We thank all the reviewers for their time to provide comments and the positive feedback from reviewer 1 and 4. Also, we appreciate all reviewers for their correct summarisation of our contributions as (1) Introduce the ice-start problem motivated by real-world examples and proposed a method to tackle it. (2) Propose a novel inference algorithm by combining amortized inference of local latent variables and a sampling approach for global weights. (3) Propose novel element-wise acquisition objectives. In the following, we will address the questions proposed by each reviewer.

> Clarify definition on two test problems (Reviewer 3) Thanks for the suggestion. We will focus on the imputation task and make the task setting more clear in the beginning of section 2 and present the prediction task as an extension. We believe that it is important to include two test tasks as these two tasks together cover the most typical real-world problems. This demonstrates the broad applicability of the proposed method. Note that we have indeed included the detailed procedures of both tasks in appendix B.2 Algorithm 3 (imputation) and 4 (active prediction). In the revised version, we will also add a problem definition section for these two tasks in main text.

> Acronyms (Reviewer 3) We will add the NLL acronym in line 80. For AUIC, we used its full name in line 205. The details about AUIC are included in algorithm 4 in appendix B.2. In the revised version, we will move it to the problem definition section in the main text.

> BELGAM related work (Reviewer 4) Thanks for pointing to further related work, which will be included in revised version. However, we would like to highlight our novel inference method for BELGAM. The previous related methods pose strong restrictions/assumptions of the model. For example, they typically require the forward model to be conjugated to the prior or use mean-field approximations for the posterior. Our proposed method uses a sampling approach, which is theoretically guaranteed to give the optimal posterior under mild conditions. In BELGAM section, we will use simple examples to help explaining the generative model for better contextualization.

> Questions about the computation of KL term. (Reviewer 4) The KL term $\mathrm{KL}(q(\theta|\boldsymbol{X})||p(\theta))$ is indeed intractable due to MCMC sampling. However, this term appears in the line 136 and is only used to train the encoder parameter ϕ. Thus, the gradient is taken w.r.t. ϕ, and the gradient contribution of this term is zero.

> Encoder for partial observations (Reviewer 1, 4) In the main text, we only briefly mentioned its structure because it is not our novel contribution. We will add a detailed introduction of this encoder in the appendix of the revised version. The term x_{io} represents the observed entries for row i. S_i is indeed a matrix. Because for each row i, we have some observed features. Each feature has a vector embedding. We concatenate each observed feature value (scalar) with its embedding (vector) to form a vectorised representation of this specific feature. Then, we group all observed features for row i and form its matrix representation S_i.

> Notation (Reviewer 4) We will add a table that summarizes the notations used in this paper in the appendix for clarity.

> Difference to basic element-wise AL and related work section (Reviewer 1) We will extend and include more details on the BALD objective in the "Data-wise active learning" subsection, and the comparison to the traditional element-wise active learning (element-wise AL) in the "Feature-wise active learning" subsection in related work. Briefly, we agree that ice-start problem is tightly related to the high-level idea of element-wise AL. However we argue that existing work of element-wise AL methods are commonly restricted to a particular application, such as classification, where a fully-observed test inputs are required; or associated with strong assumptions, such as linear missing data model (e.g. matrix completion) or heuristic acquisition objectives. Thus, they cannot handle the problems like active prediction (test task also has query budget) or imputation tasks with highly non-linear relationship in real-life applications.

>Inference algorithm description (Reviewer 1) We thank reviewer 1 for appreciating our contribution of this novel inference algorithm. We will extend this part, add relevant citations and distinctions compared to previous work in the revised version. Briefly, there is little previous work that uses Bayesian treatment over weight parameters in the context of VAE. The previous inference algorithm either relies on the conjugacy of forward and prior distribution, or strong assumptions over posterior approximations e.g. mean-field. Our proposed method does not requires strong assumptions over posterior, and is guaranteed to give accurate posterior samples under mild conditions. The better posterior estimates not only help with the prediction accuracy but will also indeed aid the acquisition because the acquisition involves the expectation over the posterior. Thus, an inaccurate posterior can result in a poor approximation of the acquisition and lead to the query of the uninformative feature.

> Questions about KL term inside expectation. (Reviewer 1) The reason for writing $\mathrm{KL}(q(\boldsymbol{z}_i|\boldsymbol{x}_i)||p(\boldsymbol{z}_i))$ inside the expectation is because we factorize the posterior in line 126 for computational efficiency. Thus, this factorization decouples the dependency of local latent variables with global weights. In addition, the local variable $\boldsymbol{Z}$ is amortized through each individual $\boldsymbol{x}_i$. The $\mathrm{KL}(q(\boldsymbol{Z}|\boldsymbol{X})||p(\boldsymbol{Z}))$ can be simply written as a summation of $\mathrm{KL}(q(\boldsymbol{z}_i|\boldsymbol{x}_i)|p(\boldsymbol{z}_i)))$ for all $\boldsymbol{x}_i$. Thus, putting it inside the expectation does not affect its value and we can write it outside in the revised version.

> Novelty and combination of acquisition functions (Reviewer 1) To the best of our knowledge, the conditional mutual information is novel in the context of element-wise AL and the previous work is more based on heuristics (e.g. feature variance and model improvement). The combination is based on our intuition as described in lines 166-168. But it indeed corresponds to an information-theoretic quantity (discussed in the paper line 192). This gives an additional warranty for its validity.