TimeDiT: General-purpose Diffusion Transformers for Time Series Foundation Model

Defu Cao [*1] **Wen Ye** [*1] **Yan Liu** [1]

Abstract

Time series modeling is critical for many real-world applications, but most existing approaches are task-specific. With the unique characteristics such as missing values, irregular sampling, multi-resolution and complex temporal dependencies, it is challenging to develop general foundation models for time series. In this paper, we introduce the Time Series Diffusion Transformer (TimeDiT) equipped with three distinct masking schemes designed to facilitate a uniform training and inference pipeline across various time series tasks. TimeDiT leverages the transformer architecture for capturing temporal dependencies and employs diffusion processes for generating high-quality candidate samples without stringent assumptions on the target distribution. Extensive experiments conducted on different datasets encompassing tasks such as forecasting, imputation, and anomaly detection demonstrate the model's effectiveness. Both in-domain and zero-shot testing scenarios confirm the potential of our model to serve as a robust foundation model for multiple time series applications.

1. Introduction

Time series analysis is pivotal in a diverse set of AI applications, such as science (Cuomo et al., 2022), sustainability (Krenn & Buffoni, 2023), health (Kaushik et al., 2020), etc. These applications root in diverse domains (Li et al., 2018; Bi et al., 2023; Cao et al., 2023b), leading to time series with various distributions (Wang et al., 2023) and a divers set of analysis tasks, such as forecasting (Nie et al., 2022), imputation (Tashiro et al., 2021), anomaly detection (Zhao et al., 2020), etc. Even though considerable progress has been made in developing specialized models optimized for specific scenarios and individual tasks, an open question remains: *Can a single time series foundation model excel across domains?* Recent initiatives have explored the possibility of universal time series models on zero-shot setting (Ansari et al., 2024; Liu et al., 2024; Gruver et al., 2024; Cao et al., 2023a), drawing inspiration from large pre-trained language models in natural language processing(NLP) and computer vision(CV), such as GPT (Radford et al., 2018), CLIP (Radford et al., 2021), which are known for their robust transfer learning capabilities. However, due to the fundamentally different semantics between text/images and time series data, the unique challenges of achieving a truly flexible and general-purpose time series model remain an open problem.

Compared with texts and images, time series exhibit unique characteristics such as *missing values* (Kollovieh et al., 2023), *irregular sampling* (Kidger et al., 2020), *multi-resolution* (Meng et al., 2022), *complex temporal dependencies*, etc. To address these challenges, a foundation model for time series must be capable of capturing long-term patterns and demonstrating flexibility across different scales to handle diverse inputs with varying distributions. Moreover, time series processes are often governed by underlying physical principles (Li et al., 2021) and can be guided by domain-specific textual information (Jin et al., 2023; Sun et al., 2023). However, integrating these diverse sources of information into a unified model poses further challenges, as the model must effectively leverage the relevant physics and textual context while adapting to the unique characteristics and distributions of each domain. Addressing each of these issues requires innovative approaches in data preprocessing, model architecture, and training strategies to create models that can seamlessly handle the diverse and complex nature of time series data.

Recently, the emergence of LLMs like GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023) suggests the potential for building time series foundation models enabling a general solution to handle multiple time series distributions. Previous attempts typically build upon the transformer backbone, which has achieved state-of-the-art performance on various time series tasks, particularly in modeling long-term dependencies. However, transformers struggle with probabilistic tasks due to their need for spe-

[*] Equal contribution with alphabetical order [1] University of Southern California, Los Angeles, USA. Correspondence to: Defu Cao <defucao@usc.edu>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

cific functional approximations to optimize a tractable lower bound to the likelihood. In contrast, diffusion models like DDPM (Ho et al., 2020) handle data generation through sequential conditional transformations, turning the density estimation task into a sequential reconstruction. A hybrid approach combining transformers' temporal modeling with diffusion models' probabilistic generation could pave the way for a versatile time series foundation model. In addition, the tokenization of time series data is especially sensitive to variations in data sources and sampling rates. Previous tokenization approaches with different schemes including token patching (Das et al., 2023; Woo et al., 2024); discretization tokens (Talukder et al., 2024) and tokens based on time series features (Yue et al., 2022; Ansari et al., 2024; Rasul et al., 2023) have either fragmented the global information or have been constructed in a manner that inherently loses important information. Moreover, most, if not all, of previous related works employ a channel independence strategy (Nie et al., 2023) or focus solely on univariate time series. Channel independence strategy, though beneficial in certain contexts, often overlooks the complex inter-temporal and cross-feature dependencies in practical applications and thus presents an opportunity for optimization (Zeng et al., 2023). In addition, there is also a lack of a unified pipeline to handle indeterminate data shapes, as well as a tendency to neglect practical challenges in favor of performing well on neatly organized benchmarks.

To tackle the aforementioned challenges, we introduce TimeDiT—a diffusion transformer-based foundation model equipped with a standardized training pipeline for different shapes of input time series and tailored for diverse distributions and downstream tasks. We apply the new unified paradigm of TimeDiT on multiple challenging time series datasets, including applications from traffic, weather, and financial domains, among others. The model is evaluated on a comprehensive set of downstream tasks, such as synthetic data generation, imputation, probabilistic forecasting, and anomaly detection. We propose a comprehensive mask mechanism for reconstruction pretraining and task-specific finetuning. TimeDiT achieves state-of-the-art or competitive results on both in-domain and zero-shot settings, demonstrating its effectiveness and efficiency across various time series applications. In addition, the results on zero-shot experiments show that our model can be used as a foundation model even without fine-tuning, although fine-tuning may be necessary in some cases. Moreover, our model is scalable, easily adaptable for downstream tasks, and can incorporate external knowledge. For example, TimeDiT can be continuously trained on multi-modal data with textual information, allowing it to leverage additional context and improve its performance on tasks involving both time series and textual descriptions. This flexibility and adaptability make TimeDiT a powerful and versatile tool for a wide range of time series applications. In summary, our contributions can be summarized as three unfolds:

- Introducing TimeDiT: a general-purpose time series foundation model that merges the strengths of diffusion and transformer models and considers practical applications of real-world time series.
- Towards Foundation Training Pipeline: We develop three distinct mask types, optimizing the training pipeline of our time series foundation model to ensure seamless adaptation to various downstream tasks using corresponding masks during inference.
- Extensive Experiments: Our model is tested across four different downstream tasks, demonstrating robust in-domain/zero-shot performance and versatility, further enhanced by integrating expert knowledge.

2. Preliminaries

Diffusion Models In recent years, diffusion models have emerged as a promising approach in generative modeling. A diffusion process is a Markov chain that incrementally adds Gaussian noise to data over a sequence of steps, effectively destroying the data structure in forward process and destroying the data structure in backward structure.

The forward process adds noise to the data $\mathbf{x}_0$ over a series of timesteps t according to a variance schedule β_t, resulting in a set of noisy intermediate variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. Each subsequent $\mathbf{x}_t$ is derived from the previous step by applying Gaussian noise:

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

The reverse process aims to denoise the noisy variables step by step, sampling each $\mathbf{x}_{t-1}$ from the learned distribution $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$. This distribution, modeled by a neural network parameterized by θ, approximates the Gaussian distribution:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \tag{2}$$

By iterating this reverse process from $t = T$ down to $t = 0$, the model gradually reconstructs the original data from noise. The reverse process learns to predict the mean and covariance of each intermediate distribution, effectively approximating the original data distribution.

3. Related Work

General Purpose Time Series Model In the past decades, researchers have developed sophisticated models for specific time series tasks. Recently, the advent of large language

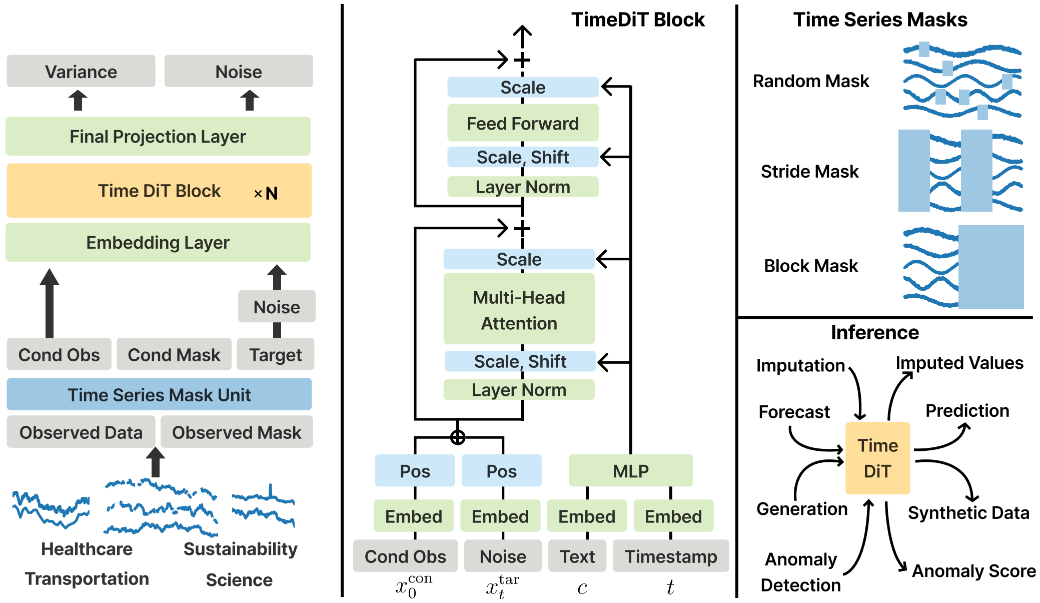

Figure 1. TimeDiT Architecture. Left: TimeDiT framework with diverse multivariate time series from different domains with multi-resolution, missing values; Middle: Structure of TimeDiT block; Right top: Illustration of masks generated by Time Series Mask Unit, reconstruction mask is neglected as it's an all-zero mask; Right bottom: Downstream tasks that TimeDiT handles during inference.

models has spurred the shift towards general-purpose and foundation time series models(Zerveas et al., 2021). While most of the approaches achieved generalization through evaluation across datasets, they mainly focused on the forecasting task(Woo et al., 2024; Das et al., 2023). Specifically, (Gruver et al., 2024) simply encoded time series as strings while (Jin et al., 2023) converted time series into language representations by alignment. (Cao et al., 2023a) further incorporated decomposition technique and prompt design and generalizes to unseen data and multimodal scenarios. (Rasul et al., 2023; Ansari et al., 2024) worked towards the foundation model from a probabilistic perspective but only considered univariate time series. Additionally, many studies started to follow a two-stage training paradigm of pretraining and finetuning (Chang et al., 2023; Dong et al., 2024; Nie et al., 2022). However, there remains substantial room for innovation in comprehensive, general task time series models rather than task-specific solutions. (Zhou et al., 2023a) employed distinct pre-trained models under GPT-2(Radford et al., 2019) structure tailored to specific downstream applications.(Talukder et al., 2024) leveraged the VQVAE tokenizer(Van Den Oord et al., 2017) to build discrete tokens for a transformer to handle time series tasks. However, this approach requires the separate construction of a pre-trained tokenizer, which may constrain the model's generalization capabilities. For more detailed literature of the general-purpose time series model, please refer to recent surveys and position paper(Liang et al., 2024; Jin et al., 2024; Jiang et al., 2024)

Diffusion models for Time Series Despite the growing interest of diffusion models in various scenarios (Peebles & Xie, 2022; Li et al., 2022a; Lu et al., 2024), the use of diffusions in time series analysis is less explored compared to pre-trained language models and transformers. Most existing studies also focused solely on forecasting and the choice of backbone model also varies among VAE(Li et al., 2022b), RNN(Rasul et al., 2021), and transformer. CSDI (Tashiro et al., 2021) utilized diffusion model for time series imputation. (Yuan & Qiao, 2024) incorporated decomposition into diffusion model to improve interoperability. Although (Kollovieh et al., 2023) build a diffusion pipeline for multiple tasks with refinement, they still train different models for each task. To the best of our knowledge, there has been no exploration of leveraging unified diffusion models for a comprehensive set of time series tasks yet. Please refer to (Yang et al., 2024) for a comprehensive literature review on diffusion models for time series analysis.

4. Methodology

4.1. 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 of the time series. Each individual entry $x_{i,j}$ represents the j-th feature at time step l, for $i \in \{1, \ldots, K\}$ and $j \in \{1, \ldots, L\}$. We define an observation mask $\mathbf{M_{obs}} = \{m_{i,j}\} \in \{0,1\}^{K \times L}$, where $m_{i,j} = 0$ if $x_{i,j}$ is missing, and $m_{i,j} = 1$ if $x_{i,j}$ is observed. Let $\mathbf{x}_0^{\text{obs}} \in X^{\text{obs}}$

denote the observed subsequence; $\mathbf{x}_0^{\text{tar}} \in X^{\text{tar}}$ denote the target subsequence (forecast target or imputation target or the whole sequence) which are subsets of the sample space $\mathbf{X}$. Let $\mathbf{x}_0^{\text{con}} \in \mathbf{X}$ denote the conditional observations. Formally, the goal of our task is to approximate the true conditional data distribution given the conditional information $q_{\mathbf{X}}\left(\mathbf{x}_0^{\text{ta}} \mid \mathbf{x}_0^{\text{con}}\right)$ with a model distribution $p_\theta(\mathbf{x}_0^{\text{tar}} \mid \mathbf{x}_0^{\text{con}})$, which can be calculated by a diffusion model.

4.2. Time Series Diffusion Transformer

Figure 1 shows the overall framework of TimeDiT. We first establish the $\mathbf{M_{obs}}$ and $\mathbf{x}_0^{\text{obs}}$ based on the given input from different distributions with multivariate sequences, missing value and multi-resolution by injecting placeholders to standardize the input shape across different time series, facilitating more efficient and consistent processing. Then, the unified time series mask unit adapts to diverse time series scenarios and builds the $\mathbf{x}_0^{\text{con}}$, $\mathbf{M}$ and $\mathbf{x}_0^{\text{tar}}$, with shape $\mathbb{R}^{B \times L \times K}$, to help TimeDiT learn robust representations in a self-supervised manner by reconstruction the noised $\mathbf{x}_0^{\text{tar}}$. After that, the embedding layer directly treats $\mathbf{x}_0^{\text{con}}$ and $\mathbf{x}_0^{\text{tar}}$ as input tokens without any reconstruction or patching, as the diffusion process is designed to handle multivariate input and operate in a continuous token space. By preserving the integrity of the input time series, TimeDiT ensures that the model can effectively capture and utilize the rich information contained within the data. The TimeDiT block's attention mechanism is designed to autonomously learn cross-channel and temporal correlations through end-to-end training.

Time Series Mask Unit. We propose a unified time series mask mechanism that includes *task-agnostic masks* and *task-specific masks*, which seamlessly integrates with the model during self-supervised pre-training and fine-tuning to cater to diverse time series scenarios. During the pre-training stage, we randomly select masks from random mask, stride mask, reconstruction mask, and block mask and train the model to recover the masked values. This step aims to improve the overall time series representation by encouraging the model to learn robust and generalizable features from the input time series. Secondly, we employ task-specific masks, including block mask, predefined imputation mask, and reconstruction mask,to adapt TimeDiT to the most common downstream time series tasks, including forecasting, imputation, anomaly detection, as well as synthetic time series generation. This strategy enables the model to adapt to the unique requirements of each task.

Metric	Methods	Sine	Stocks	Air Quality	Energy
DS	TimeDiT	**0.0075±0.004**	**0.115±0.008**	0.1778±0.004	**0.1726±0.005**
	Diffusion-TS	0.0099±0.003	0.1869±0.0159	**0.1227±0.006**	0.2301±0.006
	TimeGAN	0.1217±0.039	0.2038±0.057	0.3913±0.039	0.4969 ±0.000
	TimeVAE	0.0489±0.0562	0.1987±0.037	0.2869±0.053	0.4993±0.001
PS	TimeDiT	**0.1909±0.000**	0.0459±0.000	**0.0208±0.001**	**0.2502±0.000**
	Diffusion-TS	0.2262±0.000	**0.042±0.000**	0.022±0.002	0.2506±0.000
	TimeGAN	0.2797±0.015	0.0481±0.002	0.035±0.002	0.3305±0.003
	TimeVAE	0.2285±0.000	0.0485±0.000	0.0269±0.001	0.2878±0.001

Table 1. Synthetic Generation result on 24-length multivariate time series. We calculate discriminative and predictive score according to (Yoon et al., 2019) and results are averaged over five runs. **Bold** indicates the best performance. DS: Discriminative Score; PS: Predictive Score

Conditional Injection. During the diffusion process, to incorporate the current timestep t of the process stage, we integrate a time component following adaptive layer normalization in the TimeDiT block (Peebles & Xie, 2022). This time integration can be expressed as $\text{AdaLN}(h, t) = t_{scale}\text{LN}(h) + t_{shift}$, where h is the hidden state and t_{scale} and t_{shift} are the scale and shift parameters derived from the time embedding. Note that integrating textual information aligns well with the concept of adaptive layer normalization. In addition, given that TimeDiT utilizes a pure transformer architecture, a straightforward and intuitive approach is to include conditional frames directly as part of the input sequence. We achieve this by concatenating the latent features of the conditional time series $\mathbf{x}_0^{\text{con}}$ with the noisy frames at the token level.

5. Experiments

In this section, we experimentally demonstrate the effectiveness of our proposed framework, TimeDiT. We conduct a comparative analysis with several leading models in the field. Our baseline models include probabilistic generation-based models such as Diffusion-TS (Yuan & Qiao, 2024), TimeGAN (Yoon et al., 2019) and TimeVAE (Desai et al., 2021); deterministic models such as GPT-2 (Zhou et al., 2023b), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2022). Our comparisons include in-domain settings and zero-shot settings on multiple downstream tasks.

6. Synthetic Generation

We conduct experiments to synthesize multivariate time series and evaluate performance using the discriminative score and predictive score metrics under a 'train on synthetic test on real' experimental setting (Yuan & Qiao, 2024). Table 5 shows the result on synthetic generation where TimeDiT, in general, consistently generates higher quality synthetic samples compared to baselines, even on high-dimensional datasets such as the energy dataset. This demonstrates TimeDiT's strength in complex time series synthesis.

6.1. Imputation

We conduct experiments on six real-world datasets: ETTh1, ETTh2, ETTm1, ETTm2, Electricity, and Weather. We use random mask ratios $\{12.5\%, 25\%, 37.5\%, 50\%\}$ following previous settings (Zhou et al., 2023b). Table 2 shows the im-

Datasets	ETTh1 MSE / MAE	ETTh2 MSE / MAE	ETTm1 MSE / MAE	ETTm2 MSE / MAE	Weather MSE / MAE	Electricity MSE / MAE	1st Pl Count
DLinear	0.201/ 0.306	0.142 / 0.259	0.093 / 0.206	0.096 / 0.208	0.052 / 0.110	0.132 / 0.260	0
LightTS	0.284 / 0.373	0.119 / 0.250	0.104 / 0.218	0.046 / 0.151	0.055 / 0.117	0.131 / 0.262	0
ETSformer	0.202 / 0.329	0.367 / 0.436	0.120 / 0.253	0.208 / 0.327	0.076 / 0.171	0.214 / 0.339	0
FEDformer	0.117 / 0.246	0.163 / 0.279	0.062 / 0.177	0.101 / 0.215	0.099 / 0.203	0.130 / 0.259	0
Autoformer	0.103 / 0.214	0.055 / 0.156	0.051 / 0.150	0.029 / 0.105	0.031 / 0.057	0.101 / 0.225	0
PatchTST	0.115 / 0.224	0.065 / 0.163	0.047 / 0.140	0.029 / 0.102	0.034 / 0.055	0.072 / 0.183	0
TimesNet	0.078 / 0.187	0.049 / 0.146	0.027 / 0.107	0.022 /0.088	**0.030** / 0.054	0.092 / 0.210	1
GPT2(3)	0.069 / 0.173	0.048 / 0.141	0.028 / 0.105	**0.021 / 0.084**	0.031 / 0.056	0.090 / 0.207	2
TimeDiT	**0.051 / 0.148**	0.045 / 0.144	**0.026 /0.100**	0.031 / 0.104	0.044 / **0.045**	0.082 / **0.182**	9
Zero-Shot	0.063 / 0.169	**0.040 / 0.130**	0.047 /0.149	0.112 / 0.197	0.051 / 0.081	**0.070** / 0.183	

Table 2. Imputation result on 96-length multivariate time series averaged over the four mask ratios. We calculate MSE and MAE for each dataset. **Bold** indicates best result, Underline indicates the second best result.

Methods	TimeDiT	GPT2(6)	TimesNet	PatchTS.	ETS.	FED.	LightTS
MSL	**92.67**	82.45	81.84	78.70	85.03	78.57	78.95
SMAP	**95.49**	72.88	69.39	68.82	69.50	70.76	69.21
SWaT	**95.78**	94.23	93.02	85.72	84.91	93.19	93.33
SMD	81.06	**86.89**	84.61	84.62	83.13	85.08	82.53
PSM	**97.73**	97.13	97.34	96.08	91.76	97.23	97.15
1st Pl Count	4	1	0	0	0	0	0

Table 3. Anomaly Detection result on 100-length multivariate time series. We calculate F1 score as % for each dataset. '.' notation in model name stands for transformer. **Bold** indicates best result, Underline indicates the second best result.

putation result averaged over the four mask ratios. TimeDiT achieves the best performance on most datasets. Additionally, the zero-shot setting's performance has also achieved the best/second-best performance on ETTh1, ETTh2, and electricity datasets. Overall, TimeDiT obtained 9 first place count while the remaining baselines obtained 3 first place count overall.

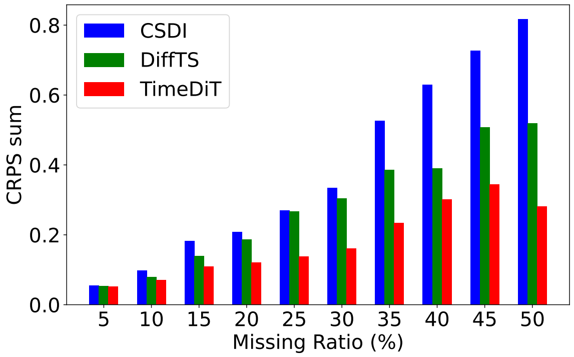

Figure 2. Forecasting results with missing value. Compared between our zero-shot TimeDiT and other diffusion-based methods.

6.2. Anomaly Detection

We conduct experiments on five real-world datasets from industrial applications: MSL, SMAP, SWaT, SMD, and PSM. Since the diffusion model excels at learning distribution. Following previous settings (Zhou et al., 2023b), we use the 99-th percentile reconstruction error to determine anomalies for MSL, SMAP, SWaT, and PSM; and use the 99.5-th percentile for SMD. We perform the standard anomaly adjustment (Xu et al., 2018) during evaluation. As shown in Table 3, TimeDiT achieves the best performance on four out of five datasets, significantly outperforming the baselines. Moreover, the zero-shot also achieved reasonably good performance on MSL, SMAP, SWaT and PSM dataset.

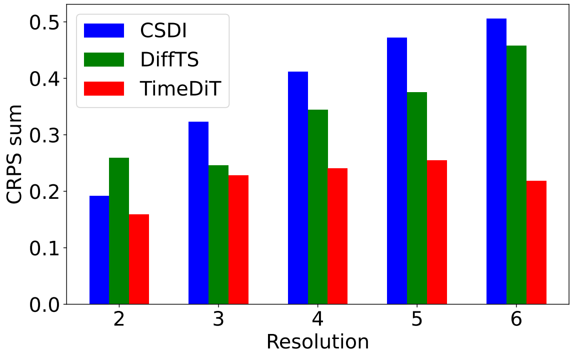

Figure 3. Forecasting results with multi resolution. Compared between our zero-shot TimeDiT and other diffusion-based methods.

6.3. Forecasting

For forecasting tasks, we build realistic missing value and multi resolution scenarios across sampling frequencies and compare our zero-shot TimeDiT with CSDI (Tashiro et al., 2021) and Diffusion-TS (Yuan & Qiao, 2024) (DiffTS in short). As shown in Figure 2 and Figure 3, TimeDiT continues to outperform the current state-of-the-art model for probabilistic time series prediction on the practical scenarios. The strong performance highlights our model's capabilities for handling prediction under real-world conditions like missing data and inherent multi resolution within the data. Even on challenging zero-shot generalizations and datasets with pervasive missing values, our approach surpasses the previous state-of-the-art for probabilistic forecasting. Note that the zero-shot TimeDiT used in Sections 6.1, 6.3 and 6.2 is trained on the uniformed datasets with varying number channels and distributions under the same model architecture.

7. Conclusion

We introduced TimeDiT, a pioneering approach to creating a versatile and robust foundation model for various time series tasks. By integrating transformer architecture with the diffusion model, TimeDiT effectively captures temporal dependencies and addresses real-world challenges unique to time series. Our innovative masking strategies: task-specific mask and task-agnostic mask allow for a consistent training framework adaptable to diverse tasks such as forecasting, imputation, and anomaly detection and synthetic data generation. Extensive experiments demonstrated the strong performance of TimeDiT on all time series tasks.

References

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Cao, D., Jia, F., Arik, S. O., Pfister, T., Zheng, Y., Ye, W., and Liu, Y. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948*, 2023a.

Cao, D., Zheng, Y., Hassanzadeh, P., Lamba, S., Liu, X., and Liu, Y. Large scale financial time series forecasting with multi-faceted model. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 472–480, 2023b.

Chang, C., Peng, W.-C., and Chen, T.-F. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *arXiv preprint arXiv:2308.08469*, 2023.

Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., and Piccialli, F. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Desai, A., Freeman, C., Wang, Z., and Beaver, I. Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

Dong, J., Wu, H., Zhang, H., Zhang, L., Wang, J., and Long, M. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Jiang, Y., Pan, Z., Zhang, X., Garg, S., Schneider, A., Nevmyvaka, Y., and Song, D. Empowering time series analysis with large language models: A survey. *arXiv preprint arXiv:2402.03182*, 2024.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Jin, M., Zhang, Y., Chen, W., Zhang, K., Liang, Y., Yang, B., Wang, J., Pan, S., and Wen, Q. Position paper: What can large language models tell us about time series analysis. *arXiv preprint arXiv:2402.02713*, 2024.

Kaushik, S., Choudhury, A., Sheron, P. K., Dasgupta, N., Natarajan, S., Pickett, L. A., and Dutt, V. Ai in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in big data*, 3:4, 2020.

Kidger, P., Morrill, J., Foster, J., and Lyons, T. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33: 6696–6707, 2020.

Kollovieh, M., Ansari, A. F., Bohlke-Schneider, M., Zschiegner, J., Wang, H., and Wang, B. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=q6X038vKgU.

Krenn, M. and Buffoni, L. Predicting the future of ai with ai: High-quality link prediction in an exponentially growing knowledge network. *Nature machine intelligence*, 2023.

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022a.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR '18)*, 2018.

Li, Y., Lu, X., Wang, Y., and Dou, D. Generative time series forecasting with diffusion, denoise, and disentanglement. *Advances in Neural Information Processing Systems*, 35: 23009–23022, 2022b.

Li, Z., Kovachki, N. B., Azizzadenesheli, K., liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., Pan, S., and Wen, Q. Foundation models for time series analysis: A tutorial and survey. *arXiv preprint arXiv:2403.14735*, 2024.

Liu, Y., Zhang, H., Li, C., Huang, X., Wang, J., and Long, M. Timer: Transformers for time series analysis at scale. *arXiv preprint arXiv:2402.02368*, 2024.

Lu, H., Yang, G., Fei, N., Huo, Y., Lu, Z., Luo, P., and Ding, M. VDT: General-purpose video diffusion transformers via mask modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Un0rgm9f04.

Meng, C., Niu, H., Habault, G., Legaspi, R., Wada, S., Ono, C., and Liu, Y. Physics-informed long-sequence forecasting from multi-resolution spatiotemporal data. In *IJCAI*, pp. 2189–2195, 2022.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Nie, Y., H. Nguyen, N., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR '23)*, 2023.

OpenAI. Gpt-4 technical report, 2023.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Rasul, K., Seward, C., Schuster, I., and Vollgraf, R. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pp. 8857–8868. PMLR, 2021.

Rasul, K., Ashok, A., Williams, A. R., Khorasani, A., Adamopoulos, G., Bhagwatkar, R., Biloš, M., Ghonia, H., Hassen, N. V., Schneider, A., et al. Lag-llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Sun, C., Li, Y., Li, H., and Hong, S. Test: Text prototype aligned embedding to activate llm's ability for time series. In *The Twelfth International Conference on Learning Representations*, 2023.

Talukder, S., Yue, Y., and Gkioxari, G. Totem: Tokenized time series embeddings for general time series analysis, 2024.

Tashiro, Y., Song, J., Song, Y., and Ermon, S. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Wang, R., Dong, Y., Arik, S. O., and Yu, R. Koopman neural operator forecaster for time-series with temporal distributional shifts. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kUmdmHxK5N.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., et al. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 world wide web conference*, pp. 187–196, 2018.

Yang, Y., Jin, M., Wen, H., Zhang, C., Liang, Y., Ma, L., Wang, Y., Liu, C., Yang, B., Xu, Z., et al. A survey on diffusion models for time series and spatio-temporal data. *arXiv preprint arXiv:2404.18886*, 2024.

Yoon, J., Jarrett, D., and Van der Schaar, M. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

Yuan, X. and Qiao, Y. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*, 2024.

Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., and Xu, B. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., and Eickhoff, C. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2114–2124, 2021.

Zhao, H., Wang, Y., Duan, J., Huang, C., Cao, D., Tong, Y., Xu, B., Bai, J., Tong, J., and Zhang, Q. Multivariate time-series anomaly detection via graph attention network. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 841–850. IEEE, 2020.

Zhou, T., Niu, P., Sun, L., Jin, R., et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023a.

Zhou, T., Niu, P., Wang, X., Sun, L., and Jin, R. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 2023b.

A. Preliminaries

Diffusion Models In recent years, diffusion models have emerged as a promising approach in generative modeling. A diffusion process is a Markov chain that incrementally adds Gaussian noise to data over a sequence of steps, effectively destroying the data structure in forward process and destroying the data structure in backward structure.

The forward process adds noise to the data $\mathbf{x}_0$ over a series of timesteps t according to a variance schedule β_t, resulting in a set of noisy intermediate variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. Each subsequent $\mathbf{x}_t$ is derived from the previous step by applying Gaussian noise:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{3}$$

The reverse process aims to denoise the noisy variables step by step, sampling each $\mathbf{x}_{t-1}$ from the learned distribution $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$. This distribution, modeled by a neural network parameterized by θ, approximates the Gaussian distribution:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \tag{4}$$

By iterating this reverse process from $t = T$ down to $t = 0$, the model gradually reconstructs the original data from noise. The reverse process learns to predict the mean and covariance of each intermediate distribution, effectively approximating the original data distribution.