15th International Conference on Information Science and Technology (ICIST 2025)
Zhanjiang, China, December 5-8, 2025

Blue Noise-based Generative Models for the Imputation of Time-series Data

Graham Bishop
Department of Mathematics and Statistics
Saint Louis University
St. Louis, MO, USA
graham.bishop@slu.edu

Tong Si
Department of Health
& Clinical Outcomes Research
Saint Louis University
St. Louis, MO, USA
tong.si@slu.edu

Haijun Gong*
Department of Mathematics & Statistics
Saint Louis University
St. Louis, MO, USA
haijun.gong@slu.edu
*Corresponding author

2025 15th International Conference on Information Science and Technology (ICIST) | 979-8-3315-1450-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICIST66592.2025.11306560

***Abstract*—Reconstructing missing data in high-dimensional time-series remains a challenging task, especially when the underlying signals exhibit complex temporal dynamics and nonlinear relationships. While most traditional approaches can not model such intricacies, generative models—particularly conditional score-based diffusion methods—have emerged as powerful alternatives, offering significant improvements in imputation accuracy. Despite their success, these models typically rely on isotropic white noise during training, which treats all frequency components uniformly and fails to preserve critical frequency-dependent correlations. Relying solely on white noise can lead to the loss of fine-scale temporal patterns, compromising the accuracy and reliability of the reconstructed data. Our recent work introduces a time-varying blue noise-based conditional score-based diffusion model for imputation (tBN-CSDI) by incorporating a time-varying blue noise schedule into the diffusion process to address the limitations of existing methods in handling missing values in time series data. Experimental results on real-world datasets demonstrate that tBN-CSDI outperforms conventional methods based on white noise schedules. We also discuss the integration of pseudotime analysis with diffusion models as a promising direction for future research, particularly for applications in dynamic biological systems where temporal ordering is critical yet uncertain.**

***Index Terms*—Missing Value Imputation, Blue Noise, Diffusion Model, Time Series Data**

I. Introduction

Missing values in time series data are a pervasive issue across many domains, and a large amount of missing data can significantly impair the performance of predictive modeling. For example, the missing rate in the single-cell RNA sequencing (scRNA-seq) data could be 50% to 90% due to the technical limitations. Accurate imputation of time-series missing values is critical for downstream analyses such as gene expression profiling, trajectory inference, and regulatory network reconstruction. Traditional imputation methods usually assume a known data distribution—a strong assumption that is often violated in real-world, high-dimensional settings where the true distribution is complex and poorly characterized.

Several generative models have been developed to impute missing values in high-dimensional time series data, achieving state-of-the-art performance. For instance, GP-VAE [1] is based on variational autoencoders (VAEs) [2] to model temporal dynamics. On the generative adversarial network (GAN)-based side, methods such as GAIN [3], sc-fGAIN [4], ImputeGAN [5], and tf-biGAIN [6] utilize adversarial training to generate realistic imputations. More recently, the Conditional Score-based Diffusion Model for Imputation (CSDI) [7] and DiffPuter [8], built upon denoising diffusion probabilistic models (DDPMs) [9], have demonstrated superior performance in imputing high-dimensional time-series data. However, these models relies on isotropic white noise during training, which treats all frequency components uniformly and fails to preserve critical frequency-dependent correlations. To overcome this limitation, our recent work [10] introduces a time-varying blue noise-based conditional score-based diffusion model (tBN-CSDI) for imputing missing values in time-series scRNA-seq data by incorporating a time-varying blue noise schedule into the diffusion process. We now present an short overview of the tBN-CSDI framework proposed in our recent work [10], followed by a discussion of potential extensions to further improve its performance and applicability in structured data imputation.

II. Method

The Denoising Diffusion Probabilistic Models (DDPMs), have been introduced in [9]. For completeness, we briefly review the core concepts of DDPMs framework in this work. DDPM consists of two key components: the forward diffusion process and the reverse denoising process. In the forward process, data is gradually transformed into noise by adding small amounts of Gaussian noise over a sequence of time steps until the original signal is nearly obliterated. The reverse process involves training a neural network to progressively denoise the data step by step recovering the underlying data structure from pure noise, so the learned reverse process enables the model to generate realistic samples.

In the forward diffusion process, noise is progressively added to the data according to the equation

$$x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, I)$ is standard Gaussian noise and $1-\alpha_t$ represents the variance at step t. The sequence of coefficients

This research was partially supported by the National Institutes of Health under Award Number R15GM148915 (HG) and President's Research Funds.

$\alpha_t \in (0,1)$ controls the rate of noise injection, gradually transforming the input data into a sample from a near-isotropic Gaussian distribution over many steps.

In the reverse denoising process, the model iteratively recovers the underlying data structure by removing noise at each step. Starting from pure noise $x_T \sim \mathcal{N}(0, I)$, the model predicts the additive noise in x_t using a neural network $\varepsilon_\theta(x_t, t)$, and progressively samples cleaner versions of the data through a Markov chain, ultimately reconstructing x_0. Mathematically, the reverse step is expressed as

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t, t)\right) + \sigma_t z,$$

β_t is the noise schedule parameter at step t, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and σ_t controls the magnitude of injected noise in the reverse step.

The training objective of DDPM is to learn a neural network ε_θ that predicts the noise added during the forward diffusion process. This is achieved by minimizing the expected squared error between the true noise ε and the model's prediction $\varepsilon_\theta(x_t, t)$, which is given by

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0, t, \varepsilon}\left[\left\|\varepsilon - \varepsilon_\theta(x_t, t)\right\|^2\right],$$

where x_0 is a sample from the data distribution, t is a randomly sampled timestep, and x_t is the noisy sample at step t obtained via the forward process.

The Conditional Score-based Diffusion Model for Imputation (CSDI) aims to learn the conditional distribution over missing values given the observed entries and a binary mask indicating which elements are observed. Compared to standard DDPMs, CSDI modifies the forward diffusion process such that only the missing entries are progressively noised, while observed entries remain untouched throughout the process. Specifically, partition x_0 into x_0^{obs} and x_0^{mis} components,

- Observed entries:

$$x_t^{\text{obs}} = x_0^{\text{obs}} \quad \text{(no noise added)}$$

- Missing entries:

$$x_t^{\text{mis}} = \sqrt{\bar{\alpha}_t}\, x_0^{\text{mis}} + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I)$$

This masking-aware diffusion strategy ensures that the model retains full information from observed data at every step. As a result, the reverse process becomes a conditional denoising task, where the model reconstructs the missing values conditioned on both the noisy version of the missing part and the clean observed data. The reverse transition is modeled as:

$$p_\theta\left(x_{t-1}^{\text{mis}} \mid x_t^{\text{mis}},\, x_0^{\text{obs}}\right),$$

where the neural network ε_θ is trained to predict the noise in the missing components, conditioned on the observed data and the timestep t.

Both DDPM and CSDI employ white Gaussian noise in the diffusion process. However, white noise assigns equal power across all frequency components, treating high- and low-frequency dynamics uniformly. This isotropic perturbation can disrupt fine-scale temporal structures and fail to preserve critical frequency-dependent correlations inherent in biological time series. As a result, reliance on white noise may lead to oversmoothing or distortion of transient but biologically meaningful patterns, ultimately compromising the fidelity of data reconstruction.

To address this limitation, our recent work [10] introduces a time-varying blue noise-based conditional score-based diffusion model for imputation (tBN-CSDI). By incorporating a time-varying blue noise schedule, where noise power increases with frequency. Our work [10] generates blue noise $\varepsilon_{\text{blue}}$ by combining Ulichney's void-and-cluster framework [11] with simulated annealing and a Cholesky-based decomposition method [12]. The procedure begins by designing a target covariance matrix Σ via a simulated annealing process applied to the dataset, then, a Cholesky decomposition is performed to obtain the lower-triangular factor L such that $\Sigma = LL^\top$. Finally, white noise $\varepsilon_{\text{white}} \sim \mathcal{N}(0, I)$ is transformed into structured blue noise via linear mapping:

$$\varepsilon_{\text{blue}} = L\, \varepsilon_{\text{white}}.$$

To balance the injection of white noise and blue noise during the diffusion process, we introduce a noise blending schedule controlled by a time-varying coefficient γ_t, defined as

$$\gamma_t = \sigma\left[\gamma_{\text{start}} + (\gamma_{\text{end}} - \gamma_{\text{start}})\left(\frac{t}{T}\right)^{\gamma_\tau}\right],$$

where t denotes the current time step, T is the total number of diffusion steps, γ_{start}, γ_{end}, $\gamma_\tau > 0$ are blending parameters which are estimated empirically. $\sigma[\cdot]$ is a saturation function—typically a sigmoid that ensures $\gamma_t \in [0, 1]$.

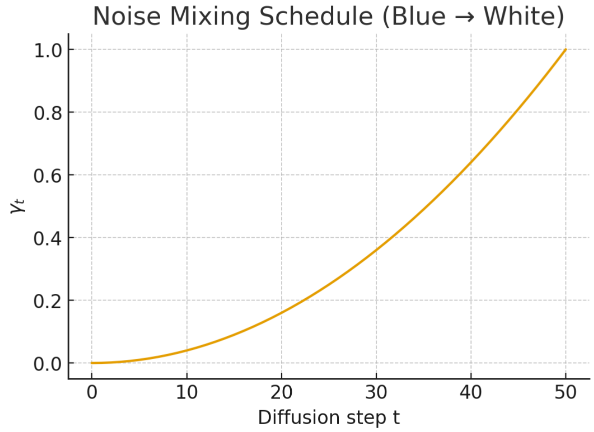

Fig. 1: Noise blending schedule coefficient changes with diffusion steps

Using the noise blending schedule coefficient γ_t, we obtain a time-varying noise $\tilde{\epsilon}_t$ that smoothly interpolates between blue and white noise:

$$\tilde{\epsilon} = \gamma_t \cdot \epsilon_{\text{white}} + (1{-}\gamma_t) \cdot \epsilon_{\text{blue}},$$

As illustrated in Fig. 1, this scheduling mechanism dynamically modulates the noise type throughout the diffusion process. During early steps ($t \ll T$), $\gamma_t \approx 0$, so $\tilde{\epsilon}_t \approx \epsilon_{\text{blue}}$, resulting in fine-scale, high-frequency perturbations that preserve local structure and texture. In contrast, during later stages ($t \to T$), $\gamma_t \to 1$, and thus $\tilde{\epsilon}_t \approx \epsilon_{\text{white}}$, enabling broader, global exploration of the sample space. This staged noise injection strategy allows the model to first capture and refine intricate details before transitioning into large-scale denoising or pattern formation, effectively balancing local fidelity with global coherence.

In the tBN-CSDI framework, the key difference from the original CSDI lies in the replacement of standard Gaussian white noise with the blended noise $\tilde{\epsilon}$ during the forward diffusion (training) process. Specifically, for the masked or missing components of the data, the noising procedure is modified as follows:

$$x_t^{\text{mis}} = \sqrt{\bar{\alpha}_t} \cdot x_0^{\text{mis}} + \sqrt{1-\bar{\alpha}_t} \cdot \tilde{\epsilon}$$

In the imputation (reverse) process, we also incorporate the blended noise $\tilde{\epsilon}$ to guide the denoising steps. Specifically, during the backward sampling phase, the missing components are updated as: $x_{t-1}^{mis} \leftarrow x_t^{mis} + \sigma_t \cdot \tilde{\epsilon}$. That is, the reverse step may take the form

$$x_{t-1}^{mis} = \frac{1}{\sqrt{\alpha_t}}(x_t^{mis} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t^{mis}, t|x_0^{obs})) + \sigma_t \cdot \tilde{\epsilon}.$$

For further details on the algorithm and the complete imputation workflow, the interested reader is referred to our recent work [10]; we include only a high-level overview here for brevity.

III. Results

We implement tBN-CSDI algorithm to impute missing values in time series data and compare its performance against CSDI and other state-of-the-art imputation methods. Our recent work [10] has analyze three datasets, including the PhysioNet Challenge dataset [13], and two single cell RNA sequencing data. Here, we provide some results that were not published before. The PhysioNet data contains clinical records from approximately 4,000 ICU patients, each covering the first 48 hours post-admission and some physiological variables (age, weight, heart rate, glucose levels, etc.) measured irregularly over time. The scRNA-seq dataset [14] contains 120 distinct monocytic THP-1 human myeloid leukemia cells at each of eight time points, totaling 960 cells. To simulate different sparsity levels, we randomly masked observed values at varying missing rates.

We perform probabilistic imputation by generating 100 independent samples for each missing value in the PhysioNet dataset. Due to the nondeterministic nature of the diffusion process, this yields a predictive distribution at each time point. As illustrated in Fig. 2–Fig. 3, the dark blue line represents the median of the predicted distribution and serves as the final imputed value. The light blue shaded area indicates the 5th to 95th percentile range of the samples, reflecting the model's uncertainty: a wider band corresponds to higher uncertainty, while a narrow region suggests greater confidence. The orange crosses (×) denote observed values, while the red dots mark the true values at missing entries (i.e., ground truth targets). Ideally, the median imputation (dark blue line) should closely align with the red dots, and the true values should fall within the light blue uncertainty band, indicating both accuracy and well-calibrated uncertainty estimation.

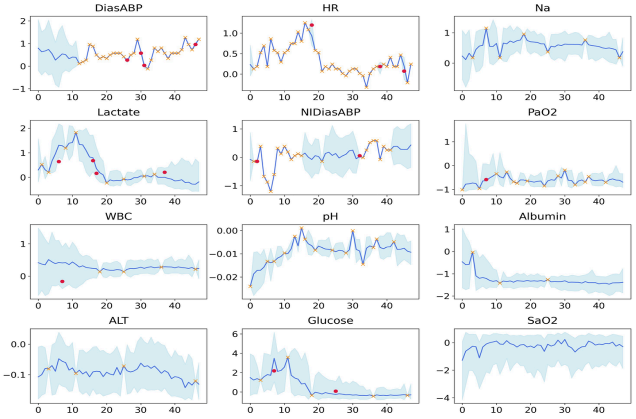

Fig. 2: Probabilistic imputation distribution for PhysioNet dataset with missing rate 0.1 using tBN-CSDI. The orange crosses (×) denote observed values, while the red dots mark the true values at missing entries.

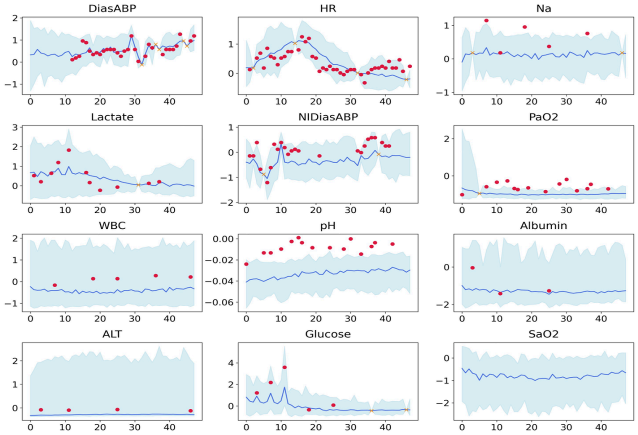

Fig. 3: Probabilistic imputation distribution for PhysioNet dataset with missing rate 0.9 using tBN-CSDI. The orange crosses (×) denote observed values, while the red dots mark the true values at missing entries.

The results in Fig. 2–Fig. 3 demonstrate that our tBN-CSDI method achieves accurate imputation with well-calibrated uncertainty estimates across varying missing rates. At a missing rate of 0.1, the model imputes missing values with high confidence, as evidenced by narrow uncertainty bands, and all true values (red dots) lie within the light blue confidence region—indicating reliable uncertainty quantification and strong imputation accuracy. Even under an extreme missing rate of

0.9, where only 10% of the data is observed, most true values remain within the predicted confidence bands. This suggests that tBN-CSDI maintains robust performance and uncertainty awareness even in highly sparse settings, enabling trustworthy imputations despite severe data incompleteness.

We also apply tBN-CSDI to impute missing values in single-cell RNA sequencing (scRNA-seq) data under varying missing rates, ranging from 0.1 to 0.9. The full experimental results are reported in [10]; here, we provide a brief summary.

Imputation performance is evaluated using the mean root mean squared error (RMSE), where lower values indicate better accuracy. All the experiments were run five times and calculated the mean values of RMSE. Traditional methods such as KNN and SoftImpute achieve RMSEs exceeding 1.0, while the original CSDI model based on white noise yields an RMSE of approximately 0.9. In contrast, our tBN-CSDI reduces the average RMSE to below 0.6, representing a relative improvement of over 30% compared to all existing approaches. Notably, tBN-CSDI maintains consistent performance across all missing rates, demonstrating insensitivity to increasing data sparsity. This robustness highlights the effectiveness of blue noise blending in preserving biological signal structure even under extreme missingness, making it particularly well-suited for challenging real-world scRNA-seq applications.

IV. Future Research

In time-series single-cell RNA sequencing (scRNA-seq) data, although cells are sampled at the same experimental time points, they often reside in different biological states, such as varying phases of the cell cycle or developmental trajectories. As a result, the observed measurement time may not reflect the true biological progression of each cell. Pseudotime analysis (e.g., Monocle [15], Slingshot [16]) has emerged as a powerful tool for inferring this latent biological order by ordering cells along a dynamic process (e.g., differentiation or cell cycle) based on their expression profiles. In contrast, our current tBN-CSDI framework performs imputation using canonical observation time, effectively ignoring the underlying pseudotemporal structure.

A promising direction for future work is to extend tBN-CSDI to operate in pseudotime rather than experimental time only. By aligning the diffusion process with inferred pseudotime trajectories, the model could leverage more biologically meaningful temporal dependencies, potentially improving imputation accuracy and preserving dynamic gene expression patterns. This would involve integrating pseudotime estimation methods (e.g., [15]) into the diffusion framework or jointly learning the latent trajectory and imputation network, enabling more accurate reconstruction of gene expression dynamics in complex developmental processes.

Another promising direction for future research is to investigate how the scheduling order of noise types—specifically blue and white noise affects imputation performance. In the current tBN-CSDI framework, we adopt a fixed schedule where structured blue noise dominates in early diffusion steps (promoting fine-scale detail preservation), while white noise is gradually introduced in later stages (enabling broad exploration). However, reversing this order, i.e., injecting white noise early and reserving blue noise for later refinement may lead to different convergence behavior or generalization properties. It remains an open question how such a reversal would influence the imputation performance.

Moreover, extending the noise spectrum beyond blue and white opens further opportunities. For instance, pink noise, which exhibits stronger low-frequency components, may better capture long-range temporal correlations. Incorporating pink noise into the blending schedule or learning the optimal spectral characteristics could enhance the model's ability to recover biologically plausible trajectories.

Exploring these alternative noise schedules and spectra will deepen our understanding of how stochastic perturbation design influences representation learning in diffusion-based imputation models.

References

[1] V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt, "Gp-vae: Deep probabilistic time series imputation," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 1651–1661.

[2] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[3] J. Yoon, J. Jordon, and M. Schaar, "Gain: Missing data imputation using generative adversarial nets," in *International conference on machine learning*. PMLR, 2018, pp. 5689–5698.

[4] T. Si, Z. Hopkins, J. Yanev, J. Hou, and H. Gong, "A novel f-divergence based generative adversarial imputation method for scrna-seq data analysis," *Plos one*, vol. 18, no. 11, p. e0292792, 2023.

[5] R. Qin and Y. Wang, "Imputegan: Generative adversarial network for multivariate time series imputation," *Entropy*, vol. 25, no. 1, p. 137, 2023.

[6] W.-S. Liu, T. Si, A. Kriauciunas, M. Snell, and H. Gong, "Bidirectional f-divergence-based deep generative method for imputing missing values in time-series data," *Stats*, vol. 8, no. 1, p. 7, 2025.

[7] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "Csdi: Conditional score-based diffusion models for probabilistic time series imputation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 804–24 816, 2021.

[8] H. Zhang, L. Fang, Q. Wu, and P. S. Yu, "Diffputer: Empowering diffusion models for missing data imputation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[10] G. Bishop, T. Si, I. Luebbert, N. Al-Hammadi, and H. Gong, "tbn-csdi: A time-varying blue noise-based diffusion model for time series imputation," *Bioinformatics Advances*, p. vbaf225, 2025.

[11] R. Ulichney, "The void-and-cluster method for dither array generation," *SPIE MILESTONE SERIES MS*, vol. 154, pp. 183–194, 1999.

[12] X. Huang, C. Salaun, C. Vasconcelos, C. Theobalt, C. Oztireli, and G. Singh, "Blue noise for diffusion models," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[13] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, "Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012," *Computing in cardiology*, vol. 39, p. 245, 2012.

[14] T. Kouno, M. de Hoon, J. C. Mar, Y. Tomaru, M. Kawano, P. Carninci, H. Suzuki, Y. Hayashizaki, and J. W. Shin, "Temporal dynamics and transcriptional control using single-cell gene expression analysis," *Genome biology*, vol. 14, pp. 1–12, 2013.

[15] C. Trapnell, D. Cacchiarelli, and et al, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature biotechnology*, vol. 32, no. 4, pp. 381–386, 2014.

[16] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC genomics*, vol. 19, no. 1, p. 477, 2018.