TimeDiT: Diffusion Transformers Foundation Model for Time Series Forecasting

Defu Cao
defucao@usc.edu
University of Southern California
Los Angeles, CA, USA

Yan Liu
yanliu.cs@usc.edu
University of Southern California
Los Angeles, CA, USA

Abstract

While foundation models have revolutionized text and video processing, time series data presents distinct challenges due to missing values and multi-resolution characteristics. Autoregressive transformers often learn deterministic dependencies while overlooking inherent uncertainties. We introduce TimeDiT, a diffusion transformer model that combines transformer-based temporal dependency learning with diffusion-based probabilistic sampling. TimeDiT employs a unified masking mechanism to harmonize training and inference across diverse tasks. Our evaluation demonstrates TimeDiT's effectiveness in zero-shot forecasting, establishing it as a foundation model that bridges the gap between general-purpose and domain-specific approaches.

ACM Reference Format:
Defu Cao and Yan Liu. 2025. TimeDiT: Diffusion Transformers Foundation Model for Time Series Forecasting . In *Proceedings of In Proceedings of the 31th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25) Symposium (Conference KDD).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Time series analysis is fundamental across natural science, sustainability, and healthcare [3, 6, 13, 30]. While specialized models like TCNs [9], LSTMs [26], GNNs [29], and Transformers [31] have advanced in specific time series tasks, their domain-specific design limits broader applicability. Inspired by pre-trained models like Deepseek [17], GPT-4 [21], LLaMA [28] and Vision Transformer [8], researchers have begun exploring universal time series forecasting models [2, 10, 18].

Recent advances in time series foundation models have established promising directions. Current approaches employ diverse tokenization strategies—such as TimeMoE [25] and TimesFM [7]'s patching—yet present opportunities for improved generalization. Channel independence strategies [19] adopted by models like Timer [18] and Chronos [2] enable efficient model scaling but suggest potential for enhanced modeling of temporal patterns and cross-channel dependencies. Moreover, conventional auto-regressive models typically learn deterministic mappings, limiting their ability to capture inherent uncertainties.

Real-world time series data poses fundamental challenges including missing values [14], multi-resolution sampling [20], and irregular temporal intervals [4] that disrupt pattern learning. Current benchmarks [1, 16, 32] often fail to reflect these complexities.

Following foundational principles of comprehensive pre-training, adaptable task-specific pipelines, and knowledge integration, we introduce TimeDiT, a foundation model designed to process practical time series data across domains, frequencies, and sampling patterns. As a diffusion transformer-based approach [22], TimeDiT combines the transformer architecture with diffusion models' capacity to explore diverse solutions, incorporating comprehensive time series mask units for both task-agnostic pre-training and task-specific inference.

Our contributions include:

- We introduce TimeDiT, a foundation model that integrates transformer-based temporal modeling with diffusion sampling capabilities. Enhanced by unified masking mechanisms, TimeDiT transcends conventional foundation models' focus on forecasting, enabling comprehensive self-supervised learning for diverse time series tasks.
- We validate TimeDiT through comprehensive zero-shot experiments aligned with our foundational principles, demonstrating state-of-the-art performance and cross-task adaptability.

2 Related Work

Recent time series foundation models like Timer [18], Chronos [2], and TimeMoE [25] primarily focus on forecasting tasks. General approaches include string encoding [10], language representation alignment [11], and decomposition techniques [5]. In diffusion models, CSDI [27] pioneered imputation while TSDiff [14] addressed multiple tasks but required separate models. TimeDiT uniquely offers a unified architecture for diverse time series tasks through a single pre-trained model with adaptive masking strategies.

3 Methodology

3.1 Problem Definition

We denote a multivariate time series as $\mathbf{X} = \{x_{i,j}\} \in \mathbb{R}^{K \times L}$, where K is the number of features and L is the length. An observation mask $\mathbf{M_{obs}} = \{m_{i,j}\} \in \{0, 1\}^{K \times L}$ indicates missing values ($m_{i,j} = 0$) or observed values ($m_{i,j} = 1$).

Conference KDD, June 03–05, 2025, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

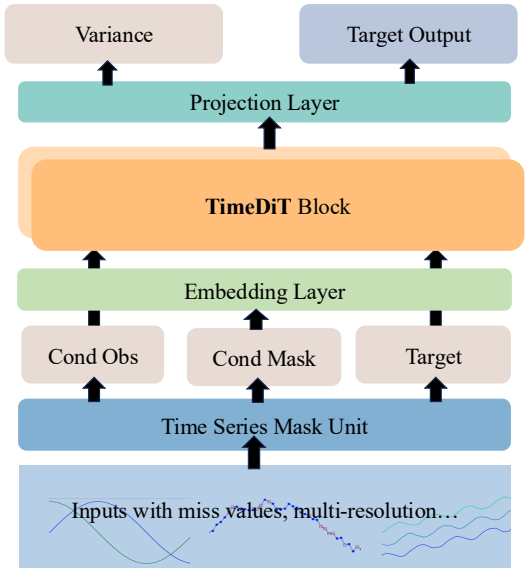

Figure 1: Overview of our proposed TimeDiT.

Let $\mathbf{x}_0^{\text{obs}} \in X^{\text{obs}}$ denote the observed subsequence; $\mathbf{x}_0^{\text{tar}}$ denote the target subsequence which could be forecast target, imputation target, or the whole sequence depending on the task. Let $\mathbf{x}_0^{\text{con}}$ denote the unmasked partial observations which act as self-conditions for the masked area $\mathbf{x}_0^{\text{tar}}$.

Using subscripts of $\mathbf{x}$ to denote diffusion timestamps (0 means no noise applied), our goal is to approximate the true conditional time series distribution: $q_X\left(\mathbf{x}_0^{\text{tar}} \mid \mathbf{x}_0^{\text{con}}\right)$ with a model distribution:

$$p_\theta\left(\mathbf{x}_{0:T}^{\text{tar}} \mid \mathbf{x}_0^{\text{con}}\right) := p\left(\mathbf{x}_T^{\text{tar}}\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1}^{\text{tar}} \mid \mathbf{x}_t^{\text{tar}}, \mathbf{x}_0^{\text{con}}\right)$$

where $\mathbf{x}_T^{\text{tar}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The mask mechanism $\mathbf{M}$ identifies the positions of $\mathbf{x}_0^{\text{con}}$ and $\mathbf{x}_0^{\text{tar}}$, enabling adaptation to tasks like forecasting, imputation, and anomaly detection.

3.2 Time Series Diffusion Transformer

As shown in Figure 1, TimeDiT establishes $\mathbf{M}_{\text{obs}}$ based on inputs with varying shapes, missing values, and multi-resolution data. The unified time series mask unit constructs masks and adapts to diverse scenarios, generating $\mathbf{x}_0^{\text{con}}$ and $\mathbf{x}_0^{\text{tar}}$ with shape $\mathbb{R}^{B\times L\times K}$ (B is batch size). This enables TimeDiT to learn robust representations by reconstructing the original sequence through denoising $\mathbf{x}_T^{\text{tar}}$.

The embedding layer maps $\mathbf{x}_0^{\text{con}}$ and noised $\mathbf{x}_0^{\text{tar}}$ into a continuous token space without vector quantization [15], preserving input integrity. The TimeDiT block autonomously learns cross-channel and temporal correlations through end-to-end training.

We train the denoising model using a weighted mean squared error loss:

$$L(\mathbf{x}_0^{\text{con}}) = \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}_t^{\text{tar}}|\mathbf{x}_0^{\text{con}})} \|\mu(\mathbf{x}_t^{\text{tar}}, \mathbf{x}_0^{\text{con}}) - \mu_\theta(\mathbf{x}_t^{\text{tar}}, t|\mathbf{x}_0^{\text{con}})\|^2$$

where $\mu(\mathbf{x}_t^{\text{tar}}, \mathbf{x}_0^{\text{con}})$ is the mean of the posterior $q(x_{t-1}^{\text{tar}}|\mathbf{x}_0^{\text{con}}, \mathbf{x}_t^{\text{tar}})$.

For conditional information of transformer, we employ adaptive layer normalization (AdaLN):

$$\text{AdaLN}(h, c) = c_{\text{scale}} \cdot \text{LayerNorm}(h) + c_{\text{shift}}$$

where h is the hidden state and c_{scale} and c_{shift} are parameters derived from $\mathbf{x}_0^{\text{con}}$. This method proved more effective than input concatenation [24], as it leverages the scale and shift of $\mathbf{x}_0^{\text{con}}$ for capturing temporal continuity.

3.3 Time Series Mask Unit

The Time Series Mask Unit incorporates multiple mask types to enhance model versatility. It generates four mask types: random mask $\mathbf{M}^{\text{R}}$, block mask $\mathbf{M}^{\text{B}}$, stride mask $\mathbf{M}^{\text{S}}$, and reconstruction mask $\mathbf{M}^{\text{Rec}}$: For random masking, we generate $\mathbf{M}^{\text{R}}(x, r) = \mathbb{1}[z_{i,j} > r]$ where $z \sim \text{Uniform}(0, 1)$ and r is the mask ratio. Block masking uses $\mathbf{M}^{\text{B}}(x, l) = \mathbb{1}[j < L - l]$ where l is the prediction length, useful for forecasting tasks. Stride masking employs $\mathbf{M}^{\text{S}}(x, n_{\text{blocks}}) = \mathbb{1}[\lfloor \frac{j}{b} \rfloor \bmod 2 = 0]$ where $b = \lceil \frac{L}{n_{\text{blocks}}} \rceil$, enabling learning from non-contiguous segments. Reconstruction masking simply uses $\mathbf{M}^{\text{Rec}} = 0$ for tasks like anomaly detection and data generation.

During pre-training, these varied masking strategies help the model develop robust representations. For task-specific applications, appropriate masks are selected based on the downstream task requirements.

4 Experiments

Table 1: Forecasting results on CRPS_{sum} for zero-shot setting.

Dataset	TimeDiT	TEMPO [5]	LagLLaMA [23]	TimeLLM [12]	Timer [18]
Solar	**0.424**	<u>0.581</u>	0.690	0.997	1.001
Electricity	**0.030**	0.081	0.065	0.303	0.301
Taxi	**0.392**	<u>0.400</u>	0.620	0.782	0.788
Exchange	**0.019**	0.030	<u>0.024</u>	0.076	0.072

Our evaluation of Zero-shot forecasting reveals TimeDiT's strong zero-shot forecasting capabilities across diverse datasets. In Table 1, TimeDiT outperforms specialized models on most benchmarks despite operating in a zero-shot environment. The model achieves state-of-the-art results on Solar, Electricity, and Taxi datasets, demonstrating its robust generalization capabilities without domain-specific fine-tuning. TimeDiT's performance is particularly notable compared to larger models like TimeLLM and Timer, which require significantly more parameters yet deliver inferior results. While TEMPO maintains competitive performance on the Traffic dataset, TimeDiT's consistent performance across varied prediction horizons suggests superior temporal pattern learning. These results establish TimeDiT as an effective foundation model for time series forecasting, balancing generalization with accurate uncertainty quantification.

5 Conclusion

TimeDiT demonstrates exceptional zero-shot forecasting capabilities across diverse datasets, consistently outperforming specialized models and establishing itself as a promising foundation model for time series analysis with robust uncertainty quantification.

References

[1] Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Trkmen, and Yuyang Wang. 2020. GluonTS: Probabilistic and Neural Time Series Modeling in Python. *Journal of Machine Learning Research* 21, 116 (2020), 1–6. http://jmlr.org/papers/v21/19-820.html

[2] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815* (2024).

[3] Manuel Burger, Fedor Sergeev, Malte Londschien, Daphné Chopard, Hugo Yèche, Eike Christian Gerdes, Polina Leshetkina, Alexander Morgenroth, Zeynep Babür, Jasmina Bogojeska, et al. 2024. Towards Foundation Models for Critical Care Time Series. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*.

[4] Defu Cao, James Enouen, Yujing Wang, Xiangchen Song, Chuizheng Meng, Hao Niu, and Yan Liu. 2023. Estimating treatment effects from irregular time series observations with hidden confounders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 6897–6905.

[5] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2023. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948* (2023).

[6] Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. 2022. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing* 92, 3 (2022), 88.

[7] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2023. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688* (2023).

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).

[9] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. 2019. Unsupervised scalable representation learning for multivariate time series. *Advances in neural information processing systems* 32 (2019).

[10] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems* 36 (2024).

[11] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[12] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Unb5CVPtae

[13] Nitin Kamra, Yizhou Zhang, Sirisha Rambhatla, Chuizheng Meng, and Yan Liu. 2021. PolSIRD: modeling epidemic spread under intervention policies: analyzing the first wave of COVID-19 in the USA. *Journal of Healthcare Informatics Research* 5, 3 (2021), 231–248.

[14] Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Bernie Wang. 2023. Predict, Refine, Synthesize: Self-Guiding Diffusion Models for Probabilistic Time Series Forecasting. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=q6X038vKgU

[15] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. 2024. Autoregressive Image Generation without Vector Quantization. *arXiv preprint arXiv:2406.11838* (2024).

[16] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations (ICLR '18)*.

[17] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[18] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2024. Timer: Transformers for Time Series Analysis at Scale. *arXiv preprint arXiv:2402.02368* (2024).

[19] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations (ICLR '23)*.

[20] Hao Niu, Guillaume Habault, Roberto Legaspi, Chuizheng Meng, Defu Cao, Shinya Wada, Chihiro Ono, and Yan Liu. 2023. Time-delayed Multivariate Time Series Predictions. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 325–333.

[21] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[22] William Peebles and Saining Xie. 2022. Scalable Diffusion Models with Transformers. *arXiv preprint arXiv:2212.09748* (2022).

[23] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, et al. 2023. Lag-llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278* (2023).

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[25] Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. 2024. Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts. arXiv:2409.16040 https://arxiv.org/abs/2409.16040

[26] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International conference on big data (Big Data)*. IEEE, 3285–3292.

[27] 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.

[28] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv* abs/2302.13971 (2023). https://api.semanticscholar.org/CorpusID:257219404

[29] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. 2020. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 753–763.

[30] Wen Ye, Yizhou Zhang, Wei Yang, Lumingyuan Tang, Defu Cao, Jie Cai, and Yan Liu. 2024. Beyond Forecasting: Compositional Time Series Reasoning for End-to-End Task Execution. *arXiv preprint arXiv:2410.04047* (2024).

[31] Yunhao Zhang and Junchi Yan. 2022. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*.

[32] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of AAAI*.