Supplementary Materials
Zero Experience Required: Plug & Play Modular Transfer Learning for Semantic Visual Navigation

Ziad Al-Halah[1] Santhosh K. Ramakrishnan[1,2] Kristen Grauman[1,2]
[1]The University of Texas at Austin [2]Meta AI
ziadlhlh@gmail.com, srama@cs.utexas.edu, grauman@cs.utexas.edu

Additional information presented in this supplementary:

1. Details on the shared implementation (Sec. 1).
2. Details of the image-goal navigation dataset (Sec. 2).
3. Detailed results on image-goal navigation across 3 levels of episode difficulties (Table 1).
4. Examples of the visual goal modalities used in target tasks (Fig. 1).
5. Dataset details for the target tasks and goal embedding space (Sec. 3).
6. Detailed results with standard deviations on target tasks (Table 2).
7. Qualitative results for our model in all tasks and goal modalities (Fig. 5).
8. Examples of failure cases for our model (Fig. 6).
9. Performance curves for all tasks and goal modalities in transfer learning setup (Fig. 2) and long-term training of Task Expert (Fig. 3).
10. Ablation on the sensor configuration used by the agent (Fig. 4)
11. Discussion of potential societal impact (Sec. 4) and limitations (Sec. 5).

1. Shared Setup

All RL methods are trained with the following setup. We use input augmentation of random cropping and color jitter for both observations and goals. The models are trained with DD-PPO [15]. We set the number of PPO epochs 2, the forward steps 128, the entropy coefficient 0.01, clipping of 0.2, and train the model end-to-end using the Adam optimizer [10]. We allocate the same number of processes and resources to all methods.

We use the Habitat simulator [13] along with the Gibson [16], Matterport3D [3], and HM3D [12] datasets. These datasets are photorealistic and scans of real-world environments with varying complexities, sizes, room layouts, types. In all our experiments, the test scenes are disjoint from those used for training to assess the agent ability to generalize to previously unseen environments.

2. Image-Goal Navigation

Dataset The training split contains 9K episodes sampled from each of the 72 Gibson training scenes. The episodes are uniformly split across 3 levels of difficulty based on the goal's geodesic distance from the start location: *easy* (1.5 - 3 m), *medium* (3 - 5 m), and *hard* (5 - 10 m). Test split A has 4.2K episodes and split B has 3K episodes. Both splits are sampled uniformly from 14 disjoint (unseen) scenes and the 3 levels of difficulty. Further, to avoid trivial straight line paths, the episodes has a minimum geodesic to euclidean distance ratio of 1.1 in A and 1.2 in B (our split B corresponds to the curved split in [9]).

Detailed Results We show in Table 1 the detailed results of all models across the three levels of episode difficulties (*easy*, *medium* and *hard*). Our model shows better performance across the different levels and in both split A and B. For a qualitative result, see Fig. 5 A.

3. Transfer Learning to Downstream Tasks

Datasets We use 29 scenes from Gibson and split them into 24 scenes for training and 5 for testing. In the following, we present the details of the datasets used for each of the downstream tasks.

- ObjectNav: We sample 24K episodes for training and 1K episodes for testing. For the sketch-goals (Fig. 1 middle), we sample 80 sketches from [7] for each object category and split them to 70 used during training and 10 for testing. For the audio-goals, we sample 12 audio clips from [4] of lengths ranging from 13 to 53 seconds and split them 50/50 for training and testing. At the start

Model	Split	Easy		Medium		Hard		Overall	
		Succ.	SPL	Succ.	SPL	Succ.	SPL	Succ.	SPL
Imitation Learning	A	18.5	17.7	8.4	8.1	2.6	2.6	9.9	9.5
Zhu *et al*. [18]	A	31.7	25.1	15.7	10.8	11.5	7.5	19.6	14.5
Mezghani *et al*. [11] w/ 90° FoV	A	17.5	11.0	8.8	6.6	0.6	0.5	9.0	6.0
DTG-RL	A	32.9	26.2	21.2	17.0	13.6	10.8	22.6	18.0
Ours	A	**39.7**	28.5	**29.6**	**22.5**	**18.2**	**13.8**	**29.2**	**21.6**
Ours (View Aug. Only)	A	37.0	**31.7**	18.5	15.9	10.7	9.0	22.0	18.8
Ours (View Reward Only)	A	32.3	22.7	24.8	18.1	16.1	11.0	24.4	17.3
Hahn *et al*. [9]	B	35.5	18.4	23.9	12.1	12.5	6.8	24.0	12.4
Ours	B	**48.0**	**34.2**	**36.0**	**25.9**	**15.1**	**10.8**	**33.0**	**23.6**
Hahn *et al*. [9] w/ noisy actuation	B	27.3	10.6	23.1	10.4	**10.5**	5.6	20.3	8.8
Ours w/ noisy actuation	B	**41.0**	**28.2**	**27.3**	**18.6**	9.3	**6.0**	**25.9**	**17.6**

Table 1. Detailed results across 3 levels of difficulties on image-goal navigation in Gibson [16].

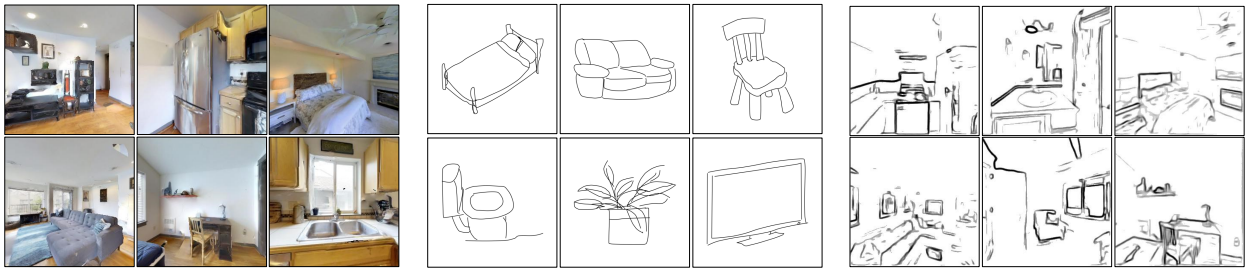

Figure 1. Examples of visual goals used for the target navigation tasks: left) ImageNav (Image), middle) ObjectNav (Sketch), right) ViewNav (Edgemap).

of each episode of the ObjectNav (Audio) task a random 4 seconds duration is sampled from the respective audio clip and split, and presented to the agent as the goal descriptor.

- RoomNav: We sample 25K episodes for training and 290 for testing.

- ViewNav: We sample 24K episodes for training and 1.5K for testing. We generate edgemaps for 300 random views per scene, and we randomly assign one of those per episode as the goal (Fig. 1 right).

Goal Embedding Space In the joint goal embedding space, we aim to learn goal encoders that are compatible to the image-goal encoder. For example, an image view from a living room with a TV detected in it will be used as the positive anchor for a sketch of a TV, a sound clip from a TV, the TV label, the living room label, and the edgemap of the view. The annotations for the sampled image view are based on model predictions from [2]. During training, the parameters of f_G^I are kept frozen, and we train the various goal encoders defined in Main/Sec.4 using the loss from Main/Sec.3.2.

Detailed Results for All Tasks In Table 2 we show the average success rate and standard deviation for all methods over 3 random seeds. Fig. 2 shows the performance of the best transfer learning methods and our approach across all tasks and goal modalities. Fig. 3 shows the Task Expert performance when trained for up to 500M steps on each of the respective tasks and in comparison to our model performance under the ZSEL setting or when it is finetuned. Furthermore, we show example navigation episodes from all tasks and goal modalities for our approach in Fig. 5. Our plug and play modular transfer learning approach enable our model to perform a diverse set of tasks effectively.

Scalability (Sensors) We evaluate our model's ability to scale across the sensor suite. Fig. 4 shows our model performance when varying the sensors' configuration in the source task (ImageNav) and evaluating on ObjectNav (Label) under the ZSEL setup. As expected, when enriching the agent sensors to include depth and pose sensors in addition to vision, we see an additional improvement in performance. More importantly, when increasing the vision sensor resolution from 128 to 256 we see a significant bump in ZSEL success rate that exceeds the one from diversifying the sensory suite. Our model seems to benefit from an en-

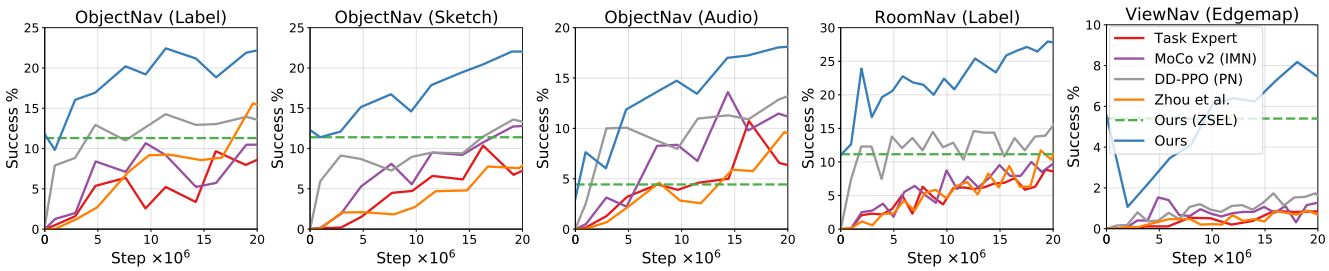

Figure 2. Transfer learning and ZSEL performance on downstream navigation tasks.

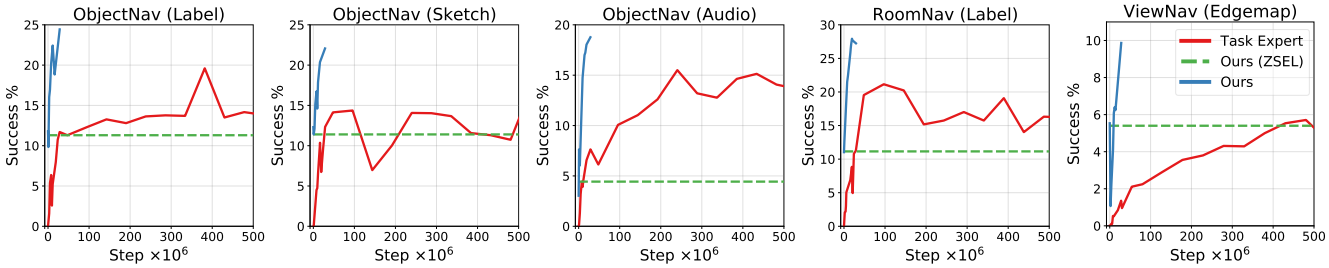

Figure 3. Long-term Task Expert training. Our model maintains it superior performance even when the Task Expert is presented with extensive experience in the target task.

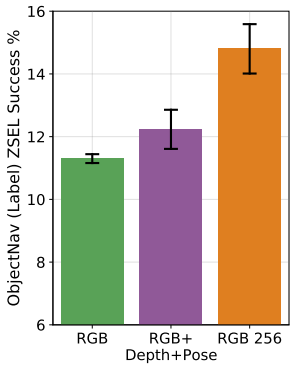

Figure 4. Scalability ablation for our model when changing the sensors configuration.

hanced vision channel as it carries the important semantic cues needed for our semantic search policy and goal embedding space.

Failure Cases We show in Fig. 6 few examples of failure cases encountered by our model. We notice that some of these failure cases are related to the type of the goal modality. For example, in ImageNav the agent sometimes finds the object described in the image however misestimate the view point the image is taken from, hence stops a bit far from the goal location (Fig. 6 A). In ViewNav (Edgemap), the goal modality lacks distinctive texture and color information which leads the agent to sometime stops at a location with similar edge structure, but it is actually not the goal (Fig. 6 B). A type of failure cases spotted in multiple tasks are the early stopping cases. In these cases, the agent fails to estimate the distance to the goal correctly and stops early resulting in an unsuccessful episode (Fig. 6 C).

4. Potential Societal Impact

Our approach's application domain is semantic visual navigation. Here, autonomous agents are trained to find semantic objects in a 3D environment. Such a technology can have positive societal impact by improving people's life, especially in domains like elder care, with robots that can aid in daily life tasks (*e.g.* find my keys, go to the bedroom and bring me my medicine). On the other side, the datasets used in this study are 3D scans of building and houses from certain geographic and cultural areas (western style houses from well-off areas). This creates certain biases in the type of building architectures, room, and object types the agent is familiar with. Consequently, this may limit the availability of this technology to a small section of the population. More diverse datasets and methods with robust adaptation to strong shifts in building layouts and object types are needed to mitigate these effects.

5. Discussion and Limitations

We propose a novel approach for modular transfer learning that enables the agent to handle multiple tasks with diverse goal modalities effectively. Our model can solve the downstream tasks out-of-the-box in zero-shot experience learning setup alleviating the need for expensive interac-

Model	Source Task	ObjectNav Label	ObjectNav Sketch	ObjectNav Audio	RoomNav Label	ViewNav Edgemap
Task Expert	-	$8.0_{\pm 0.6}$	$6.7_{\pm 1.4}$	$6.6_{\pm 0.7}$	$8.9_{\pm 0.8}$	$0.8_{\pm 0.3}$
MoCo v2 [5] (Gib.)	SSL	$10.5_{\pm 0.7}$	$9.9_{\pm 0.6}$	$8.8_{\pm 1.2}$	$9.3_{\pm 0.9}$	$1.0_{\pm 0.2}$
MoCo v2 [5] (IMN)	SSL	$7.8_{\pm 0.3}$	$12.7_{\pm 0.8}$	$11.5_{\pm 0.8}$	$9.7_{\pm 2.2}$	$1.3_{\pm 0.3}$
Visual Priors [14]	SL	$9.3_{\pm 0.1}$	$9.9_{\pm 0.7}$	$9.1_{\pm 0.8}$	$13.1_{\pm 0.9}$	$0.6_{\pm 0.1}$
Zhou *et al*. [17]	SL	$15.6_{\pm 1.0}$	$7.6_{\pm 0.3}$	$9.6_{\pm 0.8}$	$10.3_{\pm 0.9}$	$0.7_{\pm 0.1}$
CRL [6]	RL	$1.9_{\pm 0.5}$	$0.5_{\pm 0.3}$	$1.0_{\pm 0.4}$	$1.2_{\pm 0.8}$	$0.0_{\pm 0.0}$
SplitNet [8]	RL	$9.0_{\pm 1.0}$	$6.5_{\pm 0.8}$	$8.8_{\pm 1.1}$	$7.7_{\pm 1.1}$	$0.6_{\pm 0.0}$
DD-PPO (PN) [15]	RL	$13.9_{\pm 0.6}$	$13.6_{\pm 0.8}$	$12.9_{\pm 0.5}$	$13.9_{\pm 1.6}$	$1.7_{\pm 0.1}$
Ours (ZSEL)	RL	$11.3_{\pm 0.2}$	$11.4_{\pm 0.6}$	$4.4_{\pm 0.6}$	$11.2_{\pm 1.3}$	$5.4_{\pm 0.6}$
Ours	RL	$\mathbf{21.9}_{\pm 0.1}$	$\mathbf{22.0}_{\pm 0.9}$	$\mathbf{18.0}_{\pm 1.2}$	$\mathbf{27.9}_{\pm 1.9}$	$\mathbf{7.4}_{\pm 0.1}$

Table 2. Transfer learning average success rate and standard deviation on downstream semantic navigation tasks.

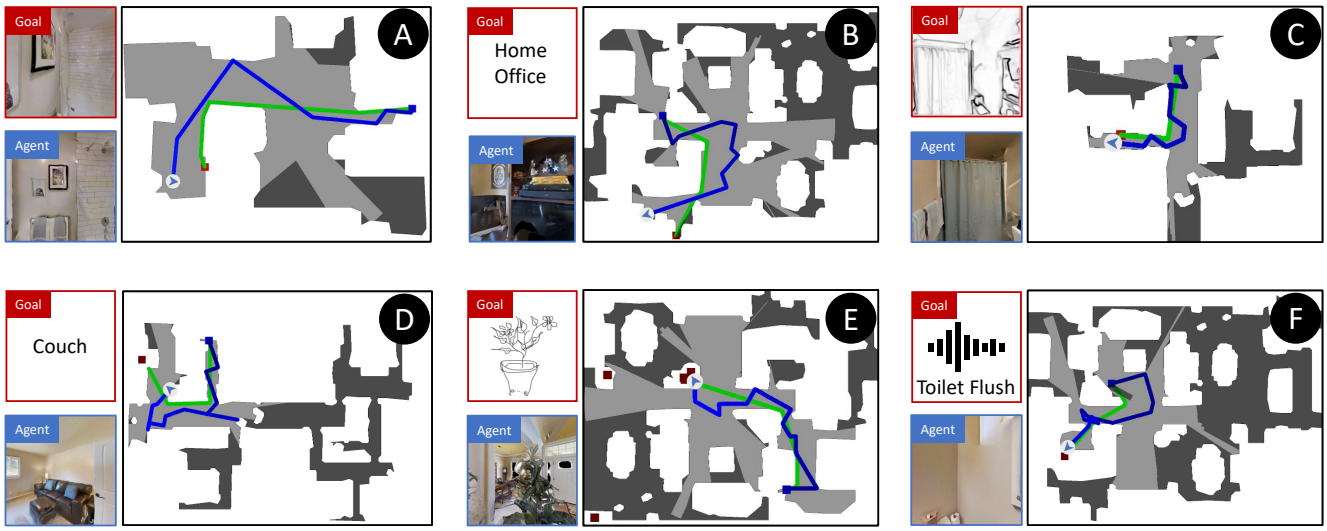

Figure 5. Qualitative results of our approach performing 6 tasks with 5 goal modalities: A) ImageNav (Image), B) RoomNav (Label), C) ViewNav (Edgemap), D) ObjectNav (Label), E) ObjectNav (Sketch), F) ObjectNav (Audio).

Figure 6. Qualitative results of failure cases in A) ImageNav (Image), B) ViewNav (Edgemap), and C) ObjectNav (Label).

tive training of the policy. Alternatively, our model can be finetuned on the downstream task to learn task-specific cues where it showed to learn faster, generalize better and reach higher performance than the baselines. While we focused in this work on semantic navigation tasks, this can be seen as a first step in this exciting direction. Additional research is needed to generalize this method to tasks that require a series of goals and a compatible policy that can plan effectively in a multi-goal setup (*e.g*. VLN [1]). Further, our results and evaluation demonstrate strong transfer learning performance for our method. However, as usual in transfer learning, there is not a theoretical guarantee that a transfer

effect will always be beneficial. Target tasks with significant differences to the source task may not benefit from transferring the accumulated experience in the source.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 4

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera. In *ICCV*, 2019. 2

[3] 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. *International Conference on 3D Vision (3DV)*, 2017. Matterport3D license available at http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf. 1

[4] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic Audio-Visual Navigation. In *CVPR*, 2021. 1

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4

[6] Yilun Du, Chuang Gan, and Phillip Isola. Curious Representation Learning for Embodied Intelligence. In *ICCV*, 2021. 4

[7] Mathias Eitz, James Hays, and Marc Alexa. How Do Humans Sketch Objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):44:1–44:10, 2012. 1

[8] Daniel Gordon, Abhishek Kadian, Devi Parikh, Judy Hoffman, and Dhruv Batra. SplitNet: Sim2Sim and Task2Task Transfer for Embodied Visual Navigation. In *ICCV*, 2019. 4

[9] Meera Hahn, Devendra Chaplot, Mustafa Mukadam, James Rehg, Shubham Tulsiani, and Abhinav Gupta. No RL, No Simulation: Learning to Navigate without Navigating. In *NeurIPS*, 2021. 1, 2

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[11] Lina Mezghani, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Karteek Alahari. Memory-Augmented Reinforcement Learning for Image-Goal Navigation. *arXiv*, 2021. 2

[12] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In *NeurIPS Datasets and Benchmarks Track (Round 2)*, 2021. HM3D license available at https://matterport.com/matterport-end-user-license-agreement-academic-use-model-data. 1

[13] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019. 1

[14] Alexander Sax, Jeffrey O Zhang, Bradley Emi, Amir Zamir, Silvio Savarese, Leonidas Guibas, and Jitendra Malik. Learning to navigate using mid-level visual priors. In *CoRL*, 2019. 4

[15] 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 *ICLR*, 2020. 1, 4

[16] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: Real-World Perception for Embodied Agents. In *CVPR*, 2018. Gibson license available at http://svl.stanford.edu/gibson2/assets/GDS_agreement.pdf. 1, 2

[17] Brady Zhou, Philipp Krähenbühl, and Vladlen Koltun. Does computer vision matter for action? *Science Robotics*, 2019. 4

[18] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning. In *ICRA*, 2017. 2