We thank the reviewers for their valuable feedback and comments. We address the concerns below:

Regarding novelty (R3, R5). R3 & R5 point out that parts of some modules are based on prior work. Firstly, note that the complete method is significantly different from prior methods ([25,37,38,41]) tackling the object goal navigation task. In fact, we propose an alternate paradigm of approaching this problem using learning-based modular models as opposed to end-to-end learning in prior methods. We show that we can still get the benefits of learning within the modules, but avoid the drawbacks of high sample complexity and memorization in end-to-end learning. The effectiveness of modular methods is not evident without our experiments which we believe the NeurIPS community benefits from knowing.

That said, even among the modules, the use of differentiable semantic mapping and a goal-driven semantic policy on top-down maps is novel. Since our goal-driven policy and map denoising operate only on top-down maps (which are domain-invariant) and our semantic mapping leverages MaskRCNN trained on real-world images, our system can be transferred from sim-to-real. This is very different from end-to-end learning policies which are known to struggle under domain shift. Novelty is also recognized by R1 ("clear algorithmic innovation") and R2 ("adds several new features").

More real-world experiments (R1, R5). All reviewers have appreciated the real-world experiments in the submission. R1 & R5 have suggested there should be more emphasis on real-world experiments. Note that prior works for this task typically conduct only simulation experiments and no real-world experiments on a mobile robot platform (for eg. [25,37,38,41]). Nevertheless, we performed more real-world experiments, testing our model and 2 baselines across 100 episodes each in two different home environments. The proposed model achieved a 62% success rate vs 46% for Active Neural SLAM [6] and 0% for the RGBD+Semantics+RL [25] baseline (either just going forward and colliding or turning on the spot for most episodes). The RL baseline reached close to the goal object in 6 episodes, but it did not take the stop action. These results highlight the difficulty of the task and the significance of our approach. Our domain-invariant modular design allows real-world transfer while end-to-end RL based baselines suffer from visual domain shift. Semantic mapping and goal-driven exploration result in improvement over [6]. Note that [6] is not designed for the object goal task, we adapt it for this task and use it as a strong baseline.

Regarding datasets and implementation details (R1, R2, R5). We will open-source complete code for training and evaluating our model including differentiable geometric computations (R1, R2), order of pretrainings and trainings (R1), map denoising (R2), dataset details (R5). Regarding datasets (R5), Gibson and MP3D are based on real-world reconstructions of indoor homes and are commonly used in visual navigation papers (for eg [6, 11, 36]). We use standard splits from the original datasets. Scenes are fairly large with an average floor area of Gibson and MP3D scenes being $368.8m^2$ and $517.3m^2$ respectively. We expect our method to perform similarly in any typical home environment. Furthermore, the method is not specific to homes or specific object categories used in our experiments, and should also work on other types of environments and objects if trained on them. We will also add all relevant details regarding datasets and implementation in the appendix including a list of scenes used (R5) and training pseudocode (R1).

Regarding error modes and analysis (R2, R5). *GT SemSeg leads to a large improvement in SPL (R2)*: due to small regions being visible and detecting objects from a greater distance, pretrained MaskRCNN often doesn't detect occluded or distant objects. *Where does the system fail most often? (R2)* (a) inefficient exploration, (b) not detecting the object when visible (quantitative numbers in Sec 5.1 and Table 2). We did not observe any local minimum in terms of exploration coverage (agent getting stuck in an area). This is probably because the explored area is explicitly marked on the semantic map. *Training performance of RL baselines (R2)*: 0.37 and 0.49 Success for RL and Semantics+RL in Gibson. *Limitations of using pretrained segmentation (R5)*: Our ablation indicates +19% success when using GT SemSeg, in practice pretrained MaskRCNN can be tuned to adapt to the environment. *More varied examples (R5)*: we provide several examples in the supplementary material and video, we will add more examples in future revisions.

Regarding segmentation and classification (R5). We do not report segmentation or classification performance as we do not propose a method for these tasks. Also, reporting segmentation performance in simulation is not very informative as sim consists of reconstruction errors not present in the real-world. We believe segmentation is not a good indicator of learning semantic priors either as the segmentation for each frame is independent of each other. The long-term goal selection is based on the complete episodic map (which includes the relative arrangement of objects across all the frames), and thus, we believe our ablation is a better indicator of learning semantic priors.

Other questions. R1: Summation over height is category-specific. Voxel representation is 4D with 4th channel being the category. We sum voxels of a particular category for the corresponding semantic map channel. Denoising network is not pretrained. **R2:** Semantic segmentation is not pretrained in Habitat experiments. Denoised map is merged with the current map using channel-wise max pooling. Removing denoising leads to a 5% drop in performance. Success in real-world experiments indicates noisy motion and sensor do not affect our method much. **R5:** Random baseline is random at low-level action. Regarding performance on the challenge leaderboard, note that our method has the best performance on both the splits in terms of success rate.