MissDDIM: Deterministic and Efficient Conditional Diffusion for Tabular Data Imputation

Youran Zhou
echo.zhou@deakin.edu.au
Deakin University
Geelong, Victoria, Australia

Mohamed Reda Bouadjenek
reda.bouadjenek@deakin.edu.au
Deakin University
Geelong, Victoria, Australia

Sunil Aryal
sunil.aryal@deakin.edu.au
Deakin University
Geelong, Victoria, Australia

Abstract

Diffusion models have recently emerged as powerful tools for missing data imputation by modeling the joint distribution of observed and unobserved variables. However, existing methods, typically based on stochastic denoising diffusion probabilistic models (DDPMs), suffer from high inference latency and variable outputs, limiting their applicability in real-world tabular settings. To address these deficiencies, we present in this paper MissDDIM, a conditional diffusion framework that adapts Denoising Diffusion Implicit Models (DDIM) for tabular imputation. While stochastic sampling enables diverse completions, it also introduces output variability that complicates downstream processing. MissDDIM replaces this with a deterministic, non-Markovian sampling path, yielding faster and more consistent imputations. To better leverage incomplete inputs during training, we introduce a self-masking strategy that dynamically constructs imputation targets from observed features—enabling robust conditioning without requiring fully observed data. Experiments on five benchmark datasets demonstrate that MissDDIM matches or exceeds the accuracy of state-of-the-art diffusion models, while significantly improving inference speed and stability. These results highlight the practical value of deterministic diffusion for real-world imputation tasks.

CCS Concepts

• **Information systems** → **Data mining**; • **Computing methodologies** → *Artificial intelligence*; **Neural networks**.

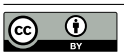
Keywords

Tabular Data Imputation; Missing Data; Generative Models; Diffusion Models; DDIM

ACM Reference Format:
Youran Zhou, Mohamed Reda Bouadjenek, and Sunil Aryal. 2025. MissDDIM: Deterministic and Efficient Conditional Diffusion for Tabular Data Imputation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3760943

1 Introduction

Missing data is a pervasive challenge in real-world applications such as healthcare [11], finance [2], recommendation systems [4], and sensor networks [5, 27]. In tabular datasets, missing values degrade model performance and introduce bias or uncertainty in downstream analysis. Effective imputation—the task of estimating missing entries from observed data—is thus a critical step in many data-centric workflows. Traditional imputation methods such as mean/mode substitution [9, 17], k-nearest neighbors [23], MICE [6], and MissForest [18] are efficient and stable, but often fail to capture feature dependencies and struggle under structured missingness. In contrast, deep generative models have shown great promise for modeling the joint distribution of observed and missing variables. Approaches based on GANs (e.g., GAIN [21], MisGAN [1]) and variational inference (e.g., MIWAE [12], HI-VAE [13]) have demonstrated stronger performance, while newer architectures like GRAPE [22] and IGRM [25] incorporate iterative or graph-based interactions for greater expressiveness. More recently, diffusion models [3, 7, 20] have emerged as powerful generative frameworks, particularly in vision and time series domains. Their gradual denoising process allows fine-grained, high-fidelity generation. In the tabular setting, several adaptations have been proposed: TabCSDI [24] employs conditional score-based diffusion; MissDiff [14] uses an unconditional formulation; and TabDDPM [10] extends diffusion to mixed-type data. Despite their modeling capacity, these methods face key limitations: (i) they rely on stochastic DDPM sampling, which incurs high inference latency and output variability; and (ii) many assume fully observed training data or lack robust conditioning on partial inputs—assumptions that rarely hold in practice [15, 26]. To address these challenges, we propose **MissDDIM**, the first framework to apply Denoising Diffusion Implicit Models (DDIM) [16] to imputation on incomplete tabular data. Unlike DDPMs, DDIM performs deterministic, non-Markovian sampling, enabling consistent outputs with significantly reduced inference cost. We further introduce a *self-masking strategy* that dynamically creates training targets from partially observed data, allowing MissDDIM to learn directly from incomplete inputs. Through extensive experiments on five real-world datasets, we demonstrate that MissDDIM achieves competitive imputation accuracy while offering substantial improvements in inference speed and output stability. By bridging the gap between expressive generative modeling and efficient, deployment-friendly inference, MissDDIM provides a practical solution for real-world tabular imputation tasks.



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3760943

2 The Proposed Method

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a tabular dataset with n samples and d features, where each sample $\mathbf{x} \in \mathbb{R}^d$ is drawn from an unknown data distribution. Given a sample $\mathbf{x}_0$ that contains missing values, we aim to generate imputation targets $\mathbf{x}_0^{\text{mis}} \in \mathcal{X}^{\text{mis}}$ by exploiting the observed

values $\mathbf{x}_0^{\text{obs}} \in \mathcal{X}^{\text{obs}}$, where $\mathcal{X}^{\text{mis}}$ and $\mathcal{X}^{\text{obs}}$ are subsets of the full feature space $\mathcal{X} = \mathbb{R}^d$.

2.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) [8] model complex data distributions by simulating a Markov chain that gradually adds noise to the data (forward process) and then learns to reverse the corruption (reverse process). Let $\mathbf{x}_0 \in \mathcal{X}$ denote a data sample drawn from the unknown data distribution $q(\mathbf{x}_0)$. The forward process defines a sequence of latent variables $\mathbf{x}_1, \ldots, \mathbf{x}_T$ in the same space $\mathcal{X}$, where T is a predefined number of time steps. At each step $t \in \{1, \ldots, T\}$, Gaussian noise is incrementally added to produce $\mathbf{x}_t$ from $\mathbf{x}_{t-1}$ according to:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}),$$

where $\{\beta_t\}_{t=1}^T$ is a variance schedule controlling the amount of noise injected at each step. This leads to a closed-form marginal:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0,\ (1-\alpha_t)\mathbf{I}),\ \text{where } \alpha_t = \prod_{i=1}^{t}(1-\beta_i). \quad (1)$$

The reverse process is modeled as a denoising procedure learned via a parameterized distribution:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1};\ \boldsymbol{\mu}_\theta(\mathbf{x}_t, t),\ \sigma_\theta^2(\mathbf{x}_t, t)\mathbf{I}), \quad (2)$$

with $\boldsymbol{\mu}_\theta$ derived from a noise prediction network ϵ_θ:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\alpha_t}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(\mathbf{x}_t, t)\right). \quad (3)$$

While DDPMs yield high-quality generations, their reliance on long sampling chains (typically hundreds of steps) results in high inference latency, making them less practical for real-time imputation tasks that require rapid and stable outputs.

2.2 MissDDIM: Efficient Imputation with Conditional DDIM

While DDPMs have demonstrated strong generative capabilities, their sequential and stochastic nature leads to computationally expensive and inherently variable inference. These limitations pose practical challenges for missing value imputation in real-world applications, where stability and speed are crucial for downstream pipelines. To address this, we propose `MissDDIM`, a conditional diffusion model that adapts Denoising Diffusion Implicit Models (DDIM) [16] for efficient and deterministic imputation on tabular data. Unlike DDPMs, DDIM enables non-Markovian, parameter-free sampling trajectories, significantly reducing the number of inference steps while preserving generation quality (see Figure 1). Despite DDIM's success in image synthesis, it has not yet been explored in the context of missing value imputation—particularly for tabular data, where heterogeneous feature types and partially observed inputs introduce unique challenges. `MissDDIM` bridges this gap by developing a conditional DDIM framework specifically tailored for imputation tasks.

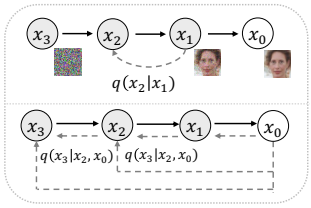


Figure 1: Different sampling processes: traditional stochastic diffusion (top) vs. deterministic non-Markovian inference (bottom).

2.2.1 Conditional DDIM. We aim to estimate the conditional distribution $p_\theta(\mathbf{x}_0^{\text{mis}} \mid \mathbf{x}_0^{\text{obs}})$, where the generative model focuses solely on missing components. To this end, we define a conditional noise prediction network

$$\epsilon_\theta(\mathbf{x}_t^{\text{mis}}, t \mid \mathbf{x}_0^{\text{obs}})$$

which predicts the noise applied to missing entries, given the observed context. This design explicitly conditions the reverse generation on known values at both training and inference stages, enabling targeted and stable imputation.

The reverse process is modified accordingly:

$$p_\theta(\mathbf{x}_{0:T}^{\text{mis}} \mid \mathbf{x}_0^{\text{obs}}) = p(\mathbf{x}_T^{\text{mis}}) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}^{\text{mis}} \mid \mathbf{x}_t^{\text{mis}}, \mathbf{x}_0^{\text{obs}}), \quad (4)$$

$$p_\theta(\mathbf{x}_{t-1}^{\text{mis}} \mid \mathbf{x}_t^{\text{mis}}, \mathbf{x}_0^{\text{obs}}) = \mathcal{N}\left(\mathbf{x}_{t-1}^{\text{mis}};\ \boldsymbol{\mu}_\theta(\mathbf{x}_t^{\text{mis}}, t \mid \mathbf{x}_0^{\text{obs}}),\ \sigma_\theta^2(t)\,\mathbf{I}\right), \quad (5)$$

where $\boldsymbol{\mu}_\theta$ is derived from the conditional noise prediction network [19] as:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t^{\text{mis}}, t \mid \mathbf{x}_0^{\text{obs}}) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t^{\text{mis}} - \frac{\beta_t}{\sqrt{1-\alpha_t}}\,\epsilon_\theta(\mathbf{x}_t^{\text{mis}}, t \mid \mathbf{x}_0^{\text{obs}})\right). \quad (6)$$

In standard DDPM-based samplers, the reverse process involves sampling from a Gaussian distribution with a learned mean and fixed variance. DDIM generalizes this by introducing a non-Markovian sampling schedule, where the level of stochasticity at each timestep is controlled by a noise parameter σ_t. When $\sigma_t > 0$, the process remains stochastic; when $\sigma_t = 0$, it becomes fully deterministic.

In `MissDDIM`, we adopt the deterministic variant by explicitly setting $\sigma_t = 0$ for all t, eliminating randomness in sampling and ensuring consistent outputs across runs—a critical property for reproducible imputation. The resulting update rule simplifies to:

$$\mathbf{x}_{t-1}^{\text{mis}} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}_t^{\text{mis}} - \sqrt{1-\alpha_t}\,\epsilon_\theta(\mathbf{x}_t^{\text{mis}}, t \mid \mathbf{x}_0^{\text{obs}})}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(\mathbf{x}_t^{\text{mis}}, t \mid \mathbf{x}_0^{\text{obs}}), \quad (7)$$

This deterministic formulation brings two key advantages: (i) it reduces inference latency by an order of magnitude due to fewer sampling steps; (ii) it ensures output consistency across runs, addressing the variability inherent in stochastic DDPM-based imputers. These properties make `MissDDIM` particularly suitable for latency-sensitive applications such as risk modeling, recommendation systems, and real-time analytics.

2.2.2 Training Objective. We adopt the standard DDPM/DDIM training strategy, adapted for conditional imputation. Given a sample with observed features $\mathbf{x}_0^{\text{obs}}$ and missing targets $\mathbf{x}_0^{\text{mis}}$, we corrupt the missing components using the forward diffusion process $\mathbf{x}_t^{\text{mis}} = \sqrt{\alpha_t}, \mathbf{x}0^{\text{mis}} + \sqrt{1-\alpha_t}, \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A conditional noise prediction network $\epsilon\theta(\mathbf{x}_t^{\text{mis}}, t \mid \mathbf{x}_0^{\text{obs}})$ is trained to recover the injected noise, i.e., $\epsilon\theta(\mathbf{x}_t^{\text{mis}}, t \mid \mathbf{x}_0^{\text{obs}}) \approx \boldsymbol{\epsilon}$. The model is optimized

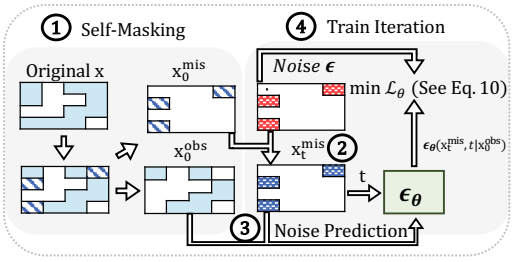


Figure 2: Illustration of the self-masking training strategy used in `MissDDIM`. Observed values are randomly partitioned into conditional inputs and pseudo-targets during training.

via a conditional denoising score matching loss:

$$\min_{\theta} \mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0,\, \boldsymbol{\epsilon},\, t}\left[\left\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t^{\text{mis}}, t \mid \mathbf{x}_0^{\text{obs}})\right\|_2^2\right], \tag{8}$$

where $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ and $t \sim \text{Unif}(\{1, \ldots, T\})$.

2.2.3 Self-masking Strategy. Most generative imputation methods rely on prior imputers (e.g., mean filling) or externally provided masking vectors to handle missing inputs. In contrast, we adopt a **self-masking** strategy that enables the model to learn directly from partially observed data without auxiliary imputations (see Figure 2). Specifically, during training, a random subset of observed entries is masked and treated as targets, while the remaining observed values serve as conditional context.

2.3 Implementation

Our implementation is based on the TabCSDI architecture [24]. Specifically, we remove the temporal transformer module and retain only the feature-wise transformer encoder and residual MLP blocks to better suit static feature spaces. Although `MissDDIM` builds upon TabCSDI, our DDIM-style sampler is model-agnostic. It depends only on the learned noise prediction network ϵ_θ, and is therefore compatible with any DDPM-based imputer, regardless of backbone. This makes `MissDDIM` a drop-in replacement for stochastic samplers, offering a general-purpose mechanism to accelerate and stabilize diffusion-based imputation pipelines.

3 Experiments

We evaluate `MissDDIM` from three perspectives: (i) **Imputation accuracy**: how well the imputed values match ground truth; (ii) **Sampling efficiency**: the trade-off between inference time and sampling steps; (iii) **Stability**: the consistency of results across runs. Code will be available at: https://github.com/echoid/MissDDIM

3.1 Experimental Setup

3.1.1 Datasets. We use five real-world datasets from the UCI Repository and Kaggle, covering both continuous-only data (*Banknote*, *California*, *Letter*) and mixed-type data (*Adult*, *Student*). Missing values are simulated at four levels (10%, 30%, 50%, 70%). We evaluate both direct reconstruction error and downstream predictive performance.

3.1.2 Baselines and Evaluation Protocol. We compare `MissDDIM` against three categories of baselines: (i) statistical methods (`Mean/Mode`, `MICE` [6], `MissForest` [18]), (ii) deep generative models (`GAIN` [21], `MIWAE` [12]), and (iii) diffusion-based approaches (`CSDI` [19], `TabCSDI` [24], `MissDiff` [14]). All diffusion-based baselines use T=100 steps with 100 stochastic samples per instance (median-aggregated). `MissDDIM` yields deterministic imputations in a single forward pass. All results are averaged over 5-fold cross-validation with 20% of data held out for testing in each fold. Continuous features are standardized before imputation. Imputation accuracy is measured using RMSE averaged over missing entries. To evaluate downstream utility, we train an XGBoost classifier for classification tasks and XGBoost regressor for regression tasks on imputed data and report weighted F1-score or MAE respectively.

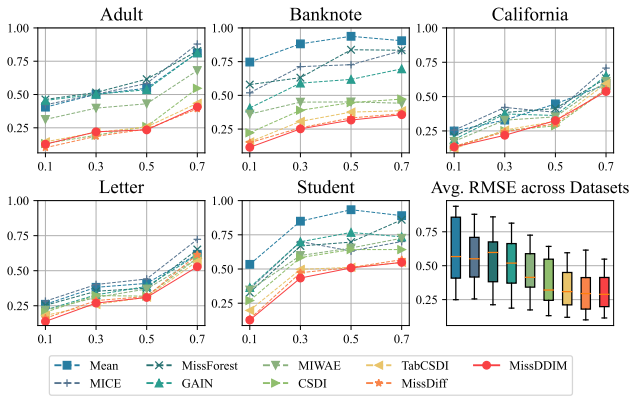


Figure 3: RMSE across five benchmark datasets under varying missing rates. The final panel presents an aggregated summary of RMSE distributions using boxplots, combining results from all datasets and missingness levels.

4 Results and Analysis

4.1 Imputation Utility

Figure 3 summarises the imputation performance of methods under varying missing rates.The performance of all models generally degrades with increasing missingness. However, `MissDDIM` demonstrates relatively stable performance across different settings. The last subplot provides a boxplot that aggregates results across all datasets and missing rates, highlighting the superior stability of our method. Furthermore, Table 1 shows that `MissDDIM` achieves strong performance in downstream predictive tasks, consistently outperforming or matching state-of-the-art baselines. Standard deviation regions are also reported to illustrate the robustness.

4.2 Sampling Time

Existing diffusion-based models such as CSDI, TabCSDI and MissDiff typically generate 100 samples and take their median to produce stable imputations, which significantly increases inference time. In contrast, `MissDDIM` adopts a deterministic sampling process that requires only a single forward pass. Table 2 reports the inference time and imputation accuracy under both standard (100-sample) and DDIM with single-sample settings for different T setted . For fairness, we evaluate all methods using the same batch size and computational environment. `MissDDIM` consistently demonstrates

Table 1: Performance under 30%, 50%, and 70% missingness across five datasets. Classification tasks are evaluated using weighted F1-score, while the *Student* dataset (regression) uses MAE. Best results are shown in bold, and second-best results are underlined.

Method	Adult			Banknote			California			Letter			Student (MAE)		
Rate	30%	50%	70%	30%	50%	70%	30%	50%	70%	30%	50%	70%	30%	50%	70%
Mean	0.4966	0.4750	0.4440	0.9000	0.7870	0.6703	0.8577	0.7595	0.6259	0.7960	0.6277	0.4046	2.2286	2.3608	2.5107
MICE	0.6734	0.6447	0.6218	0.9416	0.8414	0.7512	0.7963	0.7373	0.6103	0.7862	0.7850	0.4370	2.0147	2.1097	2.4936
MissForest	0.6971	0.6517	0.6347	0.9517	0.8727	0.7821	0.8274	0.7556	0.6300	0.8323	0.8148	0.4524	1.7569	2.0165	1.9655
GAIN	0.6048	0.5775	0.5441	0.9217	0.8725	0.7957	0.8291	0.7564	0.6094	0.8221	0.7567	0.4054	1.2574	1.4025	1.5993
MIWAE	0.6941	0.6327	0.6014	0.9226	0.8937	0.8125	0.8531	0.6905	0.6261	0.8674	0.8186	0.4253	1.2347	1.4582	1.4614
CSDI	0.6827	0.6424	0.6218	0.9358	0.8867	0.8025	0.8446	0.8163	0.6291	0.8479	0.8090	0.4249	1.0257	1.2477	1.2098
TabCSDI	0.7214	0.6851	**0.6318**	**0.9527**	0.9014	0.8657	0.8886	0.8386	0.6289	0.9049	0.8471	0.4467	0.8712	1.1371	1.1816
MissDiff	0.7046	0.6711	0.6247	0.9416	0.8867	0.8237	0.8552	0.8253	**0.6318**	0.8503	0.8242	0.4541	0.8571	1.1297	**1.1575**
MissDDIM	**0.7228**	**0.6724**	**0.6318**	**0.9527**	**0.9214**	**0.8772**	**0.8993**	**0.8608**	0.6236	**0.9117**	**0.8559**	**0.4581**	**0.8214**	**1.1143**	1.1713

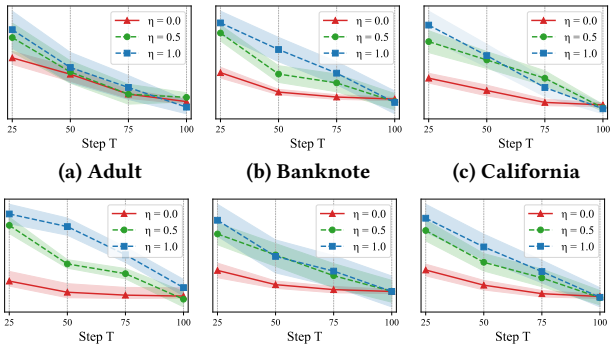
Table 2: Inference time and imputation accuracy (RMSE) comparison of generative imputation models.

#Samples	Method	Time (s) ↓	RMSE ↓
100	CSDI	820.73	0.3730
	TabCSDI	794.21	0.3319
	MissDiff	765.92	0.3163
1	MissDDIM **(T = 100)**	**775.31**	**0.3051**
	MissDDIM **(T = 50)**	385.12	0.3167
	MissDDIM **(T = 20)**	**154.61**	0.3343

Table 3: Impact of sampling stochasticity (η) and number of sampling steps (T) on RMSE under 50% missing rate.

η	Letter		California		Average	
	25	100	25	100	25	100
0.0	**0.3214**	0.2254	**0.5014**	0.3254	**0.4977**	0.3293
0.5	0.6780	**0.2054**	0.7421	0.3054	0.7412	**0.3167**
1.0	0.7510	0.2796	0.8510	**0.2996**	0.8170	0.3343

superior efficiency while maintaining competitive performance, confirming its practical advantage for real-time or large-scale imputation scenarios.



(a) Adult (b) Banknote (c) California

(d) Letter (e) Student (f) Average

Figure 4: RMSE Comparison with varying η values.

4.3 Ablation Study

To evaluate the practical utility of MissDDIM, we assess its sampling efficiency and output stability under varying inference-time configurations. We control sampling stochasticity via the DDIM noise parameter $\eta \in \{0.0, 0.5, 1.0\}$, where η=0 yields fully deterministic trajectories (MissDDIM), and larger values inject increasing levels of noise, transitioning toward standard DDPM behavior. The variance at each timestep τ_i is defined as: $\sigma_{\tau_i}(\eta) = \eta\sqrt{\frac{1-\alpha_{\tau_{i-1}}}{1-\alpha_{\tau_i}}} \cdot \sqrt{1-\frac{\alpha_{\tau_i}}{\alpha_{\tau_{i-1}}}}$, which allows us to isolate the impact of stochasticity while keeping model parameters fixed. We also vary the number of sampling steps $T \in \{25, 50, 75, 100\}$ to examine trade-offs between computational cost and imputation quality—reflecting latency-sensitive deployment scenarios. Table 3 and Figure 4 report RMSE and its standard deviation across multiple runs. As expected, MissDDIM (η=0) produces deterministic outputs with low variance. At lower T, smaller η values converge more rapidly and achieve competitive accuracy. As T increases, stochastic methods may slightly improve performance but at the cost of higher variance. Overall, MissDDIM strikes an effective balance between inference speed, output stability, and imputation fidelity.

5 Conclusion

We proposed MissDDIM, the first imputation framework that adapts deterministic DDIM sampling to tabular data. By reformulating DDIM in a conditional setting, MissDDIM supports incomplete inputs natively and enables efficient, stable imputation without repeated sampling. Unlike existing diffusion-based approaches that rely on stochastic DDPM processes, our method achieves consistent outputs with significantly reduced inference time. Experiments across diverse datasets demonstrate that MissDDIM delivers competitive or superior accuracy, while offering practical advantages in latency-sensitive and deployment-oriented scenarios. In future work, we plan to extend MissDDIM to support more complex missing data mechanisms and explore strategies for improving robustness under heterogeneous data type .

Disclosure of AI Tools

Generative AI tools were used to assist with language refinement in this manuscript. The authors manually revised all content.

Acknowledgment

This work is supported by the Air Force Office of Scientific Research under award number FA2386-23-1-4003 and Deakin University.

References

[1] Mohammed Ali Al-taezi, Yu Wang, Pengfei Zhu, Qinghua Hu, and Abdulrahman Al-Badwi. 2024. Improved generative adversarial network with deep metric learning for missing data imputation. *Neurocomputing* 570 (2024), 127062.

[2] Svetlana Bryzgalova, Sven Lerner, Martin Lettau, and Markus Pelger. 2025. Missing financial data. *The Review of Financial Studies* 38, 3 (2025), 803–882.

[3] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10850–10869.

[4] Aminu Da'u and Naomie Salim. 2020. Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review* 53, 4 (2020), 2709–2748.

[5] Jinghan Du, Minghua Hu, and Weining Zhang. 2020. Missing data problem in the monitoring system: A review. *IEEE Sensors Journal* 20, 23 (2020), 13984–13998.

[6] Sulagna Dutta and Pallav Sengupta. 2016. Men and mice: relating their ages. *Life sciences* 152 (2016), 244–248.

[7] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. 2023. Generative diffusion prior for unified image restoration and enhancement. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9935–9946.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 574, 12 pages.

[9] Rima Houari, Ahcène Bounceur, A Kamel Tari, and M Tahar Kecha. 2014. Handling missing data problems with sampling methods. IEEE, 2014 International conference on advanced networking distributed systems and applications, 99–104.

[10] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. TabDDPM: modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 725, 16 pages.

[11] Caihua Liu, Amir Talaei-Khoei, Didar Zowghi, and Jay Daniel. 2017. Data completeness in healthcare: a literature survey. *Pacific Asia Journal of the Association for Information Systems* 9, 2 (2017), 5.

[12] Pierre-Alexandre Mattei and Jes Frellsen. 2019. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*. PMLR, 4413–4423.

[13] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. 2020. Handling incomplete heterogeneous data using vaes. *Pattern Recognition* 107 (2020), 107501.

[14] Yidong Ouyang, Liyan Xie, Chongxuan Li, and Guang Cheng. 2023. MissDiff: Training Diffusion Models on Tabular Data with Missing Values. arXiv:2307.00467 [cs.LG] https://arxiv.org/abs/2307.00467

[15] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (2022), 84–90.

[16] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=St1giarCHLP

[17] Qinbao Song and Martin Shepperd. 2007. Missing data imputation techniques. *International journal of business intelligence and data mining* 2, 3 (2007), 261–291.

[18] Daniel J. Stekhoven and Peter Bühlmann. 2011. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (Oct. 2011), 112–118. doi:10.1093/bioinformatics/btr597

[19] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. CSDI: conditional score-based diffusion models for probabilistic time series imputation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 1900, 13 pages.

[20] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. 2023. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13095–13105.

[21] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*. PMLR, 5689–5698.

[22] Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. 2020. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems* 33 (2020), 19075–19087.

[23] Shichao Zhang. 2012. Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software* 85, 11 (2012), 2541–2552.

[24] Shuhan Zheng and Nontawat Charoenphakdee. 2022. Diffusion models for missing value imputation in tabular data. In *NeurIPS Table Representation Learning (TRL) Workshop*.

[25] Jiajun Zhong, Ning Gui, and Weiwei Ye. 2023. Data imputation with iterative graph reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11399–11407.

[26] Youran Zhou, Mohamed Reda Bouadjenek, and Sunil Aryal. 2024. Missing Data Imputation: Do Advanced ML/DL Techniques Outperform Traditional Approaches?. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, Albert Bifet, Tomas Krilavičius, Ioanna Miliou, and Slawomir Nowaczyk (Eds.). Springer Nature Switzerland, Cham, 100–115.

[27] Youran Zhou, Mohamed Reda Bouadjenek, Jonathan Wells, and Sunil Aryal. 2025. HI-PMK: A Data-Dependent Kernel for Incomplete Heterogeneous Data Representation. arXiv:2501.04300 [cs.LG] https://arxiv.org/abs/2501.04300