We thank all the reviewers for the constructive comments. Our response to the reviewers is given as follows.

Reviewer 1

We thank the reviewer for the detailed comments and suggestions on improving the clarity of the paper. The following clarifications are needed:

1) For backward message passing, a node also performs average aggregation on the messages from its parent nodes.

2) Distributions at Nodes include both conditional prior distribution and conditional posterior distribution. The parameters for conditional prior distribution are determined by the output of the link flow functions that connect the parent nodes. Similarly, the parameters for the conditional posterior distribution are from the outputs of the link flow functions that connect the child nodes.

3) The inference procedure is different from classical graphical models. Messages from observed nodes have to be aggregated at the root node to predict messing node values with probability.

4) Missing value inference relies on random masking training, and the inference is done via an approximate inference method.

5) Without concatenation aggregation nodes, all flows in a VFG model have the same dimension number. We can use the variable duplication trick to ensure dimension consistency at all nodes.

6) Figure 5 measures the imputation prediction mean squared error. The lower MSE, the best the method is imputing the missing values.

Reviewer 3

The results on improved ELBO are presented in section C.2.1 of the supplementary material. For one training sample x, we draw M samples to approximate the KL terms and the ELBO loss. We thank the reviewer for the suggestions on results and writing, and the issues will be addressed in the revised paper.

Reviewer 5 and Reviewer 6

We thank the reviewers for their concerns on paper clarity and writing suggestions.

1) The proposed prior distribution is a hierarchical distribution. Thus, the KL divergence between the prior and the posterior requires layer-wise KL calculation. Hence, evaluation of the KL term in each layer requires samples from both prior and posterior distributions.

2) Derivation of the log-determinant is based on the change-of-variable rule. Probability computation from one node to one of its adjacent nodes is based on the definition of normalizing flow given by equation (3).