We thank the reviewers for their detailed comments and are encouraged by their positive appraisals. We take on board all comments regarding typos, spacing, and citations [**R1** & **R3**] and will fix these appropriately in the updated manuscript.

[**R1** … *fully-available modalities* …] Although we are primarily interested in the cognitive-science motivation (§1) for all modalities being available, our formulation *can* indeed deal with missing modalities by virtue of the MoE properties—effectively amounting to assuming $\alpha_m = 0$ for missing modality m. Note however that doing so (as with [1] using the PoE), can raise questions of *consistency*. Constructing a joint encoder with a missing modality does not change the target—it simply is 'another' importance distribution—but, the decoder missing a modality changes the actual model being targetted since it is not the full generative model $p_\theta(z, x_{1:M})$ anymore! It is unclear what the right approach in this case ought to be; and in conjunction with our motivation, is why we do not explore this setting here.

[**R1** … *experimental protocol* … *results* …] We will include a more thorough discussion of the quantitative analysis for the experiments and the results. We do indeed classify digits, and use inputs from MNIST or SVHN to classify for that corresponding modality (67% & 87.3% resp.). The 94.5% corresponds to the MVAE of [1] when *both* modalities are given as inputs—the implication being that it hedges its abilities entirely on MNIST (94.42% when given just MNIST input), largely ignoring SVHN (10.5% when given only SVHN)—a consequence of the PoE formulation.

[**R1** … *more convincing experiment* … *CUB* …] Although our focus in this work is not on pixel-level generation, (as also noted by **R3**), we will include pixel-level-generation results on a downscaled version of CUB images in the supplementary as requested. We omitted results for MVAE on CUB because they were quite poor, and results on MNIST-SVHN served sufficiently to highlight its shortcomings. We will however include these results in the updated manuscript for completeness.

[**R2** … *not intrinsic generic criteria* …] We believe these criteria are still quite general in that they are typically easy to verify experimentally—the procedures in §4 aim to provide a blueprint for how to do so, and we will make this protocol more explicit in the updated manuscript.

[**R2** *Tabular data* …] Deep generative models for tabular data often focus on learning proper co-occurrence structure between different attributes, so it may be difficult to apply this method directly (which would treat each attribute as conditionally independent); although we agree this is an interesting avenue for future work.

[**R3** … *not that original* … *Tsai et al.* …] The main question across prior approaches [incl. 2,3] has been the formulation of inference. While different choices have yielded different capabilities (c.f. Fig 2, right) we believe, and show, that MoE posteriors are better. Characterising our contribution as 'not that original' does not do us justice. Thank you for the citation to Tsai et al. We note that although the motivation in latent factorisation is shared, the means are quite different—we do not explicitly structure the latent space (as originally seen in [4]), allowing the factorisation to be captured automatically (c.f. Fig 5). We also show coherent joint generation from the unconditional prior.

[**R3** … *no comparisons* … *conditional VAE* …] Our focus here is on learning *joint* generative models that additionally allows for (computationally) cheap conditional generation. Of course, we agree that if one wanted only to learn a conditional model, say from language to vision, then explicitly targetting that capability alone would be useful; but that isn't our goal—and such would not be an apples-to-apples comparison. The value of learning a joint model lies in the ability to learn, in an unsupervised manner, the abstract commonalities in the observed data—in being able to unconditionally generate (from the prior $p(z)$) data in different modalities that are related in the same way the observed data was—for MNIST-SVHN the digit (c.f. Fig 4 left) , and for CUB, more nuanced notions of language grounding (c.f. Fig 7 bottom left). The motivation comes from human perception (§1) and is not something that conditional models do.

[**R3** … *ablations* …] In comparisons against [1] we have since performed additional ablations to find out which aspect of our formulation provides the performance win—since we differ from [1] in both form of posterior (MoE vs PoE) *and* the estimator used (DReG vs ELBO)—and can show that the MoE formulation is the critical component.

[**R3** … *simple linear classifier* …] As stated in the manuscript (l258-l262) the linear classifier is used to quantify how well-separated the information is in the latent space, not simply an attempt to perform state of the art classification.

[**R3** … α_m *dynamic* …] This is indeed true—and can be handled by the DReG estimator. Note that Eq (1) computes importance weights across both modality and samples from each modality, which is similar to the dynamic weighting in multimodal fusion methods. We explicitly disallow importance weighting across modalities since different encoders can contribute to a joint representation to different degrees, which we avoid in order to explore Criteria 1 (§1). However, we will include an example in the supplement to showcase the dynamic weighting enabled by Eq (1).

[**R3** … *IWAE/ELBO* … M^2 … *multiple runs* …] See [5,6] for the tightness and posterior properties of IWAE; see Appendix A1 for linear-in-M objective. We will update the manuscript with statistics from runs across multiple seeds.

[1] Wu & Goodman, 2018 [2] Ngiam et al., 2019 [3] Blei & Jordan, 2003 [4] Bousmalis et al., 2016 [5] Burda et al., 2016 [6] Le et al., 2018