computing the parameters for z_u^i. Incorporating the autoregressive likelihood into yields:

$$p(x_u \mid x_o, b) = \left|\det \frac{dq_{x_o,b}}{dx_u}\right| \prod_{i=1}^{|u|} p(z_u^i \mid z_u^{i-1}, ..., z_u^1, x_o, b), \tag{1}$$

where $|u|$ is the cardinality of the unobserved covariates.

0.1 Training and Best Guess Objective

Since we follow a flow-based approach, we have access to the normalized conditional likelihoods; thus, we can train our model by maximizing log likelihood. During training, if we have access to complete training data, we will need to manually create binary masks b based on some predefined distribution p_b. p_b is typically chosen based on the application. For instance, Bernoulli random masks are commonly used for real-valued vectors. Given binary masks, training data x are divided into x_u and x_o and fed into the conditional model $p(x_u \mid x_o, b)$. If training data already contains missing values, we can only train our model on the remaining covariates. As before, we manually split each data point into two parts, x_u and x_o based on a binary mask b. Note that dimensions in b corresponding to the missing values are always set to 0, i.e., they are never observed during training. In this setting, we will need another binary mask m indicating those dimensions that are not missing. Accordingly, we define observed dimensions as $x_o = x\,[b]$ and unobserved dimensions as $x_u = x\,[m \odot (1-b)]$. In this setting, we optimize $p(x_u \mid x_o, m, b)$. During testing, we set $b = m$, that is, the model imputes the missing values conditioned on all the remaining covariates.

Best Guess Objective Given a trained model, multiple imputations can be easily accomplished by drawing multiple samples from the learned conditional distribution. However, certain downstream tasks may require a single imputed "best guess". Unfortunately, the analytical mean $\mathbb{E}_{p(x_u|x_o,b)}[x_u]$ is not available for flow-based deep generative models. Furthermore, getting an accurate empirical estimate could be prohibitive in high dimensional space. In this work, we propose a robust solution that gives a single best guess in terms of the MSE metric (it can be easily extended to other metrics, e.g. an adversarial one). Specifically, we obtain our best guess by inverting the conditional transformation over the mean of the latent distribution, i.e., $q_{x_o,b}^{-1}(\bar{z}) = q_{x_o,b}^{-1}(\mathbb{E}_{p_{\mathcal{Z}}(z|x_o,b)}[z])$. The mean $\bar{z}$ is analytically available for Gaussian mixture base model. To ensure that this best guess is close to unobserved values, we optimize with an auxiliary MSE loss:

$$\mathcal{L} = -\log\ p(x_u \mid x_o, b) + \lambda \| q_{x_o,b}^{-1}(\bar{z}) - x_u \|^2, \tag{2}$$

where λ controls the relative importance of the auxiliary objective. Note that we only penalize one particular point from $p(x_u \mid x_o, b)$ to be close to x_u. Hence, it does not affect the diversity of the conditional distribution. Unlike our arbitrary conditioning task, conditional generative modeling deals only with a fixed vector of external information y (such as a class label) to condition the likelihood of a fixed set of covariates x (such as pixels in an image). Typically, approaches supplement the inputs of common generative models with some encoding of the conditional information. Conditional GANs , for example, extended GANs by inputting a class label encoding into both the generator and discriminator function, allowing these models to learn multiple conditional distributions. Similarly, a conditional form of VAE was proposed by , which introduces an additional network that learns the conditional prior $p_\theta(z \mid y)$, where z is some latent encoding and y is a class label.

0.2 Arbitrary Conditioning Models

Previous attempts to learn probability distributions conditioned on arbitrary subsets of known covariates include the Universal Marginalizer , which is trained as a feed-forward network to approximate the marginal posterior distribution of each unobserved dimension conditioned on the observed ones. During sampling, they propose to use a sequential sampling mechanism by adding a newly sampled dimension to the observed sets and running the network in an autoregressive manner. However, we observe that VAEAC suffers from failure modes common in VAEs given that they optimize with respect to ELBO loss . In particular, VAEs sometimes sacrifice learning the true posterior for minimizing reconstruction loss, which often leads the model to generate blurry samples for multimodal distributions. We extend the RealNVP model by replacing all the coupling layers with our proposed arbitrary conditional alternative. For the sake of sampling efficiency, we use standard Gaussian base likelihoods here. Implementation details of and baselines are provided in Appendix . Figure. shows samples drawn from and VAEAC. More samples are available in Appendix . We notice our model can generate coherent and diverse inpaintings for all three datasets and different masks. Compared to VAEAC, our model is capable of generating sharp samples and restoring more details. Even when the missing rate is high, can still generate decent inpaintings. To quantitatively evaluate our model, we report peak signal-to-noise ratio (PSNR) and negative log-likelihoods (NLL) in Table.. We generate 5 different masks for each test image and report the average scores and the standard deviation. We note that PSNR is a metric that prefers blurry images over sample diversity . Log-likelihood measures how well the model matches the real conditional distribution, but may not correlate with visual quality . Hence, we evaluate the tradeoff between sample quality and diversity via the precision and recall scores (PRD) . Since we cannot sample from the groundtruth conditional distribution, we compute the PRD score between the imputed joint distribution $p(x_o)p(x_u \mid x_o)$ and the true joint distribution $p(x)$ via 10,000 samples from them. The PRD scores for two distributions measure how much of one distribution can be generated by another. Higher recall means a greater portion of samples from the true distribution $p(x)$ are covered by $p(x_o)p(x_u \mid x_o)$; and similarly, higher precision means a greater portion of samples from $p(x_o)p(x_u \mid x_o)$ are covered by $p(x)$. We report the $(F_8, F_{\frac{1}{8}})$ pairs in Table. to represent recall and precision, respectively.