

TimelyGPT: Extrapolatable Transformer Pre-training for Long-term Time-Series Forecasting in Healthcare

Ziyang Song, Qincheng Lu,
Hao Xu, He Zhu
School of Computer Science, McGill
University
Montreal, QC, Canada

David Buckeridge
School of Population and Global
Health, McGill University
Montreal, Quebec, Canada

Yue Li
School of Computer Science, McGill
University
Mila Quebec AI institute
Montreal, QC, Canada

ABSTRACT

Motivation: Large-scale pre-trained models (PTMs) such as BERT and GPT have recently achieved great success in Natural Language Processing and Computer Vision domains. However, the development of PTMs on healthcare time-series data is lagging behind. This underscores the limitations of the existing transformer-based architectures, particularly their scalability to handle large-scale time series and ability to capture long-term temporal dependencies.

Methods: In this study, we present Timely Generative Pre-trained Transformer (TimelyGPT). TimelyGPT employs an extrapolatable position (xPos) embedding to encode trend and periodic patterns into time-series representations. It also integrates recurrent attention and temporal convolution modules to effectively capture global-local temporal dependencies.

Materials: We evaluated TimelyGPT on two large-scale healthcare time series datasets corresponding to continuous biosignals and irregularly-sampled time series, respectively: (1) the Sleep EDF dataset consisting of over 1.2 billion timesteps collected from 197 whole-night polysomnographic sleep recordings, containing EEG, EOG, EMG, and event marker; (2) the longitudinal healthcare administrative database PopHR, comprising 489,000 patients randomly sampled from the Montreal population.

Results: Our experiments show that during pre-training, TimelyGPT excels in learning time-series representations from continuously monitored biosignals and irregularly-sampled time series data commonly observed in longitudinal electronic health records (EHRs), which can aid in healthcare time-series forecasting tasks. In forecasting continuous biosignals, TimelyGPT achieves accurate extrapolation up to 6,000 timesteps of body temperature during the sleep stage transition, given a short look-up window (i.e., prompt) containing only 2,000 timesteps. For irregularly-sampled time series, TimelyGPT with a proposed time-specific inference demonstrates high top recall scores in predicting future diagnoses using early diagnostic records, effectively handling irregular intervals between clinical records. Together, we envision TimelyGPT to be useful in a broad spectrum of health domains, including long-term patient health state forecasting and patient risk trajectory prediction.

CCS CONCEPTS

• Applied computing → Bioinformatics; • Computing methodologies → Transfer learning.

KEYWORDS

Time-series forecasting, Time-series pre-training, transfer learning, irregularly-sampled time series, biosignals, clinical diagnosis

ACM Reference Format:

Ziyang Song, Qincheng Lu, Hao Xu, He Zhu, David Buckeridge, and Yue Li. 2024. TimelyGPT: Extrapolatable Transformer Pre-training for Long-term Time-Series Forecasting in Healthcare. In *Proceedings of The 15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB '24)*. ACM, New York, NY, USA, 17 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Time-series forecasting holds significant importance in healthcare, given its potential to trace patient health trajectories and predict medical diagnoses [7, 20]. In the field of healthcare, there are two primary categories: continuously monitored and irregularly-sampled time series data. Continuous time-series, such as biosignals, have been extensively studied in various applications, including health monitoring [38], disease classification [25], and physical activity prediction [31]. Irregularly-sampled time series are commonly found in clinical records, where spontaneous updates are made due to outpatient hospital visits or inpatient hospital stays [57]. The key challenge is to extract meaningful contextualized representations from these time-series to make accurate long-term forecasting. A promising approach is to adopt transfer learning [20]. Initially, a model is pre-trained on large-scale datasets to learn contextualized temporal representations. This pre-trained model (PTM) is then fine-tuned to forecast target sequences.

The recent impressive achievements of Transformer PTMs in Natural Language Processing (NLP) and Computer Vision (CV) domains have inspired growing interest in time-series Transformer-based PTMs. Time-Series Transformer (TST) uses a mask-and-reconstruction pre-training strategy to extract contextualized representations from time series [54]. Cross-Reconstruction Transformer (CRT) learns temporal representations by dropping and reconstructing certain segments from time series [56]. Additionally, Transformer PTMs have been applied to traffic [58], tabular [22], and speech time-series [18, 19].

Transfer learning by pre-training on large time-series data followed by fine-tuning for long-term time series forecasting (LTSF) tasks is a promising avenue. However, existing studies primarily focus on training from scratch on limited data for LTSF tasks [20].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM BCB '24, Nov. 22-25, 2024, Shenzhen, Guangdong, PR China

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06

<https://doi.org/XXXXXXX.XXXXXXX>

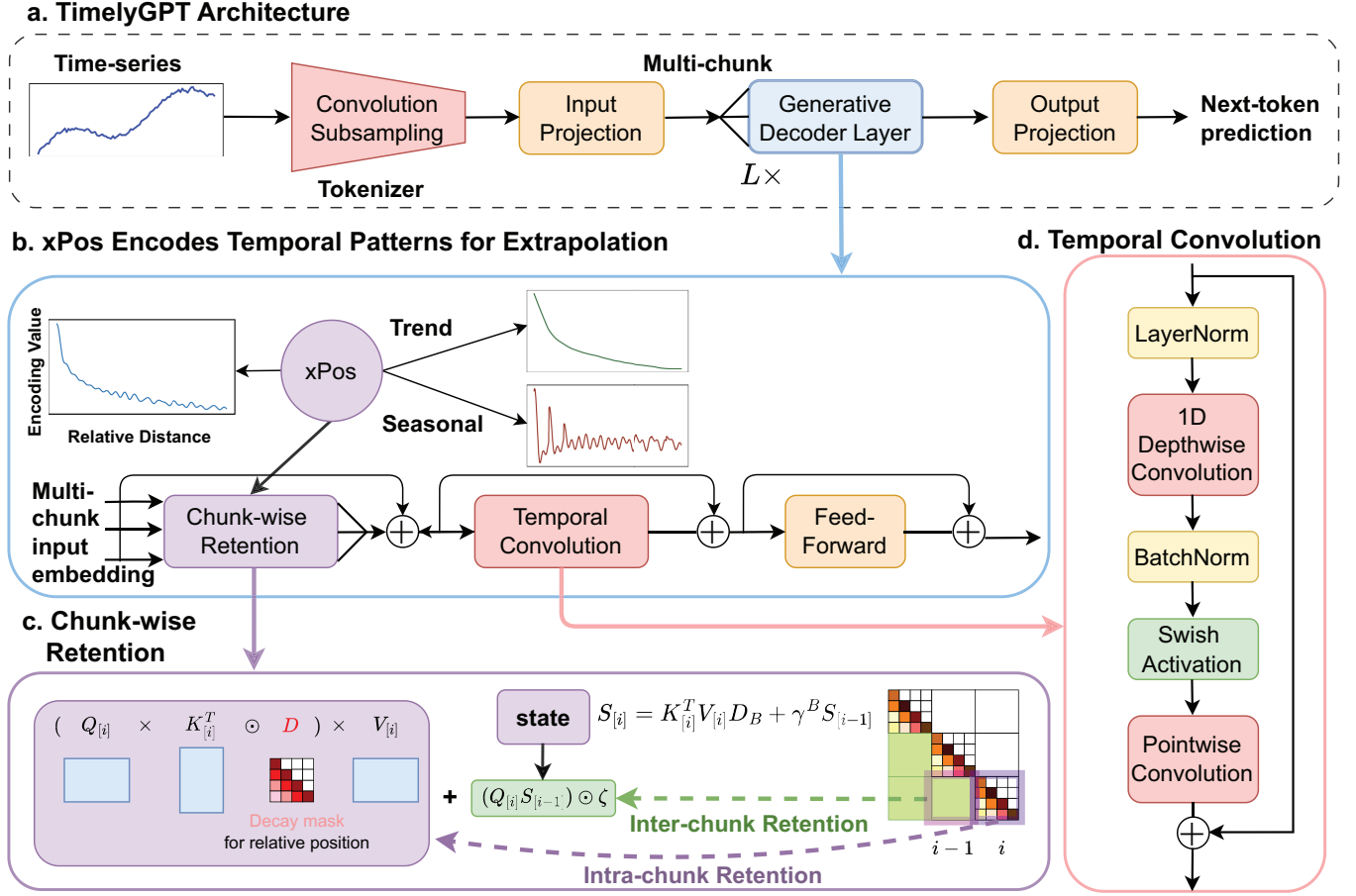


Figure 1: TimelyGPT overview. a. TimelyGPT architecture. TimelyGPT consists of a convolution-subsampling tokenizer followed by L decoder layers, with detailed overflow provided in Appendix B.3. b. Generative decoder with xPos embedding. Each decoder layer is coupled with extrapolatable position embedding (Section 3.1) that encodes trend and periodic patterns into representations, facilitating forecasting with extrapolation ability. c. Chunk-wise Retention. This module consists of parallel intra-chunk Retention and recurrent inter-chunk Retention, effectively handling long sequences in continuously monitored biosignals (Appendix B.2). d. Temporal Convolution (Section 3.3) captures nuanced local interactions from time-series representations.

These studies often introduce tailored architectures and attention modules to extract complex temporal dependencies [48, 59, 60]. However, the scalability of these transformers on large datasets for LTSTF tasks remains an open question [13]. A recent study argues that the permutation-invariant nature of self-attention causes the loss of temporal information [53]. As a result, transformers often underperform compared to convolution-based models, potentially due to their struggles with local features and multi-scale features [42, 52]. Overall, existing research on time-series transformers often lacks rigorous evaluation on large datasets and does not consistently outperform conventional approaches on small data.

In this study, we provide an in-depth analysis of existing time-series Transformer models, covering key aspects such as the attention mechanism and position embedding. We argue that the seeming inadequacy of current transformer-based models for time-series data is due to their inability to model large-scale time series.

Once these challenges are resolved, we would observe the typical scaling law found in NLP and CV domains [13, 55]. Motivated by this insight, we present a novel framework called **Timely Generative Pre-trained Transformer (TimelyGPT)** (Fig. 1) that utilizes an extrapolatable position (xPos) embedding to encode trend and periodic patterns into time-series representations [41]. TimelyGPT integrates recurrent attention (also known as Retention) and convolution modules for effectively capturing both global temporal dependencies and nuanced local interactions [10, 40].

The key contributions of our research are threefold:

- (1) We employ extrapolatable xPos embedding (Fig. 1b) to encode both trend and periodic patterns into time-series representations, facilitating long-term forecasting.
- (2) We extend recurrent attention (Fig. 1c) to handle both continuous and irregularly-sampled time series data;

- (3) We introduce convolution subsampling tokenizer (Fig. 1a) to extract features from raw time-series and temporal convolution (Fig. 1d) to sift local features among the timesteps.

Overall, our experimental results reveal that TimelyGPT effectively extrapolates temporal representations for long-term forecasting. This leads to highly effective pre-training on large-scale time-series biosignals and longitudinal EHR data, and ultimately superior task-specific fine-tuning performance compared to the existing methods.

2 RELATED WORK

2.1 Self-attention in Transformer

Transformer employs an encoder-decoder architecture composed of L layers of Transformer blocks [45]. Each block consists of a self-attention layer followed by a feed-forward layer. For an input embedding $X \in \mathbb{R}^{N \times d}$, where N is the number of tokens and d is the hidden size, the self-attention mechanism is defined as:

$$\text{Attention}(X) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V \quad (1)$$

where $Q, K, V = XW_Q, XW_K, XW_V \in \mathbb{R}^{N \times d}$ are the Query, Key, and Value matrices, respectively. The attention mechanism allows Transformer to model long-term dependencies effectively, making it extensively utilized in NLP and CV domains.

As one of the prominent time-series transformers, Conformer utilizes the self-attention mechanism to capture long-range global contexts in speech data [10]. When combined with convolution modules, Conformer enhances self-attention by exploiting fine-grained local patterns. Although widely successful, the quadratic complexity of self-attention with respect to sequence length has spurred the exploration of attention-free modules such as Multi-Layer Perceptron (MLP) [43], implicit long convolution [26], and Recurrent Neural Network (RNN) [24, 40]. In particular, RNN-based attention modules have scaled up to 14 billion parameters while maintaining competitive performance with linear training and constant inference complexities. These modules are particularly well-suited for time-series modeling by effectively capturing sequential dependencies [9]. In this study, TimelyGPT integrates the Retention mechanism and convolution modules to effectively capture both global and local contexts.

2.2 Position embedding in Transformer

Transformer relies on position embedding to capture temporal relations, since the self-attention mechanism alone does not inherently discern token order [34]. *Absolute* position embedding, which commonly employs sinusoidal functions, adds positional encoding directly to token embeddings. However, this method only encodes discrete position indexes, making it less effective for continuous timescales such as trend and periodic patterns in time-series data [53]. In contrast, speech transformers utilize *relative* position embedding to handle continuous time by encoding positional information relative to token distances [10]. Rotary Position Embedding (RoPE), prevalent in numerous large language models [2, 23, 44], applies rotation matrices to encode time information from relative distances [39]. Additionally, the RNN-based Transformer Receptance Weighted Key Value (RWKV) uses exponential decay to encode time information based on relative distance [24]. Bridging these

techniques, xPos embedding utilizes both rotation and exponential decay to effectively capture long-term dependencies [41].

One challenge for Transformer is *extrapolation*, i.e., forecasting sequences longer than those seen during training, due to the difficulty in generalizing position embeddings to unseen positions [27]. Encoder-decoder architectures often concatenate the input sequence with a zero-padded placeholder for the target sequence and predict all timesteps at once, while encoder-only models encode input sequence for forecasting [21, 59]. Both approaches struggle with extrapolation and rely heavily on their linear layer for forecasting [17], limiting their effectiveness in LTSF tasks. To address the issue, Attention with Linear Biases (ALiBi) adjusts attention with penalties linearly correlated with token distances [27]. Building on this, xPos embedding employs exponential decay to assign penalties based on relative distances [41]. Consequently, xPos can handle inference lengths up to eight times the training length while maintaining comparable performance. Our TimelyGPT extends xPos from the NLP domain to long-term forecasting in the time-series domain, focusing on exploring the underlying mechanisms that enable the temporal extrapolation.

3 TIMELYGPT METHODOLOGY

Our proposed TimelyGPT effectively pre-trains on unlabeled data using next-token prediction task to learn temporal representations (Fig. 1). It first processes time-series inputs using a convolution-subsampling tokenizer for token embedding (Fig. 1a). To extract meaningful temporal patterns, TimelyGPT integrates three technical contributions. First, TimelyGPT utilizes extrapolatable xPos embedding to encode trend and periodic patterns (Fig. 1b, Section 3.1). Second, TimelyGPT utilizes the Retention module to capture global content (Fig. 1c, Section 3.2). Third, TimelyGPT deploys the convolution module to capture the local content (Fig. 1d, Section 3.3). Integrating Retention and Convolution modules enables the modeling of interactions between global and local content.

3.1 Extrapolatable position embedding encodes temporal patterns

As our first contribution, TimelyGPT employs xPos to encode relative positional information into token embeddings based on the distance $n - m$ between token n and m [41]. Given an input embedding $X \in \mathbb{R}^{N \times d}$ for N tokens at d embedding dimensions, xPos is integrated into the n -th token embedding X_n through rotation matrix $e^{i\theta n}$ and exponential decay γ^n :

$$\begin{aligned} \tilde{Q}_n \tilde{K}_m &= X_n W_Q (\gamma e^{i\theta})^{n-m} X_m W_K = \gamma^{n-m} \hat{Q}_n \hat{K}_m \\ \text{where } \hat{Q}_n &= X_n W_Q e^{i\theta n}, \hat{K}_m = X_m W_K e^{-i\theta m} \end{aligned} \quad (2)$$

where θ and γ indicate position-dependent rotation and decay hyperparameters [39, 41]. The exponential decay γ^{n-m} determines the intensity of remembering historical information, while the rotation matrix $e^{i\theta n}$ captures the oscillation frequencies. This decay mechanism effectively attenuates the influence of distant tokens, aiding in capturing long-term dependencies and enhancing extrapolation ability [41].

While initially designed for language modeling, xPos provides a compelling way for time-series modeling, mirroring the seasonal-trend decomposition (Fig. 1c). Its exponential decay γ^{n-m} naturally

concentrates on recent times while diminishing the influence of distant times, reflecting the trend momentum of time-series. The rotation matrix $e^{i\theta(n-m)}$ captures the seasonal component of time-series through sinusoidal oscillations.

In healthcare time-series, xPos embedding effectively encodes both trend and periodic patterns crucial for modeling continuous biosignals and irregular clinical records. For continuous biosignals, trend patterns such as body temperature and vital signs are key health indicators, while electrocardiograms (ECGs) exhibit periodic patterns reflecting the physiological rhythms of the human body. In irregularly-sampled clinical records, age-related susceptibility to illnesses is observed in longitudinal population studies using administrative health data [1, 37]. Some EHRs also exhibit periodic patterns, especially for chronic diseases like COPD, which have alternating exacerbation and recovery cycles.

We hypothesize that xPos embedding can encode these trend and periodic patterns into token embeddings. By harnessing xPos, TimelyGPT can effectively model long-term dependencies essential for time-series forecasting. In Section 6.2, 6.3, and 6.4, we validated our hypothesis and explored the underlying mechanisms driving temporal extrapolation for forecasting beyond training length.

3.2 Retention for continuous and irregularly-sampled time series

As our second contribution, we adapt the Retention mechanism to effectively handle continuous time-series data [40]. The Retention mechanism based on xPos can be reformulated as an RNN to naturally model time-series data. Given the xPos embedding in Eq 2, the forward-pass of the Retention mechanism can be computed in parallel over all tokens with a linear training complexity:

$$\begin{aligned} \hat{Q}_n &= X_n W_Q e^{i\theta n}, \hat{K}_m = X_m W_K e^{-i\theta m}, V = X W_V \\ \text{Ret}(X) &= (\hat{Q} \hat{K}^\top \odot D) V, D_{nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases} \end{aligned} \quad (3)$$

where the decay matrix $D \in R^{N \times N}$ and rotation matrix $e^{i\theta(n-m)}$ encode trend and periodic patterns into token embedding, taking into account the distance between tokens $n-m$. When reformulated as an RNN, the Retention in Eq. 3 can be manifested in a recurrent forward-pass with a constant inference complexity. This reformulated RNN excels in capturing sequential dependencies from the time-series. To handle long sequences, we use chunk-wise Retention by segmenting the sequence into multiple, non-overlapping chunks (Fig. 1c). Consequently, chunk-wise Retention maintains a linear complexity for long sequences. We provide details about the three Retention forward-passes in Appendix B.2.

To accommodate irregularly-sampled time series, we modify the Retention mechanism as follows. Given N samples $\{s_1, \dots, s_N\}$, each sample s_n is represented as a tuple (x_n, t_n) , consisting of an observation x_n and a timestep t_n . Given two samples s_n and s_m , the decay mask D is adapted according to the time gap $\Delta t_{n,m} = t_n - t_m$:

$$\text{Ret}(X) = (QK^\top \odot D)V, D_{nm} = \begin{cases} \gamma^{\Delta t_{n,m}}, & t_n \geq t_m \\ 0, & t_n < t_m \end{cases} \quad (4)$$

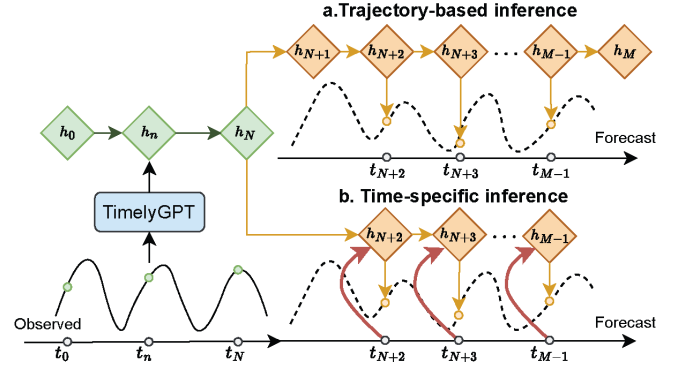


Figure 2: Two inference strategies for forecasting irregularly-sampled time series. (a) Trajectory-based inference. TimelyGPT autoregressively predicts the entire sequence at equal time intervals. The target intervals can then be taken from part of the inferred trajectory. **(b) Time-specific inference.** TimelyGPT directly predicts the target data point using his-torical hidden states and the gap between the target timestep and the last observed timestep.

For next token pre-training, the retention incorporates $\Delta t_{n,n-1}$ into the recurrent state variable $S_n \in R^{d \times d}$,

$$\begin{aligned} S_n &= \gamma^{\Delta t_{n,n-1}} S_{n-1} + K_n^\top V_n \\ \text{Ret}(X_n) &= Q_n S_n \end{aligned} \quad (5)$$

where the base case $S_1 = \mathbf{0}$ in this recurrent relation.

At inference time, to forecast irregularly-sampled time series, we consider two recurrent inference strategies, namely trajectory-based inference and time-specific inference (Fig. 2). Both strategies make predictions based on a look-up window. The former autoregressively predicts a trajectory at equal time intervals. The latter directly makes prediction at a specific time point $s_{n'} = (x_{n'}, t_{n'})$. Specifically, knowing the target timestep $t_{n'}$ and the last observed sample $s_n = (x_n, t_n)$, TimelyGPT outputs the embedding of the target token $\text{Ret}(X_{n'}) = Q_{n'} S_{n'}$, taking into account the time gap $\Delta t_{n',n} = t_{n'} - t_n$ and the recurrent state then becomes $S_{n'} = \gamma^{\Delta t_{n',n}} S_n + K_{n'}^\top V_{n'}$.

3.3 Convolution modules for local interaction

Convolution methods excel at identifying localized interactions from time series [16]. As the first part of our third contribution, we propose a **convolution-subsampling tokenizer** for feature extraction from the raw time-series input (Fig. 1a). Briefly, it uses multiple 1-D convolution layers to condense the time dimension and extract local features of the time-series. The convolution-subsampling tokenizer consists of two 1-D convolution layers with kernel size 3 and stride 2, reducing the sequence length to 1/4. Unlike the prevalent patching technique, which merely segments adjacent timesteps and features [21], the convolution tokenizer effectively captures local temporal interactions. More details are provided in Appendix B.3.

As the second part of our third contribution, we propose a **temporal convolution module** using a depth-wise separable convolution [3], sifting local temporal features from the time-series representations. As shown in Fig. 1d, this module starts with a layer normalization, followed by a 1-D depth-wise convolution and a point-wise convolution layer, with batch normalization and swish activation after the depth-wise convolution. Integrating convolution and attention allows TimelyGPT to extract global-local feature interactions [10, 49]. By stacking multiple decoder layers, each with a convolution module, TimelyGPT discerns multi-scale features that characterize patterns across varying time scales [42].

3.4 Computational complexity

TimelyGPT with its efficient Retention mechanism achieves $O(N)$ training complexity and $O(1)$ inference complexity. In contrast, BERT and GPT incur $O(N^2)$ training complexity and $O(N)$ inference complexity [14]. The vanilla attention mechanism in the Transformer, $\text{Attention}(X) = \text{Softmax}(\frac{QK^T}{\sqrt{d}})V$, introduces a training complexity of $O(N^2d)$. This quadratic computational bottleneck prevents standard Transformer models from modeling long sequences (i.e., $N \gg d$).

TimelyGPT achieves linear training complexity by following research in linear transformers [14]. In the Retention mechanism, $\text{Ret}(X_n) = Q_n S_n$, $S_n = K_n^T V_n + \gamma S_{n-1}$, both $Q_n S_n$ and $K_n^T V_n$ have $O(d^2)$ complexity. By recursively updating over N timesteps, the total complexity becomes $O(Nd^2)$. For inference, TimelyGPT proposes time-specific and trajectory-based methods. The trajectory-based inference recursively generates sequences with equally-spaced time intervals like the GPT model, incurring $O(N)$ inference complexity. In contrast, the time-specific inference directly predicts target time point with $O(1)$ complexity. Therefore, TimelyGPT achieves $O(N)$ training complexity and $O(1)$ inference complexity, making it computationally efficient and suitable for long sequences. We provided detailed discussion of computational bottleneck of Transformer and efficient linear Transformer in Appendix A.1.

4 DATA

4.1 Sleep-EDF dataset

The Sleep European Data Format (EDF) database, sourced from PhysioBank [8], contains sleep recordings from 153 healthy subjects [15]. These whole-night polysomnographic sleep recordings include 7 types of biosignals: electroencephalogram (EEG) from Fpz-Cz and Pz-Oz electrode locations, electrooculogram (EOG), submental chin electromyogram (EMG), oro-nasal airflow, rectal body temperature, and an event marker. Both EEG and EOG signals were sampled at 100 Hz (i.e., the signals were recorded at a rate of 100 samples per second), while EMG and the other features were sampled at 1 Hz (i.e., 1 sample per second). Sleep patterns (hypnograms) were manually scored by trained technicians into five sleep stages. This biosignal dataset comprises a total of 1.2 billion timesteps, segmented into 300,700 sequences of 4,000 timesteps each. It provides large-scale continuous time-series data for training large models. In our experiment, we forecast all 7 biosignals.

4.2 PopHR database

The Population Health Record (PopHR) database hosts a massive amount of longitudinal claim data from the provincial government health insurer in Quebec, Canada (Régie de l'assurance maladie du Québec, RAMQ) on health service use [32, 51]. In total, there are approximately 1.3 million participants in the PopHR database, which represents a randomly sampled 25% of the population in the metropolitan area of Montreal between 1998 and 2014. Cohort memberships are maintained dynamically by removing deceased residents and actively enrolling newborns and immigrants. We extracted irregularly-sampled time series from the patient clinical records in the PopHR database. Specifically, we converted ICD-9 diagnostic codes to phenotype codes (PheCodes) using the expert-defined PheWAS catalog [5, 6]. We selected 315 unique PheCodes each with over 50,000 token counts and excluded patients who had fewer than 50 PheCode tokens. This resulted in a dataset of 489,000 patients, averaging 112 diagnosis records each.

5 EXPERIMENTS

We first validated the scaling pattern of TimelyGPT, determining the optimal number of model parameters for different dataset sizes (Section 6.1). We then explored TimelyGPT's extrapolation capabilities for long-term forecasting up to 6,000 timesteps in Sleep-EDF's biosignal data, and analyzed extrapolation's underlying mechanism through visualization (Section 6.2). Our evaluation extended forecasting to irregularly-sampled time series (Section 6.3). Furthermore, we conducted ablation studies to evaluate the contributions of various components (Section 6.4).

5.1 Pre-training and fine-tuning

During pre-training, TimelyGPT utilizes a next-token prediction task to learn general temporal representations from unlabeled data [29]. Given a sequence with a [SOS] token, TimelyGPT predicts the subsequent tokens by shifting the sequence to the right. At the last layer, each token's output representation is fed into a linear layer for next-token prediction. The pre-training loss is Mean Squared Error (MSE) for continuous signals (e.g., biosignal) and cross-entropy for discrete signals (e.g., diagnosis codes).

Among other Transformer baselines, PatchTST adopted a masking-based approach, masking 40% of its patches as zeros [21]. CRT utilized a dropping-based pre-training, discarding up to 70% of patches [56]. For the Transformer models without established pre-training methods, we used a masking-based method by randomly masking 40% of timesteps [54]. For downstream forecasting tasks, we employ end-to-end fine-tuning on the entire model. The final linear layer is utilized for making the forecasts. All Transformer models performed 20 epochs of pre-training with MSE loss, followed by 5 epochs of end-to-end fine-tuning.

5.2 Jointly forecasting multivariate biosignals from Sleep-EDF dataset

We utilized all seven features from the Sleep-EDF dataset for a multivariate forecasting task, applying standardization as preprocessing. The Sleep-EDF dataset was split into training (80%), validation (10%), and test (10%) sets. All models were pre-trained on the entire training set and fine-tuned on a randomly chosen 20%

subset of the training data, with time-series data segmented into non-overlapping sequences. For pre-training, we chose an input length of 4,000 timesteps. For fine-tuning, we used a look-up window of 2,000 timesteps and varied forecasting windows of 720, 2,000, and 6,000 timesteps. We used MAE as a metric. We evaluated TimelyGPT against Informer [59], Autoformer [48], FEDformer [60], PatchTST [21], TimesNet [47], TS2Vec [52], and DLinear [53]. Based on the scaling law in Section 6.1, we set the model parameters for all transformers to around 18 million, with specific architectures and parameters detailed in Table S2.

5.3 Forecasting irregularly-sampled diagnostic codes from PopHR dataset

We assessed long-term forecasting task of the irregularly-sampled time series extracted from the PopHR database. We divided the dataset into training (80%), validation (10%), and testing (10%) sets. We pre-trained on the entire training set and fine-tuned on a 20% subset of training data. We used cross entropy and top- K recall to evaluate the pre-training and fine-tuning, respectively. For forecasting, we set the look-up window to be 50 timestamps and the rest as the forecasting window, containing up to more than 100 timestamps (i.e., diagnosis codes).

For our TimelyGPT, we separately evaluated the performance of trajectory-based and time-specific inferences (Section 3.2). We compared with several transformer baselines, including Informer, Fedformer, AutoFormer, and PatchTST as well as the models designed for irregularly-sampled time series, namely mTAND [35] and SeFT [11]. Given that diagnoses are discrete values, there was no need to utilize the convolution-subsampling tokenizer for TimelyGPT. Furthermore, we specified a patch size of 2 for PatchTST, indicating that every two adjacent timestamps are projected into a single patch. Based on the scaling law in Section 6.1, we set model parameters for all transformers to about 7.5 million, with specific architectures and parameters detailed in Table S2.

5.4 Model parameters

For all benchmark experiments, we tailored the architecture and parameters of TimelyGPT based on the scaling-law analysis (Section 6.1; Fig. 3). Specifically, for the Sleep-EDF dataset, TimelyGPT was configured with 18 million parameters, and for the PopHR dataset, it was configured with 7.5 million parameters. While different Transformer models may have unique optimal hyperparameters, optimizing each model's setup is computationally prohibitive with our current compute resources. For fairness of comparison, we compared TimelyGPT against all transformer baselines at the same model size (Table S2).

6 RESULTS

6.1 Scalability of TimelyGPT

We evaluated the scalability of TimelyGPT on the large-scale Sleep-EDF dataset to determine the optimal model parameters with respect to different dataset sizes [15]. We selected subsets of the Sleep-EDF dataset with timesteps ranging from 10^5 to 10^9 , splitting each dataset into training (80%), validation (10%), and testing (10%) sets. Both look-up and forecasting windows were set to 256

timesteps for this experiment. TimelyGPT's performance improves as parameter and dataset size increase (Fig. 3), which is attributed to its capacity to handle more data, known as the scaling law for Transformer [13]. We provide further discussion of scaling patterns of the existing Transformer models in Appendix A.3.

6.2 Forecasting multivariate Sleep-EDF biosignals

TimelyGPT achieved the best performance in forecasting biosignals for all windows in terms of MAE, except for window 720 (Table 1; Fig. 4a). PatchTST achieved the best MAE at 0.456, whereas TimelyGPT conferred comparable performance. DLinear was also effective for the 720-timestep forecasting window. However, as the forecasting window increased to 2,000 and 6,000 timesteps, both PatchTST and DLinear suffered performance drops due to their reliance on the linear layers and inability to extrapolate beyond the training length. In contrast, pre-trained on 4,000 timesteps, TimelyGPT consistently maintained superior performance up to 6,000 timesteps given a short look-up window (i.e., prompt) containing only 2,000 timesteps. Additionally, TimelyGPT consistently outperformed other baselines across all three forecasting windows in terms of cross-correlation performance (Fig. 4b; Table 1). TimesNet

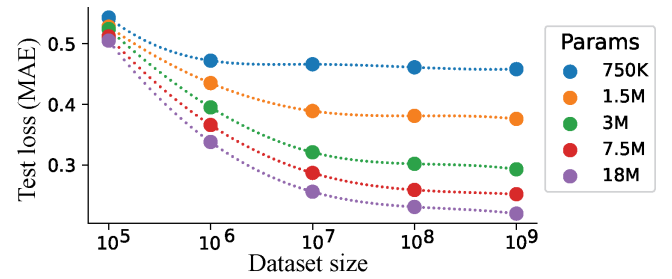


Figure 3: Test MAE of forecasting Sleep-EDF biosignals as a function of dataset sizes and parameter sizes. Both look-up and forecasting windows were set to 256 timesteps. TimelyGPT with more parameters tends to exhibit better performance when trained on larger datasets.

Table 1: Comparison of TimelyGPT as well as 7 baselines for long-term forecasting experiment on the large-scale SleepEDF dataset. Bold and underlined numbers indicate the best and second best results for each metric and window.

Window Size	MAE			Cross-Correlation		
	720	2000	6000	720	2000	6000
TimelyGPT	0.542	0.567	0.575	0.644	0.628	0.607
Informer	0.675	1.013	1.256	0.352	0.256	0.221
Autoformer	0.532	0.908	1.026	0.452	0.401	0.279
Fedformer	0.515	0.865	0.912	0.386	0.307	0.314
PatchTST	0.456	0.768	<u>0.824</u>	0.569	0.512	0.370
DLinear	0.521	0.840	0.929	0.452	0.369	0.189
TS2Vec	0.602	1.231	1.204	0.415	0.301	0.223
TimesNet	<u>0.471</u>	<u>0.742</u>	0.865	<u>0.602</u>	<u>0.573</u>	<u>0.403</u>

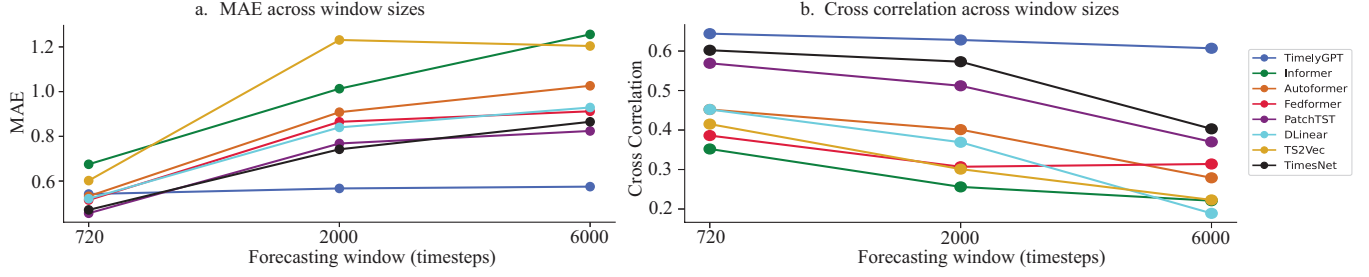


Figure 4: SleepEDF biosignal forecasting performances of TimelyGPT and seven state-of-the-art methods over various forecasting windows. a. MAE for 8 methods evaluated over 3 forecasting windows (720, 2000, and 6000 timesteps). b. Cross-correlation scores for the same methods and forecasting windows. The detailed numerical results are summarized in Table 1.

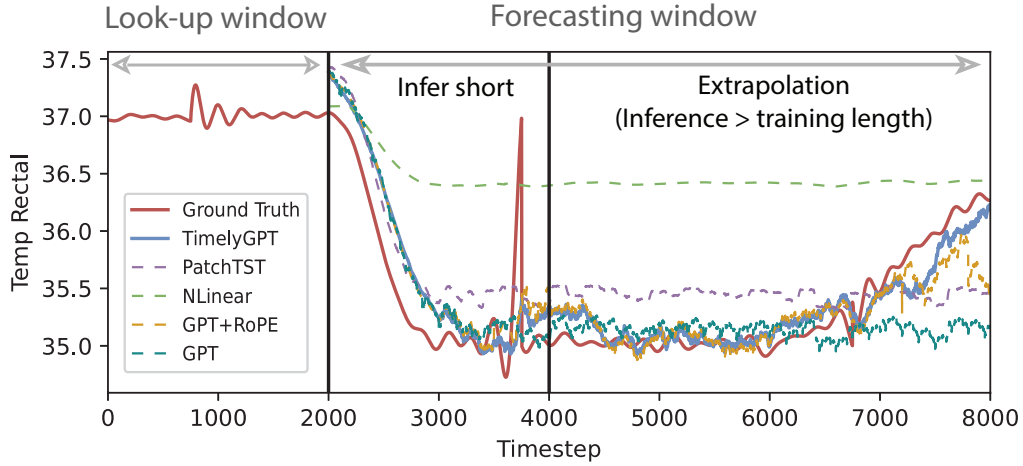


Figure 5: Predicted sequence of SleepEDF biosignals of 6,000 timesteps. Given a 2,000 look-up window, we applied TimelyGPT (blue solid line) and 4 state-of-the-art methods (dashed lines) to predict the biosignals for the next 6,000 timesteps. The groundtruth biosignals are displayed as red solid line. The two vertical lines demarcate the look-up window and the length of pre-training sequences, respectively.

was the second best performer for these windows, but declined as window size gets larger due to the extrapolation issue. These results underscore TimelyGPT's extrapolation capabilities in long-term forecasting, aligning with the findings in the NLP domain [41].

We visualized the predicted biosignals by TimelyGPT against the leading baselines (PatchTST and DLinear) and the ablated methods (GPT-2 and GPT-2 with RoPE), focusing on sleep stage transitions (Fig. 5). We utilized a 2,000-timestep look-up window and a 6,000-timestep forecasting window. Forecasting beyond 2,000 timesteps is marked as extrapolation, as it exceeds the training length. In the rectal temperature (i.e., trend signal), TimelyGPT's forecast aligned well with the groundtruth, effectively capturing distinct trend patterns. Notably, the small bump in the prompt before the 1000-th timestep is a typical indicator for temperature drop. Most models were able to capture it except for DLinear, showing the benefits of pre-training. Beyond the training length of 4000, TimelyGPT demonstrated more advantages in accurately extrapolating the rise of the rectal temperature around 7000-th timestep while PatchTST and GPT fell behind. The superior extrapolation capabilities of TimelyGPT is attributable to its ability to capture the long-term

trends with xPos embedding. In contrast, both PatchTST and vanilla GPT experienced a performance decline, likely due to the dependency on linear mapping as discussed in previous research [17]. Additionally, TimelyGPT exhibits superior extrapolation capabilities over the ablated baseline GPT+RoPE, highlighting its effective trend pattern modeling for extrapolation. We also visualized EEG periodic biosignal forecast and found a similar conclusion (Fig. S3).

6.3 Forecasting patient diagnosis trajectory

We then applied TimelyGPT and the baseline methods to forecast 315 PheCodes for 489K patients from PopHR (Section 4.2). We evaluated the performance using the average top K recall at each forecast window. TimelyGPT with time-specific inference outperformed the baselines reaching the highest recall rates of 58.65% and 70.83% at $K = 5$ and $K = 10$, respectively (Table 2). At $K = 15$, TimelyGPT ranked the second-highest recall with 82.69%. In addition, the time-specific inference outperformed the trajectory-based inference, highlighting the advantage of time decay mechanism.

Table 2: Forecasting results of TimelyGPT and 6 baselines on PopHR’s irregular-sampled time series dataset. TimelyGPT with time-specific inference achieved the highest recall at $K = 5$ and $K = 10$, and the second highest at $K = 15$, demonstrating its superior performance in long-term forecasting of irregularly-sampled time series.

Metrics	Recall @K (%)		
	$K = 5$	$K = 10$	$K = 15$
TimelyGPT (trajectory-based)	52.30	64.35	77.12
TimelyGPT (time-specific)	58.65	70.83	<u>82.69</u>
Informer	46.37	60.14	71.24
Autoformer	42.87	57.43	68.59
Fedformer	43.31	58.34	69.60
PatchTST	48.17	65.55	73.31
MTand	<u>52.59</u>	<u>70.21</u>	83.73
SeFT	49.26	68.10	79.39

We then examined the distributions of the top-5 recall rates at 3 forecast windows, comparing two inference methods of TimelyGPT with the best transformer baseline PatchTST and the leading irregular time series algorithm MTand (Fig. 6). TimelyGPT’s time-specific inference consistently outperformed trajectory-based inference as the forecasting window size increases. While both inference methods exhibited similar performance for predicting the first 50 timesteps, time-specific TimelyGPT demonstrated significantly better results beyond 50 timesteps. This improvement is likely due to time-specific inference taking into account the evolving states and the query timestep in the time decay mechanism, enhancing its ability to predict the temporal evolution of healthcare

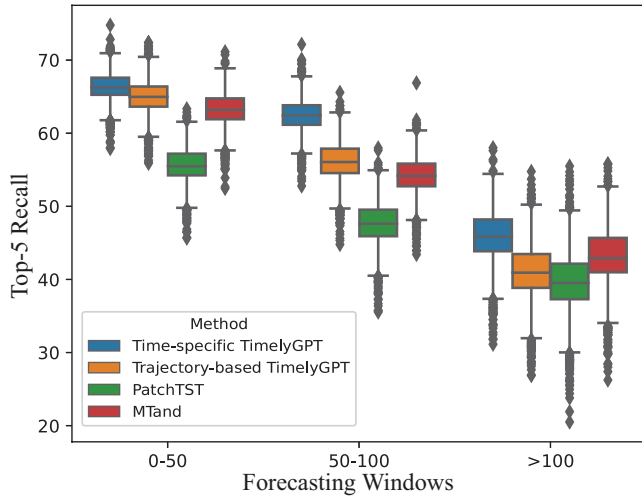


Figure 6: The distribution of top-5 recall performance for TimelyGPT with two inference methods (Time-specific and Trajectory-based), compared to PatchTST and MTand across three forecasting window sizes.

trajectories over irregular intervals. As expected, all models experienced a performance decline in predicting farther future because of the increasing uncertainties. Despite this, TimelyGPT maintained higher and more stable performance within the first 100 steps compared to PatchTST and MTand. Although MTand closely followed to time-specific TimelyGPT for the first 50 timesteps, its performance drastically declines as the forecasting window increases, reflecting its difficulty with extrapolation. These findings highlight the utility of the proposed time-specific inference in leveraging time-decay mechanism to handle irregularly-sampled time series for long-term forecasting.

We visualized the observed and predicted trajectory of a patient with neoplasm and genitourinary diseases (Fig. 7). TimelyGPT with time-specific inference produced a high top-5 recall rate of 85.7% on this patient. Indeed, most of the observed codes were among the top 5 predicted codes by the time-specific TimelyGPT. Zooming into the forecast window (Fig. 7b), TimelyGPT accurately predicted Phecodes 590.0 (Pyelonephritis) three times around the age of 61. TimelyGPT predicted PheCode 740.9 at age 61 with high probability, which appeared twice at ages 52 and 53 in the look-up window. Therefore, TimelyGPT demonstrated a promising direction to forecast patient health state despite the challenges inherent in modeling irregularly-sampled longitudinal EHR data.

6.4 Ablation study

To assess the contributions of various components in TimelyGPT, we conducted ablation studies by omitting the key components, including convolution subsampling tokenizer, temporal convolution module, exponential decay, and RoPE relative position embedding. Notably, removing all components results in a vanilla GPT-2. Since exponential decay in xPos depends on RoPE, we cannot assess the impact of exponential decay independently by removing the RoPE component. Additionally, we also ablated the pre-training strategy by training TimelyGPT from scratch on the forecasting tasks. The ablation studies focused on downstream forecasting experiments using the Sleep-EDF and PopHR datasets, corresponding to continuous biosignals and irregularly-sampled time series, respectively. We conducted the ablation on long-term forecasting of 6000 timesteps in the Sleep-EDF dataset and evaluated the top-5 recall scores in the PopHR dataset.

Table 3: Ablation results of TimelyGPT w/o specific components, showing forecasting performance for a 6,000-timestep window in the Sleep-EDF dataset and top-15 recall rate in the PopHR dataset.

Datasets	Sleep-EDF (6000)	PopHR (K=5)
TimelyGPT (with Pre-training)	0.575	58.65
w/o Convolution Subsampling	0.587	—
w/o Temporal Convolution	0.581	57.69
w/o Exponential Decay	0.715	52.50
w/o RoPE (GPT-2)	1.072	50.18
TimelyGPT (w/o Pre-training)	0.641	56.42

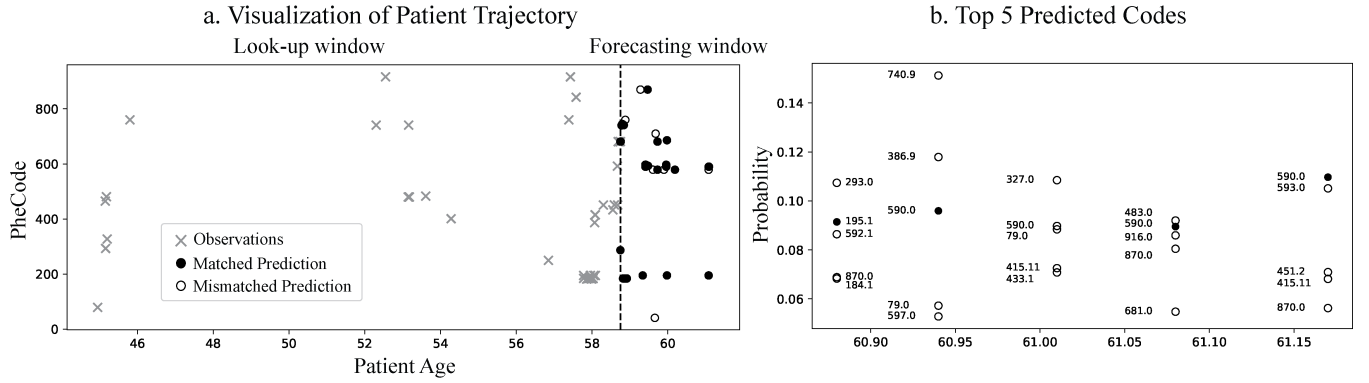


Figure 7: Visualization of a cancer patient's medical trajectory from the PopHR dataset. a. Look-up and forecast windows. Matched predictions (solid circles) were identified when the top 5 predicted PheCodes contain the groundtruth. b. The top 5 predicted PheCodes for the final 5 timesteps of the subject.

As shown in Table 3, for the Sleep-EDF forecasting task, removing the RoPE component led to the most significant performance degradation (a MAE of 0.357). The removal of exponential decay also led to increase MAE of 0.134, demonstrating its benefits of encoding trend patterns for long-term forecasting. Together, the two ablation experiments show the importance of xPos as our first main contribution (Section 3.1). The integration of convolution modules helps TimelyGPT capture local features, although the benefits were smaller compared with other components.

In the forecasting of irregularly-sampled time series, the exponential decay and RoPE components improved performance by 6.15% and 2.32%, respectively. The time decay mechanism encodes trend patterns into the modeling of patients' health trajectories, making it a promising approach for forecasting irregular clinical diagnoses. Pre-training decreased MAE by 0.066 for forecasting continuous biosignals in Sleep-EDF and increased top K recall rate by 2.21% for forecasting irregularly sampled diagnostic codes.

7 CONCLUSION AND FUTURE WORK

TimelyGPT effectively forecasts long sequences of time-series, utilizing xPos embedding, recurrent attention, and convolution modules. For continuously monitored biosignals such as Sleep-EDF, TimelyGPT can accurately extrapolate up to 6,000 timesteps given only a 2000-timestep prompt. Moreover, TimelyGPT also effectively forecasts irregularly-sampled time series by conditioning the recurrent Retention on the time. In our future work, we will perform comprehensive and in-depth analysis on the trajectory inference of the EHR data, as it may have a profound impact on the future of patient care and early intervention. TimelyGPT is a causal model with unidirectional attention [27]. This may limit its expressiveness in terms of time-series representation learning, which may be improved via a bidirectional architecture. To enhance transfer learning, we will adapt TimelyGPT for out-of-distribution biosignals, further enhancing its utility in healthcare time-series.

REFERENCES

- [1] Yuri Ahuja, Yuesong Zou, Aman Verma, David Buckeridge, and Yue Li. 2022. MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale

automatic phenotyping using the electronic health record. *Journal of biomedical informatics* 134 (2022), 104190.

- [2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]
- [3] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. arXiv:1610.02357 [cs.CV]
- [4] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXiv:1901.02860 [cs.LG]
- [5] Joshua Denny, Lisa Bastarache, Marylyn Ritchie, Robert Carroll, Raquel Zink, Jonathan Mosley, Julie Field, Jill Pulley, Andrea Ramirez, Erica Bowton, Melissa Basford, David Carrell, Peggy Peissig, Abel Kho, Jennifer Pacheco, Luke Rasmussen, David Crosslin, Paul Crane, Jyotishman Pathak, and Dan Roden. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology* 31 (11 2013). <https://doi.org/10.1038/nbt.2749>
- [6] Joshua C. Denny, Marylyn D. Ritchie, Melissa A. Basford, Jill M. Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R. Masys, Dan M. Roden, and Dana C. Crawford. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26, 9 (03 2010), 1205–1210. <https://doi.org/10.1093/bioinformatics/btq126>
- [7] Emadelddeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwah, Xiaoli Li, and Cuntai Guan. 2021. Time-Series Representation Learning via Temporal and Contextual Contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. 2352–2359.
- [8] A.L. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, R. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H.E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, 23 (2000), e215–e220. <https://doi.org/10.1161/01.cir.101.23.e215>
- [9] Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently Modeling Long Sequences with Structured State Spaces. arXiv:2111.00396 [cs.LG]
- [10] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. 5036–5040. <https://doi.org/10.21437/Interspeech.2020-3015>
- [11] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. 2020. Set Functions for Time Series. arXiv:1909.12064 [cs.LG]
- [12] Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. 2023. PDFformer: Propagation Delay-Aware Dynamic Long-Range Transformer for Traffic Flow Prediction. arXiv:2301.07945 [cs.LG]
- [13] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361 [cs.LG]
- [14] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast Autoregressive Transformers with Linear

- Attention. arXiv:2006.16236 [cs.LG]
- [15] B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Obery. 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave micro-continuity of the EEG. *IEEE Transactions on Biomedical Engineering* 47, 9 (2000), 1185–1194. <https://doi.org/10.1109/10.867928>
 - [16] Yann LeCun and Yoshua Bengio. 1998. *Convolutional Networks for Images, Speech, and Time Series*. MIT Press, Cambridge, MA, USA, 255–258.
 - [17] Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. 2023. Revisiting Long-term Time Series Forecasting: An Investigation on Linear Mapping. arXiv:2305.10721 [cs.LG]
 - [18] Andy T. Liu, Shang-Wen Li, and Hung yi Lee. 2021. TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2351–2366. <https://doi.org/10.1109/taslp.2021.3095662>
 - [19] Andy T. Liu, Shu wen Yang, Po-Han Chi, Po chun Hsu, and Hung yi Lee. 2020. Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. <https://doi.org/10.1109/icassp40776.2020.9054458>
 - [20] Qianli Ma, Zhen Liu, Zhenjing Zheng, Ziyang Huang, Siying Zhu, Zhongzhong Yu, and James T. Kwok. 2023. A Survey on Time-Series Pre-Trained Models. arXiv:2305.10716 [cs.LG]
 - [21] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. arXiv:2211.14730 [cs.LG]
 - [22] Inkit Padhi, Yair Schiff, Igor Melnyk, Mattia Rigotti, Youssef Mroueh, Pierre Dognin, Jerret Ross, Ravi Nair, and Erik Altman. 2021. Tabular Transformers for Modeling Multivariate Time Series. arXiv:2011.01843 [cs.LG]
 - [23] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv:2306.01116 [cs.CL]
 - [24] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanslaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the Transformer Era. arXiv:2305.13048 [cs.CL]
 - [25] Huy Phan, Oliver Y. Chen, Minh C. Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2021. XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. <https://doi.org/10.1109/tpami.2021.3070057>
 - [26] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena Hierarchy: Towards Larger Convolutional Language Models. arXiv:2302.10866 [cs.LG]
 - [27] Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. arXiv:2108.12409 [cs.CL]
 - [28] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [eess.AS]
 - [29] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. <https://api.semanticscholar.org/CorpusID:160025533>
 - [30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]
 - [31] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. 2019. Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. *Sensors* 19, 14 (2019). <https://doi.org/10.3390/s19143079>
 - [32] Arash Shaban-Nejad, Maxime Lavigne, Anya Okhmatovskaia, and David Buckenridge. 2016. PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data: The Population Health Record (PopHR). *Annals of the New York Academy of Sciences* 1387 (10 2016). <https://doi.org/10.1111/nyas.13271>
 - [33] Zezhi Shao, Zhao Zhang, Fei Wang, and Yongjun Xu. 2022. Pre-training Enhanced Spatial-temporal Graph Neural Network for Multivariate Time Series Forecasting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM. <https://doi.org/10.1145/3534678.3539396>
 - [34] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 464–468. <https://doi.org/10.18653/V1/N18-2074>
 - [35] Satya Narayan Shukla and Benjamin M. Marlin. 2021. Multi-Time Attention Networks for Irregularly Sampled Time Series. arXiv:2101.10318 [cs.LG]
 - [36] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. 2023. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), 710–719. <https://doi.org/10.1109/TNSRE.2022.3230250>
 - [37] Ziyang Song, Yuanyi Hu, Aman Verma, David L. Buckeridge, and Yue Li. 2022. Automatic Phenotyping by a Seed-guided Topic Model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) (*KDD '22*). Association for Computing Machinery, New York, NY, USA, 4713–4723. <https://doi.org/10.1145/3534678.3542675>
 - [38] Rachel Stirling, Mark Cook, David Grayden, and Pip Karoly. 2020. Seizure forecasting and cyclic control of seizures. *Epilepsia* 62 Suppl 1 (07 2020). <https://doi.org/10.1111/epi.16541>
 - [39] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2022. RoFormer: Enhanced Transformer with Rotary Position Embedding. arXiv:2104.09864 [cs.CL]
 - [40] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive Network: A Successor to Transformer for Large Language Models. arXiv:2307.08621 [cs.CL]
 - [41] Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benham, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A Length-Extrapolatable Transformer. arXiv:2212.10554 [cs.CL]
 - [42] Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Michael Blumenstein, and Jing Jiang. 2021. Omni-Scale CNNs: a simple and effective kernel size configuration for time series classification. In *International Conference on Learning Representations*.
 - [43] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. MLP-Mixer: An all-MLP Architecture for Vision. arXiv:2105.01601 [cs.CV]
 - [44] 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:2302.13971 [cs.CL]
 - [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]
 - [46] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. 2022. ETSformer: Exponential Smoothing Transformers for Time-series Forecasting. arXiv:2202.01381 [cs.LG]
 - [47] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.
 - [48] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2022. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. arXiv:2106.13008 [cs.LG]
 - [49] Zhanhao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. 2020. Lite Transformer with Long-Short Range Attention. arXiv:2004.11886 [cs.CL]
 - [50] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs.CL]
 - [51] Mengru Yuan, Guido Powell, Maxime Lavigne, Anya Okhmatovskaia, and David Buckenridge. 2018. Initial Usability Evaluation of a Knowledge-Based Population Health Information System: The Population Health Record (PopHR). *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2017 (04 2018), 1878–1884.
 - [52] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. TS2Vec: Towards Universal Representation of Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (Jun. 2022), 8980–8987. <https://doi.org/10.1609/aaai.v36i8.20881>
 - [53] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2022. Are Transformers Effective for Time Series Forecasting? arXiv:2205.13504 [cs.AI]
 - [54] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2020. A Transformer-based Framework for Multivariate Time Series Representation Learning. arXiv:2010.02803 [cs.LG]
 - [55] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12104–12113.
 - [56] Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. 2023. Self-Supervised Time Series Representation Learning via Cross Reconstruction Transformer. arXiv:2205.09928 [cs.LG]
 - [57] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Graph-Guided Network for Irregularly Sampled Multivariate Time Series. arXiv:2110.05357 [cs.LG]
 - [58] Liang Zhao, Min Gao, and Zongwei Wang. 2022. ST-GSP: Spatial-Temporal Global Semantic Representation Learning for Urban Flow Prediction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) (*WSDM '22*). Association for Computing Machinery, New York,

- NY, USA, 1443–1451. <https://doi.org/10.1145/3488560.3498444>
- [59] 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. arXiv:2012.07436 [cs.LG]
- [60] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. arXiv:2201.12740 [cs.LG]

A REVISITING TRANSFORMERS

A.1 Efficient attention in Transformer

Transformer models have found extensive applications in both the Natural Language Processing and Computer Vision domains [45]. In the vanilla self-attention mechanism, the query, key, value matrices are denoted as $Q, K, V \in \mathbb{R}^{N \times d}$. The output embedding for the n -th token is represented as $O_n = \frac{\sum_m^N \text{sim}(Q_n, K_m) V_m}{\sum_m^N \text{sim}(Q_n, K_m)}$, where the similarity function represents the softmax of inner-product $\text{sim}(Q_n, K_m) = \exp(Q_n K_m^\top / \sqrt{d})$. The self-attention mechanism, also known as token-mixer, aims to integrate information from every token and thus capture global-range interaction. However, computing the dot product $Q_n K_m^\top$ before the softmax operation introduces computational complexity of $O(N^2 d)$. As sequence length increases, this quadratic complexity becomes bottleneck, making it challenging to train for longer sequences. Many studies have been proposed to address the quadratic issue in self-attention mechanism. The linear attention replaces the softmax term $\text{sim}(Q_n, K_m)$ with $\phi(Q_n) \phi(K_m^\top)$ for a nonlinear kernel function $\phi(\cdot)$ [14], avoiding quadratic computation.

Recent research has explored alternatives to the token-mixer attention mechanism including Multi-Layer Perceptron (MLP) [43], convolution [26], and RNN [24, 40]. Particularly, RNN-variant models like RWKV and RetNet have successfully scaled up to more than 14 billion parameters, yielding comparable performance to conventional transformers. A fascinating connection between linear attention and RNNs has been identified [14], making RNN-based token mixer as efficient as linear attention. The output embedding from linear attention can be recast as an RNN: $O_n = \frac{\phi(Q_n) \sum_m^N \phi(K_m^\top) V_m}{\phi(Q_n) \sum_m^N \phi(K_m^\top)} = \frac{\phi(Q_n) S_n}{\phi(Q_n) Z_n}$, where $S_n = \sum_m^N \phi(K_m^\top) V_m$, $Z_n = \sum_m^N \phi(K_m^\top)$. Thus, the output embedding O_n depends on both S_n and Z_n , which are incrementally updated through cumulative sums. Thus, the RNN-based token-mixer not only competes in performance, but also offers linear training and consistent inference complexities. By employing exponential decay mechanism, it diminishes the influence of distant positions, transitioning from “token-mixing” to “time-mixing”. Considering RNN’s historical effectiveness in time-series and audio domains, it stands out as an excellent choice for temporal modeling.

A.2 Time-series Transformer

Transformers are increasingly applied in LTSF tasks, attributed to their capabilities in capturing long-term temporal dependencies [21, 46, 48, 59, 60]. Researchers have modified transformers by incorporating custom attention modules to address complex temporal dependencies [48, 59, 60]. Studies like [46, 48, 60] have introduced time-decomposition techniques into attention mechanisms to bolster modeling capability. The majority of studies focus on the encoder-decoder architecture, coupled with a one-forward prediction framework [59]. In this design, the decoder takes a concatenated input of the context (or prompt) and placeholder forecasting windows, directly generating the resulting embedding without autoregressive decoding. As a result, these models aim to avoid error accumulation seen in autoregressive frameworks, but aligning its performance closely with linear models [53]. Encoder-only models, like patchTST, use the encoded embedding for forecasting with the help of a linear layer [21]. Additionally, self-supervised representation learning techniques in time series, such as TS2Vec and TimesNet, offer valuable representation learning capabilities for forecasting tasks [47, 52].

A.3 Transformer scaling law in time-series

Despite the broad applications of transformer-based models in time-series data such as speech [10, 28], biosignals [36], and traffic flow [12, 33], their effectiveness in capturing temporal dependencies in LTSF task has been limited and often underperforms compared to linear models [53]. As Table S1 indicates, time-series transformer models often have much more parameters than the dataset size (timestep) with only two exceptions, namely large-size Conformer and CRT. Such disparities imply that many transformers may be over-parameterized, leading to highly variable performance. In Section 6.1, our study validates the Transformer scaling law in time-series domain (i.e., scaling up both model parameters and dataset size to improve performance) [13, 55]. For all benchmark experiments, our proposed TimelyGPT effectively pre-trains on large-scale data with model parameters aligned to this scaling law.

Table S1: The model parameters and utilized datasets of time-series transformers and comparison methods. These setups are sourced from papers and default implementation. Over-parameterization indicates model parameters >> dataset size (timestep).

Method	Application	Dimension	Layer	Model Parameter	Dataset Size (Timestep)	Param versus Data
Informer	Forecasting	512	3	11.3M	69.7K	Over-param
Autoformer	Forecasting	512	3	10.5M	69.7K	Over-param
Fedformer (F/W)	Forecasting	512	3	16.3/114.3M	69.7K	Over-param
PatchTST	Forecasting	128	3	1.2M	69.7K	Over-param
DLinear	Forecasting	-	1	70K	69.7K	Adequate
Conformer (L)	Classification	512	18	118.8M	55.9B	Adequate
CRT	Pre-training	128	18	8.8 M	109.2M	Adequate

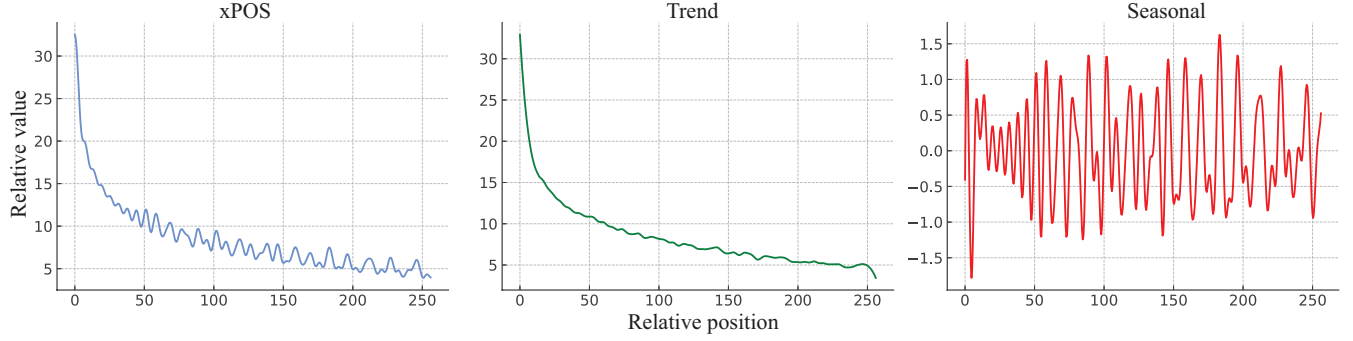


Figure S1: The xPos embedding diminishes distant temporal information according to the relative distance, enabling decomposition to capture both trend and periodic dynamics in time-series data.

B DETAILS ABOUT TIMELYGPT

B.1 From absolute to relative position embedding

Unlike RNNs or CNNs, the inclusion of positional embedding is essential for the Transformer model. Since the permutation-invariant self-attention mechanism cannot capture input order, making it challenging to differentiate tokens in various positions. The solution fall into two categories: (1) incorporate position information into the inputs, i.e., absolute position embedding; (2) modify the attention matrix to distinguish tokens at different positions, referring to relative position embedding.

In absolute position embedding, the token representation for a given token n consists of a word embedding X_n and a position embedding P_n . The self-attention mechanism is expressed as:

$$\begin{aligned} Q_n &= (X_n + P_n)W_Q, \quad K_n = (X_n + P_n)W_K, \quad V_n = (X_n + P_n)W_V \\ A_{n,m} &= \text{softmax}(Q_n K_m^\top), \quad O_m = \sum_m A_{n,m} V_m \end{aligned} \quad (6)$$

where $A_{n,m}$ is an attention score between token n and m without scaling. The inner-dot product $Q_n K_m^\top$ and output embedding O_m can be expanded as follows:

$$\begin{aligned} Q_n K_m^\top &= (X_n + P_n)W_Q ((X_m + P_m)W_K)^\top \\ &= (X_n + P_n)W_Q W_K^\top (X_m + P_m)^\top \\ &= \underbrace{X_n W_Q W_K^\top X_m^\top}_{\text{token-token}} + \underbrace{X_n W_Q W_K^\top P_m^\top}_{\text{token-position}} + \underbrace{P_n W_Q W_K^\top X_m^\top}_{\text{position-token}} + \underbrace{P_n W_Q W_K^\top P_m^\top}_{\text{position-position}} \end{aligned} \quad (7)$$

$$O_n = \sum_m \text{softmax}((X_n W_Q + P_n W_Q)(W_K^\top X_m^\top + W_K^\top P_m^\top)) (X_n + P_m)W_V \quad (8)$$

where attention arises from four types of interactions: (1) token-token interaction; (2) token-position interaction; (3) position-token interaction; (4) position-position interaction. However, absolute position embedding only incorporates fixed position information, neglecting the relative positional difference between the token n and m .

In the realm of audio processing, prevalent transformers like Conformer [10] incorporate relative positional information through the T5 position embedding [30]. Notably, the T5 model suggests a minimal interaction between tokens and positions, resulting in the exclusion of token-position and position-token terms from the attention matrix:

$$Q_n K_m^\top = X_n W_Q W_K^\top X_m^\top + \beta_{n,m} \quad (9)$$

where the position-position interaction term, $P_n W_Q W_K^\top P_m^\top$, is replaced with a trainable bias related to the position n and m . The T5 position embedding follows Transformer-XL, omitting the position term $P_m W_V$ in the attentive aggregation computation [4, 50]. As a result, the relative position embedding is only added to the dot product QK^\top :

$$O_n = \sum_m \text{softmax}(X_n W_Q W_K^\top X_m^\top + \beta_{n,m}) X_m W_V \quad (10)$$

The RoPE technique leverages the property of rotation matrix to model positional information [39]. To incorporate this relative position information into the queries Q and keys K , the method aims to identify functions $f_Q(Q, \cdot)$ and $f_K(K, \cdot)$ that satisfies this invariant criteria

about relative distance:

$$\langle Q_n, K_m \rangle = \langle f_Q(Q, n), f_K(K, m) \rangle = g(Q, K, m - n), \quad (11)$$

where g is a function that depends only on the relative distance $m - n$ and $Q = XW_Q$ and $K = XW_K$ stand for token embedding for queries and keys matrices, respectively. RoPE defines the function f involving a d -dimensional rotation matrix R :

$$f_Q(Q, n) = R_{\Theta, n}^d(X_n W_Q), \quad f_K(K, m) = R_{\Theta, m}^d(X_m W_K) \quad (12)$$

With a given hidden size d , a block diagonal matrix $R_{\Theta, n}^d$ contains multiple rotation matrices $(R_{n, \theta_1}^{(1)}, \dots, R_{n, \theta_{d/2}}^{(d/2)})$ on its diagonal:

$$R_{\Theta, n}^d = \begin{bmatrix} R_{n, \theta_1}^{(1)} & & \\ & \ddots & \\ & & R_{n, \theta_{d/2}}^{(d/2)} \end{bmatrix}, \quad R_{n, \theta_i}^{(i)} = \begin{bmatrix} \cos n\theta_i & -\sin n\theta_i \\ \sin n\theta_i & \cos n\theta_i \end{bmatrix} \quad (13)$$

where the rotation hyperparameter $\theta_i = 10000^{-2(i-1)/d}$. In RoPE, any even-dimension representation can be built by placing multiple 2-dimensional rotation matrices diagonally within the $R_{\Theta, n}^d$ matrix, expanding hidden size from 2-dimension to d -dimension. As $R_{\Theta, m-n}^d = (R_{\Theta, n}^d)^\top R_{\Theta, m}^d$, RoPE satisfies the property outlined in Eq 11:

$$\begin{aligned} \langle Q_n, K_m \rangle &= \sum_{i=1}^{d/2} \langle Q_n[2i-1:2i], K_m[2i-1:2i] \rangle \\ &= \sum_{i=1}^{d/2} R_{\Theta, m-n}^d \langle (X_n W_Q)[2i-1:2i], (X_m W_K)[2i-1:2i] \rangle \end{aligned} \quad (14)$$

In RoPE, relative position information is added to the inner product QK^\top by rotating the angles of queries and keys matrices. Recently, [41] argues that the sinusoids used in the rotation matrices do not change monotonically. Instead, they oscillate dramatically as the relative distance increases. This limitation hinders RoPE's ability to sequences of extended lengths. To address it, [41] proposes xPos that preserves the advantage of RoPE and behaves stably at long-term dependency by measuring position monotonicity [41].

B.2 Equivalence of three forward-pass Retention

According to Section 3.2, the parallel forward-pass is equivalent to the recurrent forward-pass. With the initial state variable $S_0 = 0$, the recurrent forward-pass can be expressed as follows:

$$\begin{aligned} \text{Recurrent: } S_n &= \underbrace{K_n^\top V_n}_{\text{Single-token}} + \gamma S_{n-1}, \quad \text{Ret}(X_n) = Q_n S_n \\ \implies S_n &= \sum_m^n \gamma^{n-m} K_m^\top V_m, \quad \text{Ret}(X_n) = Q_n \sum_m^n \gamma^{n-m} K_m^\top V_m \end{aligned} \quad (15)$$

where $\text{Ret}(X_n)$ calculates the Retention at single-time n by considering timestep i up to the current time. It corresponds to the n -th timestep (row) of parallel forward-pass of Retention.

$$\begin{aligned} \text{Recurrent: } \text{Ret}(X_n) &= Q_n \sum_m^n \gamma^{n-m} K_m^\top V_m \\ \implies \text{Parallel: } \text{Ret}(X_n) &= \underbrace{Q_n}_{1 \times d_{qk}} \underbrace{K_{m \leq n}^\top}_{d_{qk} \times n} \underbrace{\odot D_{m \leq n}}_{n \times n} \underbrace{V_{m \leq n}}_{n \times d_v} \end{aligned} \quad (16)$$

When the recurrent forward-pass traverses all timesteps, the parallel and recurrent forward-passes of Retention become identical. With the parallel and recurrent forward-passes of Retention, we aim to show the equivalence between the chunk-wise forward-pass and the parallel and recurrent forward-passes. The computation of chunk-wise Retention involves both parallel intra-chunk and recurrent inter-chunk computation as follows.

$$\begin{aligned}
\text{Chunk-wise: Ret}(X_{[i]}) &= \underbrace{(Q_{[i]} K_{[i]}^\top \odot D) V_{[i]}}_{\text{Intra-chunk}} + \underbrace{(Q_{[i]} S_{[i-1]}) \odot \zeta}_{\text{Inter-chunk}} \\
S_{[i]} &= \underbrace{K_{[i]}^\top (V_{[i]} \odot D_B)}_{\text{Current chunk}} + \underbrace{\gamma^B S_{[i-1]}}_{\text{Past chunk}}, \quad \zeta_j = \gamma^j
\end{aligned} \tag{17}$$

where $\zeta = [\gamma^1, \gamma^2, \dots, \gamma^B]^\top$ is a column-vector of time-decay scaling factor for inter-chunk attention between the current chunk $[i]$ and the previous chunk $[i-1]$. Specifically, γ^j is the scaling factor for the j^{th} row of chunk $[i]$ from the last row of chunk $[i-1]$ such that the bigger the j index the smaller the γ^j value. Therefore, Retention recursively aggregates information from the i -th chunk (i.e., intra-chunk embedding) and the previous chunk (i.e., inter-chunk embedding).

For the per-chunk state variable $S_{[i]}$, it computes current-chunk information as well as past-chunk information. The current-chunk information $K_{[i]}^\top V_{[i]}$ decays by D_B , which is the last row of decay matrix D . The past chunk information $S_{[i-1]}$ is decayed with respect to the chunk size B . The initial state variable $S_{[i=0]} = 0$ is computed recurrently given the chunk size B :

$$S_{[i]} = K_{[i]}^\top (V_{[i]} \odot D_B) + \gamma^B S_{[i-1]} = \sum_{m=1}^B \gamma^{B-m} K_m^\top V_m + \gamma^B S_{[i-1]} \tag{18}$$

Moreover, the update of state variable $S_{[i]}$ can be reformulated in parallel. The first term represents the information of current chunk, and the second term represented the past-chunk information decayed by the chunk size B . Consequently, $S_{[i-1]}$ represents the state information from the beginning to the $(i-1)$ -th chunk, and we represent the inter-chunk information in chunk-wise Retention:

$$\begin{aligned}
S_{[i-1]} &= \sum_{m=1}^{B*i} \gamma^{B*i-m} K_m^\top V_m = K_{1:(B*i)}^\top \odot D_{1:(B*i)} V_{1:(B*i)} \\
\underbrace{(Q_{[i]} S_{[i-1]}) \odot \zeta}_{\text{Inter-chunk}} &= (Q_{(B*i):(B*(i+1))} K_{1:(B*i)}^\top \odot D_{1:(B*i)} V_{1:(B*i)}) \odot \zeta \\
&= Q_{(B*i):(B*(i+1))} K_{1:B*i}^\top \odot D_{(B*i):(B*(i+1))} V_{1:(B*i)}
\end{aligned} \tag{19}$$

where the intra-chunk computation updates each row of the lower triangular matrix (highlighted as green in Fig. 1.c). Together, the recurrent intra-chunk computation with the parallel intra-chunk computation (highlighted as purple Fig. 1c) completes the chunk-wise forward-pass of Retention.

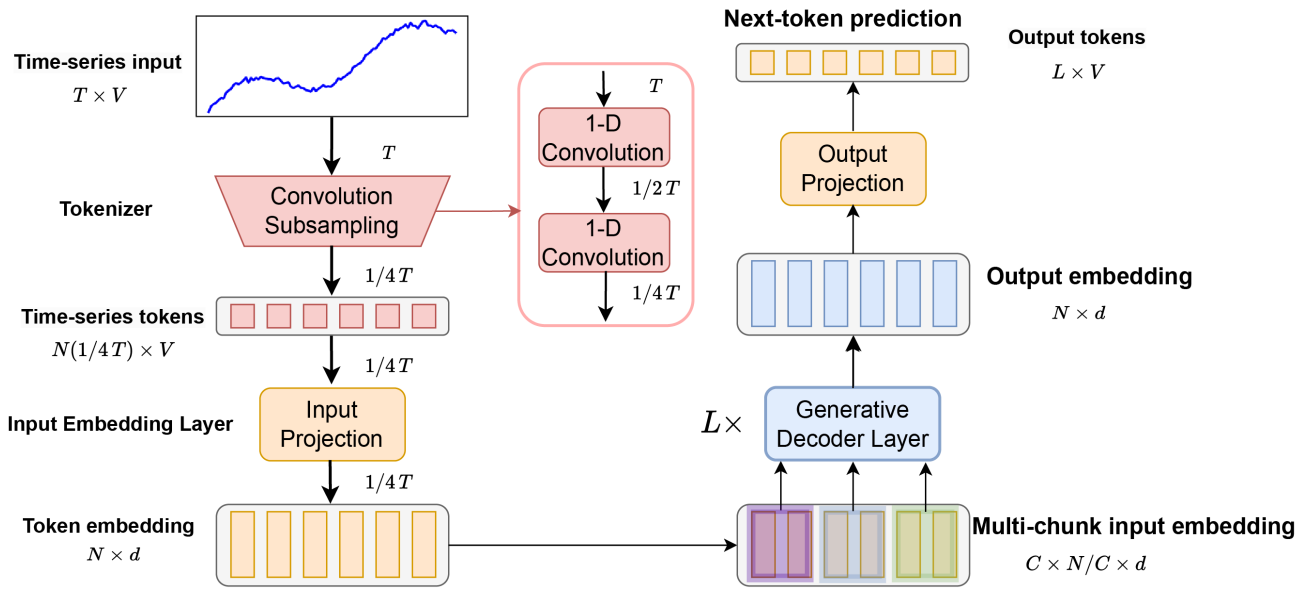


Figure S2: Schematic of the TimelyGPT Pre-Training Process

B.3 TimelyGPT pre-training overflow

For the TimelyGPT pre-training, we illustrate the full processes of input processing, model training, and next-token prediction in Fig. S2. For a time-series input with T timesteps and V variates, it is tokenized via a convolution-subsampling module. This tokenizer, typically comprising two 1-D convolution layers with a kernel size of 3 and stride of 2. It produces a sequence of tokens of the shape $N \times V$, effectively reducing the sequence length to $1/4$, i.e., $N = 1/4T$. The sequence of tokens is projected into an input embedding of the shape $L \times d$ with an linear projection layer. As a result, the input embedding is passed through L generative decoder layers, where the Retention mechanism takes segmented multiple-chunk input embedding. Finally, the output embedding of the shape $N \times d$ is passed through an output projection layer, which generate a sequence of tokens with the shape of $L \times V$ for next-token prediction.

C EXPERIMENT SUMMARY

Table S2: Configurations of TimelyGPT, transformer baselines, and recurrent models across different datasets

	Sleep-EDF	PopHR
Data Size (timesteps)	1.2B	54.9M
Model Parameters	18M	7.5M
TimelyGPT		
Decoder Layers	12	8
Heads	8	4
Dim (Q, K, V, FF)	320,320,640,640	200,200,400,400
Transformer baselines including Encoder-decoder and Encoder-only models		
Enc-Dec Layers	6 & 6	4 & 4
Encoder Layers	12	8
Decoder Layers	12	8
Heads	8	4
Dim (Q, K, V, FF)	384,384,384,1536	200,200,200,400
Recurrent Models		
Layers	12	8
Dim	384	200

We summarize the setup of model architecture for TimelyGPT and other baselines for the experiments in Table S2. Additionally, we also provide the visualization of forecasting experiment on the period signal (EEG Pz-Oz) in Fig. S3.

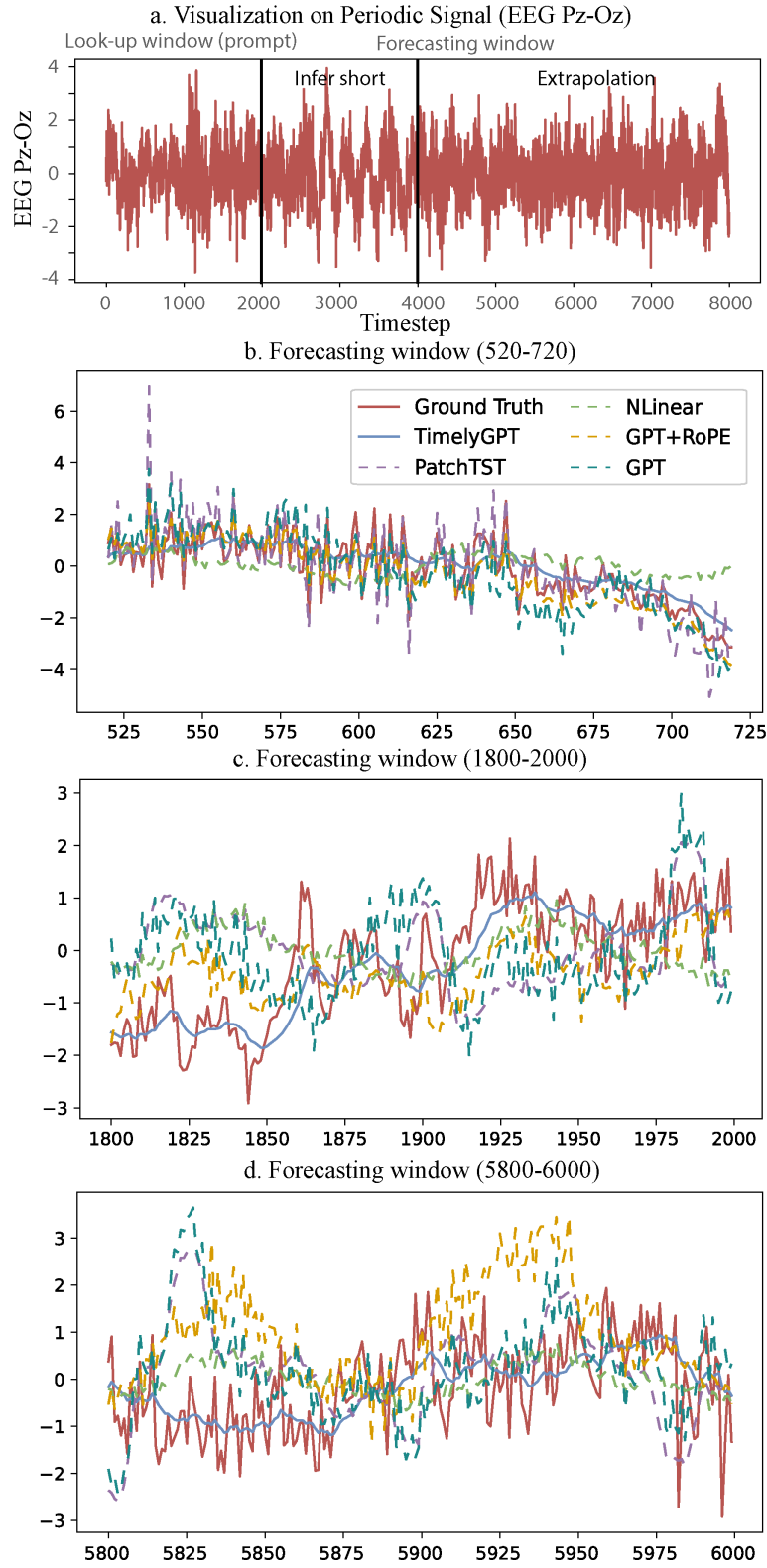


Figure S3: Example of forecasting experiments on the period signal (EEG Pz-Oz). a. the groundtruth of EEG Pz-Oz singal. Forecasting results are shown between 520 and 720 timesteps (b), 1800 and 2000 timesteps (c), and 5800 and 6000 timesteps (d). TimelyGPT is able to forecast the periodic signals up to 6000 timesteps owing to the extrapolation capabilities.