An Observed Value Consistent Diffusion Model for Imputing Missing Values in Multivariate Time Series

Xu Wang
University of Science and Technology of China
Hefei, China
wx309@mail.ustc.eud.cn

Hongbo Zhang
University of Science and Technology of China
Hefei, China
zhanghongbo@mail.ustc.edu.cn

Pengkun Wang
University of Science and Technology of China
Hefei, China
Suzhou Institute for Advanced Research, USTC
Suzhou, China
pengkun@mail.ustc.edu.cn

Yudong Zhang
University of Science and Technology of China
Hefei, China
zyd2020@mail.ustc.edu.cn

Binwu Wang
University of Science and Technology of China
Hefei, China
wbw1995@mail.ustc.edu.cn

Zhengyang Zhou
University of Science and Technology of China
Hefei, China
Suzhou Institute for Advanced Research, USTC
Suzhou, China
zzy0929@mail.ustc.edu.cn

Yang Wang*
University of Science and Technology of China
Hefei, China
Suzhou Institute for Advanced Research, USTC
Suzhou, China
angyan@ustc.eud.cn

ABSTRACT

Missing value, which is common in multivariate time series, is the most important obstacle towards the utilization and interpretation of those data. Great efforts have been employed on how to accurately impute missing values in multivariate time series, and existing works either use deep learning networks to achieve deterministic imputations or aim at generating different plausible imputations by sampling multiple noises from a same distribution and then denoising them. However, these models either fall short of modeling the uncertainties of imputations due to their deterministic nature or perform poorly in terms of interpretability and imputation accuracy due to their ignorance of the correlations between the latent representations of both observed and missing values which are parts of samples from a same distribution. To this end, in this paper, we explicitly take the correlations between observed and missing values into account, and theoretically re-derive the Evidence Lower BOund (ELBO) of conditional diffusion model in the scenario of multivariate time series imputation. Based on the newly derived ELBO, we further propose a novel multivariate imputation diffusion model (MIDM) which is equipped with novel noise sampling, adding and denoising mechanisms for multivariate time series imputation, and the series of newly designed technologies jointly ensure the involving of the consistency between observed and missing values. Extensive experiments on both the tasks of multivariate time series imputation and forecasting witness the superiority of our proposed MIDM model on generating conditional estimations.

CCS CONCEPTS

• **Mathematics of computing** → **Time series analysis**; *Variational methods.*

KEYWORDS

diffusion model, conditional generation, multivariate time series

Yang Wang is the corresponding author.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599257

ACM Reference Format:
Xu Wang, Hongbo Zhang, Pengkun Wang, Yudong Zhang, Binwu Wang, Zhengyang Zhou, and Yang Wang*. 2023. An Observed Value Consistent Diffusion Model for Imputing Missing Values in Multivariate Time Series. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3580305.3599257

1 INTRODUCTION

Massive multivariate time series data is generated in abundant real world scenarios including electricity load, traffic flow and weather condition. Fully analyzing and utilizing of these data can greatly facilitate people's lives. However, missing values resulted from various reasons such as device failures and human errors [19] are inevitable and will greatly hamper the utilization and interpretation of multivariate time series data. Therefore, time series imputation, which aims at addressing the issue of missing values by estimating missing values from observed ones, has become a recent research hotspot.

Many machine learning based time series imputation methods were firstly proposed in the past few years [20, 28, 37]. Then, with the rapid developments of deep learning technologies, many deep learning based efforts [8, 9, 13–15, 26, 36, 43, 44] were subsequently proposed to impute missing values in multivariate time series. Such recent deep learning based works can be divided into two categories, *deterministic methods* [8, 9, 13, 44] and *probabilistic methods* [14, 15, 26, 36, 43].

An intuitive implementation of deterministic model is to use Recurrent Neural Networks (RNNs) for sequence modeling [8, 9], where the missing values in input sequences are replaced with a specific token. [13] takes the correlations among different time series into consideration and combines RNN with Graph Convolution Network (GCN) together. And some other works [12, 31, 35] apply attention or self-training mechanisms to enhance the representative abilities of their models. Even though deterministic methods have achieved fair deterministic imputations, their deterministic nature determine that they are short in modeling the uncertainties of imputations. However, modeling uncertainties is of vital importance in various realistic scenarios as it allows the evaluation of the robustness of both imputation models and downstream algorithms [1, 29]. Thus, increasing efforts have been devoted to probabilistic methods in recent years.

Probabilistic models aim at generating different plausible imputations by sampling multiple noises from a same distribution and then denoising them, thus naturally address the uncertainty issue. Early attempts in probabilistic models [25, 43], which extend Generative Adversarial Networks (GANs) to the scenario of multivariate time series imputation, utilize adversarial learning to force their models to generate estimations of missing values following the distribution of training dataset. However, GAN-based methods are hard to be trained and with less interpretability. For seeking interpretability of probabilistic models, some recent works [14, 15] utilize Variational AutoEncoders (VAEs) as the cores of their models and achieves comparable performance to those GAN-based methods. These works utilize variational inference to derive the Evidence Lower BOund (ELBO) of conditional distribution of missing values with regard to given observations, thus providing theoretical supports for their interpretability. However, even compared with GAN based models, such models still perform poorly in generating more high-quality estimations of missing values. More recently, some works make attempts to extend diffusion models [1] to the problem of multivariate imputation and propose many novel models for generating conditional distribution of missing values given observations. Specifically, CSDI [36] learns conditional distribution with conditional score-based diffusion model [18, 33] by feeding observed values into denoising module of the diffusion model, and SSSD^{S4} [1] applies state space model [17] as the denoising module of Diffwave diffusion models [21] to achieve imputation. Nevertheless, existing imputation diffusion models, which borrow the idea from existing diffusion models targeting other tasks, naturally follow the diffusion processes of those existing diffusion models and **ignore the existence of the correlations between the latent representations of both observed and missing values which are parts of samples from a same distribution.** Therefore, using existing imputation diffusion models to construct conditional distribution of missing values with regard to given observed values is **necessarily unsatisfactory in approximating real conditional distribution.**

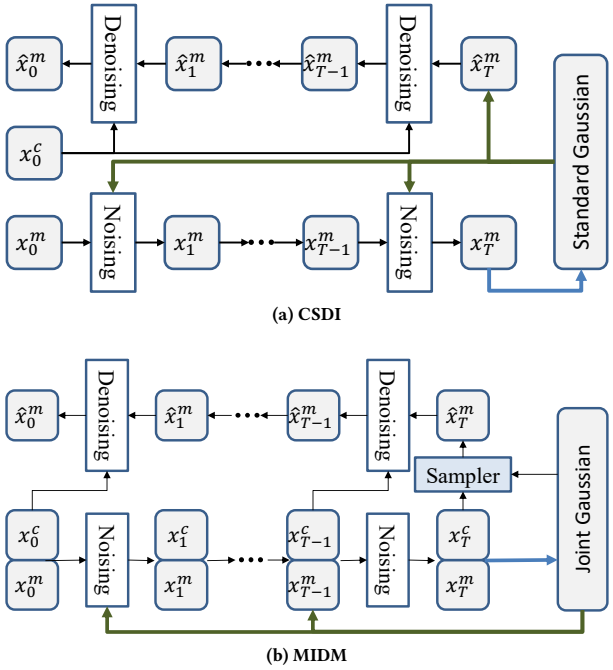


Figure 1: Comparison of diffusion processes in CSDI and MIDM. In this figure, x_t^c and x_t^m correspond to representations of observed and missing values at t-th step of diffusion process. In case $t = 0$, x_t^c and x_t^m correspond to original data. $\hat{x}_t^m$ corresponds to estimation of representations of missing values. Green arrow means sampling noise from distribution and blue arrow means following the distribution.

[1] Diffusion models [11, 34], which are built on similar theoretical basis to VAEs, are widely used in various generation tasks such as audio and image synthesis by recovering data from noise through denoising.

In this paper, considering the fact that diffusion model process data with the guidance of ELBO and previous derived ELBO [14] has never taken the consistency between observed and missing values into account, we here first theoretically re-derive the ELBO of conditional distribution of missing values in the context of involving the consistency between observed and missing values. Based on the re-derived new ELBO, we redesign the process of diffusion model and propose a novel Multivariate time series Imputation Diffusion Model (MIDM) which has different noise sampling, adding and removing processes than vanilla diffusion models. We compare the noise sampling and denoising processes of the proposed MIDM with those of vanilla diffusion models (Select CSDI [36] as the representative) in Figure 1. As illustrated, rather than sampling from a standard Gaussian distribution as in CSDI, MIDM samples noise from conditional distribution generated according to the latent representation of observed values. Similarly, at each step of noise adding process, MIDM add noise sampled from a joint Gaussian distribution with non-identical covariance matrix rather than a standard Gaussian. Note that in MIDM, the newly derived ELBO requests that the diffused observations should be used as conditional input for each step of denoising the noises corresponding to missing values. We verify the effectiveness of MIDM on several widely used multivariate imputation datasets. Note that multivariate time series forecasting is a special case of imputation where all values in the last several time steps are missing, we also evaluate our model on multivariate forecasting task. The empirical results show that MIDM achieves state-of-the-art performances on both the tasks of multivariate time series imputation and forecasting.

The main contributions of this paper can be concluded as follows,

- We derive the new theoretical Evidence Lower BOund of applying diffusion models on the problem of multivariate time series imputation by taking the consistency between observed and missing values into account.
- Based on the newly derived ELBO, we propose a brand new diffusion model equipped with novel noise sampling, adding and denoising mechanisms for multivariate time series imputation, and the series of newly designed technologies jointly ensure the involving of the consistency between observed and missing values.
- We evaluate the proposed MIDM on several widely used multivariate imputation and forecasting datasets. The empirical results show that MIDM achieves state-of-the-art performance on both multivariate time series imputation and forecasting.

2 RELATED WORK AND PRELIMINARY

2.1 Related work

Missing values are common in multivariate time series data. For fully utilizing and analyzing multivariate time series data, great efforts have been devoted to multivariate time series imputation for many years. Early works address imputation issue with machine learning methods. MICE [40] utilities chained equation to fill the missing values. Some representative autoregressive models, such as ARIMA [7] and VAR [47], can be used to impute missing values. [20] selects k nearest neighbors and uses the average of values of the neighbors to fill missing values. MF [27] factorizes the incomplete dataset into low-rank matrices and adopts the product of these two matrices to impute missing values. Those methods struggle to fit large datasets and the imputation accuracy of them is limited. With the rapid developments of deep learning, researchers tend to tackle the issue of missing values with deep models, which can be divided into deterministic models and probabilistic models.

Deterministic models generate determined imputation given observed values. [6] proposes bidirectional recurrent neural networks for efficiently estimating missing values. GRUD [9] imputes missing values by utilizing the last observation and mean of observations to represent missing patterns. BRITS [8] build a bidirectional recurrent dynamical system which directly learns the missing values. Both GRUD and BRITS take as input sequences with missing values where missing values are replaced with specific tokens, and output imputed sequences. To enhance the representative ability, [12] applies self-training mechanism for multivariate imputation. Since attention based model have achieved impressive performance on sequence model, [31, 35] proposes attention based multivariate imputation models. Considering the correlations among different variables, [13] combines RNN with Graph Convolution Network (GCN) together for generating more accurate imputation. While those deterministic methods have achieved fair performance on imputation, they fall short of modeling the uncertainties of their imputations. Considering the importance of modeling uncertainties in various realistic scenarios [1, 29], increasing efforts have been devoted to probabilistic methods in recent years.

Early attempts in probabilistic extends GAN [16, 39] to multivariate imputation. GAIN [43] proposes a general data imputation model which uses hint vectors conditioned on observed values for helping generate imputation. [24] proposes a two-stage GAN based imputation model, which consumes lots of time to train. E^2*GAN* [25] proposes a compressing and reconstructing strategy to skip the second stage in [24] and thus save the time consumption. SSGAN [26] proposes a semi-supervised classifier to tackle the issue of insufficient label. GAN-based methods are hard to be trained and with less interpretability. For seeking interpretability of probabilistic models, some recent works [14, 15] utilize Variational AutoEncoders (VAEs) to address the problem of multivariate imputation. Recently, diffusion models [11, 34] have achieved impressive performance on various generation tasks such as audio and image synthesis. Many researchers try to extend diffusion model to multivariate imputation and have achieved state-of-the-art performances. Specifically, CSDI [36] tries to learn conditional distribution with conditional score-based diffusion model [18, 33] by feeding observed values into the denoising module of their diffusion model. D^3VAE [22] equips a bidirectional VAE with diffusion model for multivariate forecasting problem, where the noise for generating imputation is sampled by the bidirectional VAE. SSSD^{S4} [1] applies recent advanced state space model [17] as the denoising module and combine it with Diffwave [21] to achieve imputation. Nevertheless, those diffusion based model ignore the nature of multivariate time series, that both observed and missing values are part of samples coming from the same data distribution, and thus fail to approximating conditional distribution of missing values given observations well.

2.2 Problem definition

In this paper, we focus on multivariate time series imputation problem, where we aim at imputing missing values in multivariate time series based on observed values of those series. The multivariate time series we are interested in contains several time series of the same length, each of which contains missing values. We here formally define the problem of multivariate time series imputation.

Definition 2.1. The multivariate time series containing N time series with length L can be denoted as $X \in \mathbb{R}^{N\times L}$*, where observed values* X^c *and missing values* X^m *are given by a mask matrix* $M \in [0,1]^{N\times L}$*, i.e.,* $X^c = \{X^{i,j}|M^{i,j}=1\}$ *and* $X^m = \{X^{i,j}|M^{i,j}=0\}$*. And the problem of multivariate time series imputation can be defined as,*

$$\max_{\theta} p_\theta(X^m|X^c) \quad where \quad (X^m, X^c) \sim p(X) \tag{1}$$

where p_θ *is a probabilistic imputation model aiming at approximating the real conditional distribution* $p(X^m|X^c)$*.*

3 METHOD

We aim at approximating conditional distribution $p(X^m|X^c)$ based on the premise that observed and missing values follow a joint distribution, i.e., $(X^m, X^c) \sim p(X)$. To achieve this goal, we first derive the evidence lower bound of $p(X^m|X^c)$, since none of existing work consider conditional distribution $p(X^m|X^c)$ with constraint as $(X^m, X^c) \sim p(X)$. With the light of derived ELBO, we are able to design the noising, sampling and denoising process for our Multivariate time series Imputation Diffusion Model (MIDM). In this section, we first derive the evidence lower bound of conditional distribution $p(X^m|X^c)$, then we detail the design of MIDM and finally we detail our error estimation model for denoising.

3.1 ELBO of conditional distribution

We start from deriving ELBO of the conditional distribution $p(X^m|X^c)$. Typically, a diffusion model noise data into a standard Gaussian distribution by gradually adding standard Gaussian noise in T steps, which can be formulated as the following Markov chain:

$$q(X_{1:T}|X_0) = \prod_{t=1}^{T} q(X_t|X_{t-1}) \tag{2}$$

where X_t corresponds to latent representation of X at t step, X_0 is the original data and equivalent to X. Our goal is to approximating conditional distribution of missing values given observed values,

$$\log p(X_0^m|X_0^c) = \log \frac{p(X_0^m, X_0^c)}{p(X_0^c)} \tag{3}$$

Since we have $(X^m, X^c) \sim p(X)$, the ELBO of the conditional distribution can be written as,

$$\begin{aligned} &\log p(X_0^m|X_0^c) = \log \frac{p(X_0)}{p(X_0^c)} \\ &\geq \mathbb{E}_{q(X_{1:T}|X_0)}[\log \frac{p(X_{0:T})}{q(X_{1:T}|X_0)} + \log \frac{q(X_{1:T}^c|X_0^c)}{p(X_{0:T}^c)}] \end{aligned} \tag{4}$$

For brevity, we denote $\mathbb{E}_{q(X_{1:T}|X_0)}$ as $\mathbb{E}$. Consider $\mathbb{E}[\log \frac{p(X_{0:T})}{q(X_{1:T}|X_0)}]$, we have,

$$\begin{aligned} &\mathbb{E}[\log \frac{p(X_{0:T})}{q(X_{1:T}|X_0)}] = \mathbb{E}[\log \frac{p(X_T)\prod_{t=1}^{T} p_\theta(X_{t-1}|X_t)}{\prod_{t=1}^{T} q(X_t|X_{t-1})}] \\ &= \mathbb{E}[\log \frac{p(X_T)p_\theta(X_0|X_1)\prod_{t=2}^{T} p_\theta(X_{t-1}|X_t)}{q(X_1|X_0)\prod_{t=2}^{T} q(X_t|X_{t-1})}] \\ &= \mathbb{E}[\log \frac{p(X_T)p_\theta(X_0|X_1)\prod_{t=2}^{T} p_\theta(X_{t-1}|X_t)}{q(X_1|X_0)\prod_{t=2}^{T} q(X_t|X_{t-1}, X_0)}] \\ &= \mathbb{E}[\log \frac{p_\theta(X_T)p_\theta(X_0|X_1)}{q(X_1|X_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(X_{t-1}|X_t)}{q(X_t|X_{t-1}, X_0)}] \\ &= \mathbb{E}[\log \frac{p(X_T)p_\theta(X_0|X_1)}{q(X_1|X_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(X_{t-1}|X_t)}{\frac{q(X_{t-1}|X_t,X_0)q(X_t|X_0)}{q(X_{t-1}|X_0)}}] \\ &= \mathbb{E}[\log \frac{p(X_T)p_\theta(X_0|X_1)}{q(X_1|X_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(X_{t-1}|X_t)}{\frac{q(X_{t-1}|X_t,X_0)q(X_t+X_0)}{q(X_{t-1}+X_0)}}] \\ &= \mathbb{E}[\log \frac{p(X_T)p_\theta(X_0|X_1)}{q(X_1|X_0)}\frac{q(X_1|X_0)}{q(X_T|X_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(X_{t-1}|X_t)}{q(X_{t-1}|X_t, X_0)}] \\ &= \mathbb{E}[\log \frac{p(X_T)p_\theta(X_0|X_1)}{q(X_T|X_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(X_{t-1}|X_t)}{q(X_{t-1}|X_t, X_0)}] \end{aligned} \tag{5}$$

Therefore, the terms in Eq 4 can be transferred as,

$$\begin{aligned} \log \frac{p(X_{0:T})}{q(X_{1:T}|X_0)} &= \log \frac{p(X_T)p_\theta(X_0|X_1)}{q(X_T|X_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(X_{t-1}|X_t)}{q(X_{t-1}|X_t, X_0)} \\ \log \frac{p(X_{0:T}^c)}{q(X_{1:T}^c|X_0^c)} &= \log \frac{p(X_T^c)p_\theta(X_0^c|X_1^c)}{q(X_T^c|X_0^c)} + \sum_{t=2}^{T} \log \frac{p_\theta(X_{t-1}^c|X_t^c)}{q(X_{t-1}^c|X_t^c, X_0^c)} \end{aligned} \tag{6}$$

And the ELBO can be further transferred as,

$$\begin{aligned} \log p(X_0^m|X_0^c) = \mathbb{E}_{q(X_{1:T}|X_0)}[&\log \frac{p(X_T)p_\theta(X_0|X_1)q(X_T^c|X_0^c)}{p(X_T^c)p_\theta(X_0^c|X_1^c)q(X_T|X_0)} \\ &+ \sum_{t=2}^{T} \log \frac{p_\theta(X_{t-1}|X_t)q(X_{t-1}^c|X_t^c, X_0^c)}{q(X_{t-1}|X_t, X_0)p_\theta(X_{t-1}^c|X_t^c)}] \end{aligned} \tag{7}$$

Applying $X_t = (X_t^c, X_t^m)$ to above equation, we finally achieve the ELBO as,

$$\begin{aligned} &\log p(X_0^m|X_0^c) \geq \mathbb{E}_{q(X_{1:T}|X_0)}[\log \frac{p(X_{0:T})}{q(X_{0:T}|X_0)}\frac{p(X_{0:T}^c)}{q(X_{0:T}^c|X_0^c)}] \\ &= \mathbb{E}_{q(X_1|X_0)}[\log p_\theta(X_0^m|X_1, X_0^c)] \\ &- D_{KL}(q(X_T^m|X_0, X_T^c)||p(X_T^m|X_T^c)) \\ &- \sum_{t=1}^{T} \mathbb{E}_{q(X_t|X_0)}[D_{KL}(q(X_{t-1}^m|X_t, X_0)||p_\theta(X_{t-1}^m|X_t, X_{t-1}^c))] \end{aligned} \tag{8}$$

3.2 Multivariate imputation diffusion model

Noising. To implement multivariate imputation diffusion model based on Eq 8, we need first determine the noising process, i.e.,

$\prod_{t=1}^{T} q(X_t|X_{t-1})$. In existing diffusion models, such process is defined as,

$$q(X_t|X_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}X_{t-1}, (1-\alpha_t)\mathbf{I}) \tag{9}$$

where α_t evolves over time according to a fixed or learnable schedule such that the distribution of the final latent $p(X_T)$ is nearly a standard Gaussian when T is large. Eq 9 is the key of noising process and is implemented by reparameterization trick as,

$$X_t = \sqrt{\alpha_t}X_{t-1} + \sqrt{(1-\alpha_t)}\epsilon \tag{10}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is a random noise term sampled from standard Gaussian. As mentioned, the prior distribution $p(X_T)$ is always a standard Gaussian, which results in independence of X_T^m and X_T^c. Noting that we have a conditional prior term $p(X_T^m|X_T^c)$ in Eq 8, such independence makes the conditional term meaningless. Therefore, we need to reconsider the noising process so that the conditional prior term $p(X_T^m|X_T^c)$ is meaningful and tractable.

A feasible way is to utilize Gaussian processes to model the latent distribution $p(X_T)$ which have been utilized in abundant studies for modeling time series [4, 5, 15, 30]. A Gaussian process is a collection of random variables, where any finite number of variables have joint gaussian distributions. To utilize Gaussian processes to model time series, the key is to build proper mean and kernel functions. Once the mean and kernel functions are determined, the conditional distribution $p(X_T^m|X_T^c)$ can be achieved. For instance, when the mean and kernel functions provide a joint distribution of X_T^m and X_T^c as,

$$\begin{bmatrix} x_T^c \\ x_T^m \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_c \\ \mu_m \end{bmatrix}, \begin{bmatrix} \Sigma_{cc} & \Sigma_{cm} \\ \Sigma_{mc} & \Sigma_{mm} \end{bmatrix}\right) \tag{11}$$

Given observation of x_T^c, we can build the conditional distribution as,

$$p(x_T^m|x_T^c) = \mathcal{N}(\mu_m + \Sigma_{mc}\Sigma_{cc}^{-1}(x_T^c - \mu_c), \Sigma_{mm} - \Sigma_{mc}\Sigma_{cc}^{-1}\Sigma_{cm}) \tag{12}$$

Since we are interested of the correlations between X_T^m and X_T^c, we simply set mean as $\mathbf{0}$ and focus on the kernel function. Actually, since we always have data with the same dimensionality, finding a proper kernel function is equal to finding proper covariance matrix. Noting in Eq 12, there is an operation of invert covariance matrix Σ_{cc}, which generally has a time complexity of $O(n^3)$, we here apply a special form of covariance matrix for seeking of efficiency. Specifically, we have,

$$\Sigma = K^T K, \quad where \quad K_{ij} = \begin{cases} k_{ij} & j \in \{i, i+1\} \\ 0 & otherwise \end{cases} \tag{13}$$

where k_{ij}s are learnable parameter and K is an upper triangular band matrix which makes Σ positive definite, symmetric, and tridiagonal. The time complexity of inverting matrices with such form can be reduced to linear [4]. So far, we have joint distribution of observed values and missing values in latent space,

$$\begin{bmatrix} x_T^c \\ x_T^m \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \Sigma = \begin{bmatrix} \Sigma_{cc} & \Sigma_{cm} \\ \Sigma_{mc} & \Sigma_{mm} \end{bmatrix}\right) \tag{14}$$

And profiting from Eq 13, we can determine the conditional distribution $p(x_T^m|x_T^c)$ with an accepted time consumption. To achieve the latent distribution as Eq 15, we only need to sample ϵ in Eq 10 from $\mathcal{N}(\mathbf{0}, \Sigma)$ rather than $\mathcal{N}(\mathbf{0}, \mathbf{I})$ at each step of the forward process of MIDM. By adding such noise as in Eq 10, we transfer original multivariate time series to a latent space, where the latent representations of the data follows a Gaussian distribution as,

$$X_T \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{15}$$

Noise sampling. As in Eq 8, the desired noise distribution is given by $p(X_T^m|X_T^c)$, which can be calculated according to Eq 12. Therefore, given observed values X_0^c, to sample noise for recovering missing values, we first transfer X_0^c to latent space to get X_T^c. Then following Eq 12, we can achieve posterior conditional distribution $p(X_T^m|X_T^c)$ and sample noise from it.

Denoising. MIDM generates estimation by gradually denoising the noise sampled from $p(X_T^m|X_T^c)$. The denoising process is supervised by the last term of Eq 8, which suggests to minimize KL-divergence between $q(X_{t-1}^m|X_t, X_0)$ and $p_\theta(X_{t-1}^m|X_t, X_{t-1}^c)$. $p_\theta(X_{t-1}^m|X_t, X_{t-1}^c)$ is the learnable estimator with parameter θ. Thus, we need to calculate the form of $q(X_{t-1}^m|X_t, X_0)$ so as to compute the KL-divergence. Due to X_{t-1}^m corresponds to missing part of X_{t-1}, calculating $q(X_{t-1}^m|X_t, X_0)$ is equivalent to calculating $q(X_{t-1}|X_t, X_0)$. And we have

$$\begin{aligned} q(X_{t-1}|X_t, X_0) &= \frac{q(X_t|X_{t-1}, X_0)q(X_{t-1}|X_0)}{q(X_t|X_0)} \\ &= \frac{\mathcal{N}(\sqrt{\alpha_t}X_{t-1}, (1-\alpha_t)\Sigma)\mathcal{N}(\sqrt{\overline{\alpha}_{t-1}}, (1-\overline{\alpha}_{t-1})\Sigma)}{\mathcal{N}(\overline{\alpha}_t X_0, (1-\overline{\alpha}_t)\Sigma)} \\ &\propto \mathcal{N}(\mu(X_t, X_0), \Sigma(t)) \end{aligned} \tag{16}$$

where $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. $\Sigma(t)$ only depends on t and is prefixed given the schedule of α_t and Σ in Eq 15. $\mu(X_t, X_0)$ can be derived as,

$$\mu(X_t, X_0) = \frac{1}{\sqrt{\alpha_t}} - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}\sqrt{\alpha_t}}\epsilon \tag{17}$$

In order to minimize the KL-divergence between $q(X_{t-1}^m|X_t, X_0)$ and $p_\theta(X_{t-1}^m|X_t, X_{t-1}^c)$, we make $p_\theta(X_{t-1}^m|X_t, X_{t-1}^c)$ a Gaussian distribution $\mathcal{N}(\mu_\theta(X_t, t, X_{t-1}^c), \Sigma(t)$ with the following specific parameterization,

$$\mu_\theta(X) = \frac{1}{\sqrt{\alpha_t}}X_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}\sqrt{\alpha_t}}\epsilon_\theta(X_t, t, X_{t-1}^c) \tag{18}$$

Therefore, the backward process can be trained by solving the optimization problem,

$$\min_\theta ||(1-M) \otimes (\epsilon - \epsilon_\theta(x_t, t))||_2^2 \tag{19}$$

where M is the mask denoting the position of missing values and $\otimes$ means element-wise product. The denoising function with parameter θ estimates the noise term ϵ added in the noising process. And during inferring, we first sample a random noise x_T from standard Gaussian, then gradually denoise x_T by removing the estimated noise $\epsilon_\theta(x_t, t)$ at each step t and finally get generated data x_0.

3.3 Error estimation model

In this part, we detail the proposed error estimation model (EEM). The proposed error estimation model takes X_t and X_{t-1}^c as input and estimates the error term in Eq 17 at step t. The architecture of proposed EEM is shown in Fig 2, EEM has three components, step embedding module, temporal embedding module and attention based estimator.

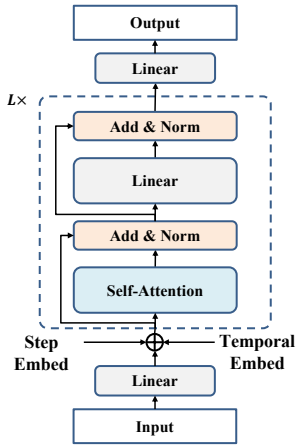


Figure 2: Architecture of error estimation model.

3.3.1 Step embedding module. Since at different steps, the values of X_t and X^c_{t-1} are quite different, EMM needs to know the current step t for generating accurate error estimations. Step embedding module is designed for denoting the steps of different inputs. Step embedding module maintains T learnable N-dimension embedding vectors for all steps, and when given step t, step embedding module outputs a N-dimension vector.

3.3.2 Temporal embedding module. The attention mechanism utilized in EEM is agnostic to the order of features in the input series, while the order information is critical in time series analysis. Thus, for more accurate estimation, we need to explicitly induce the order information to the attention module. Adding temporal embedding to the input feature is a common practice. In our temporal embedding module, fixed positional embedding is applied [38]. For an input $X \in \mathbb{L} \times \mathbb{N}$, the temporal embedding TE is calculated as,

$$\begin{aligned} TE(t, 2i) &= \sin(t/10000^{2i/d}) \\ TE(t, 2i+1) &= \cos(t/10000^{2i/d}) \end{aligned} \tag{20}$$

3.3.3 Attention based estimator. We build EEM with an attention based estimator, as attention is a promising mechanism for modeling series data. The input X_t and X^c_{t-1} are first summed and we denote the summation result as $X \in \mathbb{R}^{N \times L}$. The estimator transposes X to $L \times N$, then calculates the temporal embedding of the transposed input and sum up temporal embedding with step embedding and transposed input.

$$Z_0 = X^T + SE + TE \tag{21}$$

where SE and TE denote step embedding and temporal embedding respectively. And $Z_0 \in \mathbb{R}^{L \times N}$ is the summation result. Then Z_0 is fed into several stacked attention layers. In each layer, we apply a linear projection on Z_0 and map Z_0 to $Z_1 \in \mathbb{R}^{L \times 512}$. We further apply self-attention on Z_1, which can be formulated as,

$$\text{Att}(Z_1) = \text{softmax}(\frac{(Z_1 W_Q)(Z_1 W_K)^T}{\sqrt{d_{model}}}) Z_1 W_V \tag{22}$$

where d_{model} is set to 512. We then add $\text{Att}(Z_1)$ with Z_1 and normalize the result as,

$$Z_2 = \text{Norm}(Z_1 + \text{Att}(Z_1)) \tag{23}$$

Layer normalization [2] is utilized here. Then we have another round of residual connection and normalization.

$$Z_3 = \text{Norm}(Z_2 + \text{Linear}(Z_2)) \tag{24}$$

Z_3 is then mapped back to $\mathbb{R}^{L \times N}$ by a linear projection. Finally, we apply another transpose operation to get the error estimation $\epsilon_\theta \in \mathbb{R}^{N \times L}$.

$$\epsilon_\theta = \text{Linear}(Z_3)^T \tag{25}$$

4 EXPERIMENTS

In this section, we evaluate the performance of MIDM on multivariate time series imputation on four datasets with different settings.

4.1 Datasets and experimental settings

We apply four datasets to evaluate the proposed MIDM, including, 1) **AQI [46]** Air Quality Index dataset includes air quality data of 437 monitoring stations located in 43 Chinese cities over a period of one year (from May 2014 to April 2015). The sampling interval of air quality is one hour, and each air quality record in this dataset consists of the concentrations of six different air pollutants, where only PM2.5 is considered. There are around 26% missing values in AQI dataset. 2) **AQI-36 [41]** AQI-36 is a reduced version of AQI, which contains only records collected by 36stations in Beijing. Also, we consider only the PM2.5 pollutant. In AQI-36, there are around 13% missing values. 3) **PEMS-BAY [23]** This dataset comes from the California Department of Transportation (Caltrans) Performance Measurement System (PeMS) [10] and is collected by 325 sensors in the Bay Area over a period of 6 months from Jan 1st 2017 to May 31st 2017 with time interval 5 minutes. 0.02% of the data in PEMS-BAY is missing. 4) **CER-E [13]** CER-E is a subset of data from Irish Commission for Energy Regulation Smart Metering Project [2]. CER-E contains 485 time series sampled every 30 minutes, which record the energy consumption of small and medium-sized enterprises. 0.04% of the data in CER-E is missing.

Table 1: Statistics of dataseets.

Data	Missing Rate(%)	Variables
AQI	25.67	437
AQI(processed)	31.34	437
AQI	13.24	36
AQI(processed)	24.57	36
PEMS-BAY	0.02	325
PEMS-BAY(Point)	25.02	325
PEMS-BAY(Block)	9.1	325
CER-E	0.04	486
CER-E(Point)	25.01	486
CER-E(Block)	8.42	486

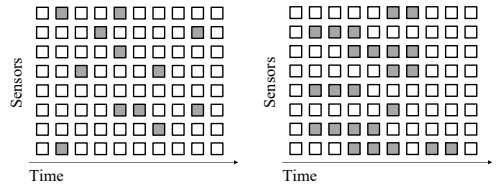
[2]https://www.ucd.ie/issda/data/commissionforenergyregulationcer/

Table 2: Imputation performance comparison on AQI and AQI-36. Performance averaged over 5 runs.

Model	AQI-36			AQI		
	MAE	MSE	MRE(%)	MAE	MSE	MRE(%)
KNN	30.21±0.00	2892.31±00.00	43.36±0.00	34.10±0.00	3471.14±00.00	51.02±0.00
MF	30.54±0.26	2763.06±63.35	43.84±0.38	26.74±0.24	2021.44±27.98	40.01±0.35
MICE	29.89±0.11	2575.53±07.67	42.90±0.15	26.39±0.13	1872.53±15.97	39.49±0.19
VAR	13.16±0.21	513.90±12.39	18.89±0.31	18.13±0.84	918.68±56.55	27.13±1.26
rGAIN	12.23±0.17	393.76±12.66	17.55±0.25	17.69±0.17	861.66±17.49	26.48±0.25
BRITS	12.24±0.26	495.94±43.56	17.57±0.38	17.24±0.13	924.34±18.26	25.79±0.20
GRIN	10.51±0.28	371.47±17.38	15.09±0.40	13.10±0.08	615.80±10.09	19.60±0.11
GP-VAE	14.11±0.24	483.91±24.36	18.43±0.45	17.84±0.16	893.27±20.39	27.46±0.19
CSDI	9.60±0.14	372.49±16.90	15.49±0.37	11.37±0.12	589.31±11.20	18.26±0.24
MIDM	**9.41±0.20**	**361.28±21.33**	**14.87±0.41**	**10.06±0.11**	**562.84±12.01**	**16.87±0.19**

For the air quality datasets AQI and AQI-36, we follow the protocol proposed by previous works [8, 41], i.e., the 3^{rd}, 3^{rd}, 3^{rd} and 3^{rd} months are used as test data and the other months as the training data. Since we have no ground truth of the original missing values in AQI and AQI-36, the protocol constructs a structured missing pattern by eliminating some observed values and the model is required to impute the eliminated values. We select time series of length $L = 24$ and $L = 36$ for AQI and AQI-36 respectively.

PEMS-BAY and CER-E contains much fewer missing values than AQI and AQI-36, which allows us to extract more flexible settings. Specifically, we mask the two datasets for evaluation by considering two different missing patterns: 1) Point missing, we randomly mask out 25% data points in the datasets. 2) Block missing, 5% of the available data is randomly dropped for each sensor at each time step. And for each missing position, there is a 0.15% probability that it is persistent and its persistent duration is sampled uniformly in the interval $[min, max]$, where min, max are set to 1 and 4 hours in PEMS-BAY, 2 hours and 2 days for CER-E. Fig 3 illustrates both missing patterns. All the dropped values are used as ground truth for training and testing. For both PEMS-BAY and CER-E, we use input sequences of 24 steps, which correspond to 2 hours and 12 hours of data respectively. We split both PEMS-BAY and CER-E into three folds, 70% for training, 10% for validation and 20% for testing. Tab 1 shows some statistics of the original datasets and processed datasets, where *Point* and *Block* correspond to point missing and block missing respectively.



(a) Point Missing **(b) Block Missing**

Figure 3: Illustration of Point Missing and Block Missing.

Three metrics are extracted for evaluation, i.e., mean absolute error (MAE), mean squared error (MSE) and mean relative error, which are defined as,

$$\begin{aligned} MAE(\mathbf{X}, \hat{\mathbf{X}}) &= mean(sum(|\mathbf{X} - \hat{\mathbf{X}}|)) \\ MSE(\mathbf{X}, \hat{\mathbf{X}}) &= mean(sum((\mathbf{X} - \hat{\mathbf{X}})^2)) \\ MRE(\mathbf{X}, \hat{\mathbf{X}}) &= mean(sum(|\frac{\mathbf{X} - \hat{\mathbf{X}}}{\mathbf{X}}|)) \end{aligned} \tag{26}$$

4.2 Baselines

We compare the proposed MIDM model with several widely-used baselines and state-of-the-art models, including, 1) **KNN [20]**, using the averaging values of the $k = 10$ most correlated variates as imputation. 2) **MICE [40]**, creating multiple imputations with chained equations. The maximum number of iterations is limited to 100 and the number of nearest features is limited to 10. 3) **MF [27]** factorizes the incomplete dataset into two low-rank matrices and adopts the product to impute missing values. 4) **VAR [47]** a Vector Autoregressive one-step-ahead predictor. 5) **rGAIN [26]** can be seen as GAIN [43] with bidirectional recurrent encoder and decoder. 6) **BRITS [8]** using recurrent dynamics to impute the missing values in multivariate time series. 7) **GP-VAE [15]** a deep probabilistic model for multivariate time series imputation, combining ideas from variational autoencoders and Gaussian processes. 8) **CSDI [36]** a recent approach to impute multivariate time series with conditional diffusion models.

4.3 Results

Tab 2 shows the experimental results on AQI and AQI-36 datasets. As shown, the proposed MIDM achieves the best imputation performance on all the three metrics on both AQI and AQI-36 datasets. Specifically, MIDM decrease MAE, MSE and MRE w.r.t. the baseline with closest performance to MIDM by $\{2.0\%, 2.7\%, 1.5\%\}$ and $\{11.5\%, 4.5\%, 7.6\%\}$ on AQI and AQI-36 respectively. Additionally, the divergence of imputations generated by MIDM over 5 runs is larger than CSDI on the two datasets. The reason is that, the denoising process of CSDI takes original observations as input which are deterministic while that of MIDM takes latent representations of observations which are probabilistic. The experimental results on AQI and AQI-36 prove the superiority of MIDM on multivariate time series imputation.

Table 3: Imputation performance on PEMS-BAY and CER-E with different missing patterns. Performance averaged over 5 runs.

Dataset	Model	Block missing			Point missing		
		MAE	MSE	MRE(%)	MAE	MSE	MRE(%)
PEMS-BAY	KNN	4.30±0.00	49.90±0.00	6.90±0.00	4.30±0.00	49.80±0.00	6.88±0.00
	MF	3.28±0.01	50.14±0.13	5.26±0.01	3.29±0.01	51.39±0.64	5.27±0.02
	MICE	2.94±0.02	28.28±0.37	4.71±0.03	3.09±0.02	31.43±0.41	4.95±0.02
	VAR	2.09±0.10	16.06±0.73	3.35±0.16	1.30±0.00	6.52±0.01	2.07±0.01
	rGAIN	2.18±0.01	13.96±0.20	3.50±0.02	1.88±0.02	10.37±0.20	3.01±0.04
	BRITS	1.70±0.01	10.50±0.07	2.72±0.01	1.47±0.00	7.94±0.03	2.36±0.00
	GRIN	1.14±0.01	6.60±0.10	1.83±0.02	0.67±0.00	1.55±0.01	1.08±0.00
	GP-VAE	2.39±0.03	14.81±0.15	4.32±0.02	1.92±0.01	12.43±0.08	3.67±0.02
	CSDI	1.16±0.01	7.02±0.09	1.96±0.01	0.83±0.00	1.79±0.00	1.42±0.00
	MIDM	**1.03±0.01**	**5.83±0.11**	**1.77±0.02**	**0.60±0.00**	**1.54±0.02**	**0.93±0.00**
CER-E	KNN	1.15±0.00	6.53±0.00	56.11±0.00	1.22±0.00	7.23±0.00	57.71±0.00
	MF	0.97±0.01	4.38±0.06	47.20±0.31	1.01±0.01	4.65±0.07	47.87±0.36
	MICE	0.96±0.01	3.08±0.03	46.65±0.44	0.98±0.00	3.21±0.04	46.59±0.23
	VAR	0.64±0.03	1.75±0.06	31.21±1.60	0.53±0.00	1.26±0.00	24.94±0.02
	rGAIN	0.74±0.00	1.77±0.02	36.06±0.14	0.71±0.00	1.62±0.02	33.45±0.16
	BRITS	0.64±0.00	1.61±0.01	31.05±0.05	0.64±0.00	1.59±0.01	30.07±0.11
	GRIN	0.42±0.00	1.07±0.01	20.24±0.04	0.29±0.00	0.53±0.00	13.71±0.03
	GP-VAE	0.76±0.01	1.80±0.02	35.92±0.23	0.81±0.00	1.63±0.02	32.99±0.14
	CSDI	0.41±0.00	1.12±0.00	19.38±0.05	0.26±0.00	0.49±0.01	13.62±0.04
	MIDM	**0.39±0.01**	**0.99±0.00**	**19.21±0.04**	**0.23±0.00**	**0.47±0.00**	**12.89±0.03**

As mentioned, we extract experiments on PEMS-BAY and CER-E with different synthesized missing patterns, i.e., block missing and point missing. The experimental results are shown in Tab 3. The proposed MIDM still achieves more accurate imputation than compared baselines on both PEMS-BAY and CER-E with respect to both missing patterns. Specifically, for PEMSE-BAY, MIDM achieves {9.6%, 11.7%, 3.3%} and {10.4%, 0.6%, 13.9%} gain on MAE, MSE and MRE than that of the best baseline under the block missing and point missing settings respectively. And for CER-E, the three metrics of MIDM, i.e., MAE, MSE and MRE, decreases 4.9%, 7.5%, 0.8% and 11.5%, 4.1%, 5.4% compared to the best baseline under the block missing and point missing settings respectively. Such results further indicates the generalizability of MIDM on dealing with various missing patterns.

4.4 Analysis

Table 4: Performance on extremely sparse data Healthcare.

Model	Healthcare		
	10% missing	50% missing	90% missing
BRITS	0.284	0.368	0.517
GP-VAE	0.413	0.592	0.926
GRIN	0.245	0.317	0.472
CSDI	0.217	0.301	0.481
MIDM	**0.206**	**0.287**	**0.449**

4.4.1 Performance on extremely sparse data. To further evaluate the robustness of MIDM, we apply one more dataset, Healthcare [32], which contains around 80% missing values. Healthcare consists of 4000 clinical time series with 35 variables for 48 hours from ICU. Following the process in previous works [8, 36], we split the dataset into hourly multivariate time series with 48 data points in each series. Since there is no ground-truth for the raw missing values, 10/50/90% of observed values are randomly chosen as ground-truth for testing. Noting that Healthcare originally contains 80% missing values, when the missing rate is set to 90%, the 90% of observed data are masked out, leading to only 2% available data. Even for missing rate as 10%, there is only 18% available data. Therefore, due to the extreme sparsity of the data, the experiments extracted in this part are quite difficult. Similarly, MAE is applied here evaluation and GP-VAE, BRITS, GRIN and CSDI are used as baselines.

Tab 4 shows the imputation performance and we can find MIDM outperformed all the baselines under all the three settings. Specifically, MIDM achieves 5.1%, 4.7%, 11.2% MAE improvements than the best MAE of all baselines under settings of 10%, 50%, 90% missing rates. Such result further proves the robustness of the proposed MIDM when being applied on extremely sparse datasets.

4.4.2 Performance on various missing rates. We first evaluate the robustness of MIDM on various missing rates. For this purpose, we apply a series of missing rates ranging from 5% to 50% and randomly mask out data points of CER-E with the missing rates. MAE is applied here for evaluation. GP-VAE, BRITS, GRIN and CSDI are used here as baselines.

As demonstrated in Fig 4, we find when the missing rate is small, GRIN and CSDI are able to achieve competitive MAE to MIDM. But when the missing value increases, the performance gap between MIDM and baselines becomes more significant. Such results indicates that when the missing rates increase, the performance of

MIDM decrease much slower than the baselines and MIDM is more robust than the baseline w.r.t. missing rates.

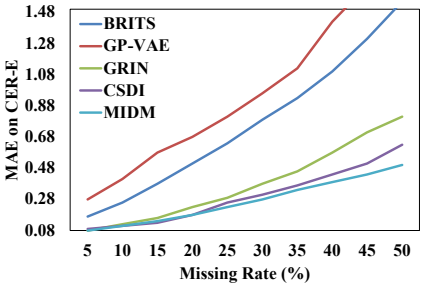


Figure 4: Imputation performance comparison on CER-E with different missing rates.

4.4.3 Probabilistic imputation. As mentioned, one advantage of applying probabilistic models for multivariate imputation is that probabilistic models are able to generate different plausible imputations, which is nice for uncertainties estimation. We here provide visual imputation examples of MIDM on AQI-36 dataset in Fig 5. The red crosses denote observed values and blue circles denote the ground truth of missing values. For generating the visualization of probabilistic imputation, we generate plausible imputations 100 times, median values of the imputations are shown as the line and 5% and 95% quantiles are shown as the shade. As shown, in most case, the median line fits original time series well and nearly all ground truth of missing values fall in the shade. Such result indicates that MIDM is able to generate accurate estimation of missing values with high confidence.

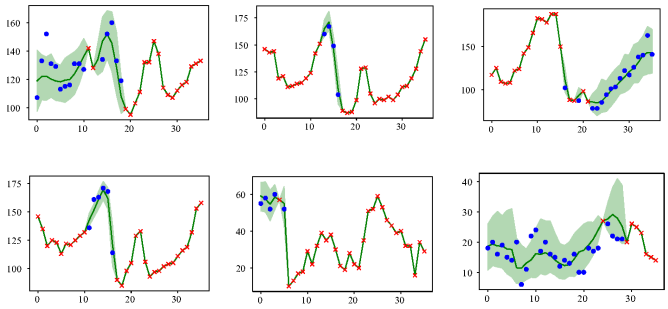


Figure 5: Examples of probabilistic time series imputation. Each subfigure shows probabilistic imputation of one chosen variable. The horizontal axis represents time and the vertical axis represents value.

4.4.4 Multivariate forecasting. Also, to show the generalizability of MIDM, we evaluate MIDM for multivariate time series forecasting task on PEMS-BAY. Actually, multivariate time series forecasting can be seen as a special case of imputation where historical data is observed and future data is missing. There exists abundant studies tackling multivariate time series forecasting evaluate their methods on PEMS-BAY, and we here follow their settings and compare MIDM with several state-of-the-art forecasting models, including 1) GMAN [45] utilizes spatial and temporal attention for forecasting. 2) AGCRN [3] proposes an adaptive graph convolution and combine it with RNN. 3) GWNET [42] introduces a covariance loss applicable to many kinds of networks. PEMS-BAY are divided into training, validation and testing set in chronological order with ratio as 7:1:2 as in [45]. Each series contains 24 data point, where the first 12 data points are used to predict the next 12 data points. MAE and MRE are applied here for evaluation.

As shown in Tab 5, MIDM can deal with multivariate forecasting well and achieves satisfying forecasting accuracy. Compared to SOTA forecasting models, MIDM achieves 1.6%, 2.4% performance gain on MAE and MRE respectively. The result demonstrates the generalizability of MIDM on multivariate forecasting problem and shows the potential of applying MIDM on other generating tasks.

Table 5: Forecasting performance comparison on PEMS-BAY.

Model	GMAN	AGCRN	GWNET	MIDM
MAE	1.86±0.02	1.97±0.03	1.91±0.02	**1.83±0.04**
MRE(%)	4.31±0.02	4.58±0.02	4.47±0.04	**4.21±0.06**

5 CONCLUSION

In this paper, we propose a novel imputation diffusion model to tackle the problem of multivariate time series imputation. In particular, we theoretically re-derive the ELBO of conditional diffusion model with considering the correlations between observed values and missing values, and we propose a multivariate time series imputation diffusion model (MIDM) by designing brand new noise sampling, adding and denoising processes based on the new ELBO, hence enabling the generation of more accurate estimations of missing values by taking advantage of the involvement of the consistency of observed and missing values. Extensive and cross-domain experiments validate the effectiveness and superiority of MIDM on both multivariate time series imputation and forecasting under various settings.

Theoretically, as long as the consistency between observation and missing values exists, the derivation of our ELBO can be extendedly utilized into generalized recovering problem in which missing data is imputed based on given partial observations and eventually recover the whole dataset. And this can promote our model to wider application scenarios, e.g., image restoration and image resolution enhancement. The potential of the proposed diffusion framework in addressing other tasks is yet to be developed.

ACKNOWLEDGMENTS

This paper is partially supported by the National Natural Science Foundation of China (No.62072427, No.12227901), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), Academic Leaders Cultivation Program, USTC.

REFERENCES

[1] Juan Miguel Lopez Alcaraz and Nils Strodthoff. 2022. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399* (2022).

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[3] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems* 33 (2020), 17804–17815.

[4] Robert Bamler and Stephan Mandt. 2017. Structured black box variational inference for latent time series models. *arXiv preprint arXiv:1707.01069* (2017).

[5] Gregory Benton, Wesley Maddox, and Andrew Gordon Wilson. 2022. Volatility Based Kernels and Moving Average Means for Accurate Forecasting with Gaussian Processes. In *International Conference on Machine Learning*. PMLR, 1798–1816.

[6] Mathias Berglund, Tapani Raiko, Mikko Honkala, Leo Kärkkäinen, Akos Vetek, and Juha T Karhunen. 2015. Bidirectional recurrent neural networks as generative models. *Advances in neural information processing systems* 28 (2015).

[7] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.

[8] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems* 31 (2018).

[9] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 6085.

[10] Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. 2001. Freeway Performance Measurement System: Mining Loop Detector Data. *Transportation Research Record* 1748, 1 (2001), 96–102. https://doi.org/10.3141/1748-12

[11] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2020. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713* (2020).

[12] Tae-Min Choi, Ji-Su Kang, and Jong-Hwan Kim. 2020. RDIS: Random drop imputation with self-training for incomplete time series data. *arXiv preprint arXiv:2010.10075* (2020).

[13] Andrea Cini, Ivan Marisca, and Cesare Alippi. 2021. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. *arXiv preprint arXiv:2108.00298* (2021).

[14] Adrian V Dalca, John Guttag, and Mert R Sabuncu. 2019. Unsupervised data imputation via variational inference of deep subspaces. *arXiv preprint arXiv:1903.03503* (2019).

[15] Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. 2020. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*. PMLR, 1651–1661.

[16] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. *arXiv preprint arXiv:1406.2661* (2014).

[17] Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. *International Conference on Learning Representations* (2022).

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[19] James Honaker and Gary King. 2010. What to do about missing values in time-series cross-section data. *American journal of political science* 54, 2 (2010), 561–581.

[20] Andrew T Hudak, Nicholas L Crookston, Jeffrey S Evans, David E Hall, and Michael J Falkowski. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment* 112, 5 (2008), 2232–2245.

[21] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).

[22] Yan Li, Xinjiang Lu, Yaqing Wang, and Dejing Dou. 2023. Generative Time Series Forecasting with Diffusion, Denoise, and Disentanglement. *arXiv preprint arXiv:2301.03028* (2023).

[23] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017).

[24] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. 2018. Multivariate time series imputation with generative adversarial networks. *Advances in neural information processing systems* 31 (2018).

[25] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. 2019. E2gan: End-to-end generative adversarial network for multivariate time series imputation. In *Proceedings of the 28th international joint conference on artificial intelligence*. AAAI Press, 3094–3100.

[26] Xiaoye Miao, Yangyang Wu, Jun Wang, Yunjun Gao, Xudong Mao, and Jianwei Yin. 2021. Generative semi-supervised learning for multivariate time series imputation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 8983–8991.

[27] Morten Morup, Daniel M Dunlavy, Evrim Acar, and Tamara Gibson Kolda. 2010. *Scalable tensor factorizations with missing data*. Technical Report. Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA ....

[28] Fulufhelo V Nelwamondo, Shakir Mohamed, and Tshilidzi Marwala. 2007. Missing data: A comparison of neural network and expectation maximization techniques. *Current Science* (2007), 1514–1521.

[29] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*. PMLR, 8857–8868.

[30] Stephen Roberts, Michael Osborne, Mark Ebden, Steven Reece, Neale Gibson, and Suzanne Aigrain. 2013. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371, 1984 (2013), 20110550.

[31] Satya Narayan Shukla and Benjamin M Marlin. 2021. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318* (2021).

[32] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. 2012. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*. IEEE, 245–248.

[33] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations*.

[35] Qiuling Suo, Weida Zhong, Guangxu Xun, Jianhui Sun, Changyou Chen, and Aidong Zhang. 2020. GLIMA: Global and local time series imputation with multi-directional attention learning. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 798–807.

[36] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* 34 (2021), 24804–24816.

[37] Stef Van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software* 45 (2011), 1–67.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[39] Pengkun Wang, Chaochao Zhu, Xu Wang, Zhengyang Zhou, Guang Wang, and Yang Wang. 2022. Inferring intersection traffic patterns with sparse video surveillance information: An st-gan method. *IEEE Transactions on Vehicular Technology* 71, 9 (2022), 9840–9852.

[40] Ian R White, Patrick Royston, and Angela M Wood. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine* 30, 4 (2011), 377–399.

[41] Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. 2016. ST-MVL: filling missing values in geo-sensory time series data. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

[42] Boseon Yoo, Jiwoo Lee, Janghoon Ju, Seijun Chung, Soyeon Kim, and Jaesik Choi. 2021. Conditional Temporal Neural Processes with Covariance Loss. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12051–12061.

[43] 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.

[44] Jinsung Yoon, William R Zame, and Mihaela van der Schaar. 2018. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering* 66, 5 (2018), 1477–1490.

[45] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. 2020. GMAN: A Graph Multi-Attention Network for Traffic Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (April 2020), 1234–1241. https://doi.org/10.1609/aaai.v34i01.5477 Number: 01.

[46] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. 2015. Forecasting Fine-Grained Air Quality Based on Big Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 2267–2276. https://doi.org/10.1145/2783258.2788573

[47] Eric Zivot and Jiahui Wang. 2006. Vector autoregressive models for multivariate time series. *Modeling financial time series with S-PLUS®* (2006), 385–429.