Thank you for the positive, constructive and in-depth reviews. We found the suggestions and comments to be very helpful. Below, we summarize the main questions and comments raised by each reviewer and provide responses.

(REVIEWER 1) Trade-off in representation power. The CGQN representation does not model agreement among the individual modalities as long as the summed representation can model the scene. Thus, it suffers if an unseen combination is given at test time. In contrast, the PoE and APoE approaches are robust to this issue as they seek the agreement through the product of experts operation. Being free from this agreement constraint, CGQN representations may be obtained by searching a broader model space than PoE-based models. However, our experiments show that having this agreement constraint (so searching the representation from a less broad space) does not hurt the scene modeling performance while solving the partial observability problem.

Probabilistic hierarchical encoder. We have not tried the hierarchical latent model version. Although this is an interesting model variant, due to space limitation we didn't prioritize exploring it. We believe that it will not be difficult to train that model as the reparameterization trick can be applied there as well. However, evaluating the advantages of having the hierarchical probabilistic encoding would require more thought.

3rd modality depth data. We agree and hope to explore this in the future work.

Why does APoE improve computation efficiency? In principle, APoE should have the same computation complexity as PoE. However, in practice we observe that our amortization helps improve the computation speed as well. It seems that the amortized ConvDraw uses GPU parallel computation more efficiently that having a ConvDraw for each modality. For better understanding, we need a more investigation on how PyTorch and CUDA actually realizes this efficiency.

Scene representation for RL. We agree that extending this work in a real world setting with a robot arm and RL policy learning is a very interesting future direction.

Other comments on the presentation. We agree on all the points and will improve it in our camera-ready.

(REVIEWER 2) In comparison to Wu & Goodman (2018) We agree that, our approach to addressing the partial observability problem extends the work of Wu and Goodman (2018) by adopting the PoE modeling approach (we will make this point clearer). However, we do not claim this as our contribution. We see our main contributions as being: (i) the formalization and demonstration of *3D modality-invariant representation learning and generation using human-like multisensory inputs under the GQN framework*, (ii) the introduction of an amortized PoE which resolves the problem that the Wu and Goodman model suffers from when it is applied to our 3D modeling problem, namely the inherent scalability problem of the PoE model due to space complexity, and (iii) the introduction of the MESE simulation environment which we believe is a significant contribution, given that there is no such environment currently available to the community. In addition, we agree that it is worth noting that the derivation of lines 165-169 is from Wu & Goodman (2018). We cite them on line 171, but we will make this point clearer.

Experiments on PoE by Wu & Goodman (2018) We indeed tried to run the PoE model of Wu and Goodman in our problem by integrating it with a CGQN. (Note that the MVAE model of Wu and Goodman cannot directly be used for our 3D modeling problem.) However, we could not run it properly after the CGQN integration due to its increased memory footprint and slowed speed. This is how we arrived at our position that the amortization contribution is indeed important. Also, in terms of accuracy, we do not think our APoE will be better than PoE. The PoE should be the upper bound of our model because our model is an amortized version of it, i.e., due to the amortization gap. So, the advantage of amortization in our case is mainly on improving efficiency and scalability not necessarily on better accuracy. Considering both reasons, we believe that it is fine not to compare to PoE in terms of accuracy—but we provided the comparison in terms of space and computation efficiency. If reviewers think it will nevertheless be helpful, we will be happy to find a way to add the PoE result in the camera-ready version.

Full-modality configuration in previous work. We fully agree on this point. There was some confusion because the way we implement the missing modality situation is somewhat more challenging in the sense that we also consider the case where the fully joint modalities are not available at all. We, however, agree that this is not a significant difference as currently described in the related work section. We will fix the description and clarify that this problem is studied in Wu & Goodman (2018).

(REVIEWER 3) Thanks for the constructive review. **More complex task.** We agree on this point. We are actually eager to extend the proposed model and investigate it in a real world setting with real objects and an robotic arm.

References

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *NeurIPS*, 2018.