Overview. We would like to thank all reviewers for their thorough reviews and helpful feedback/suggestions.

Generalization Experiments. Three of four reviewers asked about generalization to novel objects and scenes. To address this, we trained MulMON, IODINE and GQN on CLE-Aug. Then, we compared their performance on CLE-MV and 2 new datasets—Black-Aug and UnseenShape. Black-Aug contains the CLE-Aug objects but only in single, unseen colour (black). This tests the models' ability perform segmentation without colour differences/cues. UnseenShape contains only novel objects that are not in the CLE-Aug dataset—cups, cars, spheres, and diamonds. This directly tests generalization capabilities. Both datasets contain 30 scenes (more for final paper), each with 10 views. Table 1 shows that 1) all models generalize well to novel scenes, 2) MulMON still performs best for all tasks but observation prediction—where GQN does slightly better due to its more direct prediction procedure (features $\rightarrow$ layout vs. features $\rightarrow$ objects $\rightarrow$ layout), 3) MulMON can indeed understand the composition of novel objects in novel scenes—impressive novel-view predictions (observations and segmentations) and disentanglement.

Tasks	Models	CLE-Aug (train)	CLE-MV	Black-Aug	UnseenShape
Seg. (mIoU)	IODINE	$0.51 \pm 0.001^{\star}$	0.61 ± 0.002	0.50 ± 0.006	0.51 ± 0.004
	MulMON	$\mathbf{0.71 \pm 0.000}$	$\mathbf{0.71 \pm 0.004}$	$\mathbf{0.67 \pm 0.002}$	$\mathbf{0.64 \pm 0.004}$
Pred.Obs (RMSE)	GQN	0.15 ± 0.000	$\mathbf{0.15 \pm 0.001}$	$\mathbf{0.24 \pm 0.003}$	$\mathbf{0.17 \pm 0.002}$
	MulMON	$\mathbf{0.07 \pm 0.000}$	0.16 ± 0.002	0.26 ± 0.002	0.21 ± 0.006
Disent. (D,C,I)	IODINE	$0.54, 0.48, 0.21$	$0.14, 0.12, 0.26$	$0.2, 0.26, 0.27$	$0.13, 0.12, 0.26$
	MulMON	$\mathbf{0.63}, \mathbf{0.54}, \mathbf{0.68}$	$\mathbf{0.52}, \mathbf{0.48}, \mathbf{0.63}$	$\mathbf{0.55}, \mathbf{0.55}, \mathbf{0.66}$	$\mathbf{0.5}, \mathbf{0.47}, \mathbf{0.67}$
Pred.Seg (mIoU)	MulMON	$\mathbf{0.69 \pm 0.001}$	$\mathbf{0.71 \pm 0.004}$	$\mathbf{0.68 \pm 0.005}$	$\mathbf{0.60 \pm 0.005}$

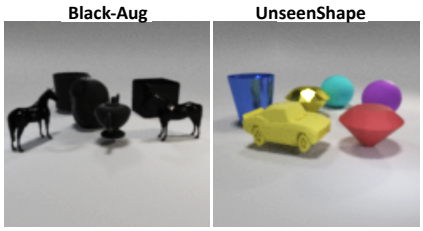

Table 1: Models' generalization performance. *Paper correction. **Figure 1**: Samples from novel-scene datasets.

Writing. We would like to thank the reviewers for feedback on writing style, including both structuring and typos, as well as related work. In particular, we would like to thank R1 for detailed suggestions to improve the abstract, introduction and related work sections. These have all been incorporated, and have greatly improved the paper's clarity.

R4: ***'the contribution in the learning method seems not enough. Most of the method are based on the previous work (IODINE)'***. IODINE's approximation of the MOSV posterior cannot maintain object correspondence or "matching" across multiple views, and thus, it is not a feasible solution to the MOMV problem. To address this, we make two contributions in the learning method itself: 1) we side-step the object matching problem by iterating over multiple views, each time using the previous iteration's approximate posterior as the new prior—see eq. 4; 2) we introduce a training procedure that forces the model to make use of these iterative updates, aggregating spatial information across multiple views. Specifically, by randomly partitioning the views of a scene into two subsets, we can ask the model to predict the views in one (novel viewpoint-queried generation) having observed the views in the other (scene learning)—see eq. 5. This forces the model to use the iterative updates to aggregate spatial information across multiple views, as it needs to form a complete 3D scene understanding in order to perform well. If instead we had asked the model to simply reconstruct the observed view, as IODINE does, it would completely overwrite the prior on each iteration—as it would not need to aggregate spatial information across multiple views in order to perform well.

R1: ***'results for scene factorization, novel view predictions and disentanglement... only 1, 2 and 3 observations'***. Figure 3 in Appendix E plots novel-viewpoint observation and segmentation results as a function of the number of views T (1-10). However, we appreciate R1's suggestion of adding baselines (IODINE, GQN) to these analyses, and also analyzing both segmentation and disentanglement results as a function of the number of views T. We will definitely include these in the final paper. ***'Can the trained model be used to generate random scenes?'*** Yes, we can generate random scenes by composing independently-sampled objects. However, to focus on forming accurate, disentangled representations of multi-object scenes, we must assume objects are i.i.d. and thus ignore inter-object correlations—e.g. two objects can appear at the same location. Nonetheless, we will include random samples in the final paper. ***'not clear ... how many views the model sees during training'***. Using a fixed number of observations could harm the model's robustness at test time. Thus, the number of observations given to the model is sampled at every training step—from the range $[1, 6]$ for the GQN-Jaco dataset, and from $[1, 5]$ for the CLEVR-based datasets. Then, the model is asked to predict the remaining, unseen views. Note that GQN-Jaco dataset has a total of 11 views, while the CLEVR-based datasets have 10. This sampling procedure is introduced as a *randomized partition of T observations into two subsets $\mathcal{T}$ and $\mathcal{Q}$* in lines 191-192 in the paper, but we will revise the text to further clarify this. ***'A video that shows segmentations when moving the camera...'*** As we cannot include external links here, we will create one for the final version.

R3: ***'include a score that ignores background, or ... discuss this difference'***. We thank R3 for pointing this out, and we agree that IODINE's poor mIoU is mostly due to its handling of the background. We will discuss this issue in the paper, and add a score (e.g. Adjusted Rand Index) to our comparisons that ignores the background. However, good models shouldn't split up such simple backgrounds—see Fig. 3 in our paper and Fig. 1 in IODINE paper. ***'In what sense is the set $\mathbf{z} = \{z_k\}$ grouped into T groups based on the views?'*** R3 is correct in that this design is just to show that we update $\mathbf{z}$ one piece (i.e. $\boldsymbol{z}^t$) at a time w.r.t. its corresponding x^t. We will add some discussions to clarify this.