We would like to thank all reviewers for their detailed, thoughtful and valuable feedback. We are encouraged that the reviewers are convinced by our motivation [R1, R2], methodical contribution [R3, R4], experimental results [R1, R2] and the thorough comparison to related work [R2, R3, R4]. We address the reviewers' comments below by first clarifying the derivation of our proposed ELBO, followed by a detailed explanation of certain aspects of our experimental setup. Lastly, we address individual questions.

Dynamic Prior and generalized JS-Divergence. [R1, R2, R4] There are two major reasons for using a PoE as distribution for the dynamic prior: 1) The KL-divergence between a Gaussian distribution and a PoE of Gaussians can be calculated in closed-form, as mentioned in Section 3.4. 2) PoE-based models [27] are able to approximate the joint posterior distribution well which we would like to utilize. Instabilities in training and overconfident experts as mentioned in [21] are mainly due to difficulties in the optimization of unimodal posterior approximations. In our case, the JS-divergence allows us to optimize the unimodal and multimodal approximation functions jointly, leading to a stable training of the PoE approach. The fact that the standard JS-divergence is defined via mixture distribution was the main reason to use the MoE in the derivations. As such, the derivation based on the MoE is a special case, which is most familiar to readers, but the result is more general and holds for any abstract mean distribution (as in [18]). We will state this more clearly in the final version of the paper.
[R2, R3] Derivation of ELBO (Eq. 9): Similar to [24], the dynamic prior is defined by a data-dependent function. In the general formulation of Eq. (6), there are not yet any additional assumptions on the prior distribution. The dynamic prior defines a valid distribution for the MoE- as well as for the PoE-variant (see appendix B.2 and B.3) and hence is a well-defined prior. This makes the first line of Eq. (9) a valid ELBO - independent of the exact formulation of the dynamic prior as long as it is a proper distribution. We would like to emphasize that Eq. (9) is not meant to be a lower bound to Eq. (6). As stated in Section 3.2 and 3.3 and proven with Lemma 2, we only claim the validity of the proposed ELBO using the JS-divergence. As proven in [18], the derivation can be generalized to any abstract mean distribution, including the geometric mean that defines the PoE. We will point this out in the final version of the paper. [R2] The references for the JS-divergence for M distributions [1,14] are given in the introduction, we will add them in Section 3 as well. The reference for the extension to generalized means in [18] is given in Section 3.4.

Experiments. [R1, R2] We highlight the advantage of modality-specific (MS) subspaces using conditional generation plots (cf. Figure 1) instead of latent traversals. For conditional generation, MS subspaces allow to mix and match different shared and MS encodings. The columns show that the shared and MS spaces disentangle (every row is a different random sample from the MS latent subspace). The shared information (digit number) is invariant per column, while the MS information is invariant per row. This gives empirical evidence that MS and shared latent spaces encode different information. We will include a visualization of low-dimensional embeddings of the shared and MS latent spaces in the Appendix. [R4] We already describe the details of all the models incl. MS spaces in the Appendix (Section C.2.2). To this, we will add the respective ELBOs utilizing MS subspaces. [R3, R4] To the best of our knowledge, we are the first to perform an experiment with three different types of modalities. In our opinion, different modalities should contain information that is specific to each modality. In [27]'s vision study, the different modalities are filtered versions of the original modality which prevents them from having true modality-specific information.
[R4] In our opinion, the quality of generated samples is only one side of the coin to evaluate multi-modal generative models. We are convinced that only its combination with the coherence of generated samples allows for a valid assessment. Although the MVAE model is able to generate high quality samples, comparing Table 3 and 4 shows that the quality of samples comes at the cost of reduced coherence accuracy (for conditional generation) which is significantly lower than MMVAE's and ours. Additionally, by introducing MS subspaces, we find a solution to generate samples of high quality which are coherent between modalities - random generation and all subsets of samples. [R3] We report the NLL-numbers for completeness as it is a de-facto standard evaluation method for VAE-based models despite the known weaknesses. [R1] The introduction of modality-specific subspaces leads to a small overhead regarding the hyperparameters (incl. priors). In our experiments, we used standard Gaussian priors - as it is common for VAE-based models - for the modality-specific subspaces which work very well in practice. [R3] A comparison of runtimes can be found in the Appendix (Table 8, Section C.2.4) which highlights the inefficiency of IWAE-based models (3x longer training time) as mentioned in the related work Section.

Further Questions. [R3] "Comparison to JMVAE": The introduction of the JS-divergence has a similar motivation as JMVAE [22] which we discuss in the related work. By using the JS-divergence the unimodal posterior approximations are automatically optimized for being close to a joint posterior distribution (the dynamic prior in our case). We are able to solve this in a scalable way while [22] have to use an exhaustive approach. [R4] "Ablation studies in the Appendix": Our hypothesis for the stable random coherence in this experiment is that the unimodal posteriors are learned in a way that their mean distribution is similar to static prior (we have no final hypothesis for the dip yet). [R4] "Calculation of coherence for random generation": The generated samples were evaluated using a pre-trained classifier for each modality. If all modalities show the same content/shared information, this is a coherent generation. From there, we calculate accuracy/precision recall.