We thank the reviewers for their consideration of our paper and their insightful suggestions that will be included in the revision. The consensus appears to be that this is a *"well written"* (R1, R2, R3, R4) paper that introduces a *"simple yet effective module"* (R1) to solve an *"important problem"* (R3) with theoretical advantages that are *"well justified"* (R2). Our approach achieves *"competitive results"* (R3), *"generalizes well"* (R1), and for set prediction tasks *"it is likely that the method will see widespread use"* (R1). We kindly address the reviewers' questions below.

Memory efficiency compared to IODINE (R2, R3)

Following our inquiry, the authors of IODINE updated their paper. We both used the same type of hardware: *"V100 GPUs with 16GB of RAM"* (Appendix A.2 in IODINE, arXiv v3).

Comparisons on realistic images (R1, R3, R4)

Our method has similar inductive biases to IODINE in the autoencoder setting (for example, we use the same type of decoder), so we believe it would perform similarly on realistic images (Figure 11 in IODINE). The strength of our method rather lies in its simplicity, efficiency, and flexibility.

Distinguishing between multiple objects of identical appearance (R2, R4)

As suggested by R2, we ran the object discovery experiment on binarized multi-dSprites to test if we can segment based on shape cues only (also requested by R4). Slot Attention achieves $69.4 \pm 0.9\%$ ARI, whereas the two baselines reported in the IODINE paper achieve $64.8 \pm 17.2\%$ (IODINE) and $68.5 \pm 1.7\%$ (R-NEM). Results on MONet were not reported. This adds further evidence that Slot Attention performs competitively even without being able to rely on color. This experiment along with the discussion will be included in the revision.

Experimental results on object discovery are marginally superior to prior arts (R2, R3)

We have comparable inductive biases to IODINE. Our improvements over IODINE mainly concern the simplicity of our approach, its computational cost, and its generality. The use of an attention mechanism in our approach significantly benefits from a learning rate schedule (to prevent early saturation or instability).

Scalability to more objects and relation between number of objects and iterations (R1)

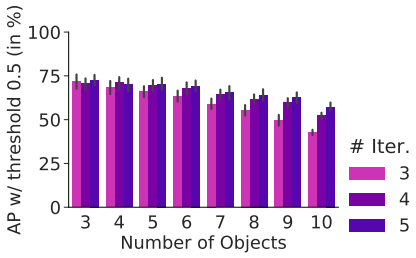

Figure 1: AP stratified per number of objects.

To illustrate the limits of Slot Attention and motivate future work on scalability we will add stratified AP results for property prediction on CLEVR using a different number of objects and iterations. An example with threshold 0.5 (where the effect is strongest) is in Figure 1. We clearly see that more complex scenes require more iterations.

Scalability improvements (such as decomposing the image into patches as in the SPACE model) are however orthogonal to our approach and would be interesting to explore in future work.

Input structured as a set and model dependencies between output elements (R2)

Concurrent work, Kosiorek et al., "Conditional Set Generation with Transformers" (2020), uses attention to directly communicate between slots, which better addresses the single vector to set example. To model dependencies between output elements one could use a graph neural network as part of the task dependent module.

Comparison with soft k-means (R4)

We ran an additional experiment on CLEVR6 where we simultaneously ablated the GRU update, LayerNorm, and the key/query/value projections in the Slot Attention module, which results in a version of soft k-means with a dot-product scoring function (as opposed to Euclidean distance). Using otherwise the same hyperparameters, this model achieves $75.5 \pm 3.8\%$ ARI as opposed to $98.8 \pm 0.3\%$ for Slot Attention. This experiment will be included in the revision.

Further additions for the final version

As suggested by the reviewers we will add the runtime details, polish the references, release the code, add details on how reconstruction masks are visualized, and update the link to the datasets.