We thank reviewers for their insightful and positive feedback! We are encouraged that they find our motivation and idea novel (R1), important (R1, R4), of value to the community (R1, R4), and has high potential significance (R1). We are glad they found our paper does not have major flaw (R2), well written (R2, R4), theoretically sound (R1, R4), our empirical evaluation well designed and thoroughly performed (R1, R2, R4), and the experimental results promising (R1, R2). We address the reviewer's comments below and will incorporate all feedback.

[R1] GNM distinguishes between purely symbolic and purely distributed representation while AIR and SPACE use it conservatively: Not true. GNM does not require purely symbolic representations. In the structured representation layer, we are actually using hybrid (symbolic +distributed) representation which is exactly the same as SPACE and AIR, i.e., distributed for $\mathbf{z}^{\text{what}}$ and symbolic for the others. Yes, the distributed part is still groundable. We will make this point clearer. **Comparison to SPACE is missing:** SPACE uses independent priors for symbolic priors and thus *by design* cannot model the distribution of global scene structure. See also a relevant answer for R4. **Reporting model sizes and hyperparameters:** Yes, we use comparable model sizes. We will include the model sizes, hyperparameters, and training curve in the supplementary material. We will also release our code. **Increasing model size of GENESIS would make it better?** In our experiments, GENESIS used 4 times more parameters (14M) than our model (3M). This is because GENESIS uses additional conv-layers followed by MLP layers for the encoder and decoder of the segmentation networks which our model does not need. **Why not compare to existing datasets, e.g. CLEVR:** We agree that to compare on existing datasets can be better if possible. However, we found that objects in existing datasets are rather independently generated without complex spatial and inter-object relations and thus do not reveal required global scene structures. This makes it difficult to evaluate the learned global structure, which is of our main interest. For follow-up works, we will release our datasets. **Why GENESIS cannot obtain digit-wise decomposition for multi-MNIST scenes:** Although we have worked hard to make GENESIS decompose the multi-MNIST scene, we could not observe the decomposition for this dataset. We believe that our result is valid. First, similar to our result, the results in Figure 3 and Figure 4 in the GENESIS paper, exhibit strong color-based segmentation bias (e.g., the circles in the wall). We actually also observed that in our Arrow-Room dataset, GENESIS often grouped different objects into the same segmentation when they have the same color. A careful hyperparameter tuning was required to prevent this. We believe that this is a reasonable thing to observe in GENESIS because it does not use the *locality* principle of objectness. Second, it is indeed interesting to know that R1 has observed the decomposition on a similar dataset. It would have been great if we were able to access the result to compare with ours. However, given that there is no published result about the multi-MNIST task using GENESIS including R1's experience, our result is the first and the only accessible result on which we can make any reasonable decision. **Strong claim that all prior methods have failed:** We fully agree and thanks for pointing this. We will tone down this and the sentence "this ability is not properly studied". **Other modeling possibilities:** We agree that it is indeed an interesting suggestion. The SPACE paper shows that the auto-regressive approach does not scale gracefully and iterative inference (e.g., of IODINE) is known to be computationally expensive. However, as pointed, we note that our proposed framework is more generally applicable beyond the SPACE representation style. **Interaction MLP** is a standard MLP that takes the CNN feature map as input and mixes (interacts) globally. Lacking this, standard ConvDRAW do not mix information between features spatially distant. **Dual representation:** both $\mathbf{z}^s$ and $\mathbf{z}^g$ are inferred from the same input $\mathbf{x}$. So they are different representations encoding the same information. **Beta parameter effect** is studied because some VAE-family models (ConvDRAW in our case) work better with a different beta. We thank all the **additional feedback** which will make our paper clearer.

[R2] Why the particular form of the symbolic representation: In this work, we have taken, as an example implementation of the symbolic inference model $q(\mathbf{z}^s|\mathbf{x})$, the SPACE model, and thus uses the same symbolic representation of theirs. However, the proposed framework is more general and not limited to the way SPACE represents an object. It can be applied to make any symbolic or hybrid (symbol + distributed) representation conform a proper distribution. **More visualization would be helpful although Figure 5 is "already really nice":** We thank you for suggesting this. We will add more visualization if we come up with an additional idea for that. **Why not categorical variable for $\mathbf{z}^{\text{what}}$:** because the categorical value cannot model within-category variation.

[R4] Need the evaluation of structured representation learning ability. Thanks for suggesting this. We initially thought that we do not need this experiment because our structured representation learning is basically the same as SPACE (with a single segment for the background modeling). However, following the suggestion, we performed the suggested experiments by comparing ours and SPACE. We indeed obtained a similar performance between ours and SPACE. Specifically, for average precision IOU on the MNIST-10 dataset, we obtained 0.459 and 0.453 for our model and SPACE, respectively. And for classification accuracy on the recognized $\mathbf{z}^{\text{what}}$ representation, we obtained 0.984 and 0.980 for ours and SPACE, respectively. The SPACE paper demonstrates its advantages against MoNET and IODINE in some settings and without loss of generality, we can extend that result to our model. We will add this result to the revision. **Manual checking data:** As suggested, we will publish the corresponding data we used for the manual investigation. **FID Score:** We will consider adding FID score. However, as agreed, it may not affect the main result of the paper. **Missing reference:** Thanks! We will add the suggested paper.