T2S: High-resolution Time Series Generation with Text-to-Series Diffusion Models

Yunfeng Ge[1,2], **Jiawei Li**[2,7], **Yiji Zhao**[3], **Haomin Wen**[4], **Zhao Li**[5], **Meikang Qiu**[6], **Hongyan Li**[1], **Ming Jin**[2*] and **Shirui Pan**[2*]
[1]School of Telecommunications Engineering, Xidian University
[2]School of Information and Communication Technology, Griffith University
[3]School of Information Science and Engineering, Yunnan University
[4]Carnegie Mellon University
[5]College of Computer Science and Technology, Zhejiang University
[6]School of Computer and Cyber Sciences, Augusta University
[7]The Hong Kong University of Science and Technology (Guangzhou)
yfge@stu.xidian.edu.cn, jli226@connect.hkust-gz.edu.cn, yjzhao@ynu.edu.cn,
haominwe@andrew.cmu.edu, lzjoey@gmail.com, qiumeikang@yahoo.com,
hyli@xidian.edu.cn, mingjinedu@gmail.com, s.pan@griffith.edu.au

Abstract

Text-to-Time Series generation holds significant potential to address challenges such as data sparsity, imbalance, and limited availability of multimodal time series data across domains. While diffusion models have achieved remarkable success in Text-to-X (e.g., vision and audio data) generation, their use in time series generation remains limit. Existing approaches face two critical limitations: (1) reliance on domain-specific captions that generalize poorly, and (2) inability to generate time series of arbitrary length, limiting real-world use. In this work, we first introduce a new multimodal dataset containing over 600,000 high-resolution text-time series pairs. Second, we propose **T**ext-**to**-**S**eries (**T2S**), a diffusion-based framework that bridges the gap between natural language and time series in a domain-agnostic manner. It employs a length-adaptive VAE to encode time series of varying lengths into consistent latent embeddings. On top of that, **T2S** effectively aligns textual representations with latent embeddings by utilizing Flow Matching and employing DiT as the denoiser. We train **T2S** in an interleaved paradigm across multiple lengths, allowing it to generate sequences of arbitrary lengths. Extensive evaluations demonstrate that **T2S** achieves state-of-the-art performance across 13 datasets spanning 12 domains.

1 Introduction

Time series generation (TSG) enables the creation of high-quality data in scenarios with limited available datasets, thus serving as a way to simulate diverse, multimodal temporal dynamics, which offers significant value in real-world applications. Generating required objects from text (Text-to-X) is a research area that meets human needs and has potential at the present time. Driven by the success of diffusion models, Text-to-X generation has made remarkable strides in domains such as image generation [Rombach *et al.*, 2022; Esser *et al.*, 2024], video generation [Girdhar *et al.*, 2023], and speech processing [Le *et al.*, 2024]. Specifically, diffusion models refine noisy data progressively through a learned process, ultimately producing high-fidelity outputs. While Text-to-X generation using diffusion models has been extensively explored in vision and audio data, their application to time series generation is still in its early stages.

Existing studies [Yang *et al.*, 2024] on TSG with diffusion models can be classified into three types based on the conditioning information. i) Label-based condition. A class-conditioned diffusion method for generating synthetic EEG signals is introduced in [Sharma *et al.*, 2023]. ii) Temporal-based condition. Sensor data synthesis using statistical adjustments is explored in [Zuo *et al.*, 2023]. Although class- and temporal-conditioned generation is well discussed [Yuan and Qiao, 2024; Liu *et al.*, 2024b], they are less flexible than text-based approaches [Gao *et al.*, 2024]. iii) Text-based condition. [Fu *et al.*, 2024] uses domain-specific metadata (e.g., location, weather) to generate synthetic energy data, while [Wang, 2024] and [Lai *et al.*, 2025; Alcaraz and Strodthoff, 2023] apply textual conditioning to synthesize sales and clinical ECG data, respectively. However, these methods are often domain-specific, their metadata captions can not support high resolution general alignment between time series and fine-grained captions.

Despite the progress made in applying diffusion models to time series modeling, two significant challenges remain in this domain. **First**, the scarcity of high-resolution general-propose text-time series caption datasets limits progress in Text-to-Time Series (T2S) generation, while existing datasets are domain-specific (e.g., healthcare [Johnson *et al.*, 2023], economics [Cortis *et al.*, 2017]). **Second**, existing TSG models [Yuan and Qiao, 2024; Desai *et al.*, 2021] typically require

[*]Corresponding author

separate training for datasets of different lengths within each domain, making it challenging to develop length-arbitrary T2S models. Most approaches rely on predefined sequence lengths tied to the training data, limiting their ability to generalize. Since real-world time series exhibit inherent variability in length due to factors such as data collection frequency or system-specific temporal dynamics, the need for length-specific training significantly hinders the scalability and practicality of these models.

To address these challenges, we introduce a new fragment-level dataset, `TSFragment-600K`, containing over 600,000 high-resolution fragment-level text-time series pairs, which serves as a foundation for exploring T2S generation. **`T2S`**, a diffusion-based model is presented that bridges the gap between natural language and time series data in a domain-agnostic manner. Specifically, **`T2S`** utilizes a length-adaptive variational autoencoder to encode time series of varying lengths into consistent latent embeddings. The model then aligns textual representations with latent embeddings using flow matching and employs a diffusion transformer as the denoiser. By training **`T2S`** in an interleaved manner across diverse datasets, the model is able to generate high-quality and semantic aligned time series of arbitrary lengths during inference, overcoming the fixed-length limitations of prior approaches. All resources have been made available[1]. This work marks three key contributions:

- We systematically explore the existing T2S datasets, and introduce a novel, high-resolution fragment-level multi-modal dataset for text-to-time series generation tasks.
- We propose the first domain-agnostic model for text-to-time series generation, which integrates flow matching and the diffusion transformer, and is capable of generating semantically aligned time series of arbitrary lengths.
- **`T2S`** sets a new state-of-the-art performance across 13 datasets from 12 domains in time series generation, consistently outperforming both diffusion-based models and those based on large language models.

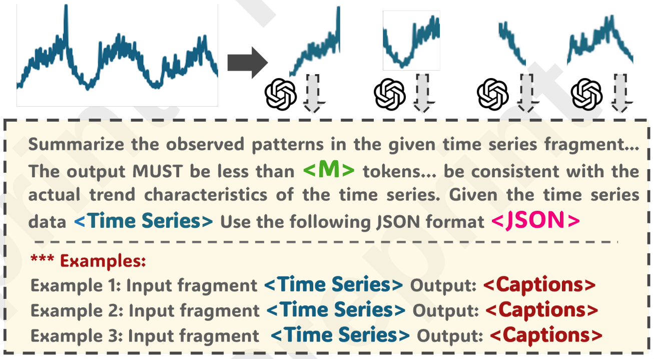
2 Definition and Dataset

2.1 Problem Definition and Notation

Let $\mathbf{x} \in \mathbb{R}^L$ denote a univariate time series of length L. Textual captions, represented as $\mathbf{T}$, provide semantic guidance across varying levels of granularity.

Definition 1 (Point-Level Description)**.** *A point-level description $\mathbf{T}_p$ provides semantic annotations for **individual time points** within the time series $\mathbf{x}$. Each point-level description $\mathbf{T}_p^{(j)}$ corresponds to the j-th time point, where $j \in [1, \dots, L]$, providing a fine-grained guidance for each point in time series $\mathbf{x}$.*

Definition 2 (Fragment-Level Description)**.** *A fragment-level description $\mathbf{T}_f$ provides semantic annotations for non-overlapping and **contiguous fragments** of the time series $\mathbf{x}$. Each fragment-level description $\mathbf{T}_f^{(j)}$ corresponds to the j-th fragment, where the length $|\mathbf{T}_f^{(j)}|$ is arbitrary and determined by the specific structure of time series $\mathbf{x}$.*



Summarize the observed patterns in the given time series fragment... The output MUST be less than <M> tokens... be consistent with the actual trend characteristics of the time series. Given the time series data <Time Series> Use the following JSON format <JSON>

*** Examples:
Example 1: Input fragment <Time Series> Output: <Captions>
Example 2: Input fragment <Time Series> Output: <Captions>
Example 3: Input fragment <Time Series> Output: <Captions>

Figure 1: Dataset Generation. GPT-4o-mini to generate high-quality natural language descriptions for each time series fragment.

Definition 3 (Instance-Level Description)**.** *An instance-level description $\mathbf{T}_i$ provides a global semantic annotation for the **entire time series** $\mathbf{x}$, providing high-level guidance encompassing all time points.*

We can use these captions to guide time series generation:

Definition 4 (T2S Generation)**.** *Given a text–time series dataset $D = \{(\mathbf{x}^{(i)}, \mathbf{T}_*^{(i)})\}_{i=1}^N$ with N samples, the task of T2S generation aims to learn a generative model G that maximizes the conditional probability $P_G(\mathbf{x} \mid \mathbf{T}_*)$, where $\mathbf{T}_*$ is the textual guidance provided at one of the following levels ($\mathbf{T}_p, \mathbf{T}_f, \mathbf{T}_i$). The generated time series $\mathbf{x}$ must semantically align with the textual guidance $\mathbf{T}_*$.*

2.2 TSFragment-600K

The proposed T2S generation task leverages textual captions at three granularity levels: point, fragment, and instance. While datasets for point-level [Liu *et al.*, 2024a] and instance-level [Kawaguchi *et al.*, 2025] descriptions are readily available, fragment-level descriptions remain an underexplored area. Fragment-level descriptions strike a balance between the granularity of point-level annotations, which may overlook broader temporal patterns, and the holistic nature of instance-level descriptions, which might obscure local dependencies. By encapsulating local temporal trends while preserving meaningful contextual relationships, fragment-level descriptions provide an ideal framework for evaluating the generative capabilities of T2S models.

To this end, we introduce `TSFragment-600K`, a novel dataset comprising over 600,000 fragment-level text-time series pairs. Each captions captures fine-grained temporal morphological characteristics, offering a rich and nuanced representation of the underlying trends. As illustrated in Figure 1, we employ GPT-4o-mini to generate high-quality natural language descriptions for each time series fragment, focusing on local trends and variations. Unlike prior approaches that rely on predefined dictionaries of time-series changes [Imani *et al.*, 2019], our captions are expressed in natural language, enhancing their interpretability and applicability.

[1] https://github.com/WinfredGe/T2S

Specifically, we propose a novel generation pipeline to construct fragment-level captions for time series data. First, a univariate time series $\mathbf{x}$ is segmented into k non-overlapping fragments,with each fragment $\mathbf{x}^{(j)} \in \mathbb{R}^{l_j}$ as a contiguous temporal segment for which textual captions are generated. Second, a seed-based prompting strategy is designed to capture high-quality captions of fragments. To accurately capture temporal dynamics, human experts curate high-quality GPT-4o descriptions for a subset of fragments, which serve as representative seed prompts. These prompts guide GPT-4o-mini in generating concise, consistent, and semantically rich captions for all fragments. A token limit $\langle M \rangle$ is applied to ensure a balance between informativeness and brevity.

Finally, five candidate captions are generated for each time series sample, and their embeddings are computed using *text-embedding-3-small*. We ensure the quality of generated textual captions by leveraging cosine similarity among their embeddings. Each caption is scored based on the average similarity of its embedding with others, and the one with the highest score is selected as the optimal text-time series pair, ensuring semantic alignment and coherence. Using this pipeline, we generate reliable fragment-level descriptions for eight classical time series datasets across diverse domains, including energy consumption, financial, exchange rates, traffic, air quality, and meteorological variables. The resulting fragment-level dataset, `TSFragment-600K`, comprises over $600,000$ samples with corresponding captions.

3 Methodology

In this section, we introduce the **`T2S`** architecture, as depicted in Figure 2. The architecture consists of two key components:

- **T2S Diffusion Transformer (T2S-DiT)**: T2S-DiT facilitates high-resolution alignment between captions and the temporal latent space. It employs flow matching [Liu *et al.*, 2022] as the diffusion backbone, the diffusion transformer module [Peebles and Xie, 2023] as the denoiser. Within this denoiser, textual information is integrated with the input features through adaptive layer normalization
- **Pretrained Length-Adaptive Variational Autoencoder (LA-VAE)**: LA-VAE encodes variable-length time series into a unified latent feature space and decodes them back to their original temporal dimensions. An interleaved training strategy is adopted to enable effective handling of varying input lengths during training.

We first employ LA-VAE to map time series of varying lengths into the latent space. The T2S-DiT module then denoises this latent space conditioned on the caption, aligning the textual and temporal features. During inference, a noise sequence of arbitrary length, encoded by LA-VAE, generates aligned time series based on the given caption.

3.1 T2S Diffusion Model

Flow Matching Framework. Inspired by [Esser *et al.*, 2024; Polyak *et al.*, 2024] and experimental comparisons with DDPM, we adopt the flow matching framework [Lipman *et al.*, 2022], specifically adopting the rectified flow approach [Liu *et al.*, 2022] using optimal transport paths. This framework offers superior generation quality and a more stable inference process compared to DDPM [Ho *et al.*, 2020], with reduced training costs.

Flow matching consists of forward and reverse processes. During training, given a time series sample in latent space $\mathbf{z}_1$, a noisy sample $\mathbf{z}_0 \sim \mathcal{N}(0, \mathbf{I})$, and a time step $t \in [0, 1]$, the forward path $\mathbf{z}_t$ is defined as:

$$p(\mathbf{z}_t \mid \mathbf{z}_1) = \mathcal{N}\left(\mathbf{z}_t; t\mathbf{z}_1, (1-t)^2\mathbf{I}\right), \tag{1}$$

where $\mathbf{z}_t$ evolves along an optimal transport path defined by:

$$\mathbf{z}_t = t\mathbf{z}_1 + (1-t)\mathbf{z}_0, \tag{2}$$

and the ground truth velocity of this transition is described by: $\mathbf{v}_t = \frac{d\mathbf{z}_t}{dt} = \mathbf{z}_1 - \mathbf{z}_0$. In the reverse process, the denoiser model aims to predict the velocity $\mathbf{u}_\theta(\mathbf{z}_t, t, \mathbf{C})$, where $\mathbf{C}$ represents the text prompt embedding generated by $\mathbf{T}_*$ in Definition 4. The model minimizes the mean squared error (MSE) between the ground truth and predicted velocities:

$$E_{\mathbf{z}_t,t,\mathbf{C}} \left|\mathbf{u}_\theta\left(\mathbf{z}_t, t, \mathbf{C}\right) - \mathbf{v}_t\right|^2. \tag{3}$$

During sampling, pure noise $\mathbf{z}_0 \sim \mathcal{N}(0, 1)$ is iteratively denoised to obtain $\hat{\mathbf{z}}_1$ by solving the ordinary differential equation (ODE) with the well-trained denoiser $\mathbf{u}_\theta\left(\mathbf{z}_t, t, \mathbf{C}\right)$. **`T2S`** adopts a classifier-free guidance framework, which does not rely on explicit class labels for conditioning. Unlike traditional class-based conditional generation methods, which use fixed condition categories to guide the generation process, classifier-free guidance provides a more flexible and effective way to balance unconditional and conditional generation. During training, the method randomly sets the condition to zero by a random ratio. During inference, the model first performs conditional generation, followed by unconditional generation, and then combines the results using a guidance scale δ. The formula is:

$$\mathbf{u}_\theta\left(\mathbf{z}_t, t, \mathbf{C}\right) = (1+\delta)\mathbf{u}_\theta\left(\mathbf{z}_t, t, \mathbf{C}\right) - \delta\mathbf{u}_\theta\left(\mathbf{z}_t, t\right), \tag{4}$$

where $\mathbf{u}_\theta(\mathbf{z}_t, t, \mathbf{C})$ is the noise estimater with input $\mathbf{z}_t$ and condition $\mathbf{C}$ and $\mathbf{u}_\theta(\mathbf{z}_t, t)$ is the conditioned noise estimater .

Diffusion Transformer. Building on the superior performance of the DiT in computer vision [Esser *et al.*, 2024] and recent advances in time series analysis [Chen *et al.*, 2024], we developed a patchified DiT denoiser that leverages DiT's fine-grained visual feature extraction and its ability to capture subtle latent temporal patterns, aligning these patterns with corresponding textual descriptions through the 2D latent representations encoded via LA-VAE. The latent space representation of time series can be conceptualized as single-channel, two-dimensional grayscale images $\mathbf{z}_t$. We first patchify $\mathbf{z}_t$ into a sequence of tokens. Then the input tokens are obtained by summing the sequence of tokens with two-dimensional positional embeddings. $\mathbf{c}_t$ and $\mathbf{z}_t$ represent the conditioning associated with time step t and input tokens. To achieve alignment between the conditioning $\mathbf{c}_t$ and the input tokens $\mathbf{z}_t$, we employ an adaptive layer normalization (AdaLN), a form of modulated layer normalization [Huang and Belongie, 2017]:

$$\mathrm{AdaLN}(\mathbf{z}_t, \mathbf{c}_t) = \gamma_t\left(\frac{\mathbf{z}_t - \mu_t}{\sigma_t}\right) + \beta_t, \tag{5}$$

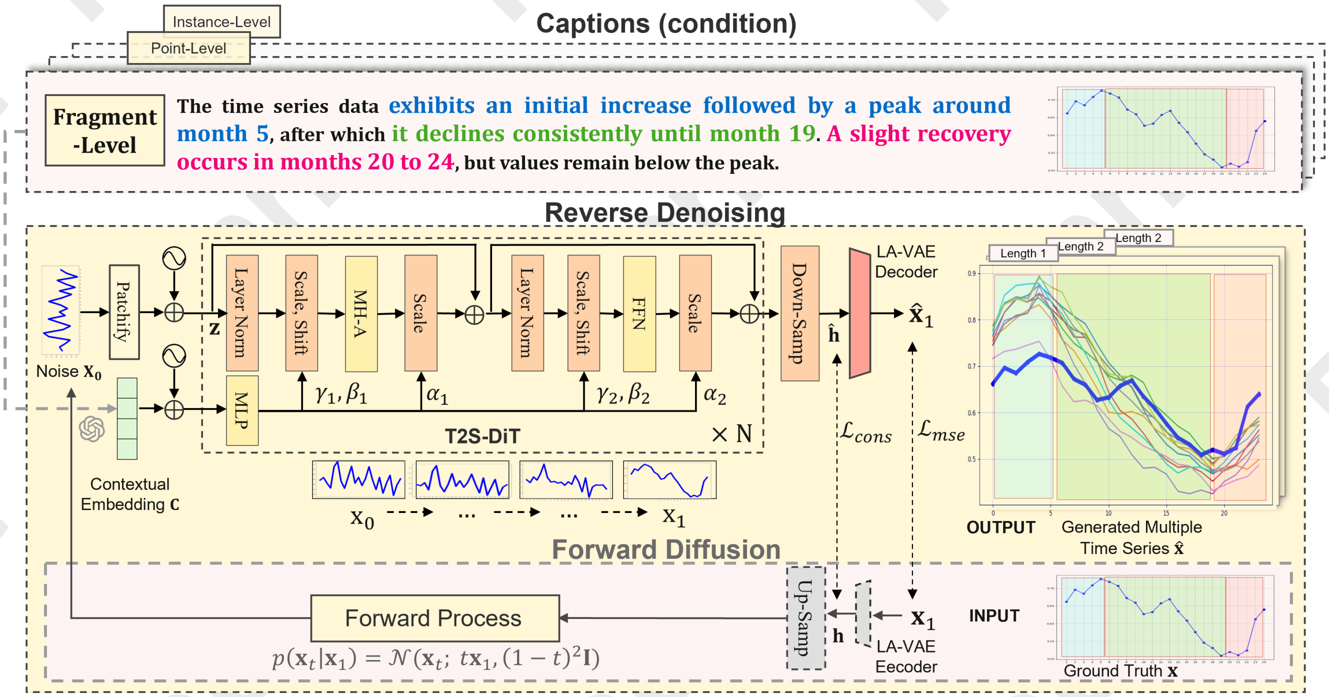


Figure 2: Overview of the **T2S** model. The framework conditions on captions for time series generation. LA-VAE encodes variable-length inputs into a latent space and decodes outputs to the original length. Forward diffusion transforms the original time series into noise, while T2S-DiT performs reverse denoising to align textual and temporal features, generating high-quality time series.

where scaling parameter $\alpha_{t,1}, \alpha_{t,2}, \gamma_{t,1}, \gamma_{t,2}$ and shifting parameter $\beta_{t,1}, \beta_{t,2}$ are chunked outputs, dynamically adjusted based on the textual information, as shown in Equation 6.

$$\gamma_{t,1}, \gamma_{t,2}, \mu_{t,1}, \mu_{t,2}, \beta_{t,1}, \beta_{t,2} = \mathrm{MLP}(\mathbf{c}_t). \tag{6}$$

The overall procedure can be formulated as follows:

$$\mathbf{z}_t^{(1)} = \gamma_{t,1}\left(\frac{\mathbf{z}_t - \mu_t^{(1)}}{\sigma_t^{(1)}}\right) + \beta_{t,1}, \mathbf{z}_t^{(2)} = \alpha_{t,1} \cdot \text{MH-A}\left(\mathbf{z}_t^{(1)}\right), \tag{7}$$

$$\mathbf{z}_t^{(3)} = \gamma_{t,2}\left(\frac{\mathbf{z}_t^{(2)} - \mu_t^{(2)}}{\sigma_t^{(2)}}\right) + \beta_{t,2}, \mathbf{z}_t^{(4)} = \alpha_{t,2} \cdot \mathrm{FFN}\left(\mathbf{z}_t^{(3)}\right), \tag{8}$$

$$\mathbf{u}_\theta\left(\mathbf{z}_t, \mathbf{t}, \mathbf{c}_\mathbf{t}\right) = \mathrm{MLP}\left(\mathrm{LayerNorm}\left(\frac{\mathbf{z}_t^{(4)} - \mu_t^{(4)}}{\sigma_t^{(4)}}\right)\right), \tag{9}$$

where $\mu_t^{(i)}$ and $\sigma_t^{(i)}$ denote the mean and variance of the input $\mathbf{z}_t$ for the i-th layer, respectively, and MH-A represents the multi-head attention. The entire process integrates adaptive layer norm into the transformer architecture, enabling it to dynamically adapt to textual conditioning. By leveraging this mechanism, the diffusion transformer aligns the input time series tokens with the contextual information, improving the generative performance of the model.

3.2 Length-adaptive VAE

T2S generation should support arbitrary-length generation to meet real-world application demands. For example, website user activity analysis and medical monitoring. However, previous works [Lai *et al.*, 2025] typically assume a fixed length, limiting their applicability. To address this limitation, we propose a pretrained LA-VAE, enabling the modeling and generation of time series with arbitrary lengths. In this study, we mixed data with lengths of 24, 48, and 96, then trained them in a unified framework. During sampling, arbitrary-length data can be generated within a specified range. Given an input time series $\mathbf{x}$, the LA-VAE encoder transforms $\mathbf{x}$ into a latent representation $\mathbf{h}$. This latent representation $\mathbf{h}$ is subsequently upsampled to a fixed-size latent embedding $\mathbf{z} = \text{Up-Samp}(\mathbf{h}_t)$, which serves as the input to the diffusion model. Through the diffusion process, a refined latent embedding $\hat{\mathbf{z}}_t$ is generated. A downsampling operation is then applied to $\hat{\mathbf{z}}_t$, yielding the latent vector $\hat{\mathbf{h}} = \text{Down-Samp}(\hat{\mathbf{z}}_t)$. Finally, the VAE decoder reconstructs $\hat{\mathbf{h}}$ into a time series with the original length. This process enables the handling of variable-length time series within a unified framework.

Consistency loss. Linear interpolation during upsampling and downsampling introduces blurriness and artifacts, as it fails to capture nonlinear features and signal curvature. To mitigate this, we introduce a latent space consistency loss term $\mathrm{MSE}(\mathbf{h}, \hat{\mathbf{h}})$ to enhance reconstruction quality:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{h}, \hat{\mathbf{h}}) = \mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda\, \mathrm{MSE}(\mathbf{h}, \hat{\mathbf{h}}), \tag{10}$$

where the first term enforces fidelity to the original time series, and the second ensures consistency in the latent space.

Algorithm 1 Interleaved Training for Mixed Datasets

```
Require: Datasets $\{D_1, \ldots, D_k\}$, LA-VAE $\phi(\cdot)$, T2S-DiT $g_\theta(\cdot)$,
    batch_size, number of iterations $N$.
$n_0 = 0, n_i \leftarrow |D_i|, i = 1, \ldots, k$
DatasetSampling():
    $n \leftarrow \sum_{i=1}^{k} n_i$
    $j \sim \text{Uniform}(1, n)$
    Find $m \in \mathbb{Z}^+$ such that $j \in \left(\sum_{i=1}^{m} n_i, \sum_{i=1}^{m+1} n_i\right]$
    Return $D_m[j - \sum_{i=1}^{m-1} n_i]$
for iter = 1 to $N$ do
    $D \leftarrow \{\}$, loss $= 0$
    for $i = 1$ to batch_size do
        $D \leftarrow D \cup$ DatasetSampling()
    end for
    for length $i = 1$ to $k$ do
        $batch_i = \{s \in D \mid \text{len}(s) = i\}$
        $\hat{\mathbf{S}} \leftarrow g_\theta(\phi(batch_i))$
        loss = loss + $\mathcal{L}$ in equation 10
    end for
    Update $\phi$
end for
```

3.3 Interleaved Training

Traditional sequential training paradigm often lead to catastrophic forgetting. To effectively train models on datasets of varying lengths within a unified framework, we propose a novel interleaved training paradigm, outlined in Algorithm 1.

Assume a domain contains d datasets with different lengths, denoted as $\{l_1, l_2, \ldots, l_d\}$. During training, we shuffle all samples across these datasets and randomly sample *batch_size* samples for each batch. Within each iteration, training is interleaved across these batches, which improves the model's generalization ability.

4 Experiments

We conduct an extensive evaluation across 13 datasets spanning 12 domains to assess the performance of the **T2S**, aiming to address the following key research questions:

- **RQ1**: How does **T2S** compare in performance to existing state-of-the-art methods given fragment-level captions?
- **RQ2**: How does **T2S** compare in performance to existing methods given point-level and instance-level captions?
- **RQ3**: How do the different components within the **T2S** affect its overall generation performance?
- **RQ4**: How sensitive is the **T2S**'s performance to key hyperparameters, and does it require additional fine-tuning?
- **RQ5**: How effective is **T2S** when trained on limited data?

4.1 Experimental Settings

Datasets. We evaluated our model on three distinct datasets: Point-Level, Fragment-Level, and Instance-Level.

- **Point-Level Dataset**: The Time-MMD dataset [Liu *et al.*, 2024a] links individual time series points with corresponding textual news, consisting of 23,618 data points across six domains, including Climate, Economy, and Social Goods. We adapted the dataset by concatenating each time series point with its associated text.
- **Fragment-Level Dataset**: `TSFragment-600K` pairs time series data with captions across seven domain, including Electricity, ETT, Exchange, and Traffic [Wu *et al.*, 2021]. It consists of over 600,000 samples.
- **Instance-Level Dataset**: SUSHI, a simulated dataset [Kawaguchi *et al.*, 2025], comprises 2,800 samples generated from 15 pre-defined functions.

Evaluation Metrics. We evaluated performance using Mean Squared Error (MSE), Weighted Absolute Percentage Error (WAPE) [Shao *et al.*, 2024], and Mean Reciprocal Rank at 10 (MRR@10)[Craswell, 2009].

- Weighted Absolute Percentage Error (WAPE):

$$\text{WAPE}(y, \hat{y}) = \frac{\sum_{i \in \Omega} |y_i - \hat{y}_i|}{\sum_{i \in \Omega} |y_i|}, \tag{11}$$

  where Ω and $\hat{\Omega}$ represent truth space and generative space, while y_i and $\hat{y}_i$ denote the corresponding i-th sample.

- Mean Reciprocal Rank at 10 (MRR@10):

$$\text{MRR@10} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \frac{1}{\text{rank}_i}, \tag{12}$$

  where $\text{rank}_i = \text{argmin}(n \mid \cos(\hat{y}_{i,n}, y_i) > \text{threshold})$.

Baselines. To evaluate **T2S**, we compared its performance against state-of-the-art fully trained models and zero-shot large language models, encompassing diverse paradigms to robustly assess its effectiveness. Among fully trained models, DiffusionTS [Yuan and Qiao, 2024] uses diffusion with text and time embeddings for time-series generation and adapts to text-conditional generation by injecting text embeddings into its encoder and decoder using AdaLayerNorm. Meanwhile, TimeVAE [Desai *et al.*, 2021] utilizes a variational autoencoder framework with caption embeddings, using dense layers with ReLU to fuse input and text for conditioning. For the zero-shot baselines, GPT-4o [OpenAI, 2023] and Llama-3.1-8b [Dubey *et al.*, 2024] are employed. To curb hallucination and ensure reliable outputs, Llama-3 is guided by domain-specific prompts, a repeat generation loop, and targeted post-processing.

4.2 Performance Comparison on Fragment-Level Descriptions (RQ1)

Table 1 presents the fragment-level performance comparison across six datasets, evaluated using WAPE, MSE, and MRR@10. **T2S** achieves top performance across all metrics, securing 14 out of 18 entries (77.8%) for MSE, significantly outperforming DiffusionTS (5.6%) and TimeVAE (11.1%). On the exchange rate dataset, **T2S** achieves an average MSE of 0.039, representing a 56.0% improvement over DiffusionTS and a 68.9% improvement over TimeVAE. These results demonstrate **T2S**'s superior ability to align textual and temporal features across diverse datasets and fragment lengths. Moreover, **T2S**'s interleaved training strategy enables cross-length training within each dataset, removing the need for length-specific training required by baseline models, thereby enhancing scalability and generalization.

* **T2S**'s interleaved training strategy enables cross-length training within each dataset, whereas other full-trained models require training and evaluation for fixed-length inputs.

Datasets	Length	**T2S**			DiffusionTS			TimeVAE			GPT-4o-mini			Llama3.1-8b		
		WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑
ETTh1	24	**0.183**	**0.008**	**0.283**	0.793	0.077	0.267	0.666	0.055	0.211	0.264	0.041	0.104	0.883	0.663	0.097
	48	**0.234**	**0.013**	0.289	1.207	0.120	**0.298**	0.647	0.055	0.286	0.414	0.080	0.100	0.923	1.260	0.086
	96	**0.229**	**0.011**	**0.291**	0.498	0.028	0.214	0.643	0.055	0.286	0.500	0.118	0.096	0.949	1.748	0.056
ETTm1	24	0.426	0.033	**0.286**	0.604	0.040	0.251	0.666	0.048	0.219	**0.244**	**0.031**	0.101	1.134	0.798	0.099
	48	0.53	**0.053**	**0.283**	1.119	0.100	0.285	0.636	0.051	0.217	**0.453**	0.112	0.097	1.074	1.496	0.079
	96	**0.414**	0.041	**0.299**	0.546	**0.031**	0.293	0.664	0.057	0.208	0.706	0.395	0.091	1.079	1.761	0.057
Electricity	24	**0.135**	**0.010**	**0.28**	0.617	0.041	0.253	0.207	0.016	0.213	0.734	0.592	0.092	0.926	1.140	0.064
	48	**0.155**	**0.013**	**0.244**	1.128	0.102	0.227	0.208	0.017	0.216	1.014	1.065	0.068	1.038	1.416	0.054
	96	0.238	0.031	**0.318**	0.545	0.032	0.247	**0.213**	**0.018**	0.257	1.024	1.210	0.059	1.085	1.740	0.034
Exchange Rate	24	**0.292**	**0.033**	**0.334**	0.791	0.077	0.272	1.165	0.105	0.252	1.072	2.060	0.052	1.258	2.052	0.045
	48	**0.259**	**0.033**	**0.315**	1.217	0.122	0.298	1.064	0.106	0.306	0.933	1.074	0.082	1.562	2.125	0.051
	96	**0.48**	**0.047**	**0.31**	0.504	0.048	0.216	0.977	0.106	0.274	1.141	1.625	0.054	1.433	1.892	0.055
Air Quality	24	0.884	**0.02**	**0.304**	0.806	0.078	0.265	2.303	0.022	0.302	**0.557**	0.379	0.093	0.878	0.697	0.085
	48	1.295	**0.044**	**0.297**	1.439	0.120	0.221	1.648	0.023	0.271	**0.791**	0.715	0.08	1.141	1.642	0.046
	96	1.377	0.049	**0.34**	**0.508**	0.028	0.304	1.270	**0.024**	0.301	0.928	1.127	0.061	1.085	1.551	0.050
Traffic	24	**0.353**	**0.005**	0.201	0.795	0.077	0.220	0.544	0.008	**0.233**	1.260	1.912	0.020	1.144	1.938	0.022
	48	**0.506**	**0.008**	**0.219**	1.202	0.120	0.188	0.594	0.011	0.211	1.189	1.928	0.011	1.138	1.988	0.004
	96	**0.543**	**0.01**	**0.262**	0.509	0.028	0.171	0.641	0.013	0.207	1.18	2.093	0.010	1.107	1.994	0.001
1^{st} count	-	12	14	16	1	1	1	1	2	1	4	1	0	0	0	0

Table 1: Performance comparison on fragment level. All results are evaluated on three different metrics WAPE, MSE, and MRR@10. For interleaved training, arbitrary lengths of $\{24, 48, 96\}$ were selected, with evaluations performed separately for each length.

4.3 Performance Comparison on Point and Instance-Level Descriptions (RQ2)

Table 2 presents the performance comparison at the point and instance levels across seven datasets. These evaluations assess models' abilities to capture fine-grained temporal annotations and generate coherent global patterns. T2S outperforms all baselines across three metrics. It consistently achieves the lowest MSE values in 17 out of 18 entries, surpassing DiffusionTS and TimeVAE, which struggle to capture fine-grained temporal variations. Similarly, T2S secures the best WAPE scores in 16 out of 18 entries, demonstrating its robustness. T2S achieves the top MRR@10 score of 0.314 on the instance-level SUSHI dataset, its WAPE and MSE results show an advantage over zero-shot models, only a slight underperformance compared to DiffusionTS in certain scenarios. In summary, T2S demonstrates state-of-the-art performance by effectively balancing fine-grained precision with high-level semantic understanding.

4.4 Ablation Study (RQ3)

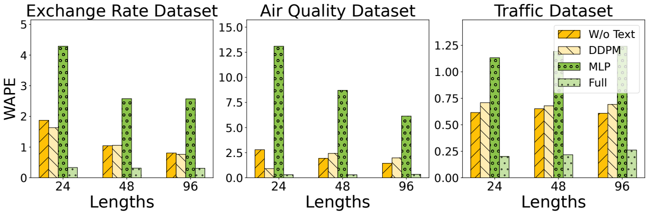


Figure 3: WAPE comparison across three datasets and generation lengths {24, 48, 96}. The legend indicates model configurations: W/o Text (no text guidance), DDPM (baseline with diffusion probabilistic modeling), MLP (denoiser replaced with Multi-Layer Perception), and Full model.

As shown in Figure 3 ,we conducted three ablation experiments on three datasets: Exchange Rate, Air Quality, and Traffic, evaluating nine different settings. First, replacing the flow matching backbone with DDPM resulted in consistent performance drops, with an average error increase of 311.00% across all datasets. Second, substituting the DiT denoiser with an MLP drastically degraded results, with errors rising by 877.67% on the Exchange Rate dataset. Lastly, removing text guidance severely impacted high-resolution generation, with average error increase of 495.13%, 327.10%, 205.23% across datasets with lengths {24, 48, 96}, respectively, highlighting the critical role of text in guiding generation quality. These findings clearly highlight the critical role of each component, demonstrating that the full model is essential for achieving high-quality generation.

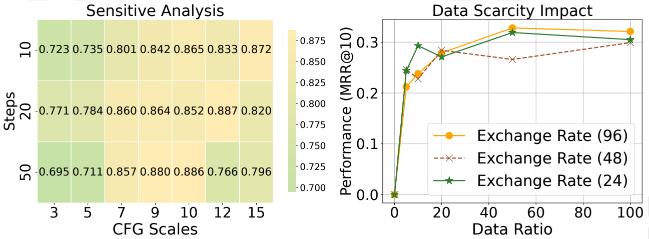


Figure 4: Parameter sensitivity analysis on the Exchange Rate dataset using MRR@10 (left). Data sparsity performance of **T2S** on different data ratios. Consistent improvements are reported with an increasing data ratio (right).

4.5 Parameter Sensitivity (RQ4)

Recent study [Li *et al.*, 2024] demonstrates the pivotal role of the inference stage in affecting the performance of diffusion models. Building on this, we explored the sensitivity of the flow matching diffusion model to key inference parameters: classifier-free guidance scales (CFG) and generation time steps,evaluated using MRR@10. Figure 4 shows a heatmap illustrating performance impact, with yellow regions yielding superior results and green areas reflecting suboptimal performance. Notably, the model achieves higher MRR@10 scores

* **T2S**'s interleaved training strategy enables cross-length training within each dataset, whereas other full-trained models require training and evaluation for fixed-length inputs.

		T2S			DiffusionTS			TimeVAE			GPT-4o-mini			Llama3.1-8b		
Datasets	Length	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑
SUSHI	2048	0.494	0.088	**0.314**	**0.407**	**0.032**	0.269	0.445	0.061	0.288	1.093	0.990	0.058	0.869	0.827	0.055
Agriculture	24	**0.183**	**0.013**	**0.661**	0.648	0.046	0.294	1.309	0.087	0.251	1.284	2.190	0.070	0.648	0.402	0.124
	48	**0.197**	**0.008**	0.209	1.422	2.306	0.256	1.165	0.106	**0.279**	1.321	2.146	0.071	1.012	0.703	0.104
	96	**0.124**	**0.014**	**0.619**	1.073	0.096	0.319	0.930	0.076	0.291	1.422	2.306	0.052	1.283	1.125	0.114
Climate	24	**0.328**	**0.016**	**0.405**	0.791	0.068	0.293	0.575	0.054	0.306	1.038	0.800	0.104	1.207	1.800	0.053
	48	**0.211**	**0.007**	**0.39**	0.554	0.037	0.305	0.513	0.051	0.335	1.014	0.997	0.092	1.284	2.124	0.046
	96	**0.294**	**0.021**	**0.476**	1.279	0.203	0.264	0.494	0.057	0.275	1.057	1.335	0.069	1.167	1.870	0.049
Economy	24	**0.118**	**0.010**	**0.561**	0.989	0.086	0.292	0.476	0.084	0.316	0.295	0.071	0.130	1.194	1.690	0.059
	48	**0.071**	**0.004**	**0.488**	1.239	0.132	0.290	0.607	0.129	0.314	0.339	0.063	0.112	0.501	0.270	0.096
	96	**0.113**	**0.01**	**0.667**	0.826	0.083	0.293	0.597	0.110	0.329	0.539	0.198	0.124	0.615	0.321	0.100
Energy	24	**0.212**	**0.005**	**0.391**	0.452	0.028	0.309	2.287	0.083	0.281	1.327	1.952	0.058	1.807	1.897	0.068
	48	**0.174**	**0.003**	**0.452**	0.373	0.031	0.295	2.012	0.086	0.268	1.408	1.949	0.056	1.139	1.935	0.065
	96	**0.372**	**0.017**	**0.450**	0.391	0.030	0.290	1.716	0.099	0.290	1.256	1.904	0.043	1.097	1.593	0.068
Health US	24	**0.328**	**0.009**	0.192	0.427	0.048	**0.224**	0.888	0.051	0.323	1.008	1.789	0.050	1.230	1.982	0.068
	48	**0.264**	**0.008**	0.129	0.424	0.052	**0.221**	0.743	0.051	0.296	1.045	1.930	0.014	1.163	1.883	0.035
	96	**0.316**	**0.012**	0.141	0.594	0.073	**0.215**	0.753	0.051	0.308	1.089	1.940	0.002	1.176	1.957	0.013
Social Goods	24	0.901	**0.024**	**0.583**	**0.640**	0.070	0.305	0.942	0.049	0.327	1.789	1.420	0.093	1.353	1.653	0.058
	48	0.721	0.082	**0.452**	**0.410**	**0.045**	0.337	0.678	0.049	0.349	1.390	1.920	0.055	1.247	1.862	0.056
	96	**0.283**	**0.020**	**0.494**	0.417	0.041	0.310	0.677	0.054	0.260	1.347	1.634	0.077	1.261	1.670	0.030
1st count	-	16	17	15	3	2	3	0	0	1	0	0	0	0	0	0

Table 2: Performance comparison on point and instance levels. All results are evaluated on three different metrics WAPE, MSE, and MRR@10. For the instance-level dataset, SUSHI is used for training and inference with a fixed length of 2048. For point-level training, arbitrary lengths of {24, 48, 96} were selected for interleaved training.

within the range of CFG scores between 7 and 10 and generation time steps between 20 and 50. This analysis underscores the importance of precise inference-stage parameter selection in optimizing the performance of Flow Matching models.

4.6 Data Scarcity (RQ5)

To explore the impact of dataset size on the model's performance, we evaluate the model on the Exchange Rate dataset under varying dataset scales. Specifically, we train and evaluate the model on subsets consisting of different proportions of the full dataset. As shown in figure 4, the model demonstrates consistent improvement with increasing dataset size. Notably, using only 50% of the dataset, the model achieves 93.8% of the full dataset performance. Similarly, for the 48- and 96-length generations, the model reaches 92.4% and 91.3% of the full dataset performance, respectively. These results highlight the model's strong generalization ability in data-scarce scenarios, showing its capacity to generate high-quality time-series even with limited training data.

5 Related Work

Text-Time Series Datasets. Existing text-time series pair datasets can be categorized into three types: instance-level, point-level, and fragment-level, based on how the caption and time series are temporally aligned. At the point level, each time series point is paired with an event description, such as financial news or clinical notes [Yu *et al.*, 2023; Cortis *et al.*, 2017; Liu *et al.*, 2024a]. At the instance level, `TRUCE` [Jhamtani and Berg-Kirkpatrick, 2021] and `SUSHI` [Kawaguchi *et al.*, 2025] utilize time series features, such as upward trends and peaks, as dictionary entries to generate coherent signals. At the fragment level, several researchers have introduced datasets tailored for time series reasoning tasks [Williams *et al.*, 2024; Chow *et al.*, 2024]. A large-scale, fine-grained, general-purpose text-time series dataset for time series generation tasks remains in its early exploration.

Text-Time Series Generation. The general text-to-time series paradigm can be achieved through contrastive learning or generative modeling. Recently, contrastive learning has been employed to facilitate text-to-time series mapping. However, these approaches are primarily focused on retrieval tasks [Ito *et al.*, 2024; Rizhko and Bloom, 2024] and cannot be directly applied to time series generation. In contrast, generative modeling, including variational autoencoders (VAEs) [Desai *et al.*, 2021; Lee *et al.*, 2023], diffusion models [Yuan and Qiao, 2024; Kong *et al.*, 2020; Wen *et al.*, 2023; Narasimhan *et al.*, 2024], and large language models [OpenAI, 2023; Dubey *et al.*, 2024], provides more versatile frameworks for generating time series conditioned on textual descriptions. Among these, conditional diffusion models [Yuan and Qiao, 2024; Cao *et al.*, 2024; Narasimhan *et al.*, 2024] show promise for text-to-time series generation due to their ability to model complex temporal dynamics and generate temporally coherent sequences. For instance, time series generation conditioned on healthcare metadata [Alcaraz and Strodthoff, 2023; Lai *et al.*, 2025] and sensor metadata [Zuo *et al.*, 2023; Fu *et al.*, 2024] has been explored. However, these methods are often domain-specific and fail to address the more general alignment between time series and their corresponding captions, limiting their broader applicability.

6 Conclusion

We proposed `TSFragment-600K`, a high-resolution fragment-level multimodal dataset for text-to-time series generation tasks, and **T2S**, the first domain-agnostic model for general text-to-time series generation. Leveraging LA-VAE and T2S-DiT, **T2S** generates semantically aligned time series of arbitrary lengths with high fidelity. Comprehensive validation across 12 diverse domains demonstrates **T2S**'s superior performance, establishing a robust foundation for text-to-time series generation.

References

[Alcaraz and Strodthoff, 2023] Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based conditional ecg generation with structured state space models. *Computers in biology and medicine*, 163:107115, 2023.

[Cao *et al.*, 2024] Defu Cao, Wen Ye, Yizhou Zhang, and Yan Liu. Timedit: General-purpose diffusion transformers for time series foundation model. *arXiv preprint arXiv:2409.02322*, 2024.

[Chen *et al.*, 2024] Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*, 2024.

[Chow *et al.*, 2024] Winnie Chow, Lauren Gardiner, Haraldur T Hallgrímsson, Maxwell A Xu, and Shirley You Ren. Towards time series reasoning with llms. *arXiv preprint arXiv:2409.11376*, 2024.

[Cortis *et al.*, 2017] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 519–535, 2017.

[Craswell, 2009] Nick Craswell. Mean reciprocal rank. *Encyclopedia of database systems*, pages 1703–1703, 2009.

[Desai *et al.*, 2021] Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Esser *et al.*, 2024] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[Fu *et al.*, 2024] Chun Fu, Hussain Kazmi, Matias Quintana, and Clayton Miller. Creating synthetic energy meter data using conditional diffusion and building metadata. *Energy and Buildings*, 312:114216, 2024.

[Gao *et al.*, 2024] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.

[Girdhar *et al.*, 2023] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[Imani *et al.*, 2019] Shima Imani, Sara Alaee, and Eamonn Keogh. Putting the human in the time series analytics loop. In *Companion proceedings of the 2019 World Wide Web conference*, pages 635–644, 2019.

[Ito *et al.*, 2024] Aoi Ito, Kota Dohi, and Yohei Kawaguchi. Clasp: Learning concepts for time-series signals from natural language supervision. *arXiv preprint arXiv:2411.08397*, 2024.

[Jhamtani and Berg-Kirkpatrick, 2021] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Truth-conditional captioning of time series data. *arXiv preprint arXiv:2110.01839*, 2021.

[Johnson *et al.*, 2023] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[Kawaguchi *et al.*, 2025] Yohei Kawaguchi, Kota Dohi, and Aoi Ito. Sushi: A system for unified semantic human interaction. https://github.com/y-kawagu/SUSHI, 2025. Accessed: 2025-01-02.

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

[Lai *et al.*, 2025] Yongfan Lai, Jiabo Chen, Deyun Zhang, Yue Wang, Shijia Geng, Hongyan Li, and Shenda Hong. Diffusets: 12-lead ecg generation conditioned on clinical text reports and patient-specific information. *arXiv preprint arXiv:2501.05932*, 2025.

[Le *et al.*, 2024] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36, 2024.

[Lee *et al.*, 2023] Daesoo Lee, Sara Malacarne, and Erlend Aune. Vector quantized time series generation with a bidirectional prior model. *arXiv preprint arXiv:2303.04743*, 2023.

[Li *et al.*, 2024] Zeyu Li, Wang Han, Yue Zhang, Qingfei Fu, Jingxuan Li, Lizi Qin, Ruoyu Dong, Hao Sun, Yue Deng, and Lijun Yang. Learning spatiotemporal dynamics with a pretrained generative model. *Nature Machine Intelligence*, 6(12):1566–1579, 2024.

[Lipman *et al.*, 2022] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[Liu *et al.*, 2022] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[Liu *et al.*, 2024a] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: A new multi-domain multimodal dataset for time series analysis. *arXiv preprint arXiv:2406.08627*, 2024.

[Liu *et al.*, 2024b] Zhen Liu, Wenbin Pei, Disen Lan, and Qianli Ma. Diffusion language-shapelets for semi-supervised time-series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14079–14087, 2024.

[Narasimhan *et al.*, 2024] Sai Shankar Narasimhan, Shubhankar Agarwal, Oguzhan Akcin, Sujay Sanghavi, and Sandeep Chinchali. Time weaver: A conditional time series generation model. *arXiv preprint arXiv:2403.02682*, 2024.

[OpenAI, 2023] OpenAI. Gpt-4o mini, 2023. Accessed: 2025-01-23.

[Peebles and Xie, 2023] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[Polyak *et al.*, 2024] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[Rizhko and Bloom, 2024] Mariia Rizhko and Joshua S Bloom. Astrom 3: A self-supervised multimodal model for astronomy. *arXiv preprint arXiv:2411.08842*, 2024.

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

[Shao *et al.*, 2024] Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, et al. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Sharma *et al.*, 2023] Gulshan Sharma, Abhinav Dhall, and Ramanathan Subramanian. Medic: Mitigating eeg data scarcity via class-conditioned diffusion model. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

[Wang, 2024] Lijun Wang. Multi-modality conditional diffusion model for time series forecasting of live sales volume. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2675–2679. IEEE, 2024.

[Wen *et al.*, 2023] Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–12, 2023.

[Williams *et al.*, 2024] Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, et al. Context is key: A benchmark for forecasting with essential textual information. *arXiv preprint arXiv:2410.18959*, 2024.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

[Yang *et al.*, 2024] Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. A survey on diffusion models for time series and spatio-temporal data. *arXiv preprint arXiv:2404.18886*, 2024.

[Yu *et al.*, 2023] Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*, 2023.

[Yuan and Qiao, 2024] Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*, 2024.

[Zuo *et al.*, 2023] Si Zuo, Vitor Fortes Rey, Sungho Suh, Stephan Sigg, and Paul Lukowicz. Unsupervised statistical feature-guided diffusion model for sensor-based human activity recognition. *arXiv preprint arXiv:2306.05285*, 2023.