Time-series Transformer using Linear Attention for Representation Learning and Health Forecasting

COMP565 Machine Learning in Genomics and Healthcare

Ziyang Song
ziyang.song@mail.mcgill.ca

School of Computer Science
McGill University, Canada

November 19th, 2024

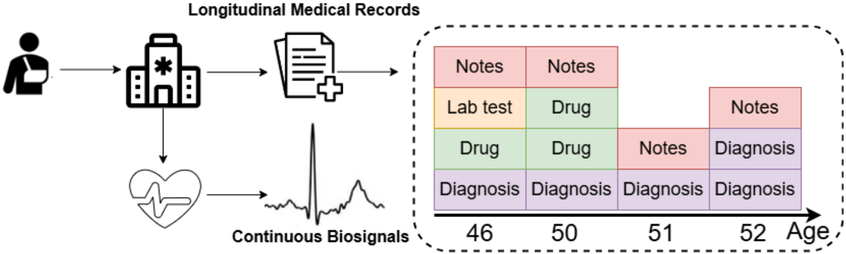
Outline

Outline

Continuous and Irregular-sampled Time Series

Continuous time series

- Data is collected at equal intervals, such as biosignals (e.g., heart rate).
- Classic time-series models, like RNNs and Transformers, apply to this category.



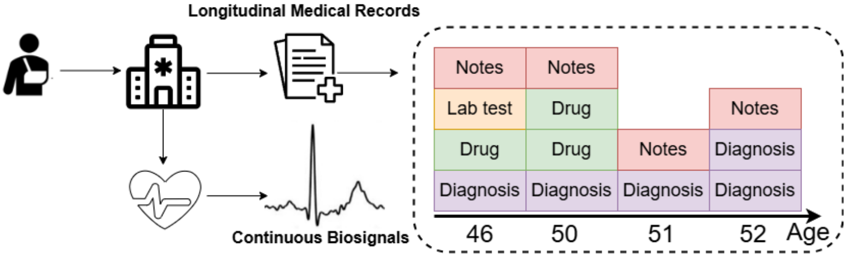
Continuous and Irregular-sampled Time Series

Continuous time series

- Data is collected at equal intervals, such as biosignals (e.g., heart rate).
- Classic time-series models, like RNNs and Transformers, apply to this category.

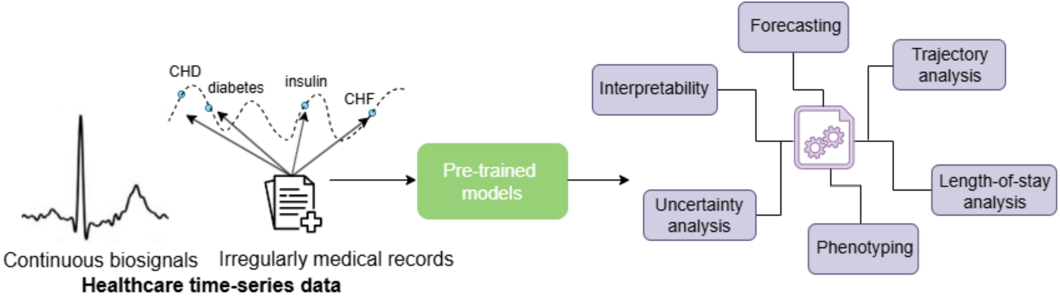
Irregularly-sampled time series

- Observations are collected at uneven intervals, as seen in patient medical records.
- Specialized models are needed to effectively handle this category of data.



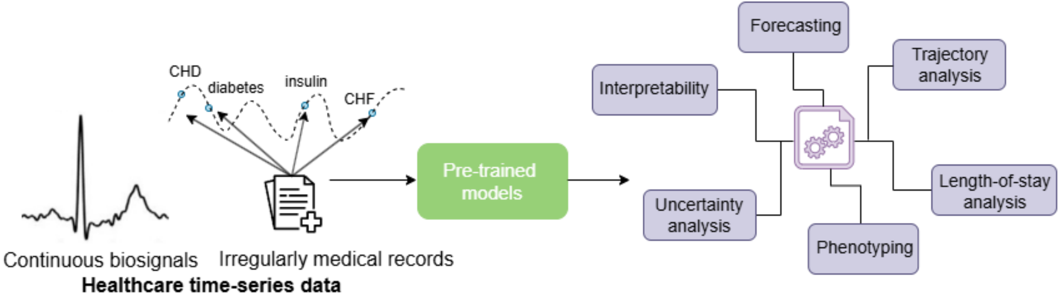
Time Series Representation Learning

- Pre-trained models (PTMs) learn generalizable temporal patterns.



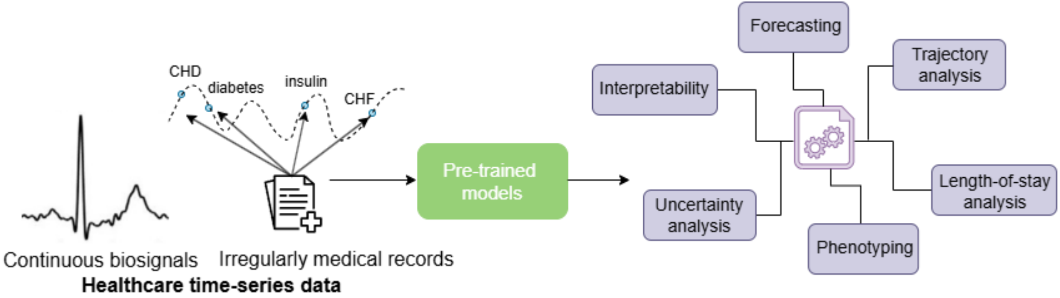
Time Series Representation Learning

- Pre-trained models (PTMs) learn generalizable temporal patterns.
- PTMs could provide strong zero(few)-shot learning for various tasks.



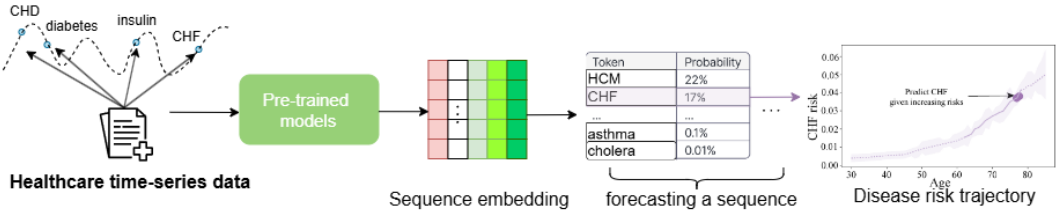
Time Series Representation Learning

- Pre-trained models (PTMs) learn generalizable temporal patterns.
- PTMs could provide strong zero(few)-shot learning for various tasks.
- How to handle both continuous and irregularly-sampled time series?



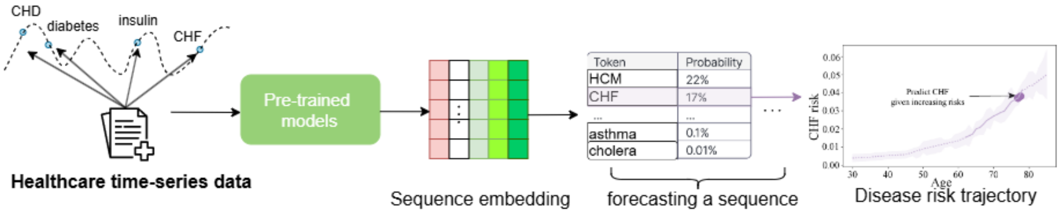
Forecasting and Trajectory Analysis

- PTMs take output sequence representation to forecast a target sequence with a specific inference method.



Forecasting and Trajectory Analysis

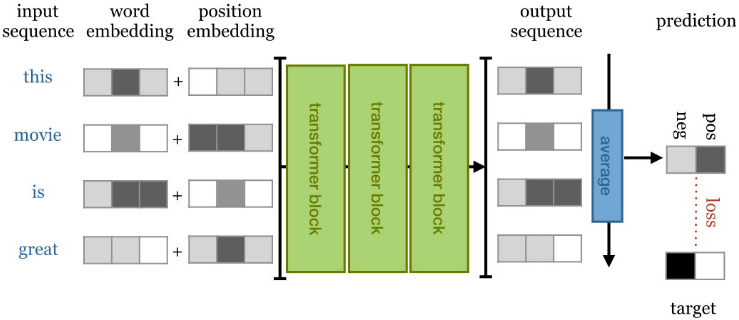
- PTMs take output sequence representation to forecast a target sequence with a specific inference method.
- Compute probabilities for a specific token over time to generate a disease risk trajectory, aiding in interpretable analysis.



Outline

Absolute Position Embedding (PE) for Transformer

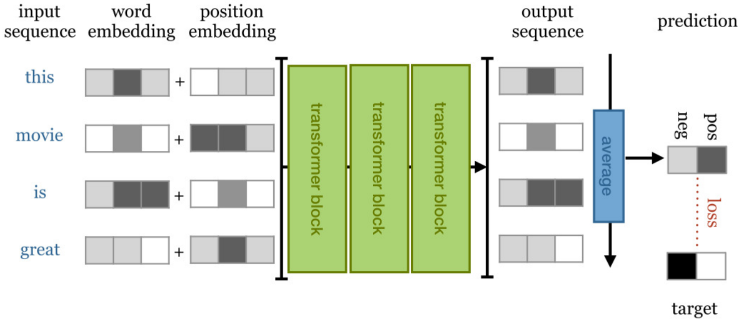
- Self-attention is position-invariant: $\text{Softmax}(QK^\top)V$



Absolute Position Embedding (PE) for Transformer

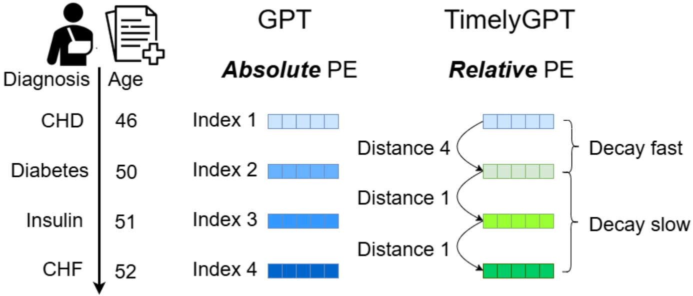
- Self-attention is position-invariant: $\mathrm{Softmax}(QK^\top)V$
- For each token n, the input embedding consists of token embedding X_n and position embedding P_n:

$$Q_n = (X_n + P_n)W_Q, \quad K_n = (X_n + P_n)W_K, \quad V_n = (X_n + P_n)W_V$$



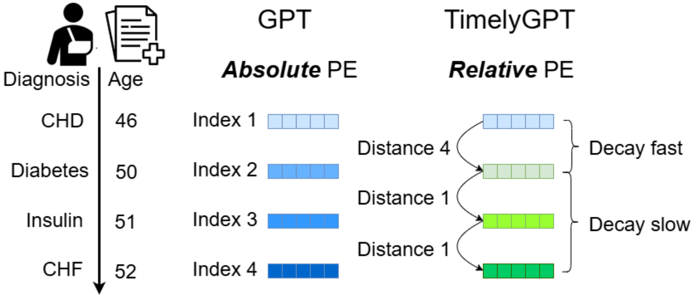
PE for Irregularly-sample Time Series

- Absolute PE only assigns discriminable embedding for the token positions.



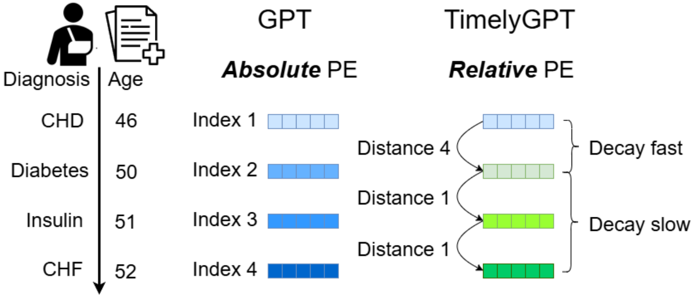
PE for Irregularly-sample Time Series

- Absolute PE only assigns discriminable embedding for the token positions.
- Absolute PE fails to capture varying **time interval**.



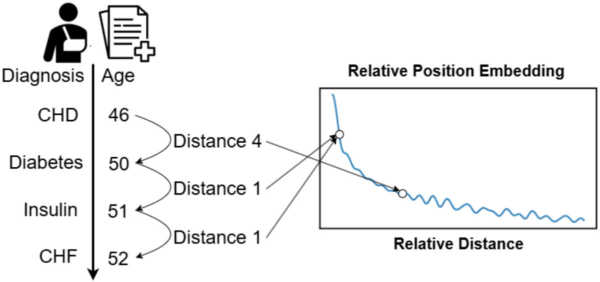
PE for Irregularly-sample Time Series

- Absolute PE only assigns discriminable embedding for the token positions.
- Absolute PE fails to capture varying **time interval**.
- Time intervals reflect underlying health dynamics; it has a high sampling rate during a poor health state.



Relative PE Handles Irregularity

- Relative PE (e.g., RoPE, xPos) encodes positional information based on relative distance between two tokens (e.g., age).

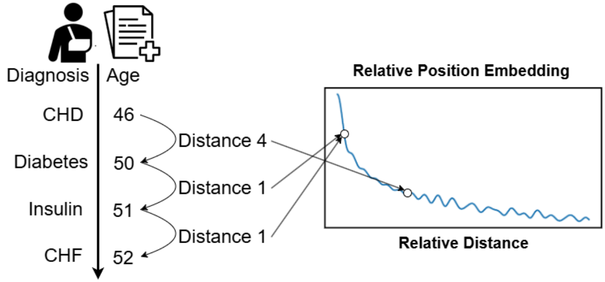


Relative PE Handles Irregularity

- Relative PE (e.g., RoPE, xPos) encodes positional information based on relative distance between two tokens (e.g., age).
- RoPE defines the following positional encoding:

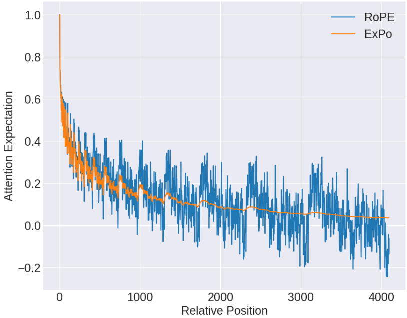
$$Q_n = X_n W_Q e^{i\theta t_n}, \quad K_m = X_m W_K e^{-i\theta t_m}, \quad Q_n^\top K_m = W_Q^\top X_n^\top e^{i\theta(t_n - t_m)} X_m W_K$$

where $t_n - t_m$ represents the difference in age between two tokens.



Extrapolatable Position (xPos) Embedding

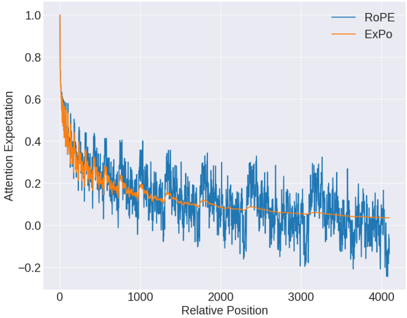
- Transformers face challenges of length extrapolation, leading to performance decline if inference length exceeds training length.



Extrapolatable Position (xPos) Embedding

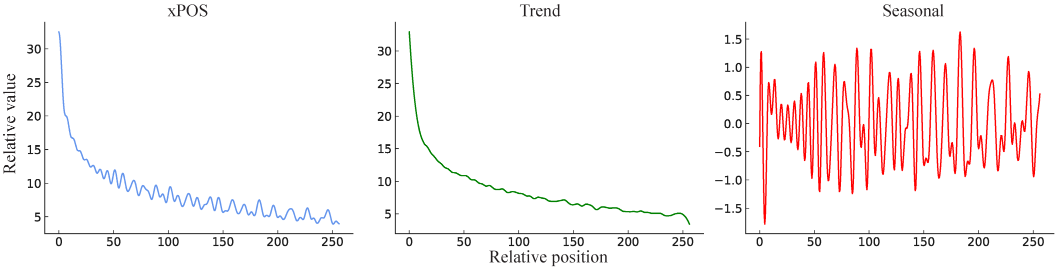
- Transformers face challenges of length extrapolation, leading to performance decline if inference length exceeds training length.
- xPos incorporates both decay γ^n and rotation $e^{i\theta n}$ for extrapolation:

$$Q_n^\top K_m = W_Q^\top X_n^\top (\gamma e^{i\theta})^{n-m} X_m W_K$$



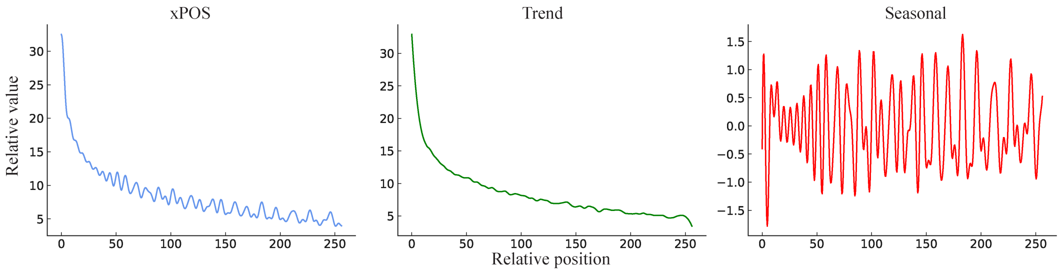
xPos Makes Extrapolation using Trend Patterns

- xPos mirrors the seasonal-trend decomposition in time series.



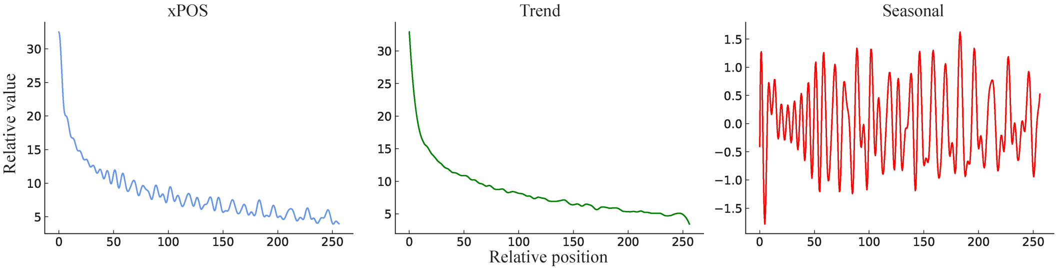
xPos Makes Extrapolation using Trend Patterns

- xPos mirrors the seasonal-trend decomposition in time series.
- The exponential decay γ^n and rotation matrix $e^{i\theta n}$ in the xPos embedding correspond to trend and periodic patterns, respectively.



xPos Makes Extrapolation using Trend Patterns

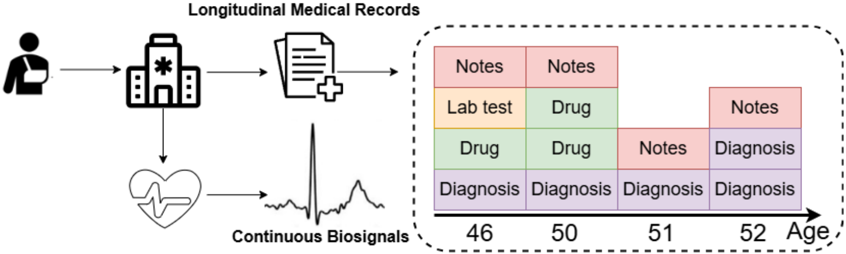
- xPos mirrors the seasonal-trend decomposition in time series.
- The exponential decay γ^n and rotation matrix $e^{i\theta n}$ in the xPos embedding correspond to trend and periodic patterns, respectively.
- Transformer with xPos embedding can make extrapolation by extending trend patterns over longer sequences.



Trend and Periodic Signal in Healthcare Time Series

Continuous time series (e.g., biosignals)

- Trend patterns: body temperature reflect human indicators.
- Periodic patterns: ECGs reflect physiological rhythms of the human body.



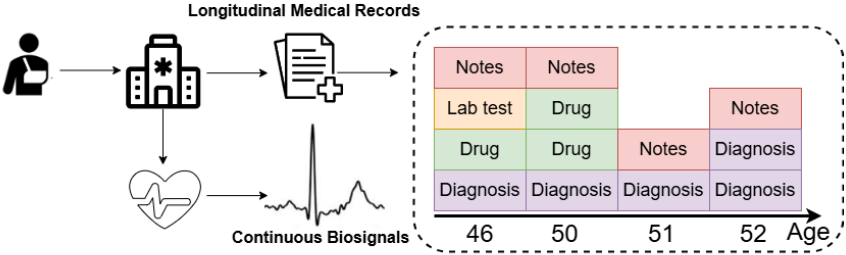
Trend and Periodic Signal in Healthcare Time Series

Continuous time series (e.g., biosignals)

- Trend patterns: body temperature reflect human indicators.
- Periodic patterns: ECGs reflect physiological rhythms of the human body.

Irregularly-sampled time series (e.g., Longitudinal EHR)

- Trend patterns: the age-related susceptibility to illnesses.
- Periodic patterns: the alternating exacerbation and recovery cycles



Outline

Retention Handles Irregularity

- Transformer with xPos can be rewritten as a recurrent attention (Retention), with both parallel and recurrent forms.

Retention Handles Irregularity

- Transformer with xPos can be rewritten as a recurrent attention (Retention), with both parallel and recurrent forms.
- The parallel form:

$$Q_n = X_n W_Q e^{i\theta n}, \quad K_m = X_m W_K e^{-i\theta m}, \quad V = X W_V,$$

$$O = (QK^\top \odot D)V, \quad D_{nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases}$$

Retention Handles Irregularity

- Transformer with xPos can be rewritten as a recurrent attention (Retention), with both parallel and recurrent forms.
- The parallel form:

$$Q_n = X_n W_Q e^{i\theta n}, \quad K_m = X_m W_K e^{-i\theta m}, \quad V = XW_V,$$
$$O = (QK^\top \odot D)V, \quad D_{nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases}$$

- The recurrent form involves a state variable S_n:

$$O_n = Q_n S_n, \quad S_n = \gamma S_{n-1} + K_n^\top V_n$$

Retention Handles Irregularity

- Transformer with xPos can be rewritten as a recurrent attention (Retention), with both parallel and recurrent forms.
- The parallel form:

$$Q_n = X_n W_Q e^{i\theta n}, \quad K_m = X_m W_K e^{-i\theta m}, \quad V = XW_V,$$
$$O = (QK^\top \odot D)V, \quad D_{nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases}$$

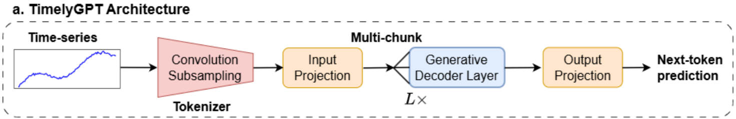
- The recurrent form involves a state variable S_n:

$$O_n = Q_n S_n, \quad S_n = \gamma S_{n-1} + K_n^\top V_n$$

- Retention handles irregularity by adapting $n - m$ to the time interval.

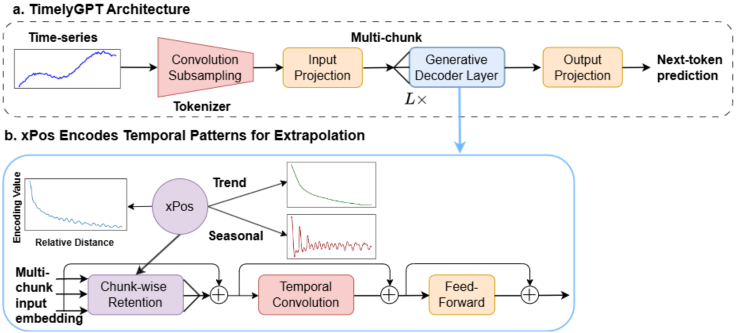
TimelyGPT Architecture

- TimelyGPT processes continuous biosignal data using a convolution subsampling module.
- For irregularly-sampled time series, TimelyGPT simply uses a learnable embedding layer to project a discrete token to an embedding.



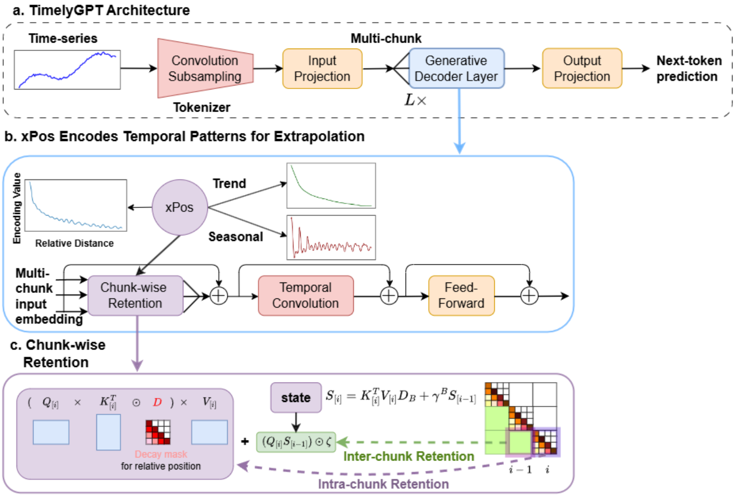
TimelyGPT Architecture

- xPos encoding trend and seasonal patterns with respect to distances into the token embedding.
- xPos can handle both continuous and irregular time series.



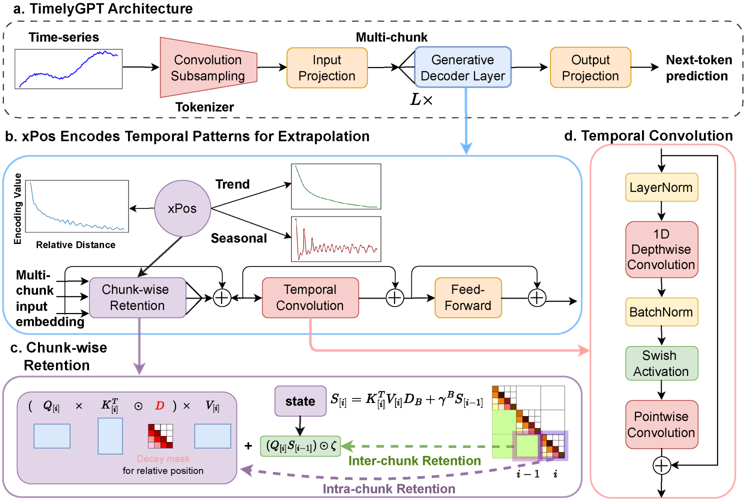
TimelyGPT Architecture

- Chunk-wise Retention processes long sequences with linear complexity.
- TimelyGPT with xPos falls short in modeling nuanced local patterns.



TimelyGPT Architecture

- A temporal convolution module in each layer captures local patterns.
- It models multi-scale features by stacking multiple layers.



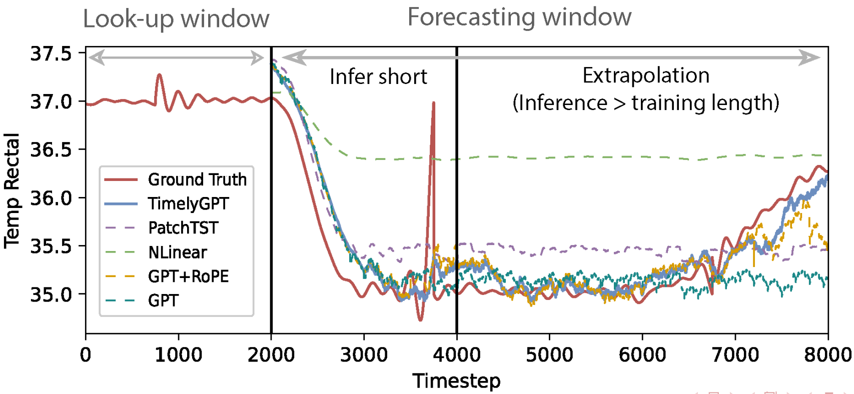
$(Q_{[i]} \times K^T_{[i]} \odot D) \times V_{[i]}$

$S_{[i]} = K^T_{[i]} V_{[i]} D_B + \gamma^B S_{[i-1]}$

$+ (Q_{[i]} S_{[i-1]}) \odot \zeta$

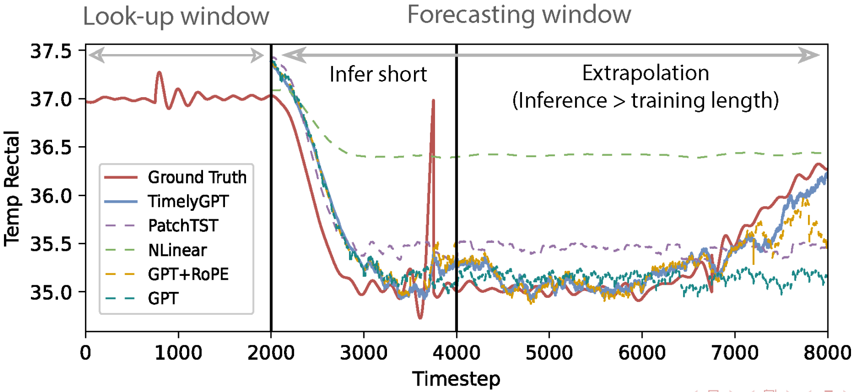
Forecasting Biosignals with a Case Study

- Pre-train on 4000 timesteps and infer on 8000 timesteps (with a 2000-step look-up window and a 6000-step forecasting window).



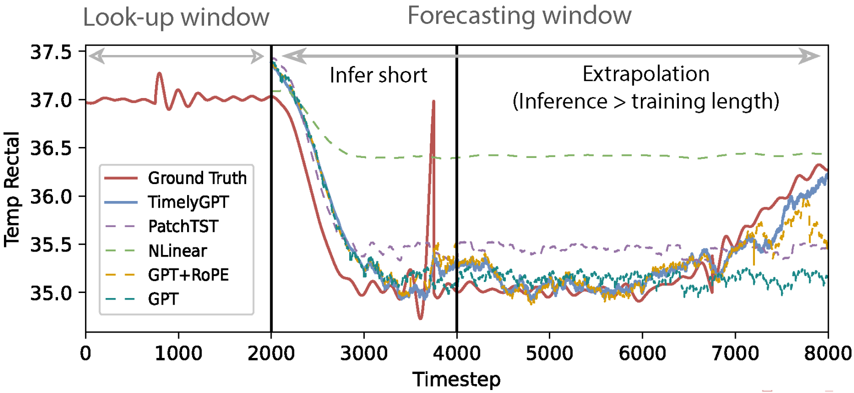
Forecasting Biosignals with a Case Study

- Pre-train on 4000 timesteps and infer on 8000 timesteps (with a 2000-step look-up window and a 6000-step forecasting window).
- Forecast beyond 2000 timesteps is considered as extrapolation.



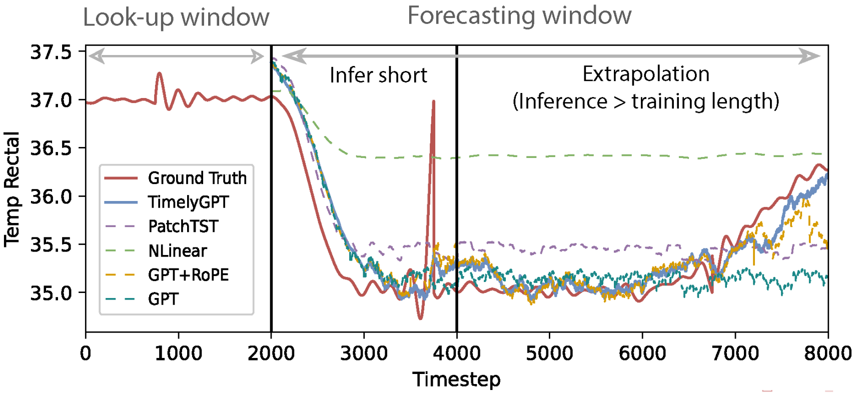
Forecasting Biosignals with a Case Study

- All pre-trained models can forecast the temperature drop by identifying the bump at 1000 timesteps, a typical indicator for temperature drop.



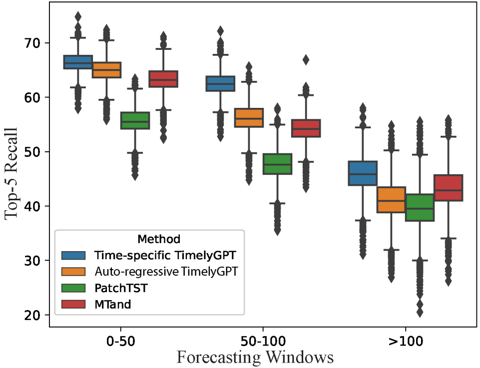
Forecasting Biosignals with a Case Study

- All pre-trained models can forecast the temperature drop by identifying the bump at 1000 timesteps, a typical indicator for temperature drop.
- TimelyGPT can make inference beyond the training length (4,000), indicating strong extrapolation capabilities.



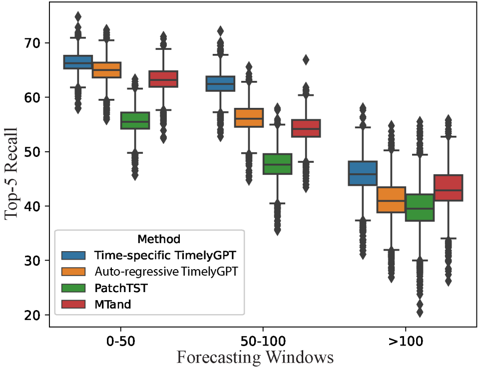
Forecasting Irregular Diagnoses on Varied Windows

- Given a 50-step look-up window, we evaluate top-5 recall distributions over three forecasting windows.



Forecasting Irregular Diagnoses on Varied Windows

- Given a 50-step look-up window, we evaluate top-5 recall distributions over three forecasting windows.
- TimelyGPT with a time-specific inference can avoid error accumulation, achieving better and stable prediction.



Outline

- Background of Healthcare Time Series and Transformer
- Relative Position Embedding for Continuous and Irregular Time Series
- TimelyGPT: Extrapolatable Transformer Pre-training for Long-term Time-Series Forecasting in Healthcare
- TimelyGPT with Time Specific Inference in an Ordinary Different Equation Framework
- Other and Future Works
- Reference

Ordinary Differential Equations (ODEs)

- ODE is a differential equation with only one independent variable time t:

$$\frac{dz(t)}{dt} = f_\theta(z(t))$$

where it models a continuous dynamics.

Ordinary Differential Equations (ODEs)

- ODE is a differential equation with only one independent variable time t:

$$\frac{dz(t)}{dt} = f_\theta(z(t))$$

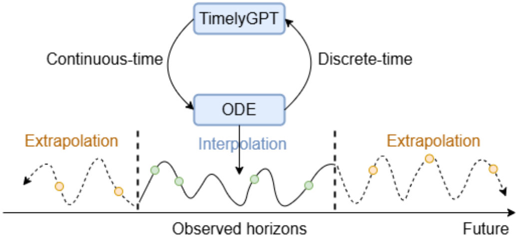
  where it models a continuous dynamics.

- Given t_0, an ODE can predict t_1 by integrating over its interval:

$$z(t_1) = z(t_0) + \int_{t_0}^{t_1} f_\theta(z(t))$$

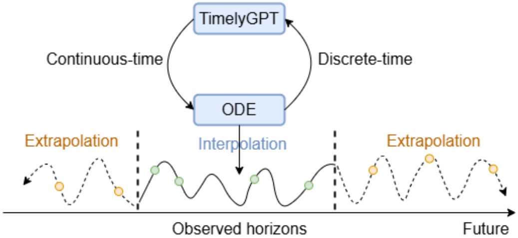
Connection to ODEs

- Our recurrent attention module can be viewed as a discretized ODE.



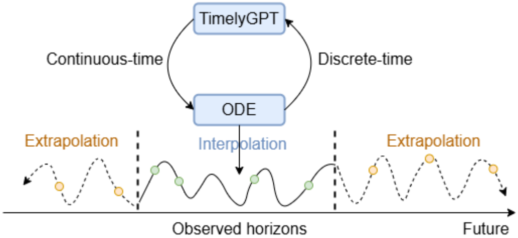
Connection to ODEs

- Our recurrent attention module can be viewed as a discretized ODE.
- It learns the continuous dynamics and handles irregular data through discretization with varying step sizes.



Connection to ODEs

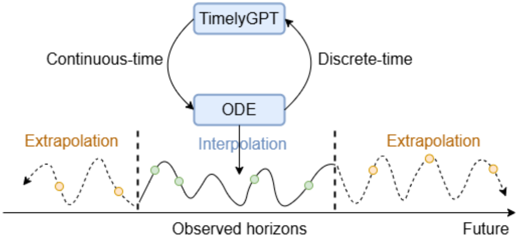
- Our recurrent attention module can be viewed as a discretized ODE.
- It learns the continuous dynamics and handles irregular data through discretization with varying step sizes.
- Interpolation evolves the dynamics within the observed timeframe using a unit discretization step size.



Connection to ODEs

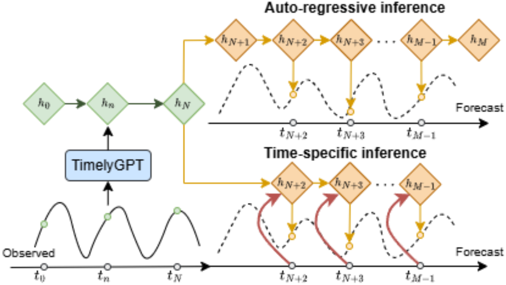
McGill UNIVERSITY

- Our recurrent attention module can be viewed as a discretized ODE.
- It learns the continuous dynamics and handles irregular data through discretization with varying step sizes.
- Interpolation evolves the dynamics within the observed timeframe using a unit discretization step size.
- Extrapolation evolves the dynamics forward or backward in time beyond the observed timeframe.



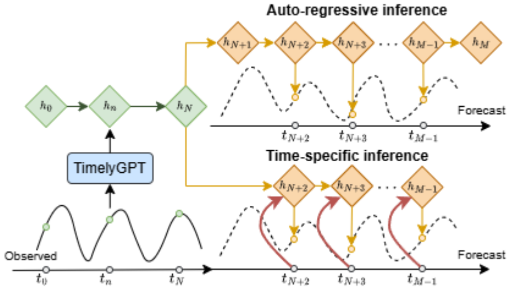
Time-specific Inference

- Leverage extrapolation technique, it forecasts observations at arbitrary timesteps.



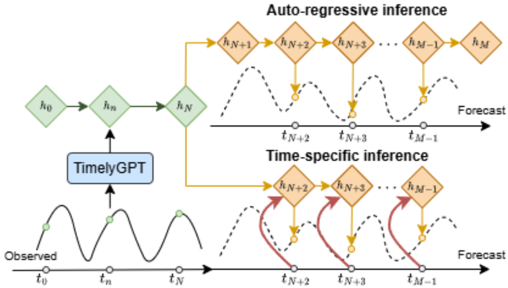
Time-specific Inference

- Leverage extrapolation technique, it forecasts observations at arbitrary timesteps.
- To forecast a target point at t_{N+2}, it uses both the target timestep t_{N+2} and the last hidden states h_N to estimate the corresponding observation h_{N+2}.



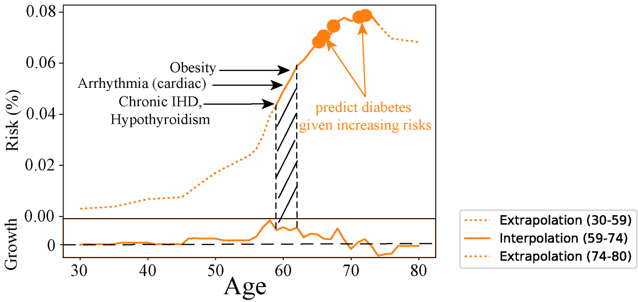
Time-specific Inference

- Leverage extrapolation technique, it forecasts observations at arbitrary timesteps.
- To forecast a target point at t_{N+2}, it uses both the target timestep t_{N+2} and the last hidden states h_N to estimate the corresponding observation h_{N+2}.
- Time-specific inference reduces computational steps and error accumulation, leading to better forecasting performance.



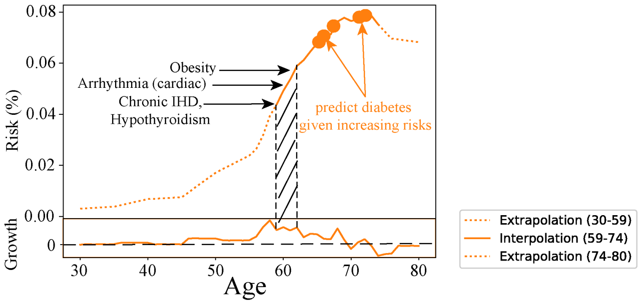
Risk Trajectory Analysis for a Diabetic Patient

- We compute the probabilities of the diabetes token over time as a risk trajectory, highlighting the crucial points with significant risk growth.



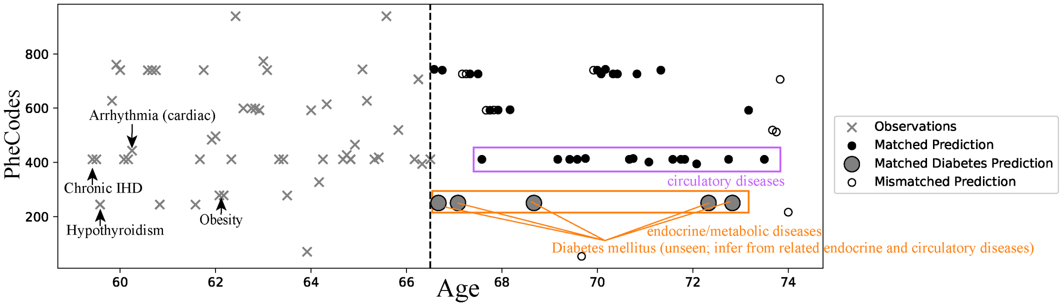
Risk Trajectory Analysis for a Diabetic Patient

- We compute the probabilities of the diabetes token over time as a risk trajectory, highlighting the crucial points with significant risk growth.
- Our model effectively captures clear trends of increasing risks with age, reflecting age-related vulnerability to chronic diseases.



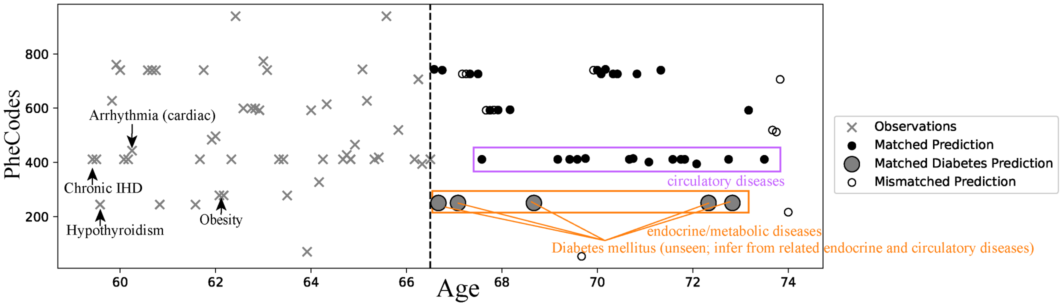
Disease Trajectory Analysis for a Diabetic Patient

- We generate a disease trajectory over irregular timestamps, where a prediction is considered correct if the top-10 predictions match the ground truth.



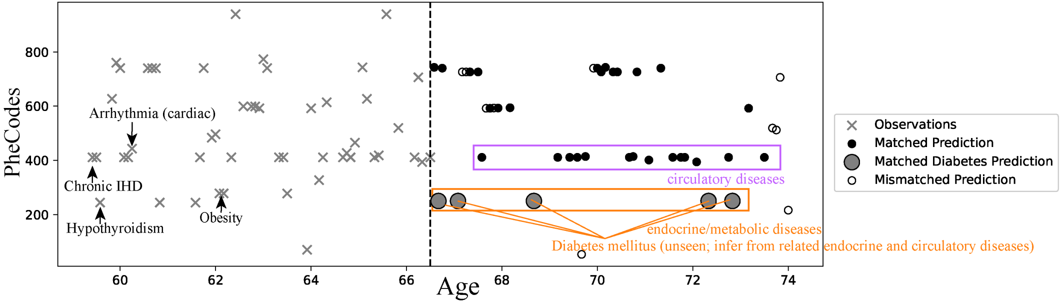
Disease Trajectory Analysis for a Diabetic Patient

- We generate a disease trajectory over irregular timestamps, where a prediction is considered correct if the top-10 predictions match the ground truth.
- Our model can predict diabetes even if it is not observed in the look-up window.



Disease Trajectory Analysis for a Diabetic Patient

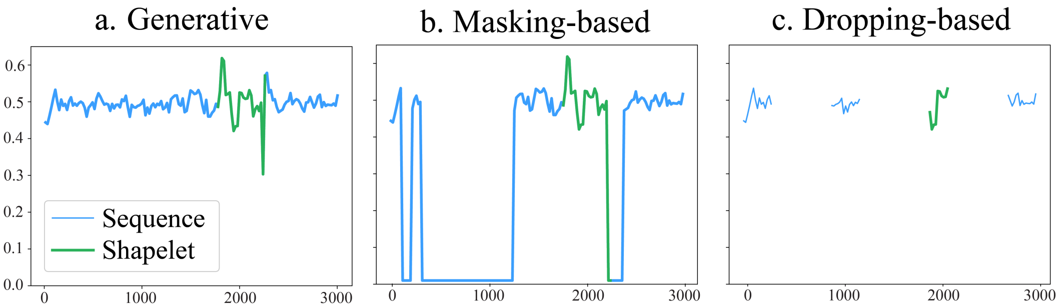
- We generate a disease trajectory over irregular timestamps, where a prediction is considered correct if the top-10 predictions match the ground truth.
- Our model can predict diabetes even if it is not observed in the look-up window.
- Correlated phenotypes within the look-up window contribute to the increased risk of diabetes.



Outline

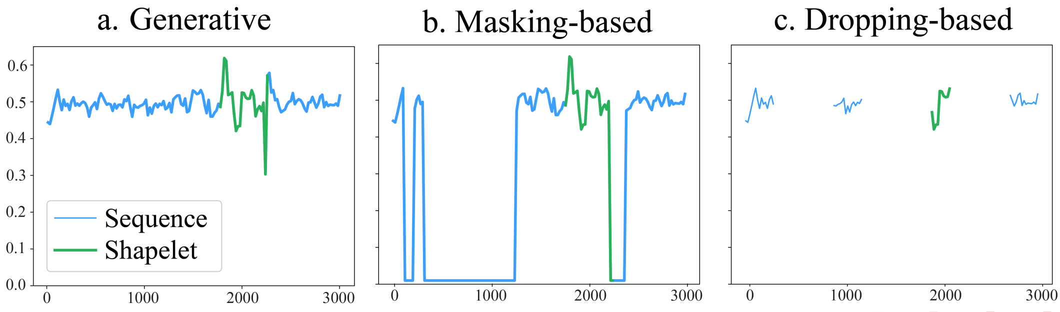
BiTimelyGPT for Discriminative Tasks

- In time-series data, only short segments (a shapelet) are crucial for prediction.



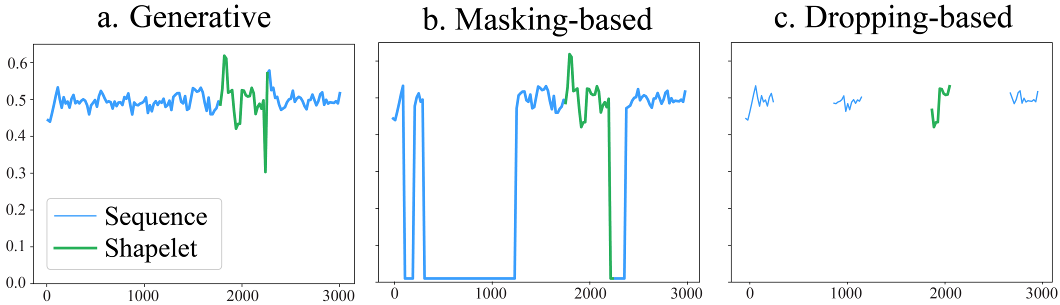
BiTimelyGPT for Discriminative Tasks

- In time-series data, only short segments (a shapelet) are crucial for prediction.
- Masking-based or dropping-based pre-training will disrupt the informative shapelet, hindering its prediction performance.



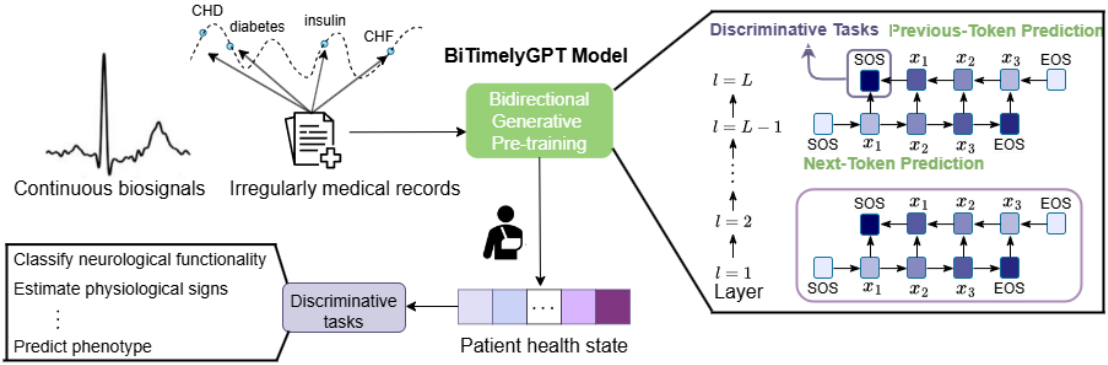
BiTimelyGPT for Discriminative Tasks

- In time-series data, only short segments (a shapelet) are crucial for prediction.
- Masking-based or dropping-based pre-training will disrupt the informative shapelet, hindering its prediction performance.
- Generative pre-training preserves the time-series shapelet, while only learn unidirectional contexts.



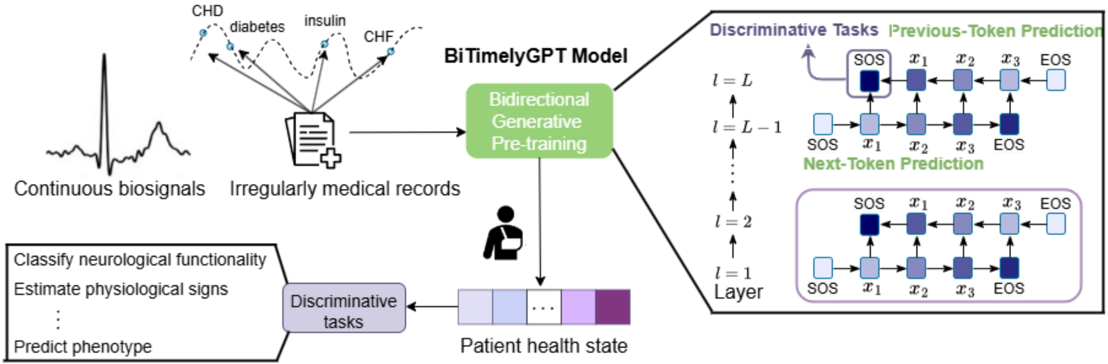
BiTimelyGPT with Bidirectional Generative Pre-training

- BiTimelyGPT alternates forward and backward attention mechanisms across layers, learning bidirectional contexts for discriminative prediction.



BiTimelyGPT with Bidirectional Generative Pre-training

- BiTimelyGPT alternates forward and backward attention mechanisms across layers, learning bidirectional contexts for discriminative prediction.
- The last two layers perform next-token prediction and previous-token prediction. This generative pre-training task preserves time-series shapelet.



Outline

Reference

- Song et al. (2024). TimelyGPT: Extrapolatable Transformer Pre-training for Long-term Time-Series Forecasting in Healthcare. ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB).
- Song et al. (2024). TrajGPT: Healthcare Time-Series Representation Learning for Trajectory Prediction. NeurIPS 2024 Workshop Time Series in the Age of Large Models (TSALM). Submitted to ICLR 2025, under review.
- Song et al. (2024). Bidirectional generative pre-training for improving healthcare time-series representation learning. Machine Learning for Healthcare (MLHC) and Proceeding of Machine Learning Research (PMLR).