Impute Missing Entries with Uncertainty

Jaesung Lim[1*], **Seunghwan An**[2*], **Jong-June Jeon**[3†]

[1]Department of Statistics and Data Science, University of Seoul, S. Korea
[2]Department of Information and Telecommunication Engineering, Incheon National University, S. Korea
[3]Department of Statistics, University of Seoul, S. Korea
js.lim@uos.ac.kr, sh.an@inu.ac.kr, jj.jeon@uos.ac.kr

Abstract

Missing data presents a widespread challenge in real-world data collection. In this paper, our goal is to impute missing entries while accurately reflecting the uncertainty associated with them. We introduce U-VAE, a method that employs a non-parametric distributional learning strategy to parameterize the likelihood of missing values. To address the infeasibility of directly estimating the underlying conditional distributions due to data incompleteness, we incorporate stochastic re-masking and un-masking techniques during training. Specifically, we replace the conventional reconstruction loss with the continuous ranked probability score (CRPS), a strictly proper scoring rule, and theoretically demonstrate that the discrepancy between the underlying conditional distribution and our imputer is upper-bounded. We evaluate the performance of U-VAE on 11 real-world datasets, showing its effectiveness in both single and multiple imputations, while also enhancing post-imputation performance and supporting valid statistical inference.

Code — https://github.com/Optim-Lab/U-VAE

Appendix — https://github.com/Optim-Lab/U-VAE/blob/main/Appendix(U-VAE).pdf

1 Introduction

Incomplete tabular data is a widespread issue in real-world data collection. For instance, in the social sciences, missing entries often occur when participants skip questions or choose not to respond (Little 1988), while in the medical domain, patient records frequently contain gaps due to incomplete documentation or unperformed tests (Barnard and Meng 1999). In this context, substantial progress has been made in the theory, methodology, and software for handling missing data over the past decades (van Buuren 2012).

Missing values are *unknown* and inherently uncertain. As a result, point estimation tasks, such as inferring the population mean, are not statistically valid when incomplete data is given (Rubin 1975; Gibson, Little, and Rubin 1989; Manski 2003). Although single imputation methods are simple and have practical advantages, they fail to account for this uncertainty and are thus inadequate for such statistical inference tasks (Troyanskaya et al. 2001; Stekhoven and Bühlmann 2012; Khan and Hoque 2020).

Q: Parameterization of uncertainty?		
None	*Additive noise*	*Non-parametric*
Mean kNNI missForest SoftImpute	VAEAC MIWAE not-MIWAE GAIN ReMasker	**U-VAE** (ours)

Table 1: Categorization of imputation methods by their parameterization of uncertainty. Our proposed method, **U-VAE**, models uncertainty non-parametrically.

To mitigate this limitation in the presence of missing data, Rubin and Schenker (1986) introduced *multiple imputation*, a framework that generates several plausible values to reflect uncertainty. It enables, beyond point estimation, valid statistical inference of evaluating bias, coverage rate, and confidence intervals (Rubin 1989; van Buuren and Groothuis-Oudshoorn 2011; Murray 2018).

Supporting multiple imputation necessitates estimating the underlying distribution, conditioned on arbitrary patterns of observed entries (van Buuren 2012; Ivanov, Figurnov, and Vetrov 2019; Nazábal et al. 2020). Deep generative models, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models, have been adopted to estimate the conditional distribution of incomplete tabular data and to support multiple imputation (Mattei and Frellsen 2019; Jarrett et al. 2022; Wen et al. 2024). Recently, masked autoencoder (MAE) frameworks have shown promise in imputation performance (He et al. 2022; Du, Melis, and Wang 2024; Kim, Lee, and Park 2024). Specifically, ReMasker (Du, Melis, and Wang 2024) leverages the self-attention mechanism of Transformers to capture inter-feature dependencies.

However, a key limitation of these methods is that their imputation models capture only a partial aspect of the inherent uncertainty associated with missingness. For instance, VAE-based imputation methods (Ivanov, Figurnov,

[*]These authors contributed equally.
[†]Corresponding author.

and Vetrov 2019; Mattei and Frellsen 2019; Nazábal et al. 2020) typically assume Gaussianity and estimate only the conditional expectation of missing values. MAE-based imputation methods (Du, Melis, and Wang 2024; Kim, Lee, and Park 2024) rely on point-wise reconstruction of masked features given the observed ones in a deterministic manner.

Therefore, we propose **U-VAE** (Uncertainty-VAE), an uncertainty-aware VAE-based imputation method capable of estimating the conditional distribution of missing entries given observed values. Our main contribution is that **the uncertainty of the imputed values is modeled non-parametrically beyond additive noise, and this is theoretically supported by showing that the KL divergence between our imputation distribution and the underlying conditional distribution can be upper bounded**. To achieve this, we adopt the continuous ranked probability score (CRPS), a strictly proper scoring rule (Matheson and Winkler 1976; Gneiting and Raftery 2007) for our reconstruction loss. Moreover, to account for the partial observability of real-world datasets—where the true values of the missing entries are inherently unknown—we incorporate stochastic re-masking and unmasking during training. This enables the model to learn from diverse conditioning set-target combinations.

We substantiate the effectiveness of our proposed method by evaluating its performance in both single and multiple imputations across 11 real-world tabular datasets. Furthermore, we include the metrics of bias, coverage, and confidence interval width to evaluate whether the imputation model supports valid statistical inference (van Buuren 2012; Lee and Huber 2021; Zhang et al. 2023).

Notations We assume that the incomplete dataset consists of n i.i.d. samples $\{(\mathbf{x}^{(i)}, \mathbf{r}^{(i)})\}_{i=1}^n$, where the missingness pattern of an observation is described by a binary indicator vector $\mathbf{r} \in \{0,1\}^p$, where $\mathbf{r}_j = 0$ indicates that the jth variable $\mathbf{x}_j$ is missing. It indicates that $\mathbf{x}$ is a p-dimensional observation vector including missing entries, where each element $\mathbf{x}_j$ denotes either a continuous or a categorical variable. We define the multi-delta distribution as $p^*(\mathbf{x_r}) := \frac{1}{n}\sum_{i=1}^n \delta\big(\mathbf{x_r} - \mathbf{x}_\mathbf{r}^{(i)}\big)$, where $\delta(\cdot)$ indicates the Dirac delta function and $\mathbf{x}_\mathbf{r}^{(i)} := \mathbf{x}^{(i)} \odot \mathbf{r}^{(i)}$ ($\odot$ is the element-wise multiplication and missing entries are replaced with zero). Let I_C and I_D be index sets for continuous and categorical variables, respectively, such that $I_C \cup I_D = \{1, \cdots, p\}$. For each categorical variable $\mathbf{x}_j$ with $j \in I_D$, let C_j denote the number of categories.

2 Proposal

2.1 Modeling of Uncertainty

Distributional learning The ultimate goal for faithful multiple imputation that captures uncertainty is to find the imputer such that minimizing the following distributional learning objective (Rubin 1996):

$$\min_{\theta,\eta} \quad D\big(p^*(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x_r}) \,\big\|\, p(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x_r}; \theta, \eta)\big), \tag{1}$$

where $p(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x_r}; \theta, \eta)$ is an imputer and D is a well-defined divergence between two distributions (in this paper, we use the KL-divergence for D).

To achieve this, under the variational inference (Kingma and Welling 2014), our proposed objective is given by

$$\min_{\theta,\eta,\phi} \quad \mathbb{E}_{p^*(\mathbf{x_r})p^*(\mathbf{x}_{1-\mathbf{r}}|\mathbf{x_r})}\big[\mathcal{L}^*(\mathbf{x_r}, \mathbf{x}_{1-\mathbf{r}}; \theta, \eta, \phi)\big], \tag{2}$$

where

$$\begin{aligned}&\mathcal{L}^*(\mathbf{x_r}, \mathbf{x}_{1-\mathbf{r}}; \theta, \eta, \phi) \\ &:= \mathbb{E}_q\Bigg[\sum_{\substack{j \in I_C: \\ \mathbf{r}_j = 1}} \underbrace{\int_0^1 \rho_\alpha\Big((\mathbf{x}_{1-\mathbf{r}})_j - Q_j(\alpha, \mathbf{z}, \mathbf{x_r}; \theta_j)\Big) d\alpha}_{\textit{CRPS}}\Bigg] \\ &\quad + \beta \cdot D_{\mathrm{KL}}\left(q(\mathbf{z} \mid \mathbf{x_r}, \mathbf{x}_{1-\mathbf{r}}; \phi) \| p(\mathbf{z} \mid \mathbf{x_r}; \eta)\right).\end{aligned}$$

Here, $\mathbb{E}_q$ denotes the expectation with respect to $q(\mathbf{z} \mid \mathbf{x_r}, \mathbf{x}_{1-\mathbf{r}}; \phi)$, $\rho_v(u) = u(v - \mathbb{I}(u < 0))$, and $Q_j(\cdot,\cdot,\cdot;\theta) : [0,1] \times \mathbb{R}^d \times \mathbb{R}^p \mapsto \mathbb{R}$, which is a neural network parameterized with θ. $p(\mathbf{z} \mid \mathbf{x_r}; \eta)$ and $q(\mathbf{z} \mid \mathbf{x_r}, \mathbf{x}_{1-\mathbf{r}}; \phi)$ denote the conditional prior and posterior, respectively, both of which are assumed to be multivariate normal distributions.

In (2), our reconstruction loss is the *continuous ranked probability score (CRPS)*, which is a strictly proper scoring rule (Gneiting and Raftery 2007; An and Jeon 2023). **This implies that we adopt a non-parametric approach (i.e., the quantile function estimation) for parameterizing uncertainty, i.e., the likelihood of the missing entries.** Moreover, (2) can be derived from the upper bound of (1), by utilizing an infinite mixture of asymmetric Laplace distributions (see Appendix for a detailed derivation).

Theoretical rationale To ensure the well-posedness of the inverse mapping, we impose Assumption 1 on the conditional quantile function. In addition, Assumption 2 is adopted, which states that the dependencies among missing entries can be captured in a low-dimensional latent space.

Assumption 1. *For all $j \in I_C$, $\mathbf{z} \in \mathbb{R}^d$, and $\mathbf{x} \in \mathbb{R}^p$, $Q_j(\cdot, \mathbf{z}, \mathbf{x}; \theta_j)$ is invertible and differentiable.*

Assumption 2. $\{\mathbf{x}_j : \mathbf{r}_j = 0\}$ *are conditionally independent given the latent variable $\mathbf{z}$ and the other observed variables $\{\mathbf{x}_j : \mathbf{r}_j = 1\}$.*

Proposition 1 (Validity of imputer)**.** *For any $\varepsilon > 0$, suppose that there exist ϕ and η such that: $D_{\mathrm{KL}}\left(q(\mathbf{z} \mid \mathbf{x_r}; \phi) \,\|\, p(\mathbf{z} \mid \mathbf{x_r}; \eta)\right) < \varepsilon$. Given such ϕ and η, we define $\hat{\theta}$ as the minimizer of the objective function in* (2)*. Then, under Assumptions 1 and 2, we have*

$$D_{\mathrm{KL}}\big(p^*(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x_r}) \,\big\|\, p(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x_r}; \hat{\theta}, \eta)\big) < \varepsilon.$$

Proposition 1 states that if the learned prior distribution $p(\mathbf{z} \mid \mathbf{x_r}; \eta)$ is sufficiently close to the aggregate posterior $q(\mathbf{z} \mid \mathbf{x_r}; \phi)$, then the estimated conditional distribution, i.e., our imputer $p(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x_r}; \hat{\theta}, \eta)$ can closely approximate the ground-truth conditional distribution $p^*(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x_r})$. Note that the second KL divergence in Proposition 1 corresponds to the ultimate objective in (1), while the KL divergence term in our objective (2) serves as an upper bound of $D_{\mathrm{KL}}\left(q(\mathbf{z} \mid \mathbf{x_r}; \phi) \,\|\, p(\mathbf{z} \mid \mathbf{x_r}; \eta)\right)$, and is minimized during training.

Therefore, our imputer not only captures uncertainty of missing entries through a non-parametric quantile function (see Table 1) but is also theoretically justified by Proposition 1.

In practice, we parameterize Q_j using a linear spline quantile function, which yields a closed-form loss (Gasthaus et al. 2019) and satisfies Assumption 1*. The detailed formulation is provided in the Appendix.

Remark 1. *In existing VAE-based imputation methods (Ivanov, Figurnov, and Vetrov 2019; Nazábal et al. 2020; Mattei and Frellsen 2019; Ipsen, Mattei, and Frellsen 2021), the decoder is typically modeled as a parametric distribution, most commonly a Gaussian. This limits the modeling capacity to capture uncertainty in the form of additive noise. In contrast, we shift to full distribution estimation, allowing for a more accurate approximation of the inherent uncertainty in the incomplete data.*

Un- and re-masking However, because the given incomplete dataset contains missing values, it is not feasible to directly estimate (2) from observed data. In other words, the underlying conditional distribution $p^*(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x}_\mathbf{r})$ is *unknown*. To address this, following (Ivanov, Figurnov, and Vetrov 2019; Du, Melis, and Wang 2024), we adopt a *re-masking* and *un-masking* strategy that explicitly accounts for incompleteness of the dataset during distributional learning.

Let $\mathbf{m} \in \{0,1\}^p$ be a binary vector indicating which observed variables are artificially masked during training. Specifically, $\mathbf{m}_j = 0$ represents that the jth column is masked (i.e., treated as missing), while $\mathbf{m}_j = 1$ indicates that it is unmasked and used as input. Given a masking vector $\mathbf{m} \in \{0,1\}^p$, the un-masking and re-masking observations are defined as

- (Un-masking) $\mathbf{x}_\mathbf{m} \coloneqq \mathbf{x}_\mathbf{r} \odot (\mathbf{m} \wedge \mathbf{r})$
- (Re-masking) $\mathbf{x}_{1-\mathbf{m}} \coloneqq \mathbf{x}_\mathbf{r} \odot ((1-\mathbf{m}) \wedge \mathbf{r})$,

where $\odot$ denotes element-wise multiplication and $\wedge$ denotes the element-wise minimum operation.

This transformation introduces a new challenge: to learn the conditional distribution for imputation, distributional learning must be performed that can cover nearly all possible conditioning set-target combinations induced by the given missing entries. To achieve this, we introduce sufficient variability in the masking patterns applied to the observed data, where the sampling scheme for the masking vector $\mathbf{m}$ is formally defined in Definition 1. Note that the combination of $\mathbf{x}_{1-\mathbf{m}}$ and $\mathbf{x}_\mathbf{m}$ reconstructs the full observed vector $\mathbf{x}_\mathbf{r}$.

Definition 1 (Masking distribution)**.** *Let $u \sim \mathrm{U}(0,1)$ be a random variable of the masking probability. The masking distribution $p(\mathbf{m})$ is defined as:*

$$p(\mathbf{m}) \coloneqq \mathbb{E}_{u\sim \mathrm{U}(0,1)} \left[\prod_{j=1}^{p} u^{\mathbf{m}_j} (1-u)^{1-\mathbf{m}_j} \right].$$

*The linear isotonic spline is non-differentiable only at a finite number of points, where the underlying measure has no point masses.

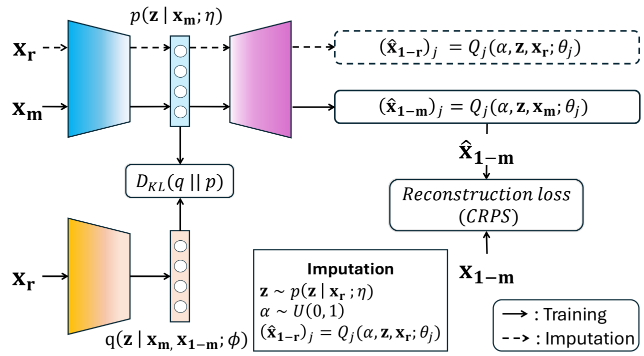

Figure 1: Overall procedure of U-VAE.

Our modified practical objective function incorporating the masking strategy is given by:

$$\min_{\theta,\eta,\phi} \quad \mathbb{E}_{p^*(\mathbf{x}_\mathbf{r})p(\mathbf{m})}\big[\mathcal{L}(\mathbf{x}_\mathbf{r}, \mathbf{m}; \theta, \eta, \phi)\big], \tag{3}$$

where

$$\begin{aligned} &\mathcal{L}(\mathbf{x}_\mathbf{r}, \mathbf{m}; \theta, \eta, \phi) \\ &\coloneqq \mathbb{E}_q\left[\sum_{\substack{j \in I_C: \\ \mathbf{r}_j = 1, \\ \mathbf{m}_j = 0}} \int_0^1 \rho_\alpha\Big((\mathbf{x}_{1-\mathbf{m}})_j - Q_j(\alpha, \mathbf{z}, \mathbf{x}_\mathbf{m}; \theta_j) \Big) d\alpha \right] \\ &+ \beta \cdot D_{\mathrm{KL}}\left(q(\mathbf{z} \mid \mathbf{x}_\mathbf{m}, \mathbf{x}_{1-\mathbf{m}}; \phi) \| p(\mathbf{z} \mid \mathbf{x}_\mathbf{m}; \eta) \right). \end{aligned}$$

In our objective (3), the expectation with respect to $p^*(\mathbf{x}_\mathbf{r})p^*(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x}_\mathbf{r})$ is replaced by the expectation with respect to $p^*(\mathbf{x}_\mathbf{r})p(\mathbf{m})$.

Minimizing the log-likelihood only for the entries that are both observed and re-masked serves as a valid proxy for learning the ideal objective in (2). This is because, by the definition of $p(\mathbf{m})$—which has full support over $\{0,1\}^p$ (Ivanov, Figurnov, and Vetrov 2019; Du, Melis, and Wang 2024)—the model is exposed to all possible conditioning–target configurations. In other words, there can exist a mask vector $\mathbf{m}$ such that $\mathbf{r}^{(i)} = \mathbf{m} \wedge \mathbf{r}^{(j)}$, allowing the model to learn the conditional distribution given $\mathbf{r}^{(i)}$ using the missingness pattern of the jth observation. $\mathbb{E}_{p(\mathbf{m})}$ is approximated using Monte Carlo and ancestral sampling.

Proposition 2 ((An et al. 2024))**.** *Assume that the missingness mechanism in $\mathbf{x}$ follows the MAR. Then, since $\mathbf{r} \perp\!\!\!\perp \mathbf{m} \mid \mathbf{x}$ and $\mathbf{m} \perp\!\!\!\perp \mathbf{x}$ by Definition 1, the joint behavior of $\mathbf{r}$ and $\mathbf{m}$ depends only on the observed components $\mathbf{x}_\mathbf{r}$:*

$$(\mathbf{r}, \mathbf{m}) \perp\!\!\!\perp \mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x}_\mathbf{r}.$$

Proposition 2 demonstrates that the missingness pattern in $\mathbf{x}_\mathbf{m}$ inherits the MAR property of $\mathbf{x}_\mathbf{r}$. Therefore, this result allows us to perform distributional learning using only observed and re-masked entries in a theoretically valid manner (Ivanov, Figurnov, and Vetrov 2019; Du, Melis, and Wang 2024). However, as shown in Section 4, our model also demonstrates strong empirical performance even in cases where the MAR assumption does not strictly hold.

Algorithm 1: Missing data imputation

Input: Incomplete observation: $(\mathbf{x}, \mathbf{r})$
Output: Imputed observation: $\widetilde{\mathbf{x}}$

```
1: x̂ = [0, 0, ··· , 0]
2: z ~ p(z | x_r; η)
3: for j = 1, 2, ··· , p do
4:     if j ∈ I_C then                ▷ inverse transform sampling
5:         x̂_j ← Q_j(α, z, x_r; θ_j), α ~ U(0, 1)
                              # Uniform sampling of quantile level
6:     if j ∈ I_D then                ▷ Gumbel-Max sampling
7:         G_1, G_2, ..., G_{C_j} ~_{i.i.d.} Gumbel(0, 1)
8:         x̂_j ← argmax_{c=1,···,C_j} {log π_jc(z, x_r; θ_j) + G_c}
9: x̃ ← x ⊙ r + x̂ ⊙ (1 − r)
```

2.2 Imputation with Uncertainty

Indicator method The definition of the un-masked and re-masked observations can be seen as the *zero imputation* technique (Mattei and Frellsen 2019; Nazábal et al. 2020; Ipsen, Mattei, and Frellsen 2021). However, when we input $\mathbf{x_m}$ or $\mathbf{x_r}$ in the neural networks, the masked (missing) values cannot be discriminated from actual zero values. Therefore, to distinguish between actual 0 values and masked (missing) values, we further input additional information as follows:

$$\mathbf{x_m} \leftarrow \big[\mathbf{x_m} : 1 - (\mathbf{m} \wedge \mathbf{r})\big], \quad \mathbf{x_r} \leftarrow \big[\mathbf{x_r} : 1 - \mathbf{r}\big],$$

where [:] is concatenate operation. Note that this approach can be seen as the indicator method (or missing indicator), which is shown to be beneficial in downstream task (Stene and Miettinen 1987; Barata et al. 2019; Morvan and Varoquaux 2025).

Incorporating categorical variables Thus far, we have focused on the scenario in which the observations consist solely of continuous random variables, in order to streamline the explanation of our proposed model. However, incorporating categorical variables into the learning process is straightforward. To handle categorical columns, we add a cross-entropy loss (i.e., a classification loss) to the reconstruction loss term in (3). The corresponding reconstruction loss is given by:

$$\sum_{\substack{j \in I_C: \\ \mathbf{m}_j=0, \mathbf{r}_j=1}} \int_0^1 \rho_\alpha\Big((\mathbf{x}_{1-\mathbf{m}})_j - Q_j(\alpha, \mathbf{z}, \mathbf{x_m}; \theta_j)\Big) d\alpha$$

$$- \sum_{\substack{j \in I_D: \\ \mathbf{m}_j=0, \mathbf{r}_j=1}} \sum_{c=1}^{C_j} \mathbb{I}(\mathbf{x}_j = c) \cdot \log(\pi_{jc}(\mathbf{z}, \mathbf{x_m}; \theta_j)),$$

where $\pi_j(\cdot, \cdot; \theta) : \mathbb{R}^d \times \mathbb{R}^p \mapsto \Delta^{C_j-1}$ is a neural network parameterized with θ, where Δ^{C_j-1} is the $(C_j - 1)$-dimensional simplex, and the subscript c refers to the cth element of the output π_j.

Missing Data Imputation Our imputation procedure, outlined in Algorithm 1, addresses the uncertainty of deciding which value to impute for both continuous columns and categorical columns as follows:

1. (continuous) sampling a quantile level α from a continuous uniform distribution (line 6 of Algorithm 1), which serves as the input for the inverse transform sampling method, and
2. (categorical) Gumbel-Max sampling (line 11 of Algorithm 1) (Gumbel 1954).

Although Algorithm 1 is based on single imputation, it's important to note that our imputation algorithm can be readily extended to a multiple imputation approach. This can be achieved by sampling several quantile levels and performing multiple Gumbel-Max samplings.

3 Related Works

Classical imputation methods include statistical heuristics such as mean imputation (Farhangfar, Kurgan, and Pedrycz 2007), k-nearest neighbors imputation (kNNI) (Altman 1992), and missForest (Stekhoven and Bühlmann 2012), as well as iterative approaches such as Expectation-Maximization (Dempster, Laird, and Rubin 1977), MICE (van Buuren and Groothuis-Oudshoorn 2011), and MIRACLE (Kyono et al. 2021).

In the realm of deep generative model-based imputation methods, they have been proposed to learn joint or conditional distributions for imputation. GAIN (Yoon, Jordon, and van der Schaar 2018) employs adversarial training, while VAE-based methods such as VAEAC (Ivanov, Figurnov, and Vetrov 2019) and MIWAE (Mattei and Frellsen 2019), which is similar to our approach, leverage variational inference with random masking. Additionally, methods such as ReMasker (Du, Melis, and Wang 2024) and PMAE (Kim, Lee, and Park 2024) employ the MAE framework for imputation tasks. However, these methods formulate reconstruction loss using MSE, which fails to properly capture the inherent uncertainty in the imputed values (Schulz et al. 2024).

4 Experiments

4.1 Overview

We empirically evaluate U-VAE under diverse missing data mechanisms across a range of benchmark datasets. Our experiments[†] are designed to address the following key research questions[‡]:

RQ1. *Does U-VAE achieve state-of-the-art performance in single imputation tasks?*

RQ2. *Can U-VAE support statistically valid multiple imputation by capturing uncertainty in the imputed values?*

RQ3. *How robust is U-VAE to varying missingness rates and patterns in sensitivity analyses?*

[†]All experiments are conducted on an NVIDIA RTX A6000 GPU using implementations in `PyTorch` and `scikit-learn`.

[‡]We conduct an additional experiment to evaluate the congeniality of U-VAE in Subsection 4.3.

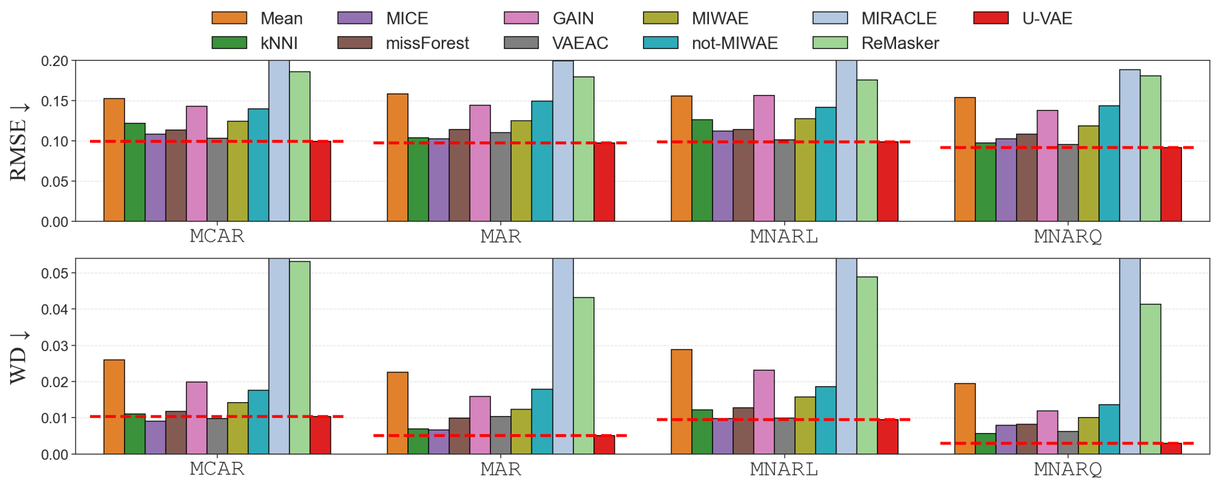

Figure 2: **Imputation fidelity** at 0.3 missingness rate. The corresponding missingness mechanism is indicated below the figure. The red dashed line indicates the performance of U-VAE under each missingness mechanism. The means across 11 datasets and 5 random seeds are reported. ↓ denotes that lower is better.

Datasets Similar to several recent studies (Ipsen, Mattei, and Frellsen 2021; Mattei and Frellsen 2019; Nazábal et al. 2020; Muzellec et al. 2020; Zhao et al. 2023; Jarrett et al. 2022), we evaluate our model on 11 real-world tabular datasets from UCI and Kaggle repositories[§], covering a wide range of sizes and variable types. Each dataset is randomly split into 80% training and 20% testing sets. Detailed statistics are summarized in the Appendix.

In addition, following prior work (Muzellec et al. 2020; Jarrett et al. 2022; Zhao et al. 2023), we simulate missingness using four common mechanisms: `MCAR`, `MAR`, `MNARL`, and `MNARQ`. The detailed procedures for generating missing values under each mechanism are described in the Appendix.

Baselines In this experiment, we investigate our proposed model, U-VAE, and 10 imputation methods.

- *Single imputation*: Mean (Little and Rubin 2002), kNNI (Troyanskaya et al. 2001), missForest (Stekhoven and Bühlmann 2012), MIRACLE (Kyono et al. 2021), and ReMasker (Du, Melis, and Wang 2024).
- *Multiple imputation*: MICE (van Buuren and Groothuis-Oudshoorn 2011), and GAIN (Yoon, Jordon, and van der Schaar 2018), VAEAC (Ivanov, Figurnov, and Vetrov 2019), MIWAE (Mattei and Frellsen 2019), not-MIWAE (Ipsen, Mattei, and Frellsen 2021).

4.2 Evaluation Metrics

RQ1 Similar to recent works (Mattei and Frellsen 2019; Nazábal et al. 2020; Muzellec et al. 2020; Ipsen, Mattei, and Frellsen 2021; Jarrett et al. 2022; Zhao et al. 2023; Du, Melis, and Wang 2024), we evaluate the quality of imputation from two complementary perspectives: imputation fidelity and imputation utility.

[§]https://archive.ics.uci.edu/, https://www.kaggle.com/datasets/

For imputation fidelity, we evaluate how well the imputed values recover the ground-truth data using two metrics: root mean square error (RMSE), which measures point-wise accuracy, and Wasserstein distance (WD), which assesses distributional similarity to the true data.

For imputation utility, we assess how well the imputed data support downstream tasks. Specifically, we use symmetric mean absolute percentage error (SMAPE) for regression, accuracy (Acc) for classification, and feature selection consistency (Feat), which is adopted from (Hansen et al. 2023). A detailed description of these metrics and evaluation protocols is provided in the Appendix.

RQ2 We assess the effectiveness of multiple imputation using interval-based inference for the population mean, as originally proposed by (Rubin and Schenker 1986). Since the true population mean is generally unobservable, we instead use the column-wise mean of the complete dataset as a proxy for the parameter of interest (Lee and Huber 2021; Zhang et al. 2023). Specifically, we report the raw bias (Bias), the empirical coverage rate (Coverage), and the average width of the confidence interval (Width) computed from 100 multiple imputations across all continuous columns and datasets (van Buuren 2012). A nominal 95% confidence level is used to evaluate Coverage and Width. The detailed evaluation procedure for multiple imputation is provided in the Appendix.

Remark 2. *To account for imputation uncertainty, we follow (Rubin and Schenker 1986; van Buuren 2012) and apply Rubin's rule to aggregate estimates and variances from multiple imputations.*

4.3 Results

RQ1: Imputation fidelity As shown in Figure 2, U-VAE outperforms all baselines by achieving the lowest RMSE and Wasserstein Distance (WD) across all missingness mechanisms. In total, U-VAE ranks first on 32 out of 40 fidelity

model	MCAR SMAPE ↓	MCAR Acc ↑	MCAR Feat ↑	MAR SMAPE ↓	MAR Acc ↑	MAR Feat ↑	MNARL SMAPE ↓	MNARL Acc ↑	MNARL Feat ↑	MNARQ SMAPE ↓	MNARQ Acc ↑	MNARQ Feat ↑
Mean	$.254_{\pm.022}$	$.680_{\pm.032}$	$.689_{\pm.025}$	$.226_{\pm.019}$	$.669_{\pm.032}$	$.621_{\pm.028}$	$.256_{\pm.021}$	$.653_{\pm.034}$	$.649_{\pm.026}$	$.225_{\pm.018}$	$.718_{\pm.030}$	$.696_{\pm.026}$
kNNI	$.223_{\pm.021}$	$.681_{\pm.031}$	$.702_{\pm.023}$	$.190_{\pm.017}$	$.678_{\pm.031}$	$.716_{\pm.024}$	$.222_{\pm.020}$	$.655_{\pm.031}$	$.649_{\pm.026}$	$.190_{\pm.017}$	$.730_{\pm.030}$	$.762_{\pm.021}$
MICE	$.220_{\pm.021}$	$.684_{\pm.032}$	$.775_{\pm.022}$	$.193_{\pm.018}$	$.678_{\pm.030}$	**$.756$**$_{\pm.021}$	$.220_{\pm.021}$	$.661_{\pm.032}$	$.713_{\pm.026}$	$.194_{\pm.017}$	$.725_{\pm.030}$	**$.802$**$_{\pm.019}$
missForest	$.228_{\pm.020}$	$.683_{\pm.032}$	$.724_{\pm.025}$	$.204_{\pm.017}$	$.674_{\pm.032}$	$.678_{\pm.024}$	$.228_{\pm.020}$	$.657_{\pm.033}$	$.662_{\pm.025}$	$.207_{\pm.018}$	$.726_{\pm.030}$	$.735_{\pm.021}$
GAIN	$.229_{\pm.020}$	$.714_{\pm.029}$	$.701_{\pm.021}$	$.209_{\pm.018}$	$.719_{\pm.028}$	$.687_{\pm.023}$	$.251_{\pm.022}$	$.711_{\pm.028}$	$.675_{\pm.020}$	$.216_{\pm.019}$	$.748_{\pm.027}$	$.730_{\pm.018}$
VAEAC	$.216_{\pm.020}$	$.771_{\pm.025}$	$.763_{\pm.020}$	$.200_{\pm.017}$	**$.765$**$_{\pm.025}$	$.750_{\pm.018}$	$.217_{\pm.019}$	**$.765$**$_{\pm.025}$	$.738_{\pm.020}$	$.198_{\pm.017}$	$.780_{\pm.024}$	$.791_{\pm.018}$
MIWAE	$.219_{\pm.019}$	$.688_{\pm.026}$	$.656_{\pm.023}$	$.205_{\pm.018}$	$.656_{\pm.022}$	$.664_{\pm.022}$	$.221_{\pm.019}$	$.648_{\pm.023}$	$.623_{\pm.025}$	$.204_{\pm.018}$	$.730_{\pm.026}$	$.705_{\pm.022}$
not-MIWAE	$.233_{\pm.020}$	$.674_{\pm.033}$	$.699_{\pm.024}$	$.218_{\pm.019}$	$.654_{\pm.031}$	$.658_{\pm.024}$	$.238_{\pm.021}$	$.652_{\pm.034}$	$.636_{\pm.027}$	$.217_{\pm.019}$	$.715_{\pm.030}$	$.709_{\pm.021}$
MIRACLE	$.235_{\pm.019}$	$.672_{\pm.032}$	$.706_{\pm.023}$	$.210_{\pm.016}$	$.673_{\pm.032}$	$.660_{\pm.029}$	$.236_{\pm.019}$	$.655_{\pm.033}$	$.636_{\pm.030}$	$.211_{\pm.016}$	$.720_{\pm.031}$	$.730_{\pm.025}$
ReMasker	$.241_{\pm.021}$	$.676_{\pm.034}$	$.688_{\pm.030}$	$.220_{\pm.019}$	$.663_{\pm.033}$	$.662_{\pm.029}$	$.246_{\pm.021}$	$.655_{\pm.035}$	$.629_{\pm.030}$	$.227_{\pm.019}$	$.716_{\pm.031}$	$.749_{\pm.026}$
U-VAE	**$.209$**$_{\pm.021}$	**$.778$**$_{\pm.029}$	**$.786$**$_{\pm.022}$	**$.188$**$_{\pm.019}$	$.761_{\pm.029}$	$.747_{\pm.025}$	**$.208$**$_{\pm.021}$	$.758_{\pm.030}$	**$.743$**$_{\pm.024}$	**$.187$**$_{\pm.019}$	**$.805$**$_{\pm.028}$	$.798_{\pm.021}$

Table 2: **Imputation utility** at 0.3 missingness. The means and their standard errors across 11 datasets and 5 random seeds are reported. ↑ (↓) denotes that higher (lower) is better. The best result is **red bolded**, and the second best is blue underlined.

model	MCAR Bias↓	MCAR Coverage	MCAR Width↓	MAR Bias↓	MAR Coverage	MAR Width↓	MNARL Bias↓	MNARL Coverage	MNARL Width↓	MNARQ Bias↓	MNARQ Coverage	MNARQ Width↓
MICE	.018(5.8)	$.642_{\pm.039}$	**$.038$**$_{\pm.003}$	.012(3.5)	$.755_{\pm.028}$	$.038_{\pm.003}$	.020(6.1)	$.595_{\pm.044}$	$.038_{\pm.003}$	.011(3.6)	$.767_{\pm.027}$	$.038_{\pm.003}$
GAIN	.044(11.2)	$.314_{\pm.037}$	**$.038$**$_{\pm.003}$	.032(7.8)	$.536_{\pm.029}$	**$.037$**$_{\pm.003}$	.053(13.8)	$.269_{\pm.037}$	**$.037$**$_{\pm.003}$	.020(5.9)	$.641_{\pm.030}$	**$.037$**$_{\pm.003}$
VAEAC	.017(5.3)	$.653_{\pm.036}$	**$.038$**$_{\pm.003}$	.015(4.1)	$.696_{\pm.025}$	$.038_{\pm.003}$	.018(5.3)	$.670_{\pm.035}$	$.038_{\pm.003}$	.011(3.6)	$.756_{\pm.026}$	$.038_{\pm.003}$
MIWAE	.012(3.8)	$.769_{\pm.033}$	$.043_{\pm.003}$	.009(2.5)	$.857_{\pm.025}$	$.042_{\pm.003}$	.014(4.2)	$.750_{\pm.033}$	$.044_{\pm.003}$	.008(2.6)	$.840_{\pm.023}$	$.040_{\pm.003}$
not-MIWAE	.015(4.2)	$.701_{\pm.038}$	$.044_{\pm.004}$	.014(3.6)	$.772_{\pm.027}$	$.042_{\pm.004}$	.017(4.7)	$.646_{\pm.040}$	$.044_{\pm.004}$	.011(3.1)	$.781_{\pm.026}$	$.041_{\pm.004}$
U-VAE	**.009**(2.6)	**$.867$**$_{\pm.024}$	**$.038$**$_{\pm.003}$	**.008**(2.0)	**$.889$**$_{\pm.018}$	$.038_{\pm.003}$	**.010**(2.9)	**$.839$**$_{\pm.024}$	$.038_{\pm.003}$	**.005**(1.4)	**$.908$**$_{\pm.014}$	$.038_{\pm.003}$

Table 3: **Multiple imputation performance** at 0.3 missingness rate. The means and their standard errors across 11 datasets and 5 random seeds are reported. The values in parentheses next to Bias represent the percent bias. ↓ denotes lower is better. Coverage close to 0.95 indicates better performance. The best result is **red bolded**, and the second best is blue underlined.

scores (2 metrics × 4 mechanisms × 5 missing rates), as detailed in the Appendix.

In particular, U-VAE consistently achieves the lowest RMSE and WD under MNARL and MNARQ, demonstrating strong robustness to non-ignorable missingness patterns. This performance gap becomes more pronounced under these challenging mechanisms, where many baselines suffer from substantial errors and distributional divergence.

The superior performance of U-VAE in these settings underscores its strength in modeling conditional distributions even when missingness depends on unobserved or partially observed data. This is attributed to its quantile-based objective and CRPS loss, which encourage learning the full predictive distribution rather than point estimates. Additional results across different missingness rates are provided in the Appendix.

RQ1: Imputation utility As shown in Table 2, U-VAE consistently achieves strong performance across all utility evaluation metrics under all four missingness settings. In total, U-VAE ranks first on 43 out of 60 utility scores (3 metrics × 4 mechanisms × 5 missing rates), as detailed in the Appendix. Notably, it ranks within the top-2 in 56 out of 60 cases, consistently demonstrating superior accuracy and feature selection consistency across diverse imputation settings.

Unlike fidelity metrics that directly assess the closeness of imputed values to the ground truth, utility metrics evaluate how well the imputed data supports subsequent modeling tasks. These results highlight U-VAE's effectiveness in producing imputations that not only match the original data distribution, but also retain the predictive and structural integrity necessary for downstream analysis. Additional results across missingness rates and per-dataset utility scores are provided in the Appendix.

RQ2 As shown in the Table 3, U-VAE consistently achieves the best overall performance across all mechanisms, with the lowest bias (all below 3%), highest coverage (84%–91%), and the narrowest or second-narrowest confidence intervals. In fact, U-VAE ranks first in 46 out of 60 inferential validity scores (3 metrics × 4 mechanisms × 5 missing rates), as detailed in the Appendix.

Specifically, U-VAE shows the lowest bias in all settings (e.g., 1.4% under MNARQ), clearly satisfying the 5% criterion, while attaining the highest coverage across the board, indicating reliable uncertainty quantification. The interval widths of U-VAE are among the most compact, avoiding overly conservative inference. In contrast, models such as GAIN and MICE exhibit either inflated bias (up to 13.8%) or low coverage (as low as 27%), undermining their inferential reliability despite having narrow intervals. MIWAE, while achieving acceptable bias and good coverage, suffers from relatively wider intervals, which may lead to conservative conclusions.

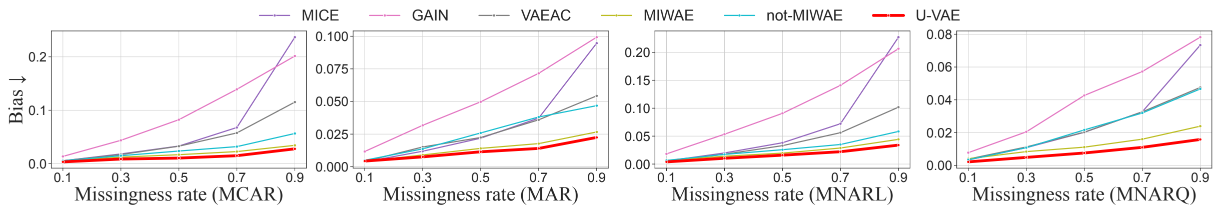

Figure 3: **Sensitivity analysis** for missingness rates. Imputation performance comparison of various imputation models under different missing data mechanisms (`MCAR`, `MAR`, `MNARL`, `MNARQ`) and missingness rates (0.1, 0.3, 0.5, 0.7, 0.9) using Bias. ↓ denotes that lower is better.

dataset	`anuran`		`default`	
model	$\|\mathbf{w}-\widetilde{\mathbf{w}}\|_1 \downarrow$	$\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2 \downarrow$	$\|\mathbf{w}-\widetilde{\mathbf{w}}\|_1 \downarrow$	$\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2 \downarrow$
Mean	$.354_{\pm .006}$	$.270_{\pm .010}$	<u>.283</u>$_{\pm .008}$	<u>.166</u>$_{\pm .009}$
kNNI	$.404_{\pm .012}$	$.356_{\pm .020}$	$.302_{\pm .011}$	$.198_{\pm .016}$
MICE	$.675_{\pm .027}$	$1.232_{\pm .086}$	$.785_{\pm .062}$	$3.535_{\pm .656}$
missForest	$.357_{\pm .007}$	<u>.268</u>$_{\pm .010}$	$.286_{\pm .009}$	$.167_{\pm .009}$
GAIN	$.521_{\pm .017}$	$.658_{\pm .045}$	$.426_{\pm .026}$	$.549_{\pm .071}$
VAEAC	<u>.340</u>$_{\pm .019}$	$.270_{\pm .026}$	$.382_{\pm .025}$	$.394_{\pm .050}$
MIWAE	$.597_{\pm .012}$	$.759_{\pm .022}$	$.439_{\pm .013}$	$.438_{\pm .026}$
not-MIWAE	$.360_{\pm .007}$	$.270_{\pm .010}$	$.308_{\pm .009}$	$.190_{\pm .010}$
MIRACLE	$.641_{\pm .026}$	$1.032_{\pm .078}$	$.544_{\pm .052}$	$2.436_{\pm .707}$
ReMasker	$.392_{\pm .010}$	$.327_{\pm .016}$	$.308_{\pm .012}$	$.215_{\pm .021}$
U-VAE	**.338**$_{\pm .007}$	**.244**$_{\pm .011}$	**.270**$_{\pm .008}$	**.158**$_{\pm .009}$

Table 4: **Congeniality** at 0.3 missingness rate under `MAR`. The means and their standard errors across 5 random seeds are reported. ↓ denotes that lower is better. The best result is **red bolded**, and the second best is <u>blue underlined</u>.

These results demonstrate that U-VAE not only produces accurate point estimates but also enables valid and efficient statistical inference, balancing low bias, proper coverage, and a reasonable interval.

RQ3 As shown in Figure 3, U-VAE demonstrates strong robustness across increasing missingness rates under all mechanisms. Specifically, while other methods, such as not-MIWAE, GAIN, and MICE, exhibit rapidly growing bias as the missingness rate increases, U-VAE maintains consistently low bias. This suggests that U-VAE effectively handles high missingness levels without substantial degradation in inference validity. The stable behavior of U-VAE, especially in Bias, highlights its reliability in uncertainty-aware multiple imputation even under challenging scenarios. Additional results under various metrics used in this experiment are provided in the Appendix.

Congeniality The congeniality of an imputation model refers to its ability to preserve the relationship between features and the target variable after imputation (Meng 1994; Burgess et al. 2013; Deng et al. 2016; Yoon, Jordon, and van der Schaar 2018). To evaluate this, we use two real-world datasets, `anuran` and `default`. For each dataset, we first estimate the logistic regression coefficients $\mathbf{w}$ from the fully observed data. We then compute $\widetilde{\mathbf{w}}$ by imputing the missing entries in an incomplete version of the dataset and fitting the same model. To quantify the discrepancy between $\mathbf{w}$ and $\widetilde{\mathbf{w}}$, we report (1) the mean L_1 distance $\|\mathbf{w}-\widetilde{\mathbf{w}}\|_1$, which captures the overall deviation in coefficient values, and (2) the mean squared L_2 distance $\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2^2$, which measures the magnitude of distortion in parameter estimates.

As shown in Table 4, U-VAE outperforms all competing baselines by achieving the lowest values in both coefficient deviation metrics on both datasets. For instance, on the `anuran` dataset, U-VAE yields 0.338 and 0.244 in absolute and squared deviations, respectively, outperforming all other baselines. Similarly, on the `default` dataset, U-VAE reports the lowest deviations of 0.270 and 0.158. These results highlight U-VAE's superior ability to retain the original feature-target relationship post-imputation, ensuring high congeniality.

5 Conclusion and Limitations

In this paper, we introduce U-VAE, a novel imputation framework, which is capable of parameterizing the conditional likelihood of missing entries non-parametrically learned. By leveraging the stochastic re-masking and unmasking strategy, we circumvent the limitation that the conditional distributions estimable from observed incomplete datasets are inherently restricted. Moreover, we theoretically justify that the discrepancy between the underlying conditional distribution of the missing entries and our imputer is upper-bounded, and empirically demonstrate that U-VAE outperforms existing baselines in terms of fidelity, utility in single imputation, and inferential validity in multiple imputation. In particular, U-VAE achieves correct coverage rates with low bias and narrow intervals, satisfying principled criteria for multiple imputation.

However, our approach has several limitations. First, our theoretical results are based on the unknown conditional distribution $p^*(\mathbf{x}_{1-\mathbf{r}} \mid \mathbf{x}_\mathbf{r})$. While this provides a theoretical guarantee, we acknowledge that the applicability of our results is limited in practice due to the inaccessibility of the true conditional distribution. Second, we need to explore alternative parameterizations that satisfy Assumption 1 while allowing efficient computation of the CRPS reconstruction loss. Lastly, if the posterior distribution—i.e., the parameter ϕ—does not satisfy Assumption 2, the validity of our imputer cannot be guaranteed. For future work, we plan to incorporate temporal dependencies into the imputation process by extending U-VAE for time-series data.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) (No. RS-2022-NR068754, No. RS-2025-23525436). The authors acknowledge the Urban Big data and AI Institute of the University of Seoul supercomputing resources (http://ubai.uos.ac.kr) made available for conducting the research reported in this paper.

References

Altman, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3): 175–185.

An, S.; and Jeon, J.-J. 2023. Distributional Learning of Variational AutoEncoder: Application to Synthetic Data Generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

An, S.; Woo, G.; Lim, J.; Kim, C.; Hong, S.; and Jeon, J.-J. 2024. Masked Language Modeling Becomes Conditional Density Estimation for Tabular Data Synthesis. In *AAAI Conference on Artificial Intelligence*.

Barata, A. P.; Takes, F. W.; van den Herik, H. J.; and Veenman, C. J. 2019. Imputation Methods Outperform Missing-Indicator for Data Missing Completely at Random. *2019 International Conference on Data Mining Workshops (ICDMW)*, 407–414.

Barnard, J.; and Meng, X. L. 1999. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Stat Methods Med Res*, 8(1): 17–36.

Burgess, S.; White, I. R.; Resche-Rigon, M.; and Wood, A. M. 2013. Combining multiple imputation and meta-analysis with individual participant data. *Statistics in Medicine*, 32: 4499 – 4514.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM - algorithm plus discussions on the paper.

Deng, Y.; Chang, C.; Ido, M. S.; and Long, Q. 2016. Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Scientific Reports*, 6.

Du, T.; Melis, L.; and Wang, T. 2024. ReMasker: Imputing Tabular Data with Masked Autoencoding. In *The Twelfth International Conference on Learning Representations*.

Farhangfar, A.; Kurgan, L. A.; and Pedrycz, W. 2007. A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(5): 692–709.

Gasthaus, J.; Benidis, K.; Wang, Y.; Rangapuram, S. S.; Salinas, D.; Flunkert, V.; and Januschowski, T. 2019. Probabilistic forecasting with spline quantile function RNNs. In *The 22nd international conference on artificial intelligence and statistics*, 1901–1910. PMLR.

Gibson, M. G.; Little, R. J. A.; and Rubin, D. B. 1989. Statistical Analysis with Missing Data. *The Statistician*, 38: 82.

Gneiting, T.; and Raftery, A. E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102: 359 – 378.

Gumbel, E. J. 1954. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office.

Hansen, L.; Seedat, N.; van der Schaar, M.; and Petrovic, A. 2023. Reimagining Synthetic Tabular Data Generation through Data-Centric AI: A Comprehensive Benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Ipsen, N. B.; Mattei, P.-A.; and Frellsen, J. 2021. not-{MIWAE}: Deep Generative Modelling with Missing not at Random Data. In *International Conference on Learning Representations*.

Ivanov, O.; Figurnov, M.; and Vetrov, D. 2019. Variational Autoencoder with Arbitrary Conditioning. In *International Conference on Learning Representations*.

Jarrett, D.; Cebere, B. C.; Liu, T.; Curth, A.; and van der Schaar, M. 2022. HyperImpute: Generalized Iterative Imputation with Automatic Model Selection. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 9916–9937. PMLR.

Khan, S. I.; and Hoque, A. S. M. L. 2020. SICE: an improved missing data imputation technique. *Journal of Big Data*, 7.

Kim, J.; Lee, K.; and Park, T. 2024. To Predict or Not To Predict? Proportionally Masked Autoencoders for Tabular Data Imputation. arXiv:2412.19152.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kyono, T.; Zhang, Y.; Bellot, A.; and van der Schaar, M. 2021. MIRACLE: Causally-Aware Imputation via Learning Missing Data Mechanisms. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Lee, J. H.; and Huber, J. C. 2021. Evaluation of Multiple Imputation with Large Proportions of Missing Data: How Much Is Too Much? *Iranian Journal of Public Health*, 50: 1372 – 1380.

Little, R.; and Rubin, D. 2002. *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley. ISBN 9780471183860.

Little, R. J. A. 1988. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6: 287–296.

Manski, C. F. 2003. Partial Identification of Probability Distributions.

Matheson, J. E.; and Winkler, R. L. 1976. Scoring Rules for Continuous Probability Distributions. *Management Science*, 22: 1087–1096.

Mattei, P.-A.; and Frellsen, J. 2019. MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. In *International Conference on Machine Learning*.

Meng, X.-L. 1994. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9: 538–558.

Morvan, M. L.; and Varoquaux, G. 2025. Imputation for prediction: beware of diminishing returns. In *The Thirteenth International Conference on Learning Representations*.

Murray, J. S. 2018. Multiple imputation: a review of practical and theoretical findings.

Muzellec, B.; Josse, J.; Boyer, C.; and Cuturi, M. 2020. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, 7130–7140. PMLR.

Nazábal, A.; Olmos, P. M.; Ghahramani, Z.; and Valera, I. 2020. Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, 107: 107501.

Rubin, D. B. 1975. Inference and Missing Data. *Psychometrika*, 1975: 19.

Rubin, D. B. 1989. Multiple imputation for nonresponse in surveys.

Rubin, D. B. 1996. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91: 473–489.

Rubin, D. B.; and Schenker, N. 1986. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association*, 81: 366–374.

Schulz, A.; Vetter, J.; Gao, R.; Morales, D.; Lobato-Ríos, V.; Ramdya, P.; Gonçalves, P. J.; and Macke, J. H. 2024. Modeling conditional distributions of neural and behavioral data with masked variational autoencoders. *bioRxiv*.

Stekhoven, D. J.; and Bühlmann, P. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1): 112–118.

Stene, J. K.; and Miettinen, O. S. 1987. Theoretical Epidemiology: Principles of Occurrence Research in Medicine. *Biometrics*, 43: 482.

Troyanskaya, O. G.; Cantor, M. N.; Sherlock, G.; Brown, P. O.; Hastie, T. J.; Tibshirani, R.; Botstein, D.; and Altman, R. B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17 6: 520–5.

van Buuren, S. 2012. Flexible Imputation of Missing Data.

van Buuren, S.; and Groothuis-Oudshoorn, K. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3): 1–67.

Wen, Y.; Wang, Y.; Yi, K.; Ke, J.; and Shen, Y. 2024. Diffimpute: Tabular Data Imputation with Denoising Diffusion Probabilistic Model . In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. Los Alamitos, CA, USA: IEEE Computer Society.

Yoon, J.; Jordon, J.; and van der Schaar, M. 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 5689–5698. PMLR.

Zhang, N.; Liu, C.; Steiner, S. J.; Colletti, R. B.; Baldassano, R. N.; Chen, S.; Cohen, S.; Kappelman, M. D.; Saeed, S. A.; Conklin, L. S.; Strauss, R.; Volger, S.; King, E. C.; and Lo, K. H. 2023. Using multiple imputation of real-world data to estimate clinical remission in pediatric inflammatory bowel disease. *Journal of Comparative Effectiveness Research*, 12.

Zhao, H.; Sun, K.; Dezfouli, A.; and Bonilla, E. V. 2023. Transformed Distribution Matching for Missing Value Imputation. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 42159–42186. PMLR.