The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

Context-Aware Diffusion for Telemetry Time Series with Permutation-Stable Feature Modeling (Student Abstract)

Giovanni B. Esposito[1], Daniele Cesarini[2], Andrea Bartolini[1]
[1]University of Bologna, Italy
[2]CINECA, Italy
g.esposito@unibo.it, d.cesarini@cineca.it, a.bartolini@unibo.it

Abstract

We present a context-aware diffusion model for multivariate time series generation in dynamic and partially observed environments, with applications to data-center computing node's telemetry and beyond. The model integrates pretrained textual embeddings to represent feature semantics, enabling flexible, context-guided generation and improved adaptability to unseen or re-ordered input features. Built on a transformer architecture, it employs both time-wise and feature-wise masking to support missing data during training and inference. We show that the model is robust to permutations with respect to the feature dimension, mantaining stable performance in settings where input configurations vary. Empirical evaluations on HPC sensor data illustrate the model's versatility across generation and imputation tasks. This work introduces a modular and generalizable framework for time series modeling in complex, high-dimensional systems which can serve as a digital-twin for data-center's compute node telemetry.

Introduction

High-performance computing (HPC) systems underpin modern data-intensive applications, producing vast streams of multivariate telemetry from diverse sensors measuring CPU load, memory usage, I/O, temperature, and power. Accurate modeling of such data is crucial for predictive maintenance, anomaly detection, and digital twins, yet remains challenging due to high dimensionality, missing or reconfigured sensors, and lack of semantic guidance. Traditional time series models assume fixed, fully observed features and treat signals as anonymous vectors, limiting adaptability to dynamic environments.

We propose a context-aware diffusion model for multivariate time series generation that addresses these constraints. Diffusion models, originally introduced by Sohl-Dickstein et al. (2015) and later made popular by Ho, Jain, and Abbeel (2020), have recently demonstrated remarkable performance in high-dimensional data synthesis tasks, including time-series modeling (Lin et al. 2024). Our diffusion model is built on transformer-based backbone that incorporates time and feature-wise masking to handle partial observations, while conditioning on statistical properties and textual feature descriptions to exploit semantic information. The model is explicitly designed to be robust to feature permutation and evolving sensor layouts, enabling generalization across heterogeneous HPC systems.

Methodology

We propose a transformer-based diffusion model that integrates denoising diffusion with modular attention to handle noisy, partially observed multivariate time series.

Diffusion Framework

Given a multivariate time-series $X \in \mathbb{R}^{N\times P}$, with N timesteps and P features, the forward process of the diffusion corrupts it with Gaussian noise:

$$X_t = \sqrt{\bar{\alpha}_t} X + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where t is randomly drawn from a discrete $U(1,T)$, with T fixed, and α is a parameter that controls the noise injection. A denoiser f_θ is trained to predict the added noise $\hat{\epsilon} = f_\theta(X_t, t, C)$ conditioned on context C.

Context Encoding

Each feature channel is enriched with a context vector C, made of: (i) semantic embeddings of their textual descriptions, extracted with Sentence-BERT and projected into the model space; (ii) statistical descriptors of the series (mean, quantiles, skewness, kurtosis). Both are concatenated with the noisy input X_t, yielding $Z_t \in \mathbb{R}^{(N+2)\times P\times H}$.

Denoising Architecture

A transformer backbone alternates between *temporal* and *feature-wise* self-attention blocks, enabling modeling of intra-feature dynamics and cross-feature dependencies under masking. The diffusion timestep t is encoded as a sinusoidal embedding and added to all tokens. After processing, context tokens are discarded and a projection layer outputs $\hat{\epsilon} \in \mathbb{R}^{N\times P}$.

Training

The model is trained by minimizing the error of the noise prediction $\hat{\epsilon}$:

$$\mathcal{L}_{MSE} = \mathbb{E}_{X,t,\epsilon}\big[\|\epsilon - f_\theta(Z_t, t, C)\|_2^2\big]. \tag{2}$$

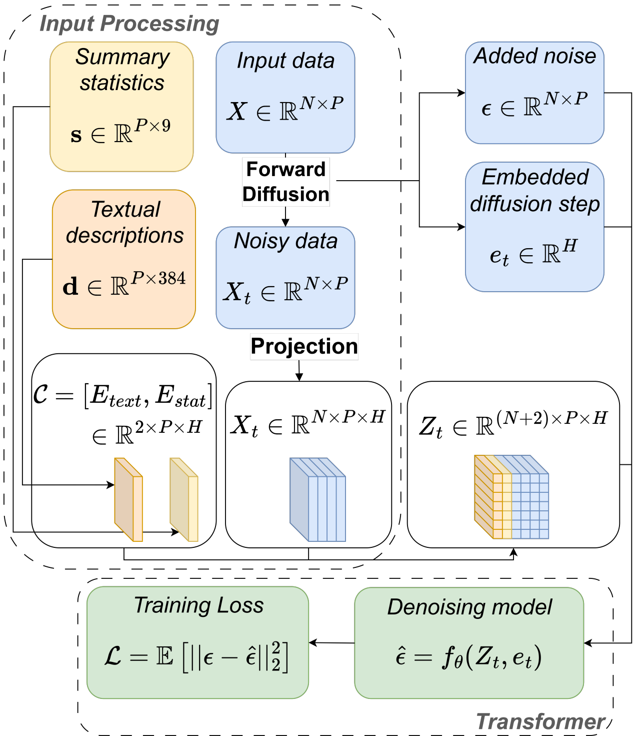

Figure 1: Model overview

To improve robustness, we apply stochastic augmentations at training time: *feature permutations*, *timestep masking*, *feature masking*, or no augmentation. Only unmasked values contribute to the loss.

Conditioned Generation

At inference, a sample is drawn from a Gaussian distribution and iteratively denoised:

$$X_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(X_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_t\right) + \sigma_t z, \quad z \sim \mathcal{N}(0, I), \quad (3)$$

with conditioning provided by feature semantics, statistics, and attention masks.

This design, explained in detail in Figure 1 enables flexible, permutation-robust generation under dynamic sensor configurations.

Experimental Results

Dataset and Setup

We evaluate the model on the M100 dataset (Borghesi et al. 2023), a real-world HPC telemetry comprising 88 sensor signals per timestep, aggregated with 15-minute pooling. Each input sequence has length N with $P = 88$ standardized features. Sensor descriptions were encoded via Sentence-BERT to provide fixed semantic embeddings. Data were split into non-overlapping training, validation and test sets.

Evaluation

We evaluated generation accuracy under different perturbations (fully observed, shuffled, masked features/timesteps). As reported in Table 1, results show stable performance across contexts, with graceful degradation under masking and strong invariance to feature permutations.

Setting	**MAE** ↓ (mean ± std)
No perturbations	0.0225 ± 0.0363
Shuffled features	0.0225 ± 0.0364
Masked features	0.0622 ± 0.0362
Masked timesteps	0.0613 ± 0.0365

Table 1: Model performances in different scenarios. Features and timesteps are masked with random probabilites $p \sim U(0.1, 0.5)$

We also explored with ablation studies the contribution of (i) semantic embeddings and (ii) the use of perturbations during training phase. Results are summarized by Table 2. The ablation results highlight that semantic embeddings alone provide a significant improvement over the baseline, while the introduction of augmentations happen to degrade performance. In particular, the use of random shuffling and masking during training appears to distort the underlying structure of the input sequences, introducing noise that hinders the model's ability to learn stable representations. This suggests that embeddings may already provide enough robustness for the model itself. This will be explored in future works.

Setup		**MAE**			
Emb.s	**Aug.s**	**Gen.**	**Shuf.**	**F. Mask**	**T. Mask**
No	No	0.441	0.441	0.418	0.432
No	Yes	0.154	0.154	0.156	0.155
Yes	Yes	0.119	0.119	0.125	0.108
Yes	No	0.022	0.022	0.062	0.061

Table 2: Ablation Study of Model Components

Conclusion and Future Directions

This work presents a promising step toward the design of context-aware generative models for data center telemetry. Our experiments show that the proposed architecture is robust to feature ordering, enabling consistent performance even when sensors are missing, reconfigured, or dynamically evolving. Furthermore, the model shows good imputation capabilities, maintaining stable accuracy under both feature-level and temporal masking. These properties can make it well-suited for real-world HPC environments where sensor layouts and observability are inherently dynamic. Future developments of this work will focus on benchmarking it against state-of-the-art diffusion models, extending it to other telemetry domains, and exploring lightweight training and adaptive conditioning for scalable deployment. Together, these expansions will help establish a foundation for general-purpose generative models of complex, partially observed systems.

Acknowledgments

The activities of this work have been supported by the EU - NextGenerationEU with funds made available by the National Recovery and Resilience Plan (NRRP) Mission 4, Component 2, Investment 3.3 (M.D. 117/2023) and by CINECA

References

Borghesi, A.; Di Santi, C.; Molan, M.; Ardebili, M. S.; Mauri, A.; Guarrasi, M.; Galetti, D.; Cestari, M.; Barchi, F.; Benini, L.; et al. 2023. M100 exadata: a data collection campaign on the cineca's marconi100 tier-0 supercomputer. *Scientific Data*, 10(1): 288.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Lin, L.; Li, Z.; Li, R.; Li, X.; and Gao, J. 2024. Diffusion models for time-series applications: a survey. *Frontiers of Information Technology & Electronic Engineering*, 25(1): 19–41.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.