We thank all reviewers for their valuable feedback. In particular, the reviewers have supported our:
research and ideas (R4-promising research direction & appealing, R5-novel, R2-fresh and convincing),
experiments (R2-thorough and well-conducted, R5-carefully designed) as well as
presentation (R1-clear, R2-well-written, R5-well written and easy to understand).

We address all of the concerns below. In particular, we provide *new comparisons* to past works, address the question of performing this *research in simulation* and place this within the context of past works in *interactive perception*.

R1: **Segmentation and mass are not the end goal of interaction.** We totally agree. However, addressing more complex tasks requires obtaining reasonable performance on these building blocks, which itself is very challenging in a self-supervised setting. Complex downstream tasks including rich manipulations are part of our ongoing research.

R1: **Comparison with [2, 45].** We now provide results for a new baseline [45] - augmenting Mask-RCNN with their Robust Set Loss using their public implementation. Across the suite of metrics, this helps a little beyond Table-1(i), but still inferior to our approach, e.g. 22.3 (theirs) vs 28.06 (ours) on NovelSpaces Box $AP^{0.5}$, row (e). [2] addresses the problem of poking an object from one point to another – very different from us, and not valid to compare to. However, we now provide results for our model trained on their dataset (we do not poke, but use their pokes instead) and with almost no hyperparameter tuning obtain 39.1 for Box $AP^{0.5}$ – validating that our method shows promise on real world data. The higher number (39 vs 28) also indicates the relative complexity of the THOR scenes.

R1: **All in a simulated environment and simple scenes** Please see *R4 Research/Assumptions in simulation* below.

R2: **Rule-based decisions.** These were made to make the learning process more manageable. Future extensions of this work will address more generic architectures and learning paradigms, to remove some these handcrafted designs.

R2: **Prioritized replay and curriculum learning.** Prioritized replay and curriculum learning encourage sampling high-loss and low-loss examples, respectively. Combining these two approaches discourages sampling medium-loss examples. This resulted in high gains for our method.

R2: **Fully supervised Mask-RCNN for novel objects.** It is trained with groundtruth mass and segmentation masks (obtained from THOR) for seen objects (no category information is used). Novel categories are not in training images.

R4: **Research in simulation.** The end goal of this research is to train interactive agents in the real world. Using a rich, large scale and variable simulator allows us to have faster research iterations via fast training (no mechanical constraints), improve generalization (scene and object variability compared to [2, 45]) whilst ensuring safety. Future work includes (a) deployment on a Locobot robot (b) exploring simulation-to-real transfer via real world fine tuning, along the lines of navigation research in embodied AI (e.g., Habitat ICCV19, CARLA CORL17, iGibson ICRA20).

R4: **Assumptions in simulation.** Our setup presently makes some assumptions. However, past real world works also use highly simplified setups. E.g. [45] assumes: (1) objects on a flat surface (2) fixed distance of surface to camera, (3) fairly uniform background, (4) objects going out of the field of view after interaction, (5) fixed camera view, (6) no exogenous motion. Our work relaxes (1)-(4) by providing variable objects, surfaces, rooms and backgrounds.

R4: **Interactive perception baselines.** One can cluster past works in this area into two parts: (a) passive methods that use objects motion to segment them (b) active approaches that jointly learn to act and segment. Papers in **a** cannot be thought of as baselines – they are orthogonal to our method. When our agent learns to interact with objects, the resulting motion can be exploited by these works to result in better segmentation. Adding such techniques on top of our proposed method is very interesting and we leave it for future work. Also, papers mentioned by R4 use *supervised* components. Papers in **b** are very relevant, and we provide new comparisons above. Please see response to R1. However, our method is designed to not just provide object masks but also object attributes – an improvement over past algorithms.

R4: **Additional supervision decreases the quality.** Without supervision the masks and interaction points are noisy. We believe this serves as a form of data augmentation for our method and helps the model to better generalize.

R4: **Majority vote is > 33%.** The metric is mean per class accuracy. So if everything is predicted as the majority class, the accuracy is 100% for that class and 0 for the other 2 classes. Hence, majority vote accuracy is (100+0+0)/3 = 33.3.

R4: **Intuition for mass prediction & how to use motion cues.** The model probably learns to associate the size and texture of the objects to the mass. To incorporate motion cues, we can enable our training pipeline during inference. It can predict the interaction points and infer the mass based on the motion caused by interaction.

R4: **Class imbalance & Objectness prior.** We circumvent the class imbalance issue by learning the objectness score by focal loss only on pixels selected for interaction (line 238). This improves successful interactions from 4% to 60% during training. We did not use objectness priors (e.g., [56,33]) since they use supervised data.

R4: **Mass prediction.** It is per object. We predict an interaction point for objects and a mass map for the entire image. The mass for an object is the mass we predict for the pixel corresponding to the interaction point of that object.

R5: **Non-rigid objects & force directions.** Great challenging suggestions and will be considered for future work.

R5: **Forces.** L112-4 is about the point that the force is applied to. L184-5 is about the direction of the force vector, which is not necessarily in the direction of the ray from the camera center to the point.