Appendix

A. SHapley Additive exPlanations (SHAP)

In iSEE, we apply SHAP to identify which GRU units were relevant in the prediction of a concept of interest. Let f denotes the model (Gradient boosted tree) that is trained to predict the concept from hidden units. Let S denote a subset of GRU units, then $f_x(S) \approx E[f(x) \mid x_S]$ is the estimated expectation of the model's output conditioned on the set S of GRU units. Then, relevance of a GRU unit i (Shapley value: $\phi_i(f, x)$) in predicting the concept for a given example is given by

$$\phi_i(f, x) = \sum_{R \in \mathcal{R}} \frac{1}{M!} \left[f_x(P_i^R \cup i) - f_x(P_i^R) \right] \tag{1}$$

where $\mathcal{R}$ is the set of all unit orderings, P_i^R is the set of all GRU units that come before unit i in ordering R, and M is the number of GRU units. In simple words, we calculate the change in outcome of model f when the unit i is added in the model with existing GRU units subset P_i^R. Then by averaging the change over all possible subsets, we get the Shapley value of unit i for a given example thus providing the local importance for that example. To obtain the global importance of unit i, we aggregate Shapley values of unit i over multiple examples from the validation trajectories.

Although computing exact Shapley values in model agnostic setting has exponential time complexity, Lundberg et al. [23] came up with an elegant algorithm for GBTs that allows computing Shapley values in polynomial time. We refer the reader to [23] for more details about the efficient SHAP algorithm for GBTs.

B. Concept prediction by OBJECTNAV agent

We report the prediction results of concepts that were not reported in the main text in Figure A1. We observe that reachability at radius = 4 × $gridSize$ (Figure A1a) and radius = 6 × $gridSize$ (Figure A1b) shows a pattern similar to radius = 2 × $gridSize$ as reported in the main text. Reachability of angles in front (around 0 degrees) is more predictable than angles in back (around 180 degrees) of the agent. The agent's position (R_a) with respect to its spawn location is only slightly more predictable than baselines while orientation (θ_a) prediction is almost equal to baselines (Figure A1c). The prediction of collision event is also not much better than baselines (Figure A1d).

C. Concept prediction by POINTNAV agent

We report the prediction results of concepts that were not reported in the main text in Figure A2. We observe that when GPS sensor (target distance and orientation) is available all models can predict target distance and orientation well. The agent's position with respect to its spawn location can be predicted by POINTNAV agents but not the baselines when GPS sensor is available. Further, when the GPS sensor is removed, the agent's position and orientation can not be predicted. Visited history and collision events are also not predictable. Surprisingly, visited history and collision events are slightly more predictable for baselines than POINTNAV trained models when only target information is used. A possible explanation for above observation could be that the position and orientation of a location with respect to target will be same if a location is visited (or in case of a collision event) and therefore can be predicted using the GPS sensor. However, since the trained POINTNAV models do not predict visited history and collision events , this information might not be relevant to solving POINTNAV task.

D. Visualization of R_a and R_t units of SC_{PN}

In Figure A3 a, we visualize the SHAP plots for top-4 units most relevant for predicting agent's position with respect to spawn location (R_a) and target's position with respect to agent's current location (R_t) as these were most predictable concepts from SC_{PN}'s hidden state. From Figure A3 b and c we observe that top R_a unit is constant throughout the episode while top R_t unit's response increases as the agent moves closer to the target suggesting that R_t unit's response is negatively correlated with target distance. As top R_a unit's response was almost constant throughout the episode we further investigated its variance in all the validation episodes and observed that it was extremely low (4.9×10^{-9}). This result was surprising as we expected it to be correlated with agent's progress towards the target. One possibility is that the change in hidden's unit response is extremely low and it is not possible to visualize its change with respect to other units. Another possibility is that this unit might not be encoding R_a independently but combined with other units (e.g. R_t units). The main focus of iSEE was on identifying individual units relevant for predicting a concept and in current form it can not explain how multiple units together predict a concept.

E. SHAP value distribution across units

In the main text, we visualized the distribution of SHAP values across individual examples in the validation set for units most relevant for predicting a concept. In Figure A4, we visualize the distribution of aggregate (average of absolute SHAP values across examples) SHAP values across units for the most predictable concepts by OBJECTNAV (Figure A4a) and POINTNAV (Figure A4b) agents. We observe that generally for all concepts, aggregate SHAP values show a sharp drop suggesting only a few units are rele-

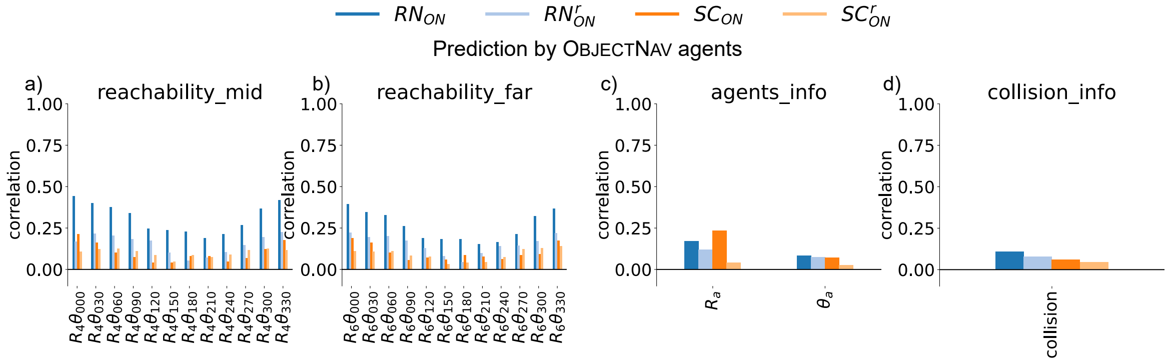

Figure A1. **Metadata prediction by OBJECTNAV GRU units:** a) Reachability mid (R = 4 × $gridSize$) b) Reachability far (R = 6 × $gridSize$) c) Agent information d) Collision

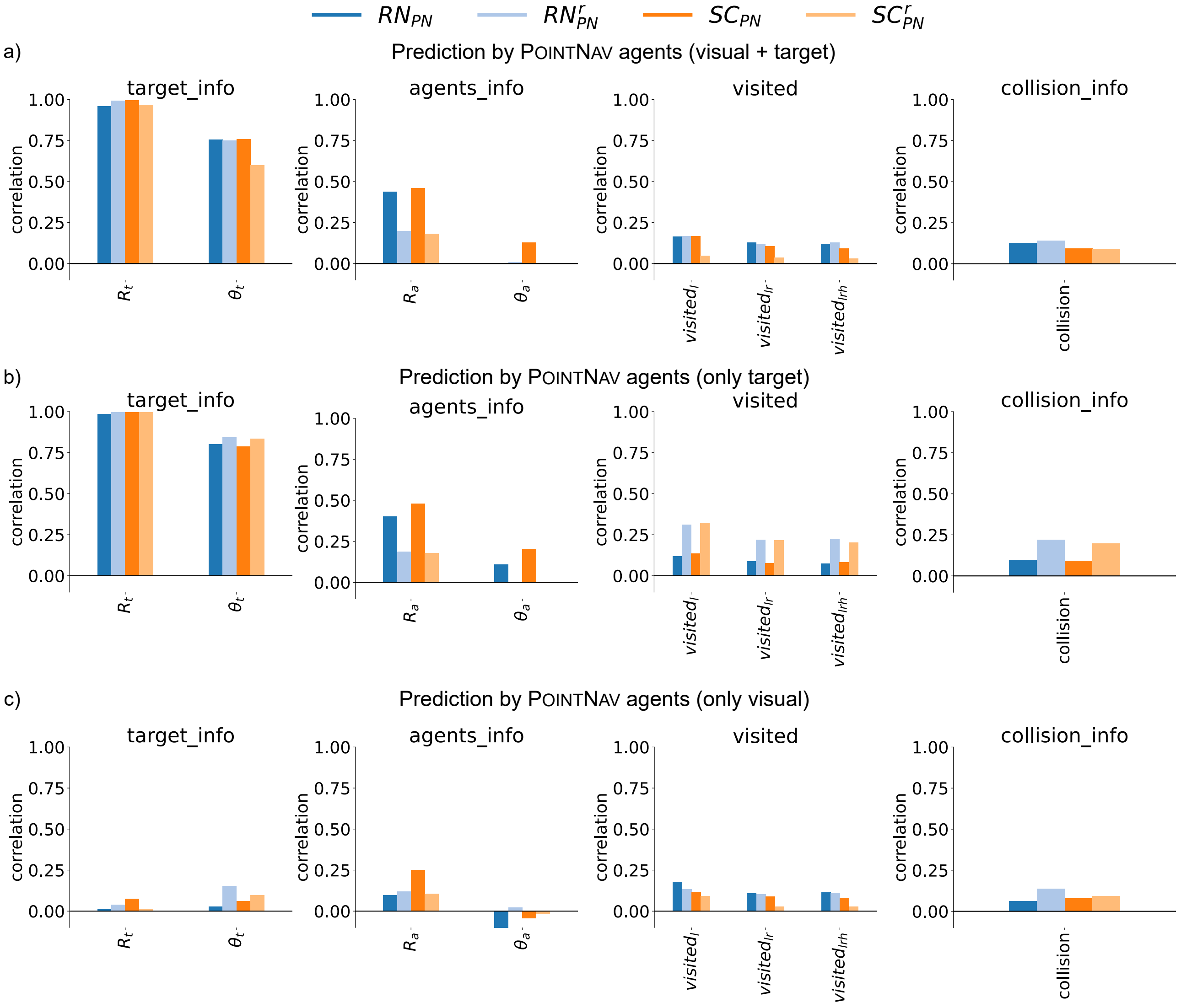

Figure A2. **Metadata prediction by POINTNAV GRU units:** using a) both visual and target sensor b) only target sensor c) only visual sensor

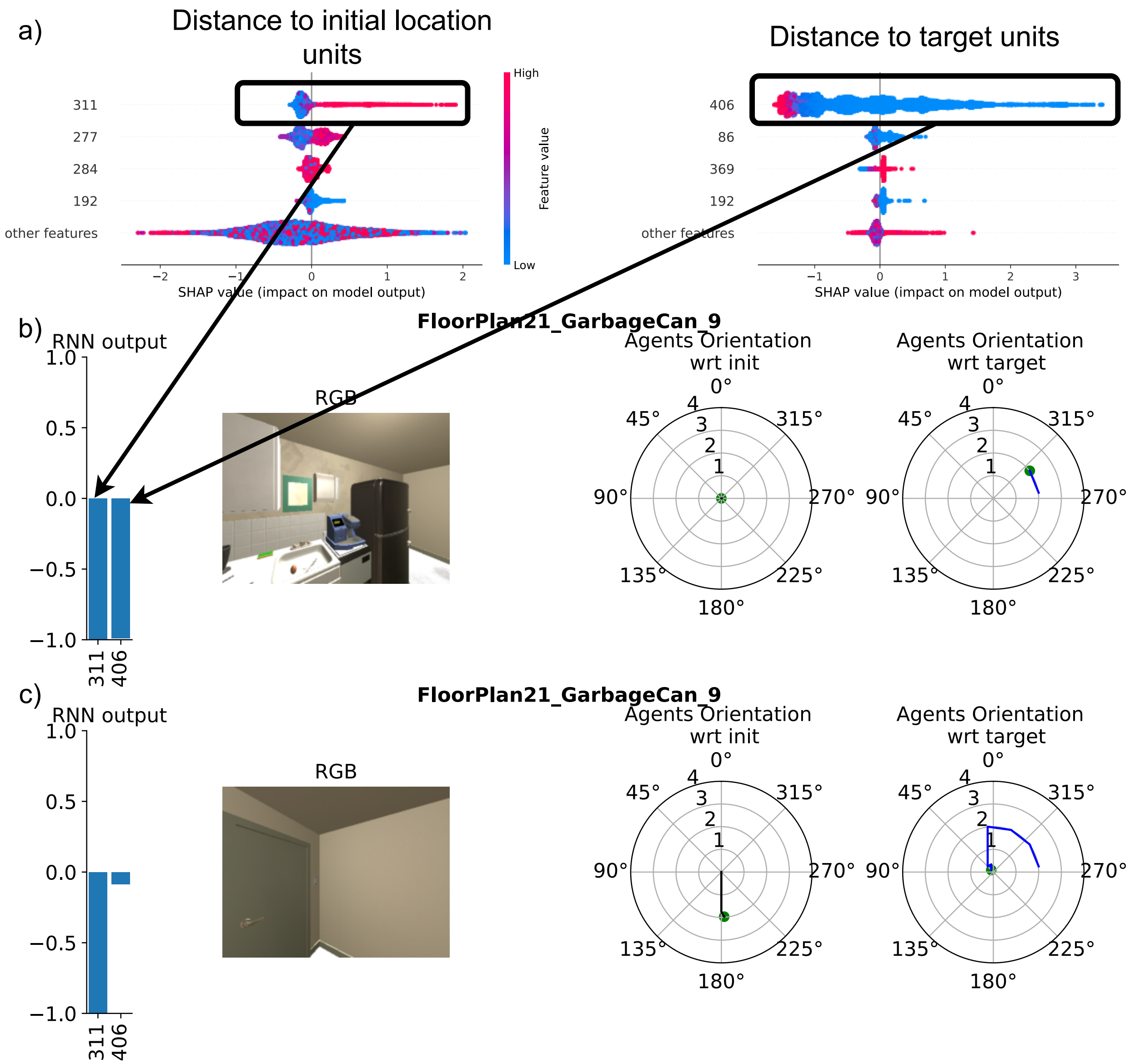

Figure A3. **Visualization of PointNav hidden units** a) Top-4 most relevant hidden units for prediction of distance to agent's initial location (R_a) and target location (R_t) b) The bar plot on left shows response of unit 311 (R_a unit) and unit 406 (R_t unit). The image at the center is agent's current observation. The polar plots on right shows the distance (in meters) and orientation of the agent (in degrees) with respect to agent's initial location (third column) and with respect to target (fourth column). In this case, the agent is at around 2 meters away from the target and is oriented around 315 degrees. The response of both the units is negative. c) In this case, the agent is now closer to target and unit 406's (R_t unit) response increases significantly suggesting that unit 406 is negatively correlated to R_t. Unit 311's response remains almost constant throughout the episode. We further found that unit 406 was in 11^{th} most relevant unit for predicting R_a suggesting that a constant unit like unit 311 together with a R_t unit (e.g. 406) is predicting R_a.

vant for predicting a concept.

F. Concept prediction vs. ObjectNav training progress

We evaluate which concepts investigated in this work were better predicted as the training progressed. In Figure A5, we observe that prediction of target visibility, visited history and reachability improves significantly as compared to other concepts suggesting their importance for ObjectNav task.

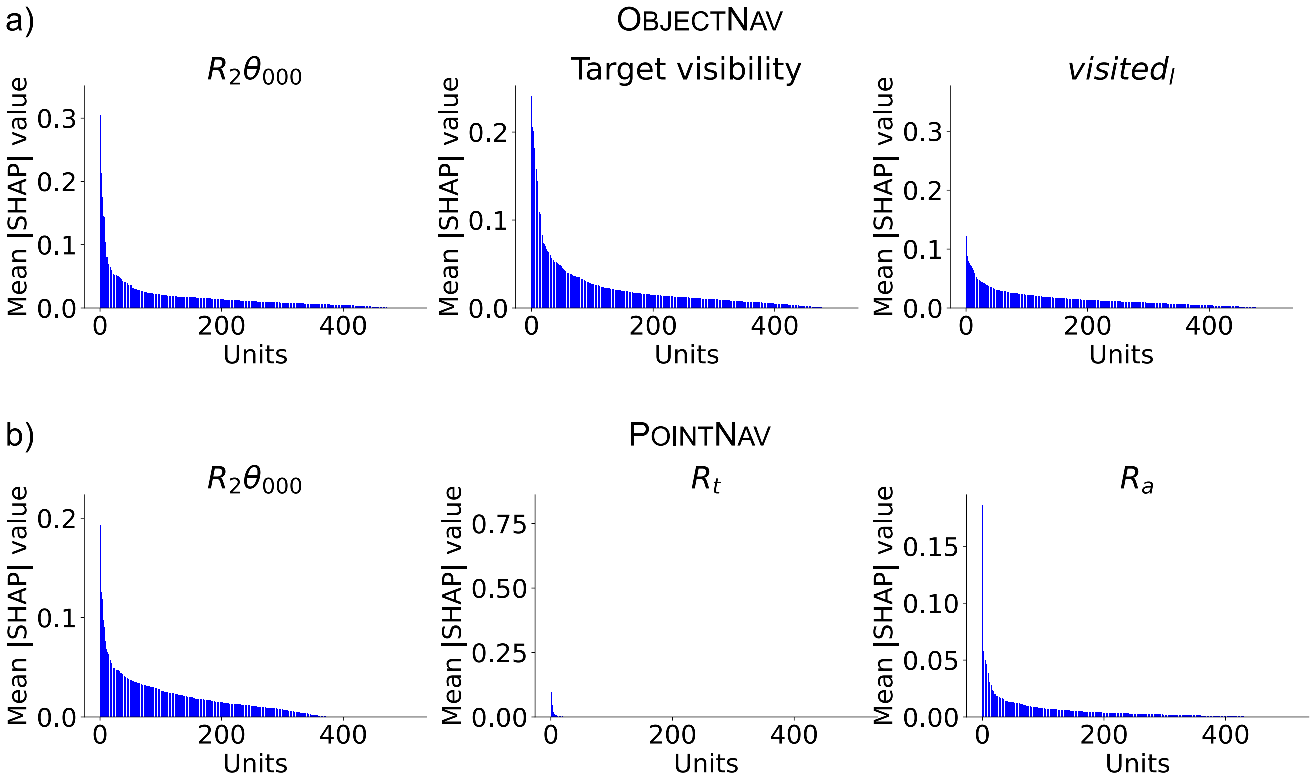

Figure A4. **Distribution of aggregate SHAP values:** for a) concepts learned by OBJECTNAV agent and b) concepts learned by POINTNAV agent. The units are ordered in the decreasing order of the aggregate SHAP values.

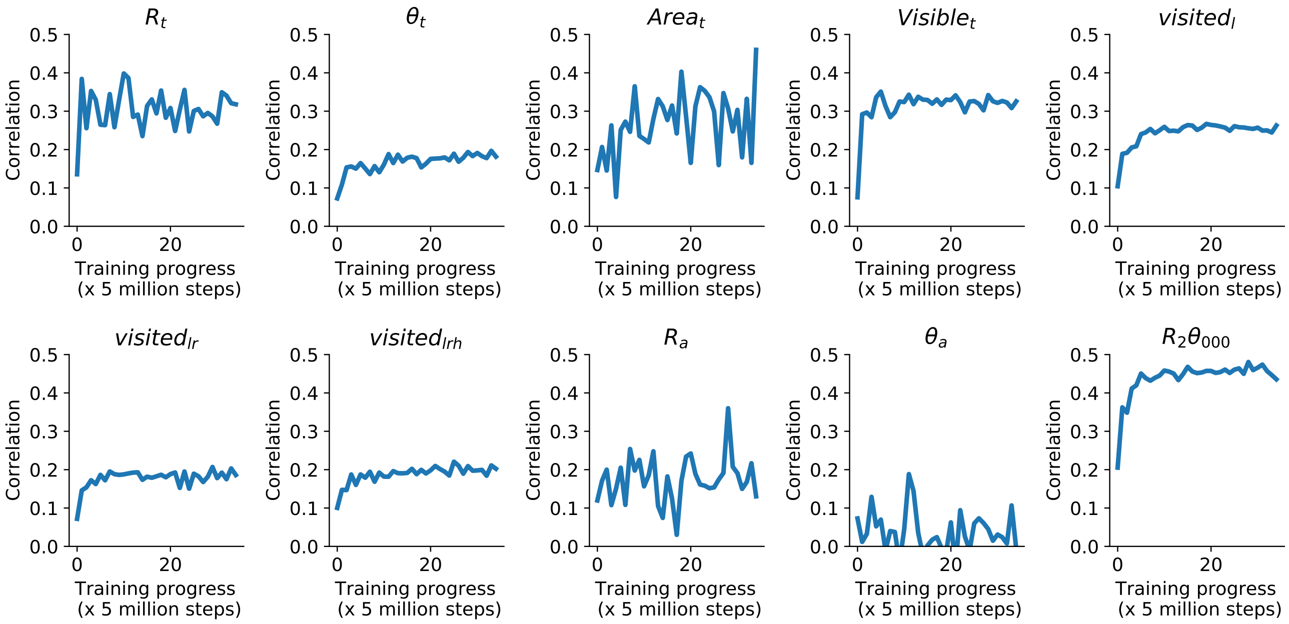

Figure A5. **Concept prediction vs.** RN_{ON} **training progress**

G. Irrelevant units ablation

We investigated removing units that were not relevant for predicting any learned concept in ObjectNav (RN_{ON}) and PointNav (SC_{PN}). The table below shows the impact of removing 25% and 50 % of the irrelevant units. As we can observe removing 25% units does not impact performance significantly in ObjectNav. In PointNav, although the success is close even after dropping 25% of the irrelevant units SPL drops significantly. Dropping more units (50%) significantly drops performance in both ObjectNav and PointNav tasks suggesting that we did not exhaustively investigate all possible concepts that were relevant for performing these tasks and the ablated units might be encoding those missing concepts.

	ObjectNav		**PointNav**	
Ablated units	SPL	Success	SPL	Success
0%	0.227	0.455	0.714	0.879
25%	0.225	0.445	0.659	0.865
50%	0.204	0.419	0.314	0.578

H. Limitations and Future directions

This study has investigated several human interpretable concepts such as target visibility, reachability, etc. However, one can continue to broaden the set of concepts considered in such a study. We also restrict this study to navigation agents, but future studies should consider agents performing interactive tasks. Furthermore, we study RNN neurons individually. However, there is also a chance that multiple neurons together can encode some interesting property. We leave this study for future work.

The insights gained from our paper can also benefit future works: 1. Sparse representation of the target and ablation experiments suggest that irrelevant neurons can be assigned to another task leading to an efficient multitask agent or removed to reduce the size. 2. We found PointNav agents rely less on RGB information. That could be the reason why they perform well in unseen rooms. ObjectNav agents reliance on RGB information could be a weakness and it might help to design separate modules for target identification and navigation to be more robust to room changes. 3. We found that during training, early models (lower performance) do not predict reachability, target visibility, and visited history as well as saturated models (higher performance) suggesting their importance for ObjectNav tasks.

Licenses for assets In this work we use three publicly available assets:

- AI2Thor[1]: Apache 2.0 License
- Allenact[2]: MIT License
- shap[3]: MIT License
- xgboost[4]: Apache-2.0 License

[1] https://github.com/allenai/ai2thor/blob/main/LICENSE

[2] https://allenact.org/LICENSE/

[3] https://github.com/slundberg/shap/blob/master/LICENSE

[4] https://github.com/dmlc/xgboost