We thank the reviewers (R) for their insightful comments. We acknowledge that the reviewers highlighted the importance of our motivation (R1, R2, R3,R4), the significance and simplicity of our methodology (R1,R2,R4), and the soundness of empirical evaluations (R1,R2,R4). We address the concerns of each reviewer below.

R1 & R2: Do marginal VAEs scale to high dim data? Yes. The marginal VAE training is highly scalable to high dim data since the marginal VAEs can be trained in parallel very easily, using simple vectorization tricks. This is how we implemented our experiments. Moreover, the network size in each marginal VAE is very small as it only need to learn one dimensional marginal distribution for each feature. We will open-source our code upon acceptance.

R1 & R3: Does two stage training introduce suboptimiality? The suboptimality, if any, comes from the error induced by marginal VAEs. As pointed out by R1, this should not be a big problem since they are fit to one-dimensional variables. In appendix D.2, we have evaluated the approximation quality of each marginal VAEs, which is indeed very high. *Also, we have introduced a new baseline where the model is trained jointly. This will be presented later in this rebuttal*

R3 : Is VAEM novel in comparison with HI-VAE? And what does "uniformity" mean? Indeed, VAEM highly relates to HI-VAE *which is one of our baselines (dubbed as VAE-HI)* in our experiments. In all the experiments, we have shown that VAEM has a very significant improvement over HI-VAE, confirming the novelty of our contribution. The reviewer may have missed this baseline due its naming, which we will change from VAE-HI to HI-VAE and clarify accordingly.

To understand the novelty of VAEM, we would like to point out that: **1**, in HI-VAE, the first layer latent representation z_{nd} are *deterministic*, while in VAEM they are stochastic. **2**, unlike HI-VAE (trained end-to-end), VAEM is trained in two-stage. Therefore, the *marginal statistics and inter-variable dependencies are separated.* Meanwhile, it's now possible to introduce prior terms $p(z)$ (Appendix A.2). Thanks to $p(z)$, the marginal distribution for z_{nd} is enforced to be standard Gaussian, so that the dependency network only has to model random variables that are of the same statistical type and with homogeneous marginal distributions. This *"Gaussianization" (acknowledged by R4, also referred as "uniformity" in our paper) does not happen with the HI-VAE (nor other more general hierarchical VAE methods).*

R3 & R4: Comparison with two-layer VAE baseline? We acknowledge this suggestion. For completeness, we ran two-latent-layer VAE (trained jointly, with matching latent dimensions) as baseline on data generation tasks. Other training hyperparameters are consistent with other baselines. The nllh performance (Bank: 1.678±.05, Boston: -0.629±.01, MIMIC: -0.394±.00, Avocado: -0.137±.00, Energy: -1.46±.01) is generally worse than our method.

R3: Missing description of the heterogeneity in the dataset? We have indeed presented this information in Appendix C (mentioned in Section 5). All sources of heterogeneity are presented in the data-sets used. You can also see it from the ground truth data distribution in Appendix E. We will add further information regarding each feature in each dataset.

R3: In VAE-adaptive baseline, a single minibatch is not sufficient to compute the scaling factor. In our VAE-adaptive baseline, the scaling factors are indeed fine-tuned *using the entire dataset* as suggested. We mentioned the "mini-batch" approach in Appendix C.1.2. only because it is more general and scalable. We will clarify this. Also, we have tried to fine-tune each scaling factor manually, which yields similar results, and VAEM still outperforms this baseline significantly. We did not include this result as it is similar to the VAE-adaptive.

R3: Comparison to more complicated models such as Ladder VAE? Our two-stage VAEM approach is in principle compatible with any VAE decoders and could also be applied to Ladder VAEs. Other advances in VAE can be applied to VAEM in the same way as in VAE. To further address the reviewer's concern, we would like to point out that our HI-VAE baseline has a similar structure/parameter numbers/model complexity compared with VAEM. Also, HI-VAE is trained end-to-end. Hence, HI-VAE already serves as a baseline for ablation study in this case.

R3 : Why is SAIA used? Besides its high application impact, SAIA is highly relevant since it quantifies how well the model fits the data and how good the inference is. SAIA can be treated as an extension of imputation tasks, since it assesses the overall imputation performance of the model without specifying a certain ratio of missing data at test time. *In a heterogeneous data setting*, if a model cannot handle heterogeneity well, it might favor certain types of features, resulting in poor performance in the SAIA task.

R3 :Is negative NLLHs a sign of overfitting? Not necessarily. Negative NLLHs are perfectly possible when there are many continuous variables with highly peaked densities. This is indeed the case in our datasets (Appendix C and E).

R4: relationship to (Valera et al., 2017) and necessity of VAE in the first stage. We have discussed the mentioned work (Valera et al., 2017) in our paper. It is orthogonal to our work since it addresses the problem of automatic type discovery in a traditional LVM setting. In our scenario (all types are known), VAEs are particularly useful, since: **1**, VAEs are very efficient and scalable in practice, and **2**, we need a probabilistic model for down-stream tasks such as SAIA, since this will enable efficient quantification of information gains.

R4: Likelihood is not sufficient. Why not try missing value imputation tasks? We have included imputation tasks in Sec. 5.3, and provided imputation error metric in Appendix D.