To Predict or Not to Predict? Proportionally Masked Autoencoders for Tabular Data Imputation

Jungkyu Kim, Kibok Lee*, Taeyoung Park*

Department of Statistics and Data Science, Yonsei University, South Korea
{kim.normal, kibok, tpark}@yonsei.ac.kr

Abstract

Masked autoencoders (MAEs) have recently demonstrated effectiveness in tabular data imputation. However, due to the inherent heterogeneity of tabular data, the uniform random masking strategy commonly used in MAEs can disrupt the distribution of missingness, leading to suboptimal performance. To address this, we propose a proportional masking strategy for MAEs. Specifically, we first compute the statistics of missingness based on the observed proportions in the dataset, and then generate masks that align with these statistics, ensuring that the distribution of missingness is preserved after masking. Furthermore, we argue that simple MLP-based token mixing offers competitive or often superior performance compared to attention mechanisms while being more computationally efficient, especially in the tabular domain with the inherent heterogeneity. Experimental results validate the effectiveness of the proposed proportional masking strategy across various missing data patterns in tabular datasets.

Introduction

Tabular data often contain missing values in real-world scenarios, posing significant challenges for the deployment of machine learning algorithms (Donders et al. 2006). Inspired by the recent success of masked autoencoders (MAEs) in representation learning across domains such as computer vision (CV) (He et al. 2022) and natural language processing (NLP) (Devlin et al. 2018), Du, Melis, and Wang (2024) proposed adapting MAEs for tabular data imputation. However, we argue that naively applying the uniform random masking strategy from MAEs to tabular data results in suboptimal performance due to the intrinsic heterogeneity of tabular data. Unlike images or word tokens, which are relatively homogeneous and semantically invariant to spatial shifts, tabular data are inherently heterogeneous. That is, each column contains distinct information, making spatial shifts meaningless. Such heterogeneity also extends to the distribution of missing values, which can vary across columns. Thus, uniform random masking can unintentionally disrupt these distributions by omitting critical variables that are essential for predicting others (Wilms et al. 2021), thereby leading to suboptimal performance (Wu et al. 2024). These challenges underscore the need for a masking strategy specifically designed to account for the heterogeneity of tabular data.

Regarding the architecture of MAEs, while Transformers (Vaswani et al. 2017) have shown strong performance, their self-attention mechanisms typically focus globally on all columns, which can hinder their ability to effectively capture the local group interactions that are often characteristics of tabular data (Yan et al. 2023). This limitation makes Transformers less suited for capturing the complex relationships between columns, particularly in the presence of missing values. In contrast, we argue that MLP-Mixers (Tolstikhin et al. 2021) are better suited for the tabular domain, as their fully-connected layers equipped with activation functions are inherently more capable of capturing such multiple group interactions.

We also address the challenge of evaluating imputation performance using a single metric. Tabular data typically consists of discrete categorical variables and continuous numerical variables, each requiring distinct evaluation metrics. However, recent studies (Du, Melis, and Wang 2024) often rely solely on the root mean square error (RMSE) across all column types, which fails to adequately assess categorical variables. Specifically, categorical variables are encoded as uniformly distributed values between 0 and 1, despite lacking inherent relative similarities. For example, if a categorical variable is encoded as 0, 0.5, and 1, predicting 0 as 0.5 or 1 is equally incorrect, yet RMSE penalizes the latter more, leading to skewed evaluations. This underscores the need for a more intuitive and appropriate evaluation metric to accurately assess imputation performance in tabular data. While accuracy is commonly used for categorical variables, it is not directly comparable to RMSE because it is not normalized. To overcome this, we propose a unified evaluation metric that combines accuracy for categorical variables and the coefficient of determination (R^2) for numerical variables.

The main contributions of our work are as follows:

- We propose an MAE with a novel masking strategy based on observed proportions, coined **P**roportionally **M**asked **A**uto**E**ncoder (**PMAE**), specifically designed to address the challenge of tabular data imputation.
- We reveal the effectiveness of MLP-Mixers over Transformers in tabular data imputation tasks through extensive experimental studies.

*Co-corresponding authors

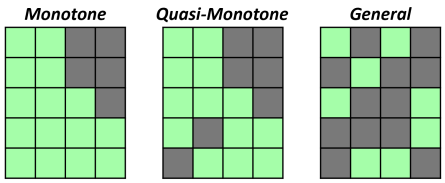


Figure 1: The simplest missing pattern is *Monotone*, where some columns are fully observed, common in longitudinal studies. As the missing pattern becomes *General* (with all columns prone to missing with varying ratios) imputation becomes more challenging. Practitioners need strategies to address these patterns, which are currently under explored.

- We demonstrate the efficacy of the proposed **PMAE** across various missing data distributions and patterns that practitioners commonly encounter. Our experimental results indicate that **PMAE** outperforms the state-of-the-art method (Du, Melis, and Wang 2024) by up to 34.1% in the most challenging *General* pattern (see Figure 1 for an illustration of missing patterns).

Related Work

Imputation Methods Simple imputation (Hawthorne and Elliott 2005) replaces missing data using summary statistics or KNN-based averages (Troyanskaya et al. 2001). Iterative approaches, such as Expectation-Maximization (Dempster, Laird, and Rubin 1977), MICE (Shah et al. 2014), MissForest (Stekhoven and Bühlmann 2012), MIRACLE (Kyono et al. 2021), and Hyperimpute (Jarrett et al. 2022), iteratively refine estimates conditioned on observed data, relying on distributional assumptions. Optimal transport-based approaches, such as TDM (Zhao et al. 2023), impute data in the latent space by matching similar incomplete data batches but may fail to handle categorical data effectively. Generative methods, such as GAIN (Yoon, Jordon, and Schaar 2018) and MIWAE (Mattei and Frellsen 2019), require fully observed data for initialization, potentially introducing bias. Graph-based approaches, such as IGRM (Zhong, Gui, and Ye 2023), impute data by modeling sample-wise relationship but do not scale effectively with large sample sizes. ReMasker (Du, Melis, and Wang 2024), which is closely related to our approach, adapts MAE for imputation tasks but applies uniform masking across all columns. This strategy may be less effective for handling complex missing patterns, such as *non-monotone* missing patterns (Sun et al. 2018).

Propensity Score Weighting Approaches Our masking function design is inspired by the inverse propensity score weighting method, originally introduced in causal inference (Rosenbaum and Rubin 1983) and later adapted for addressing missing data problems. This approach aims to produce unbiased estimates by using only observed data, assigning higher weights to underrepresented samples (Seaman and White 2013). However, inaccurate estimation of the propensity score can result in high variance and increased generalization errors (Guo et al. 2021; Li et al. 2023). While Li et al. (2022) proposed a stabilized approach to mitigate these issues, it requires the joint learning of separate models and extensive parameter tuning, which complicates the optimization process.

Deep Learning Architectures for Tabular Data Transformer-based architectures have been widely studied for tabular domain (Huang et al. 2020; Gorishniy et al. 2021; Somepalli et al. 2021; Zhang et al. 2023). However, as Yan et al. (2023) noted, interactions in tabular data often exist within discrete groups, making soft combinations in self-attention less efficient. Motivated by the effectiveness of the approach proposed by Tolstikhin et al. (2021), we investigate an alternative design to address this issue.

Preliminaries

Incomplete Data Let $\mathbf{x}_i = (x_{i1}, ..., x_{id})^T \in \mathbf{R}^d$ be the i-th tabular data with d columns, sampled from a data distribution $f(\mathbf{x})$. Without loss of generality on the order of columns, let $\mathbf{x}_i = (\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{mis}})_i$ be a decomposition, where $\mathbf{x}_{\text{obs}}$ and $\mathbf{x}_{\text{mis}}$ represent the *observed* and *missing* columns of the data, respectively. For each scalar entry x_{ij}, let $\delta_{ij} := I(x_{ij} \text{ is observed})$ be its associated *missing indicator*. Then, an incomplete data and its corresponding observed mask are defined as follows:

1. The scalar entry of an *incomplete data* is expressed as
$$\tilde{x}_{ij} := x_{ij} \cdot I(\delta_{ij} = 1) + \text{nan} \cdot I(\delta_{ij} = 0). \quad (1)$$
2. The *observed mask* $\mathbf{m}_i \in \mathbf{R}^d$ is the realization of a random missing indicator $\boldsymbol{\delta}_i$.

Missing Mechanism Three types of missing mechanisms commonly occur in the real world (Little and Rubin 2019):

1. **M**issing **C**ompletely **A**t **R**andom (MCAR) occurs when the missingness does not depend on the data, i.e., $\forall \mathbf{x}$, $P(\boldsymbol{\delta}|\mathbf{x}) = P(\boldsymbol{\delta})$.
2. **M**issing **A**t **R**andom (MAR) occurs when the missingness depends only on the observed data, i.e., $P(\boldsymbol{\delta}|\mathbf{x}) = P(\boldsymbol{\delta}|\mathbf{x}_{\text{obs}})$.
3. **M**issing **N**ot **A**t **R**andom (MNAR) occurs when the missingness does not depend only on the observed data, i.e., $P(\boldsymbol{\delta}|\mathbf{x}) \neq P(\boldsymbol{\delta}|\mathbf{x}_{\text{obs}})$.

Missing Patterns We propose a categorization of missing data patterns commonly encountered in practice, as illustrated in Figure 1. When a value x_{ij} is missing for a particular variable j, the pattern can be classified as follows: (i) *Monotone*, when there is a rearrangement of columns such that all subsequent variables x_{ik} for $k > j$ are also missing for the same observation i (Molenberghs et al. 1998), (ii) *Quasi-Monotone*, which is similar to *Monotone*, but allowing a few *exceptions* instead of requiring strict *equality* in the sequence of missing data, and (iii) *General*, when no specific structure exists.

Imputation Task Given an *incomplete dataset* $\mathcal{D} := \{(\tilde{\mathbf{x}}_i, \mathbf{m}_i)\}_{i=1,...,n}$, we aim to obtain plausible estimates for inputs $\tilde{\mathbf{x}}_i$ by learning an imputation function $\hat{f}(\tilde{\mathbf{x}}_i, \mathbf{m}_i; \hat{\Theta})$ that can best approximate the true value of *missing data*: $\hat{x}_{ij} = \hat{f}(\tilde{\mathbf{x}}_i, \mathbf{m}_i; \hat{\Theta}) \cdot I(\delta_{ij} = 0) \approx x_{ij} \cdot I(\delta_{ij} = 0)$.

Motivation

In this section, we formalize the application of MAEs in the tabular domain.

Masking Function Suppose an additional missing mask is applied to the raw data. After this additional masking, the observed mask $\mathbf{m}_i$ can be expressed as

$$\mathbf{m}_i = \mathbf{m}_i^+ + \mathbf{m}_i^-, \tag{2}$$

where $\mathbf{m}_i^+$ is an indicator vector representing the parts that *remain observed* after applying the additional mask, with entries set to 1 for observed parts, and $\mathbf{m}_i^-$ is an indicator vector with entries set to 1 for *additionally masked* parts (see Fig 2). The process of *additional masking* is as follows:

- Draw a uniform random variable $u_{ij} \sim U(0,1)$,
- Given that $\delta_{ij} = 1$, mask according to the following:

$$m_{ij}^- := \begin{cases} 1 & \text{if } u_{ij} < M_j(\cdot), \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $M_j(\cdot) \in \mathbb{R}$ denotes the masking function applied to the column j, indicating the extent to which the additional parts of data is masked.

MAE A masked autoencoder, $h = d \circ g$, is an encoder-decoder architecture designed to predict the entries with the *additional missing mask*, $\mathbf{m}_i^-$, applied to the raw data. It operates on partial input information, $\mathbf{x}_i' = \mathbf{x}_i \odot \mathbf{m}_i^+$, which is provided due to masking.

Tabular MAE Loss While typical MAE in CV/NLP focus on prediction tasks, reconstruction is also important in the tabular domain as they are semantically complex and non-redundant (Du, Melis, and Wang 2024). We formally define the *general loss* for the standard MAE of *tabular data* for a sample i in an arbitrary batch B of the j-th column as:

$$l_{ij} := \frac{((h(\tilde{\mathbf{x}}_i \odot \mathbf{m}_i^+)_j - x_{ij}) \cdot m_{ij})^2}{\sum_{i \in B} m_{ij}}, \tag{4}$$

where the j-th column of $\tilde{\mathbf{x}}_i$ is masked or unmasked based on the value of m_{ij}^- in (3).

Since randomness exists in the encoder input through $\mathbf{m}_i^+$ over all j, we begin our analysis by fixing all but column j. Then, along with $\mathbf{m}_i^+$, we focus on the randomness in the j-th entry of the following expression:

$$\mathbf{m}_i^0 := \mathbf{m}_i^+ - m_{ij}^+ \mathbf{e}_j, \tag{5}$$

where $\mathbf{e}_j$ is the one-hot vector. Let $l_{ij}^0 := \frac{(h(\tilde{\mathbf{x}}_i \odot \mathbf{m}_i^0)_j - x_{ij})^2}{\sum_{i \in B} m_{ij}}$ be a *prediction* loss and $l_{ij}^+ := \frac{(h(\tilde{\mathbf{x}}_i \odot \mathbf{m}_i^+)_j - x_{ij})^2}{\sum_{i \in B} m_{ij}}$ be the *reconstruction* counterpart. Then, (4) can be re-written as:

$$l_{ij} := m_{ij}^- l_{ij}^0 + m_{ij}^+ l_{ij}^+. \tag{6}$$

Note that the MAE will solve for *prediction* task if $m_{ij}^- = 1$ and *reconstruction* task if $m_{ij}^- = 0$. Now, since the randomness only exists for m_{ij}^- (or equivalently for $m_{ij}^+ = 1 - m_{ij}^-$, where observed mask $m_{ij} = 1$, if $\delta_{ij} = 1$), we can compute the expectation of the loss for column j defined in (6):

$$\begin{aligned} \mathbb{E}_{m_{ij}^-}[l_{ij}] &= \mathbb{E}_{m_{ij}^-}[m_{ij}^-] l_{ij}^0 + \mathbb{E}_{m_{ij}^-}[m_{ij}^+] l_{ij}^+ \\ &= M_j(\cdot) l_{ij}^0 + (1 - M_j(\cdot)) l_{ij}^+. \end{aligned} \tag{7}$$

The Impact of Uniform Random Masking Applying uniform random masking at a constant ratio (i.e., $M_j(\cdot) \overset{\forall j}{=} 0.5$) in (7) essentially assigns equal prediction importance to all columns. However, since tabular data are heterogeneous and exhibit complex relationships between columns, such equal weighting may not be ideal. Moreover, this approach may mask fully observed columns just as likely as partially observed ones. If the inadvertently masked entries are critical for predicting values in other columns, this could lead to *omitted variable bias* (Wilms et al. 2021), leading to sub-optimal model performance (Wu et al. 2024). For columns without any missing data, it may be more effective to avoid masking entirely and, consequently, refrain from predicting values that are already fully observed.

Balancing with Inverse Propensities Moreover, naively computing the loss over only the complete cases may lead to biased estimation. Let us focus on the randomness in the missingness of $\mathbf{x}$; recall that the random variable δ_{ij}, indicates whether x_{ij} is observed. Consider the naive empirical loss for the j-th column, $\hat{l}_j^{\text{naive}} := \frac{1}{|B|} \sum_{i \in B} \delta_{ij} \cdot l_{ij}$. Then,

$$\begin{aligned} \mathbb{E}_\delta \left[\hat{l}_j^{\text{naive}} \right] &= \mathbb{E}_\delta \left[\frac{1}{|B|} \sum_{i \in B} l_{ij} - (1 - \delta_{ij}) \cdot l_{ij} \right] \\ &= \hat{l}_j^* - \mathbb{E}_\delta \left[\frac{1}{|B|} \sum_{i \in B} (1 - \delta_{ij}) \cdot l_{ij} \right]. \end{aligned} \tag{8}$$

Since δ is a random variable that may depend on $\mathbf{x}$, the second term in the last equality generally depends on $\mathbf{x}$ unless $\delta \perp\!\!\!\perp \mathbf{x}$. Consequently, $\hat{l}_j^{\text{naive}}$ may be a biased estimator of l_j^* due to the second term in (8).

However, if we have access to the probability of an entry being observed given the realized data, or the *propensity score function*, $\pi_{ij}(\mathbf{x}_i; \boldsymbol{\phi}) := P(\delta_{ij} = 1 | \mathbf{x}_i; \boldsymbol{\phi})$, we can balance the loss, using only the complete cases. Let $\hat{l}_j^{\text{IPS}} := \frac{1}{|B|} \sum_{i \in B} \frac{\delta_{ij} \cdot l_{ij}}{\pi_{ij}(\mathbf{x}_i; \boldsymbol{\phi}^*)}$ be the empirical loss weighted by the inverse propensity score. Then,

$$\begin{aligned} \mathbb{E}_\delta \left[\hat{l}_j^{\text{IPS}} \right] &= \frac{1}{|B|} \sum_{i \in B} \mathbb{E}_\delta \left[\frac{\delta_{ij}}{\pi_{ij}(\mathbf{x}_i; \boldsymbol{\phi}^*)} \right] l_{ij} \\ &= \frac{1}{|B|} \sum_{i \in B} l_{ij} = \hat{l}_j^*. \end{aligned} \tag{9}$$

Intuitively, this approach assigns higher importance to less frequently observed samples to balance the overall loss (Seaman and White 2013; Kim and Shao 2021). This ensures that the model adequately accounts for underrepresented samples, which might otherwise have a minimal impact on the overall loss.

However, the practical application of this approach relies on accurately estimating the propensity score function, $\pi_{ij}(\mathbf{x}_i; \boldsymbol{\phi})$. Incorrect estimation could result in unbounded loss values, compromising model performance (Li et al. 2023). To address this challenge, we propose a more implicit masking strategy guided by the following core principles: (i) the *design of the masking function* determines which samples are assigned higher prediction loss, (ii) masks should

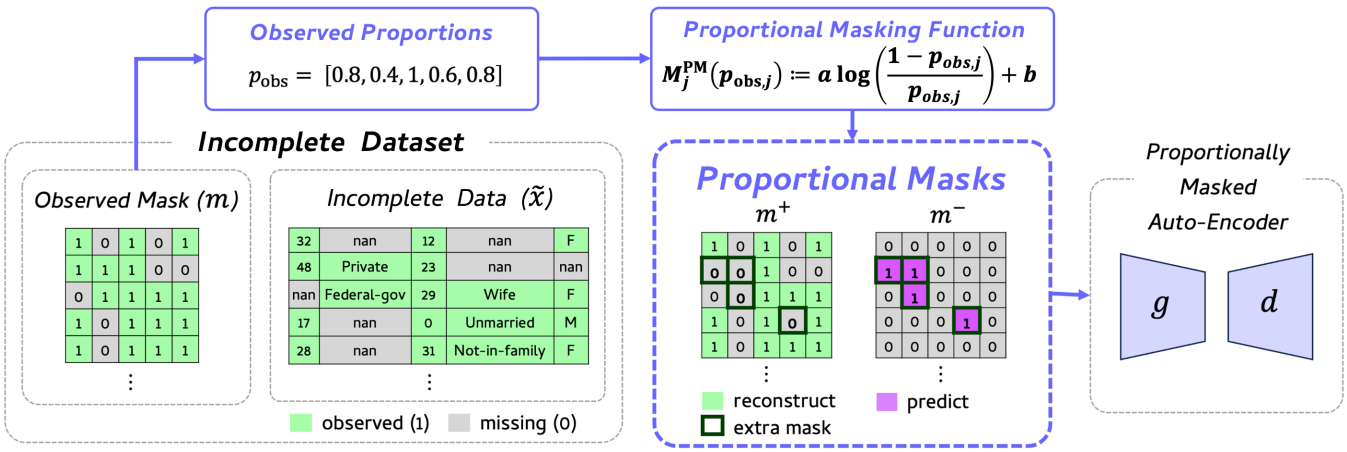
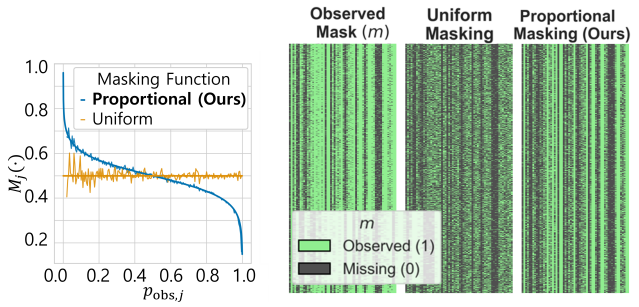


Figure 2: **PMAE**. Given the observed mask $\mathbf{m}$, we calculate the observed proportions and apply an additional mask, $\mathbf{m}^-$, where the extra masking probabilities are inversely proportional to the observed proportions.



(a) Proportional masking (b) Comparison of $\mathbf{m}^+$ after masking

Figure 3: Masking data inversely to observed proportions, prioritizing *prediction* when data is sparse and *reconstruction* when more data is observed (the last column of (b)).

be applied in proportions that align with the statistical characteristics of the dataset to avoid omitted variable bias, and (iii) as shown in (9), the loss can be balanced by assigning weights inversely to the observed proportions.

Method

Proportional Masking

We propose to guide MAEs to emphasize prediction or reconstruction in specific columns using a well-designed masking scheme by leveraging the observed proportions of columns as prior information, defined as:

$$\mathbf{p}_{\text{obs}} := \frac{1}{|B|} \sum_{i \in B} \mathbf{m}_i \in \mathbf{R}^d. \tag{10}$$

Since $\delta_{ij} \sim \text{Ber}(\pi_{ij})$, (10) serves as an MLE estimate for the column-wise average of the unknown propensity score function, i.e., $p_{\text{obs},j} \approx \mathbb{E}_{\mathbf{x}}[\pi_{\cdot j}(\mathbf{x}; \phi)]$.

Our proposed masking function takes the *logit-transformed* value of (10), where we verify the effectiveness of this design choice in Table 3:

$$M_j^{\text{PM}}(p_{\text{obs},j}; a, b) := a \log\left(\frac{1-p_{\text{obs},j}}{p_{\text{obs},j}}\right) + b, \tag{11}$$

where we clip the output to $[0, 1]$ to prevent undefined behavior as $p_{\text{obs},j}$ approaches 0 or 1, and a and b are hyperparameters. Note that this mask is not applied to columns that are fully observed across all rows, i.e., when $p_{\text{obs},j} = 1$.

We aim to achieve two objectives with this design: (i) to provide neural networks with signals corresponding to the average propensity score function within the batch, which implicitly serve as inductive biases; and (ii) to emphasize samples inversely proportional to their observed proportions. Intuitively, a controls the sensitivity to changes in the observed proportion $p_{obs,j}$ (or the logit), and b determines the base masking ratio regardless of $p_{obs,j}$. While the parameters could be learnable, we fix them at $a = 0.05$ and $b = 0.5$ (behavior shown in Fig. 3 (a)), as our grid search across validation datasets indicates these values perform well (Fig. 6).

Parameter Choice We layout the following design principles for selecting parameters in the masking function:

1. (Choice of a):
   - *Sign*. Ensure $a > 0$ to emphasize samples inversely to their observed proportions.
   - *Magnitude*. The value of a should be carefully balanced: (i) If a is *too small*, it leads to uniform random masking, and (ii) if a is *too large*, it enforces a hard decision rule where the model *always predicts* when $p_{\text{obs},j}$ is smaller than b, and *always reconstructs* otherwise.
2. (Choice of b): b represents a prior weight that determines the baseline importance of prediction. In the absence of prior knowledge, setting $b = 0.5$ is recommended for an equal balance between prediction and reconstruction.

Architecture: MLP-Mixers vs. Transformers

Transformer-based architectures have been prevalent in many MAE designs in the CV/NLP domains, where stacks

of Self-Attention (SA) and feed-forward blocks are utilized for learning intricate pairwise (and higher-order) relationships between different columns. Nonetheless, as discussed by Yan et al. (2023), interactions may exist in multiple groups, and we postulate that this makes the imputation problem harder to solve; learning such relations in a pairwise manner may be inefficient. Thus, we propose to utilize the MLP-Mixer architecture (Tolstikhin et al. 2021), where token-mixing MLPs with L^2 regularization can disconnect combinations of columns through activation functions. Given the tokenized data (positional information and [CLS] token appended: $\mathbf{x} = q_{\text{tok}}(\tilde{\mathbf{x}} \odot \mathbf{m}^+) \in \mathbf{R}^{n\times(d+1)\times c}$), we feed it to encoder-decoder architecture (h), where the following basic blocks are stacked in multiple layers:

- **Transformer Blocks**

$$\mathbf{x} \leftarrow \mathbf{x} + \text{LN}(\mathbf{x} + \text{SA}_d(\text{LN}(\mathbf{x}))),$$
$$\mathbf{x} \leftarrow \mathbf{x} + \text{LN}(\mathbf{x} + \text{MLP}_c(\text{LN}(\mathbf{x}))). \quad (12)$$

- **Mixer Blocks**

$$\mathbf{x} \leftarrow \mathbf{x} + \text{LN}(\mathbf{x} + \text{MLP}_d(\text{LN}(\mathbf{x}))),$$
$$\mathbf{x} \leftarrow \mathbf{x} + \text{LN}(\mathbf{x} + \text{MLP}_c(\text{LN}(\mathbf{x}))). \quad (13)$$

The subscript in SA/MLP denotes the dimension in which attention/mixing is applied. Only the token-mixing part (in the d dimension) differs where the MLP replaces SA.

Experiments

Experimental Setup

Semi-synthetic Missing Pattern Generation Given a complete dataset, we specify *missing patterns* as follows:

1. Monotone Missing
   - Generate $\mathcal{M}^{\text{m}}$ with $p^{\text{col}} \in \{0.3, 0.6\}$.
   - Generate missing entries with a fixed probability of 0.5 (i.e., $P(\delta_{ij} = 1) = 0.5$ for all j).
2. Quasi-Monotone Missing
   - Set $p^{\text{col}} = 0.6$ for $\mathcal{M}_1^{\text{q}}$, and let $\mathcal{M}_2^{\text{q}} := (\mathcal{M}_1^{\text{q}})^c$.
   - $\forall j \in \mathcal{M}_1^{\text{q}}$, $p_j \sim U(0.95, 0.99)$, $P(\delta_{ij} = 1) = p_j$.
   - $\forall j \in \mathcal{M}_2^{\text{q}}$, $p_j \sim U(0.2, 0.8)$, and $P(\delta_{ij} = 1) = p_j$.
3. General Missing (or Non-Monotone Missing)
   - Set $p^{\text{col}} = 1$ for $\mathcal{M}^{\text{g}}$.
   - $\forall j \in \mathcal{M}^{\text{g}}$, $p_j \sim U(0.2, 0.8)$, and $P(\delta_{ij} = 1) = p_j$.

Note that $\mathcal{M} := \{j | P(j \in \mathcal{M}) = p^{\text{col}}\}$ denotes the column indices that will have missing data, with the missing proportion p^{col}. The propensity score function $\pi_{ij}(\mathbf{x}_i; \boldsymbol{\phi}) = P(\delta_{ij} = 1 | \mathbf{x}_i)$ is specified with the *missing mechanism* and the generated p_j. In all settings, we apply the most challenging MNAR. Details are provided in the Appendix.

Datasets We evaluate PMAE along with the baselines using a semi-synthetic setup on nine real-world benchmark datasets (Asuncion, Newman et al. 2007).

Dataset	# Samples	# Categorical	# Numerical
Diabetes	442	1	9
Wine	1,599	11	0
Obesity	2,111	8	8
Bike	8,760	3	9
Shoppers	12,330	8	10
Letter	20,000	16	0
Default	30,000	3	9
News	39,644	2	46
Adult	48,842	9	6

Table 1: Benchmark tabular datasets of varying sizes and data types across different domains.

Baselines and Evaluation

Baseline Methods We compare our model with the following baselines: Naive (numerical:mean and categorical:mode), KNN (Troyanskaya et al. 2001) EM (Dempster, Laird, and Rubin 1977), MissForest (Stekhoven and Bühlmann 2012), MiceForest (Shah et al. 2014), MIWAE (Mattei and Frellsen 2019), GAIN (Yoon, Jordon, and Schaar 2018), MIRACLE (Kyono et al. 2021), HyperImpute (Jarrett et al. 2022), TDM (Zhao et al. 2023), and ReMasker (Du, Melis, and Wang 2024).

Parameter Setting Our implementation mostly follows ReMasker with the same optimization procedures, but introduces (i) a *new loss* formulation in (4) and (ii) a novel *masking function* $M_j^{\text{PM}}(p_{\text{obs},j}; 0.05, 0.5)$, which are applied to (iii) *Transformer* and *MLP-Mixer* architecture.

Imputation Accuracy Instead of the widely used RMSE, we evaluate imputation performance using a metric we call *Imputation Accuracy*, a weighted average of Accuracy (for categorical columns) and R^2 (for numerical columns):

$$\text{Acc}_j = \frac{\sum_{i\in I_j^{\text{mis}}} I(\hat{x}_{ij} = x_{ij})}{|I_j^{\text{mis}}|}, \quad (14)$$

$$R_j^2 = 1 - \frac{\sum_{i\in I_j^{\text{mis}}} (\hat{x}_{ij} - x_{ij})^2}{\sum_{i\in I_j^{\text{mis}}} (\bar{x}_{ij} - x_{ij})^2}, \quad (15)$$

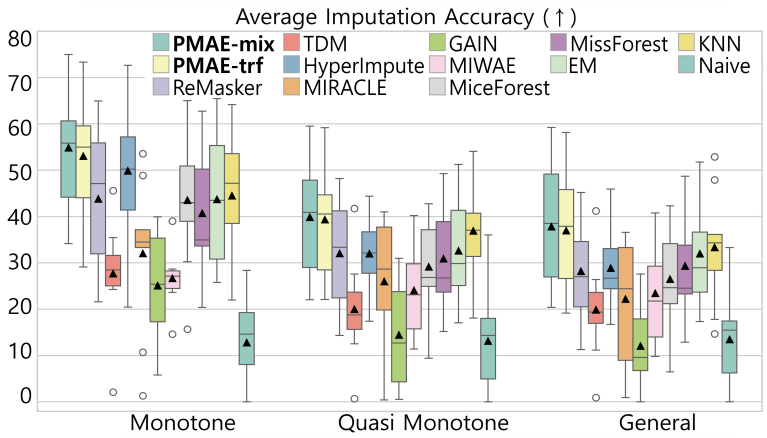
$$\text{Imp. Acc} := \frac{1}{|D^{\text{mis}}|} \sum_{j\in D^{\text{mis}}} (\text{Acc}_j \cdot I(j \in D_c^{\text{mis}}) + R_j^2 \cdot I(j \notin D_c^{\text{mis}})), \quad (16)$$

where D_c^{mis} and I_j^{mis} denote missing categorical column indices and missing row indices for column j respectively. Since the values are adapted to each data type and normalized to 1, this evaluation serves as an intuitive and holistic measure of imputation performance for tabular data.

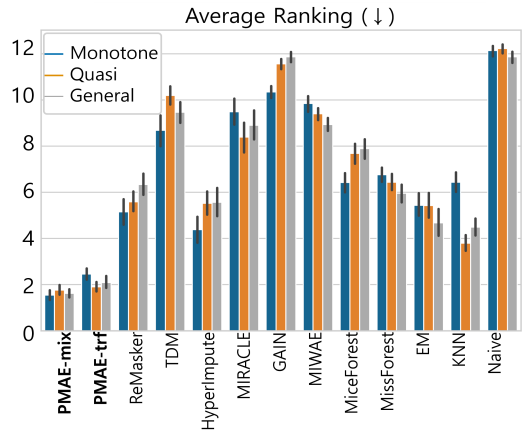
Other Measures While not the primary focus of this study, we also evaluate the following:

	Imp. Acc. (↑)		R^2 (↑)		Acc. (↑)		RMSE (↓)		DT (cls)(↑)		DT (reg)(↑)	
Method	Avg.	Rank	Avg.	Rank	Avg.	Rank	Avg.	Rank	Avg.	Rank	Avg.	Rank
Naive	13.2 ± 3.4	12.1	0.0 ± 0.0	12.1	36.4 ± 8.9	11.1	27.9 ± 4.5	10.5	88.3 ± 2.7	8.1	23.4 ± 1.8	7.9
KNN	38.3 ± 6.0	4.5	26.5 ± 7.7	5.0	54.8 ± 7.5	4.6	20.8 ± 2.2	5.8	88.5 ± 2.2	6.6	24.1 ± 2.4	7.4
EM	36.1 ± 5.1	5.4	26.6 ± 6.1	4.8	49.1 ± 6.5	6.8	20.8 ± 2.0	5.1	89.1 ± 2.0	6.0	24.7 ± 2.2	5.4
MissForest	33.7 ± 5.0	6.6	23.4 ± 6.0	6.2	49.5 ± 5.8	7.3	21.0 ± 2.3	5.7	88.5 ± 2.1	7.1	24.5 ± 1.3	5.9
MiceForest	33.1 ± 6.6	7.1	20.3 ± 8.7	8.0	53.9 ± 7.6	5.5	23.7 ± 2.5	9.2	87.4 ± 2.4	9.3	22.4 ± 3.2	9.3
MIWAE	24.5 ± 4.4	9.3	9.3 ± 4.1	9.6	48.8 ± 7.2	8.6	25.8 ± 3.0	9.4	85.7 ± 3.1	10.6	21.8 ± 3.9	7.8
GAIN	17.2 ± 3.8	11.4	4.8 ± 4.3	11.1	39.6 ± 7.3	10.6	29.6 ± 2.7	10.7	88.1 ± 1.9	7.6	23.4 ± 2.8	8.0
MIRACLE	26.8 ± 7.0	8.6	14.4 ± 7.4	8.9	42.9 ± 8.2	8.0	29.0 ± 3.7	10.2	88.5 ± 2.9	7.4	21.8 ± 1.9	9.6
HyperImpute	36.9 ± 6.8	5.1	26.2 ± 9.3	5.2	53.6 ± 8.2	5.4	21.3 ± 2.5	6.2	88.1 ± 2.0	6.4	23.9 ± 1.6	6.8
TDM	22.5 ± 4.2	9.5	7.7 ± 4.0	9.2	45.2 ± 8.3	9.2	22.8 ± 2.5	7.3	88.1 ± 2.4	8.1	24.5 ± 2.1	6.1
ReMasker	34.7 ± 6.1	6.1	25.4 ± 9.4	5.5	50.6 ± 7.2	6.4	19.8 ± 2.2	4.3	89.1 ± 2.0	5.7	24.3 ± 3.0	6.4
PMAE-trf	43.1 ± 5.8	2.3	34.6 ± 7.4	2.2	56.1 ± 8.3	3.4	18.6 ± 2.1	2.8	**90.1** ± 1.3	**3.2**	**25.2** ± 2.3	**4.5**
PMAE-mix	**44.2** ± 5.9	**1.8**	**36.0** ± 7.7	**1.7**	**56.4** ± 8.4	**2.8**	**18.4** ± 1.9	**2.5**	90.1 ± 1.2	3.6	24.8 ± 1.8	5.2

Table 2: Summary result of 13 state-of-the-art methods on 9 datasets, three different *missing patterns*, applied with *MNAR* mechanism, repeated 10 times; *Avg.* are average of 9 (dataset) × 3 (pattern) × 10 (seed) = 270 runs, and *Rank* are average of 270 ranked values. For better readability, *Avg.* values are scaled (×100). Details for each datasets can be found in the Appendix.



(a) **Imputation Accuracy (average)**. Every point in each boxplot corresponds to the average value across datasets. Mean value are indicated with (▲).



(b) **Imputation Accuracy (rank)**. *PMAE* outperforms all other state-of-the-art methods across all patterns.

Figure 4: Imputation accuracy of state-of-the-art methods across 9 benchmark datasets on missing patterns (*Monotone*, *Quasi Monotone*, and *General*) under NMAR mechanism. Methods are arranged such that the most recent is on the left.

- *Downstream Task (DT) Performance*: The utility of imputed data measured by evaluating the supervised learning performance; we report the average of test set performances of XGBoost and Linear model (regression, reg: R^2, classification, cls: AUROC).
- *RMSE* as in Du, Melis, and Wang (2024) for comparison.

Experimental Results

Performance Comparison Table 2 presents the average values and ranks of imputation methods, while Figure 4 (a) shows performance variability across datasets/patterns, and Figure 4 (b) compares the methods' relative performances across patterns. We summarize key empirical findings:

- PMAE-mix achieves the highest imputation accuracy, improving by *at least* 17.5% (38.3 → 44.2).
- With our *novel masking function*, we observed an overall improvement of 27.3% (34.7 → 43.1) compared to the ReMasker counterpart overall (at least 7.4% on *shoppers*, up to 82.3% on *news* dataset).
- Across different missing data patterns, PMAE outperforms ReMasker by 25.1% on *Monotone* up to 34.1% for *General* pattern.
- Although the performance of all methods declines as the pattern moves from *Monotone* to *General*, ours maintain consistent relative rankings (as seen in Figure 4 (b)).
- Across data types, we improved R^2 by at least 37.6%

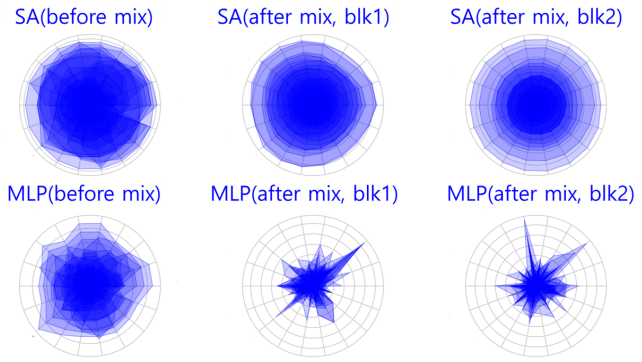


Figure 5: Magnitude comparison of representations after token mixing via. Self-Attention vs. MLP on *Shoppers* dataset. Absolute value is applied for the sake of analysis.

(26.6 → 36.0), and Accuracy by at least 2.9% with 11.4% gain for ReMasker (50.6→56.4).

- With RMSE as the main metric, the relative rank of PMAE are superior and consistent; however, the average rank for different methods shift, placing ReMasker as the second best (6.1 → 4.3).
- Our methods (PMAE-trf, PMAE-mix) have superior DT performances over all baselines. Also, Transformer architecture is more suitable than MLP-Mixer architecture for these tasks; perhaps, MLPs, being more flexible, generate overfitted predictions compared to SAs.

Impact of MLPs on Token Mixing Figure 5 illustrates the representation vector of SA-based and MLP-based mixing on *shoppers*, highlighting changes in relative magnitudes before and after each mixing block. This shows that MLPs, with activation functions, offer more flexible token mixing compared to SAs, potentially capturing discrete group interactions between columns more effectively.

Ablation Study: Masking Function Design The results in Table 3 support our initial motivations: (i) model performance declines without balancing reconstruction and prediction tasks (*No recon./pred.*); (ii) masking more observed data reduces performance (*Reversed*); (iii) increasing prediction weights for less observed data is beneficial (*Piecewise*); (iv) the concavity of the masking function is critical as performance improves with a convex function for $p_{\text{obs},j} \leq 0.5$ and concave function for $p_{\text{obs},j} > 0.5$, but worsens if concavities are reversed (*Sigmoid-like* vs. *Logit*).

Ablation Study: Loss, $M_j(\cdot)$, Architecture ReMasker uses the following loss in their implementation: $l_{ij}^{\text{Re}} := \frac{(h(\tilde{\mathbf{x}}_i \odot \mathbf{m}_i^+))_j - x_{ij})^2 \cdot m_{ij}^+}{\sum_{i \in B} m_{ij}^+} + \frac{((h(\tilde{\mathbf{x}}_i \odot \mathbf{m}_i^+))_j - x_{ij})^2 \cdot m_{ij}^-}{\sum_{i \in B} m_{ij}^-}$, which is equivalent to $m_{ij}^+ l_{ij}^0 + m_{ij}^- l_{ij}^+$ after multiplying by some constant. Then, unlike our formulation in (6), *prediction loss* l_{ij}^0 is weighted by the unmasked parts m_{ij}^+. Moreover, ReMasker applies $M(\cdot) = 0.5, \forall_{p_{\text{obs},j}}$. Table 4 shows performance gains from adjusting the main loss, masking function, and encoder/decoder blocks.

Description	$M_j(\cdot)$	**Perf.**	**Gain** (%)
Const.	$0.5 \cdot I(p_{\text{obs},j} < 1)$	33.4	0
No recon.	1.0	7.7	-76.9
No pred.	0.0	8.6	-74.3
Linear	$1 - p_{\text{obs},j}$	34.8	4.09
Reversed	$p_{\text{obs},j}$	31.9	-4.6
Piece-wise	$(1 - p_{\text{obs},j}) \cdot I(p_{\text{obs},j} < 0.5) + 0.5 \cdot I(p_{\text{obs},j} \geq 0.5)$	33.6	0.69
Sigmoid-like	$\frac{1}{\exp(-10(0.5 - p_{\text{obs},j}) + 1)}$	33.7	0.78
Logit (Ours)	$0.05 \cdot \log\left(\frac{1 - p_{\text{obs},j}}{p_{\text{obs},j}}\right) + 0.5$	**35.4**	**5.83**

Table 3: **Masking function design**. Results of other masking functions averaged on *Shoppers*, *Wine*, and *Diabetes*.

(a) MLP-Mixer (b) Transformer

Figure 6: **Grid search results.** Imputation accuracy across different a, b averaged on *Shoppers*, *Wine*, and *Diabetes*.

Method	**Changes**	**Perf.**	Δ(%)
ReMasker	-	34.7	0.0
ReMasker	Loss: l_{ij}^{Re} → (6)	36.2	+4.3
PMAE-trf	$M_j(\cdot)$: 0.5 → (11)	43.1	+20.0
PMAE-mix	Architecture: trf → mix	44.2	+3.0

Table 4: Ablations on loss, $M_j(\cdot)$, and architecture.

Conclusion

Tabular data is inherently heterogeneous, with each column having distinct characteristics and often exhibiting complex patterns of missing values. To address this challenge, we propose PMAE, a simple yet effective strategy that employs a logit-based masking function. This method preserves the distribution of missingness while prioritizing data inversely to their observed proportions. When tested across diverse set of missing patterns, PMAE demonstrated robust performance, consistently surpassing state-of-the-art methods

Limitations and Future Work This study does not examine the relationship between a dataset's covariance structure and the proportional masking scheme, which could provide deeper understanding and broader applicability. Capturing covariance in the presence of missing data, however, remains a challenge. Addressing these issues may enable application of PMAE to high-dimensional tasks like image inpainting.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2020R1A2C1A01005949, 2022R1A4A1033384, RS-2023-00217705, RS-2024-00341749), the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program (RS-2023-00259934) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), and the Yonsei University Research Fund (2024-22-0148).

References

Asuncion, A.; Newman, D.; et al. 2007. UCI machine learning repository.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Donders, A. R. T.; Van Der Heijden, G. J.; Stijnen, T.; and Moons, K. G. 2006. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10): 1087–1091.

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

Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2021. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34: 18932–18943.

Guo, S.; Zou, L.; Liu, Y.; Ye, W.; Cheng, S.; Wang, S.; Chen, H.; Yin, D.; and Chang, Y. 2021. Enhanced doubly robust learning for debiasing post-click conversion rate estimation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 275–284.

Hawthorne, G.; and Elliott, P. 2005. Imputing cross-sectional missing data: Comparison of common techniques. *Australian & New Zealand Journal of Psychiatry*, 39(7): 583–590.

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.

Huang, X.; Khetan, A.; Cvitkovic, M.; and Karnin, Z. 2020. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.

Jarrett, D.; Cebere, B. C.; Liu, T.; Curth, A.; and van der Schaar, M. 2022. Hyperimpute: Generalized iterative imputation with automatic model selection. In *International Conference on Machine Learning*, 9916–9937. PMLR.

Kim, J. K.; and Shao, J. 2021. *Statistical methods for handling incomplete data*. Chapman and Hall/CRC.

Kyono, T.; Zhang, Y.; Bellot, A.; and van der Schaar, M. 2021. Miracle: Causally-aware imputation via learning missing data mechanisms. *Advances in Neural Information Processing Systems*, 34: 23806–23817.

Li, H.; Xiao, Y.; Zheng, C.; Wu, P.; and Cui, P. 2023. Propensity matters: Measuring and enhancing balancing for recommendation. In *International Conference on Machine Learning*, 20182–20194. PMLR.

Li, H.; Zheng, C.; Zhou, X.-H.; and Wu, P. 2022. Stabilized doubly robust learning for recommendation on data missing not at random. *arXiv preprint arXiv:2205.04701*.

Little, R. J.; and Rubin, D. B. 2019. *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Mattei, P.-A.; and Frellsen, J. 2019. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, 4413–4423. PMLR.

Molenberghs, G.; Michiels, B.; Kenward, M. G.; and Diggle, P. J. 1998. Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52(2): 153–161.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Seaman, S. R.; and White, I. R. 2013. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3): 278–295.

Shah, A. D.; Bartlett, J. W.; Carpenter, J.; Nicholas, O.; and Hemingway, H. 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology*, 179(6): 764–774.

Somepalli, G.; Goldblum, M.; Schwarzschild, A.; Bruss, C. B.; and Goldstein, T. 2021. Saint: Improved neural networks for tabular data via row attention and contrastive pretraining. *arXiv preprint arXiv:2106.01342*.

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

Sun, B.; Perkins, N. J.; Cole, S. R.; Harel, O.; Mitchell, E. M.; Schisterman, E. F.; and Tchetgen Tchetgen, E. J. 2018. Inverse-probability-weighted estimation for monotone and nonmonotone missing data. *American journal of epidemiology*, 187(3): 585–591.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34: 24261–24272.

Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; and Altman, R. B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6): 520–525.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *NeurIPS*.

Wilms, R.; Mäthner, E.; Winnen, L.; and Lanwehr, R. 2021. Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, 5: 100075.

Wu, Z.; Dadu, A.; Tustison, N.; Avants, B.; Nalls, M.; Sun, J.; and Faghri, F. 2024. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*.

Yan, J.; Chen, J.; Wu, Y.; Chen, D. Z.; and Wu, J. 2023. T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10720–10728.

Yoon, J.; Jordon, J.; and Schaar, M. 2018. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, 5689–5698. PMLR.

Zhang, H.; Zhang, J.; Srinivasan, B.; Shen, Z.; Qin, X.; Faloutsos, C.; Rangwala, H.; and Karypis, G. 2023. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv preprint arXiv:2310.09656*.

Zhao, H.; Sun, K.; Dezfouli, A.; and Bonilla, E. V. 2023. Transformed distribution matching for missing value imputation. In *International Conference on Machine Learning*, 42159–42186. PMLR.

Zhong, J.; Gui, N.; and Ye, W. 2023. Data imputation with iterative graph reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11399–11407.