We thank all reviewers for their valuable and constructive comments. Below, we address the detailed comments. In particular, we clarify some potential misunderstandings from R#3 and provide extra experiments as suggested by R#3.

To R#1: Q1: Generality, constraint design and heavy bias. Thanks for acknowledging our novelty. Indeed, ASR is a general framework to properly incorporate structural knowledge into DGMs without heavy bias as long as the knowledge can be quantitatively represented. It is shown that PR can be extended to "selectively" incorporate uncertain knowledge (e.g., with noise) represented by the general language of first-order logic [*1], where highly uncertain knowledge will be dropped according to the faithfulness of fitting the given data; ASR extends PR to an amortized version for structured generation, thereby inheriting the generality in a principled manner. As for the concrete example about bounding boxes with varying sizes, one possible way is to define a probabilistic model over sizes and use ASR to regularize its posterior by constraining some statistics (e.g., mean (i.e., average size) and variance). This idea can be generally applied to other constraints in a soft manner without introducing heavy bias. As discussed in L46-49, ASR has several advantages over a structural prior on incorporating useful structural knowledge; it would be valuable for many vision tasks (e.g., the indoor scene of furniture arrangement [*2]). Finally, it is feasible to use a NN F to present the constraints, in which case the differential condition can be met as in L131 and the end-to-end training strategy still applies under a similar regularization form as in Eqn.(10). We'll make this clearer in the final version.

Q2: Fig 3 and Fig 4. The odd columns are real data and even ones are the reconstruction results. The bounding boxes in real data are inferred by ASR which is used to demonstrate the quality of inference and boxes in reconstructions are used to highlight the each individual reconstructed object. It was a fault to miss the 8-th column (i.e., the reconstruction results of the data in 7-th column). The training and testing data are selected randomly which results in the fact that the examples picked are different. We'll fix these issues for better presentation.

Q4: Compare to SQAIR. We mainly focus on generating multi-object images, but ASR can be applied to sequential models. For example, SQAIR may suffer from the issue of swapping inference order as pointed in Appendix G of SQAIR [19]. ASR provides a possible solution to this issue by adding extra regularization, e.g. minimizing the distance between the appearance latent variables for the same object in different time steps. As SQAIR requires several days to converge, we'll add the results of ASR regularized SQAIR in the final version.

To R#2: Thanks for your comments. We'll update the derivation about J' in L140.

To R#3: Q1: Definition of "structure". Structure mainly refers to some regularity (e.g., size, shape) of an object or the relationship among objects in an image. Under PR, we generally refer to the posterior constraints that consider the regularity or relationship. We'll make this clearer.

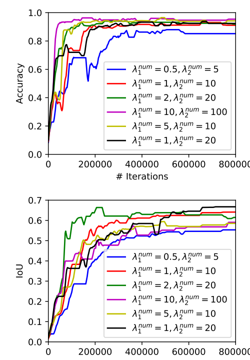

Figure A: The sensitivity analysis about λ in AIR-ASR-13.

Q2: Novelty. Our main contribution is on on extending PR to DGMs for structured generation, as agreed by R#1 and R#2. ASR provides a general solution, which is more flexible than previous efforts on designing structured priors (See L22-33). Although PR is a well-known technique, it is nontrivial to apply for DGMs. Specifically, the variational distribution in the vanilla PR is typically of a simple form and sample-specific (See Sec.2.2). For DGMs, we extend PR to an amortized version (See Sec.3.2), which can be trained in an end-to-end manner under a regularization formulation (i.e., problem (10)).

Q3 & Q4: $q(Z)$ and the penalty term. In fact, $q(Z)$ is NOT a prior distribution; it is the variational distribution to approximate the target posterior $p(Z|X)$. In Line141, we develop an amortized version of PR, i.e., using a recognition model to explicitly define $q(Z)$ as $q(Z|X;\phi)$ with parameters ϕ. Then we get problem (9) with constraints of $q \in Q$. In order to train the model in an end-to-end manner efficiently, we further turn the constraints into a penalty term R in Eq.(10). For sufficiently large λ_i, problem (10) is equivalent to (9); in general it is a relaxed form. We present two examples of the penalty term in Sec.4. The effect of the penalty term will be answered in Q5. We'll make it clearer.

Q5: Experiments. As for **ablation study**, the effect of the penalty terms was already verified in Fig. 3 and Fig. 4 and the quantitative results were reported in Tab. 1 and Tab. 2. Without ASR, AIR-pPrior (i.e., AIR with learnable parameterized prior) tends to stuck in the trivial local optimum where the whole image is treated as a single object and the underlying structures are ignored, as discussed in Sec.6.1 and Sec.6.2. With the penalty terms which represent the structural constraints, ASR can successfully regularize AIR to escape the local optimum and help AIR capture the underlying structures. We further provide **sensitivity analysis** for ASR on the number of objects with 1 or 3 objects on Multi-MNIST. The accuracy of the inferred number of objects and mIoU are reported at the top and bottom plots of Fig. A correspondingly. As we can see, ASR is robust to the hyperparameter λ. Finally, as for **dataset**, we adopted the widely-used datasets mainly for a fair comparison to previous state-of-the-art models. Our empirical results indeed demonstrate that ASR is an effective approach to embedding human knowledge into DGMs, as agreed by R#1 and R#2. Nevertheless, ASR can be applied to more complicated datasets (e.g., the 3D scene images used in [3, 11]) by exploring its generality (See our response to Q1 of R#1), which is our future work. We'll make this clearer in the final version.

[*1] Mei, et al., Robust RegBayes: selectively incorporating first-order logic domain knowledge into Bayesian models, ICML 2014.
[*2] D. Ritchie, K. Wang, Y. Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. CVPR, 2019.