room apartment was mistaken for a bed several times in the limited field of view of the agent and spare jugs for the water cooler in the cafeteria were mistaken as a vase. To generate the scenes, only six 3D models of vases were used. Thus, some semantic confusion is perhaps unsurprising, and PHONE2PROC with more visual diversity might be used to even greater effect.

References

[1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 1

[2] 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. *Conference on Neural Information Processing Systems*, 2022. 1

[3] 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. 1

[4] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022. 1

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[6] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *ArXiv*, abs/2109.08238, 2021. 1

[7] Karmesh Yadav, Santhosh Kumar Ramakrishnan, John Turner, Aaron Gokaslan, Oleksandr Maksymets, Rishabh Jain, Ram Ramrakhya, Angel X Chang, Alexander Clegg, Manolis Savva, Eric Undersander, Devendra Singh Chaplot, and Dhruv Batra. Habitat challenge 2022. https://aihabitat.org/challenge/2022/, 2022. 1