Dynamic Semantic Tokenization for Time Series via Elastic Sampling on Physics-aware Perception

Huaizhang Liao[1], Zhixiong Yang[1] *, Jingyuan Xia[1] †, Yuheng Sun[1], Yue Zhang[2], Shengxi Li[2], Yongxiang Liu[1]

[1]College of Electronic Science and Technology, National University of Defense Technology
[2]College of Electronic and Information Engineering, Beihang University
{lhz17, yzx21, j.xia10, sunyuheng185}@nudt.edu.cn, yue_zhang@buaa.edu.cn, shengxili2014@gmail.com, lyx_bible@sina.com

Abstract

Despite the remarkable success of semantic token learning in NLP and vision domains, token-level representation mechanisms face fundamental challenges when extended to continuous time series analysis. We identify a core limitation lies in the intrinsic absence of semantically meaningful tokenization boundaries within time-series, which differs substantially from discrete text tokens and presents unique complexities compared to spatially coherent image patches. While existing works mechanically apply fixed-length partitioning, recent evidence from time series foundation models reveals performance ceilings in prediction tasks under such paradigms. This paper introduces a novel tokenization framework known as physics-aware tokenization (PATK), designed to implement adaptive time-frequency tokenization via distribution-sensitive sampling strategies. Key innovations include: 1) A Rate-of-Variation (RoV) distribution is meticulously structured to encompass multi-scale temporal dynamics in the time domain, alongside a Spectral Energy Intensity (SEI) distribution devised to reveal global seasonal patterns within the frequency domain; 2) A physics-aware hidden Markov modeling (PA-HMM) is then established to adaptively breaks down continuous time-series into distinct tokens with elastic lengths, responding to physics-aware probabilities sampled from RoV and SEI distributions. The proposed PATK allows steady integration with both conventional Transformers and advanced large-scale time series models (including LLM-transferred methods and pretrained time series foundation models). Simulations across various datasets demonstrate that PATK excels in classification and forecasting tasks, showing notable adaptability to model long-term dependencies, strengthening resilience against disturbances, and robustness to missing data events.

Code — https://github.com/XYLGroup/PATK

Introduction

Time series analysis is widely applied in fields such as financial analysis (Hamilton 2020; Sezer, Gudelek, and Ozbayoglu 2020; Wu et al. 2021), anomaly detection (Xu et al. 2021; Li and Jung 2023), and climate monitoring (Dimri, Ahmad, and Sharif 2020; Afrifa-Yamoah et al. 2020). The recent emergence of Transformer-based architectures (Zeng et al. 2023; Tang and Matteson 2021; Nie et al. 2022) and large foundation models (Zhou et al. 2023; Liu et al. 2025; Jin et al. 2023) has catalyzed a paradigm shift in time series processing, driving unprecedented interest in semantic-aware representation learning for time series. However, despite the remarkable success of semantic tokenization mechanisms in NLP (e.g., BERT's subword tokenization (Yang et al. 2020; Lan et al. 2019; Devlin et al. 2018; Liu et al. 1907)) and computer vision (e.g., ViT's patch embedding (Jiang, Benge, and King 2022; Li et al. 2019; Arnab et al. 2021; Han et al. 2022)), their direct extension to continuous time-series has revealed fundamental limitations. As evidenced by (Tan et al. 2024) findings, even state-of-the-art time series foundation models achieve merely marginal performance gains compared to conventional deep learning approaches in forecasting tasks. This limitation largely stems from a fundamental representational mismatch: unlike the discrete tokens with inherent semantic boundaries that underpin the success of LLMs (Yang et al. 2020; Lan et al. 2019; Devlin et al. 2018; Liu et al. 1907), continuous time-series data lacks explicit, naturally occurring segmentation points. The intrinsic semantic discontinuity of raw time series hinders the direct application of tokenization strategies successful for time series.

The core challenge in adapting large models to time series stems from a fundamental representational gap: continuous, physically-grounded temporal evolution lacks the explicit, discrete semantic units inherent to language. Contemporary approaches to bridge this gap predominantly adopt three strategies: 1) mechanical partitioning through fixed lengths (Nie et al. 2022), 2) hybrid feature engineering combining multiscale temporal patches and frequency components (Eldele et al. 2024; Bian et al. 2024), and 3) image transformation paradigms that convert time series into 2D representations (Chen et al. 2024; Li, Li, and Yan 2023). Although methods such as NHiTs (Challu et al. 2023) illustrate that multi-rate signal sampling can effectively capture intricate frequency patterns, and vision-based approaches like VisionTS (Chen et al. 2024) demonstrate the transferability

*Huaizhang Liao and Zhixiong Yang are contributed equally to this work.
†Jingyuan Xia is the corresponding author.

of image patching techniques to time series, these remain constrained by segmentation or representation schemes not inherently rooted in the physical dynamics of the data. Crucially, time series fundamentally lack the codebook-based semantic abstraction that underpins successful tokenization in language (vocabularies). A critical shortcoming across these approaches is their tendency to overlook intrinsic physical awareness, such as localized volatility, global seasonality, and event signatures, that governs time-series evolution. Rigid segmentation or artificial transformation often fails to adapt to the intrinsic spatio-temporal structures of the data due to the absence of physics-informed sampling mechanisms. This persistent limitation underscores the critical need for a novel physics-aware tokenization strategy explicitly designed for the unique properties of time series data.

We introduce the **P**hysics-**A**ware **T**o**K**enization (PATK) framework to bridge this gap, pioneering two fundamental innovations: physical distribution-guided semantic extraction and sampling-based elastic tokenization. Initially, PATK introduces a paradigm shift by leveraging the intrinsic physical dynamics to define semantically meaningful token boundaries, specifically, local volatility (governed by gradient dynamics) and prominent periodic patterns (revealed by spectral energy concentrations). Here, the Temporal Rate-of-Variation (RoV) distribution characterizes multi-scale volatility through an examination of local gradient dynamics, whereas the Spectral Energy Intensity (SEI) distribution isolates key seasonal elements using frequency-domain energy assessments. These two distributions serve as physics-informed anchors, where RoV intensities dictate token granularity in volatile regions and SEI concentrations prioritize periodic patterns. Second, a Physics-Aware Hidden Markov Model (PA-HMM) implements elastic tokenization through two neural-parameterized mechanisms: lightweight networks dynamically adjust token lengths based on RoV thresholds, while network-based samplers recalibrate segment initial points priorities by learning SEI energy intensities. By jointly optimizing the elastic-length networks (controlling token resolutions) and priority-weight networks (guiding initial points selection), PATK achieves physics-aware adaptation. This dynamic mechanism enables cross-domain generalization, automatically reconfiguring tokenization strategies for medical signals, industrial sensors, and traffic flows. Our main contributions are fourfold:

- We introduce the first tokenization framework for time series explicitly grounded in the physics of signal evolution (volatility and periodicity). It achieves elastic, meaningful, and adaptive tokenization by establishing physics-anchored partition boundaries that dynamically align with signal volatility and spectral seasonality.
- We establish dual-domain distribution models: the RoV quantifies multi-scale volatility via local gradient dynamics, while the SEI profiles frequency-domain energy concentrations to capture periodicity. These models jointly enable semantically meaningful tokenization by physics-anchored probabilities for adaptive discretization.
- We introduce a PA-HMM that consolidates RoV and SEI parameters using learnable stochastic processes, achieved through dual neural strategies concerning token lengths and initialization points. This amalgamation facilitates adaptive tokenization by progressively engaging with time-frequency across features and downstream objective semantics.
- PATK demonstrates universal compatibility with Transformer architectures and large-scale time series models (including LLM-transferred methods and pretrained time series foundation models), achieving superior performance across 15+ benchmarks in classification and forecasting tasks while showing robustness to noise injection, missing data, and long-term dependencies.

Related Work

Pre-trained Time Series Models

Existing approaches to pre-trained time series models can be categorized into three paradigms: end-to-end deep architectures, time-frequency hybrid methods, and series tokenization strategies. We critically analyze their limitations.

End-to-End Deep Architectures. Traditional models (1D-CNNs(Rizvi 2022; Tang et al. 2020)/RNNs(Hewamalage, Bergmeir, and Bandara 2021; Ma, Li, and Cottrell 2020)/Transformers(Zeng et al. 2023; Tang and Matteson 2021)) process raw time series via sequential/convolutions but lack semantic awareness. Autoformer (Wu et al. 2021) improves this modeling through fixed-scale autocorrelation yet fails to capture multi-scale dynamics, reflecting broader limitations in bridging time series with physical semantics.

Time-Frequency Methods. Existing methods (Woo et al. 2022; Zhou et al. 2022; Wang et al. 2024) fuse time-frequency features for signal enhancement but face critical limitations: NHiTs (Challu et al. 2023) employs multi-rate signal decomposition with static temporal windows insensitive to transients/seasonality, while TimeDiT's (Cao et al. 2024) spectral energy profiling remains decoupled from tokenization boundaries—passively informing features rather than dynamically guiding segmentation. These approaches underscore the necessity for physics-aware tokenization where spectral energy intensity directly governs adaptive token granularity.

Series Tokenization Strategies. Tokenization methods adapted from NLP/vision domains impose fixed windowing on time series: PatchTST (Nie et al. 2022) and its variants (Eldele et al. 2024; Bian et al. 2024) employ non-overlapping segments that suppress transient detection; Moirai-MoE (Liu et al. 2024) utilizes sparse expert mixtures with static patch resolution; TOTEM (Talukder, Yue, and Gkioxari 2024) relies on predefined codebooks misaligned with temporal dynamics. VisionTS (Chen et al. 2024) applies a time-series-to-image transformation, processing converted 2D representations with vision models. While demonstrating transferability of image patching techniques, this paradigm distorts temporal causality and obscures localized physical patterns due to non-invertible transformations. These limitations stem from rigid protocols

inherited from discrete modalities, fundamentally conflicting with time series' continuous physics-driven evolution.

Tokenization in NLP and CV

Tokenization advancements (Devlin et al. 2018; Zhang et al. 2022; Cao et al. 2023) in NLP (BERT (Devlin et al. 2018; Lan et al. 2019; Liu et al. 1907), GPT (Achiam et al. 2023; Liu et al. 2023), Llama (Touvron et al. 2023a,b)) and vision (ClusterMIM (Du, Wang, and Wang 2023)) have transitioned from static vocabulary-based segmentation to dynamic semantic clustering, enabling feature learning through self-supervised token amalgamation. This evolution liberates frameworks from manual label constraints, balancing granularity with adaptability, inspiring physics-aware tokenization strategies for time series' continuous dynamics.

Proposed Method

In this section, we demonstrate the overall architecture of our PATK, which is shown in Figure 1. We introduce in detail our physical prior distribution in the time-frequency domain, present the construction of the PA-HMM simulation, and describe the implementations of PATK on the transformer structures and large-scale time series models.

Dual-Domain Physical Prior Distribution

A statistical framework is first constructed to establish physics-aware probability distributions to quantify temporal volatility and spectral seasonality in time and frequency domains, respectively. Let $\boldsymbol{X} = [\boldsymbol{x}^1, \ldots, \boldsymbol{x}^L]$ represent the input time series, the Rate of Variation (RoV) and Spectral Energy Index (SEI) distributions are briefly introduced with workflow as follows.

Rate-of-Variation Distribution. The Rate-of-Variation (RoV) distribution probabilistically characterizes localized temporal dynamics by modeling the instantaneous variation rate of time series through differential operators. Formally, given a time-series $\boldsymbol{X} \sim \mathcal{P}$ sampled from the unknown distribution $\mathcal{P}$, we define the RoV distribution $\mathcal{P}_{\mathrm{R}}(\boldsymbol{X}_\Delta|\boldsymbol{X})$ as

$$\mathcal{P}_{\mathrm{R}}(\boldsymbol{X}_\Delta|\boldsymbol{X}) = \mathcal{R}(\lim_{\Delta t \to 0} \frac{\boldsymbol{x}^{l+\Delta t} - \boldsymbol{x}^l}{\Delta t}, l), \tag{1}$$

where $\mathcal{R}$ denotes the distribution with the parameters of variation and the timestamp l, $\boldsymbol{X}_\Delta = [\Delta\boldsymbol{x}^1, ..., \Delta\boldsymbol{x}^L]$ denotes the RoV intensity vector aligned with $\boldsymbol{X}$'s time-dimension, and $\Delta\boldsymbol{x}^l$ is the rate of variation at t^{th} timestamp. This establishes a probabilistic mapping of temporal volatility. The resulting distribution provides region-aware variational weights for adaptive tokenization, where segments are discretized proportionally to their local RoV intensities.

Spectral Energy Intensity Distribution. The Spectral Energy Intensity (SEI) distribution characterizes global seasonal patterns, revealing dominant periodic behaviors through energy proportion analysis in the frequency domain. Let $F(\omega)$ denote the frequency series obtained via the Fourier transform. Then, based on the Parseval theorem (Kelkar, Grigsby, and Langsner 2007), the SEI distribution $\mathcal{P}_{\mathrm{S}}(F(\omega)|\boldsymbol{X})$ can be further defined as

$$\mathcal{P}_{\mathrm{S}}(F(\omega)|\boldsymbol{X}) = \mathcal{S}(\|F(\omega^k)\|_2^2, k), \tag{2}$$

where $F(\omega^k)$ is the $k-$th frequency component, and $\|\cdot\|$ is the modulo operation, and $\mathcal{S}$ is the distribution with the parameter of the energy intensity and the index of frequency component k. The SEI distribution offers frequency-aware seasonal weights to guide adaptive tokenization, dynamically prioritizing tokens aligned with spectral energy concentrations that encapsulate global periodic patterns.

The established RoV and SEI distributions serve as physics-grounded anchors that explicitly encode multi-scale temporal volatility and spectral seasonality patterns. These dual-domain priors first provide explicit physical domain priors for the subsequent sampling-based tokenization process. More descriptions are given in the Supplementary A.1.

The Physics-Aware Sampling

Building upon the foundational statistical distributions, we propose a physics-aware adaptive sampling mechanism that synergistically integrates time-frequency domain characteristics with tokenization operators, which adaptively learns to select elastic tokens concerning their ability to reveal volatility and seasonality while maintaining semantic-constrained sampling prioritization governed by the downstream task objectives. This process is constructed through a hidden Markov modeling, named PA-HMM, which is composed of physics-aware initialization sampling, length adaptation and token significance re-weighting. The details are demonstrated as below.

Physics-aware Initialization Sampling. Based on the modeled RoV and SEI distribution in the time and the frequency domains, the PA-HMM first randomly sample the initialization point with a moderate length from the time-series. Let $t = 1, \cdots, T$ represent the training iteration with maximum number T, $c = 1, \cdots, C$ represent the sampling index with maximum sampling times C, and $\mathcal{I} = \{\mathrm{R}, \mathrm{S}\}$ be the set of subscripts to simplify the descriptions, the sampling process is defined as follows:

$$p(\boldsymbol{z}_{\mathrm{d}}^{t,c}|\theta_{\mathrm{d}}^{t,c}) \propto \mathcal{P}_{\mathrm{d}}, \mathrm{d} \in \mathcal{I}, \tag{3}$$

where $\boldsymbol{z}_{\mathrm{d}}^{t,c} = [\boldsymbol{x}_{\mathrm{d}}^{\theta_{\mathrm{d}}^{t,c}}, ..., \boldsymbol{x}_{\mathrm{d}}^{\theta_{\mathrm{d}}^{t,c}+h}]$ is the sampled segments with fixed length h at c^{th} sampling process, the initial position $\theta_{\mathrm{d}}^{t,c}$ is sampled from the corresponding physical distribution $\mathcal{P}_{\mathrm{d}}$, and $p(\boldsymbol{z}_{\mathrm{d}}^{t,c}|\theta_{\mathrm{d}}^{t,c})$ indicates the sampled slices $\boldsymbol{z}_{\mathrm{d}}^{t,c}$ are determined by the initial position $\theta_{\mathrm{d}}^{t,c}$. Here we note that the probability of sampling the initialization $\theta_{\mathrm{d}}^{t,c}$ is directly determined by the corresponding probabilities of t-th index on the distributions $\mathcal{P}_{\mathrm{R}}$ and $\mathcal{P}_{\mathrm{S}}$. Here we note that the initialization probability satisfies:

$$p(\theta_{\mathrm{d}}^{t,c}) \propto \underbrace{\sum_{i=l}^{l+h} (\Delta\boldsymbol{x}^i)^2}_{\text{temporal volatility}} \cdot \underbrace{\int_{\omega^k-\epsilon}^{\omega^k+\epsilon} |F(\omega)|^2 d\omega}_{\text{spectral energy}}, \tag{4}$$

where $p(\theta_{\mathrm{d}}^{t,c})$ is the sampling probability, $\Delta\boldsymbol{x}^i = (\boldsymbol{x}^{i+\Delta t} - \boldsymbol{x}^i)/\Delta t$, ϵ is the bandwidth radius with the center frequency ω^k. This mechanism essentially implements the empirical principle that both sharp temporal fluctuations and spectrally

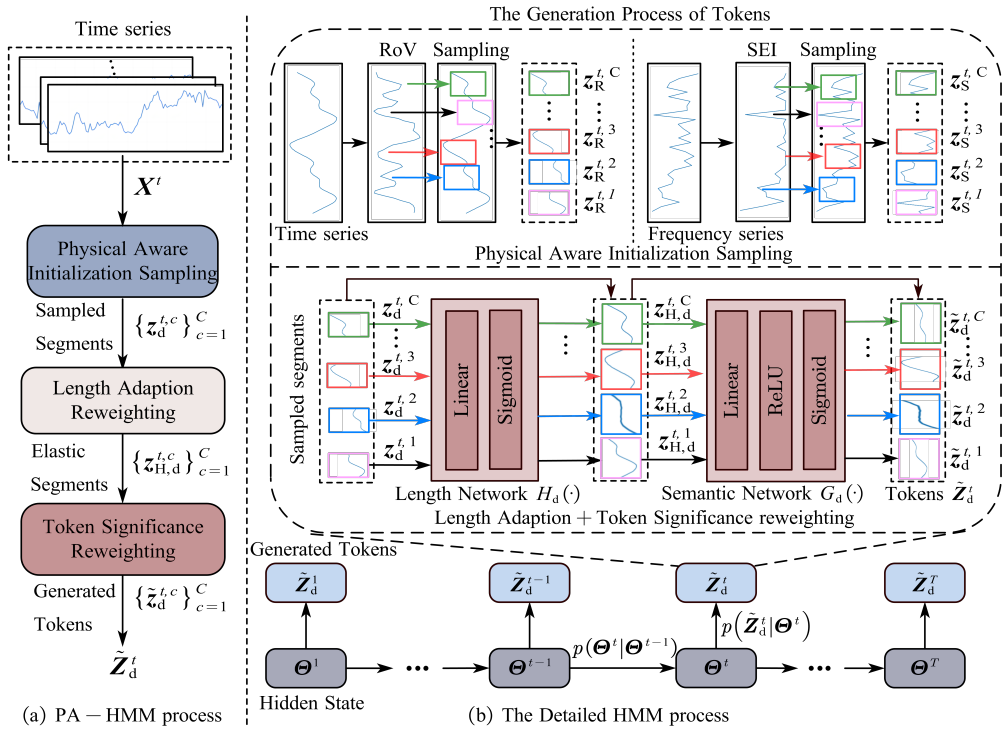

Figure 1: The overall framework of PATK.

prominent components inherently encompass richer semantic content, ensuring physically-grounded semantic fidelity in time-series tokenization processes.

Length Adaptation Re-weighting. The discretely sampled segments $\boldsymbol{z}_{\mathrm{d}}^{t,c}$ undergo elastic length reweighting through a learnable length adaptation operator $H_{\mathrm{d}}(\cdot;\boldsymbol{\theta}_{H_{\mathrm{d}}}^{t})$ denoted by

$$p(\boldsymbol{z}_{\mathrm{H,d}}^{t,c}|\boldsymbol{z}_{\mathrm{d}}^{t,c},\boldsymbol{\theta}_{H_{\mathrm{d}}}^{t}) \propto \boldsymbol{z}_{\mathrm{H,d}}^{t,c} = \mathcal{E}(h_{\mathrm{d}}^{t,c},\boldsymbol{z}_{\mathrm{d}}^{t,c}), \mathrm{d} \in \mathcal{I}, \quad (5)$$

where $\boldsymbol{z}_{\mathrm{H,d}}^{t,c} = [\boldsymbol{x}_{\mathrm{d}}^{\theta_{\mathrm{d}}^{t,c}}, ..., \boldsymbol{x}_{d}^{\theta_{\mathrm{d}}^{t,c}+h_{\mathrm{d}}^{t,c}}]$ is the resampled segment with re-weighted length $h_{\mathrm{d}}^{t,c} = hH_{\mathrm{d}}(\boldsymbol{z}_{\mathrm{d}}^{t,c};\boldsymbol{\theta}_{H_{\mathrm{d}}}^{t})$, $\mathcal{E}$ is the resample operator with linear interpolation. The parameterized posterior distribution $p(\boldsymbol{z}_{\mathrm{H,d}}^{t,c}|\boldsymbol{z}_{\mathrm{d}}^{t,c},\boldsymbol{\theta}_{H_{\mathrm{d}}}^{t})$ explicitly encodes the dependency between resampling operations and network parameters $\boldsymbol{\theta}_{H_{\mathrm{d}}}^{t}$, which are ultimately optimized through downstream losses. Specifically, the physics-aware loss will serve as the critical regularizer, enforcing two essential consistency conditions: i) temporal saliency preservation by penalizing least local volatility intensity, and ii) spectral fidelity maintenance by preserving dominant frequency energy.

Token Significance Re-weighting. Then, a secondary weighting operating on the physics-prioritized tokens $\boldsymbol{z}_{\mathrm{H,d}}^{t,c}$ from the length adaptation stage is employed via a learnable semantic saliency evaluator $G_{\mathrm{d}}(\cdot;\boldsymbol{\theta}_{G_{\mathrm{d}}}^{t})$ as follows

$$p(\tilde{\boldsymbol{z}}_{\mathrm{d}}^{t,c}|\boldsymbol{z}_{\mathrm{H,d}}^{t,c},\boldsymbol{\theta}_{G_{\mathrm{d}}}^{t}) \propto \tilde{\boldsymbol{z}}_{\mathrm{d}}^{t,c} = G_{\mathrm{d}}(\boldsymbol{z}_{\mathrm{H,d}}^{t,c};\boldsymbol{\theta}_{G_{\mathrm{d}}}^{t})\boldsymbol{z}_{\mathrm{H,d}}^{t,c}, \mathrm{d} \in \mathcal{I}, \quad (6)$$

where $\tilde{\boldsymbol{z}}_{\mathrm{d}}^{t,c}$ denotes the re-weighted token at the c-th sampling index. The parameters $\boldsymbol{\theta}_{G_{\mathrm{d}}}^{t}$ are discriminatively trained to amplify tokens whose feature representations maximally reduce the downstream task loss referring to a contrastive learning loss which will be illustrated soon.

Hidden Markov Modeling

The above three operators with respect to temporal volatility patterns, spectral dominance characteristics, and task-specific semantics, are dynamically coupled by a hidden Markov framework. Within the PA-HMM framework, the hidden state variables $\boldsymbol{\Theta}_{\mathrm{d}}^{t} = \{\{\theta_{\mathrm{d}}^{t,c}\}_{c=1}^{C}, \boldsymbol{\theta}_{H_{\mathrm{d}}}^{t}, \boldsymbol{\theta}_{G_{\mathrm{d}}}^{t}\}$ aggregate the physics-aware initialization, length adaptation parameters, and semantic weighting parameters as defined in Eq. 3 to 6, while the corresponding observed states $\tilde{\boldsymbol{Z}}_{\mathrm{d}}^{t} = [\tilde{\boldsymbol{z}}_{\mathrm{d}}^{t,1}, ..., \tilde{\boldsymbol{z}}_{\mathrm{d}}^{t,C}], \mathrm{d} \in \mathcal{I}$ represent the optimized tokens generated through the coupled time-frequency-semantic operations. Thus, the probability of the emission process is formulated as

$$p(\tilde{\boldsymbol{Z}}_{\mathrm{d}}^{t}|\boldsymbol{\Theta}_{\mathrm{d}}^{t}) = \underbrace{p(\tilde{\boldsymbol{z}}_{\mathrm{d}}^{t,c}|\boldsymbol{z}_{\mathrm{H,d}}^{t,c},\boldsymbol{\theta}_{G_{\mathrm{d}}}^{t})}_{\text{token significance}} \underbrace{p(\boldsymbol{z}_{\mathrm{H,d}}^{t,c}|\boldsymbol{z}_{\mathrm{d}}^{t,c},\boldsymbol{\theta}_{H_{\mathrm{d}}}^{t})}_{\text{length adaption}} \underbrace{p(\boldsymbol{z}_{\mathrm{d}}^{t,c}|\theta_{\mathrm{d}}^{t,c})}_{\text{PA sampling}}, \mathrm{d} \in \mathcal{I}. \quad (7)$$

The $p(\tilde{\boldsymbol{Z}}_{\mathrm{d}}^{t}|\boldsymbol{\Theta}_{\mathrm{d}}^{t})$ indicates that the output of PA-HMM is determined by the hidden states within tokenization process. Thus, the Markov Chain between the outputs $\tilde{\boldsymbol{Z}}_{\mathrm{d}}^{t}$ and $\tilde{\boldsymbol{Z}}_{\mathrm{d}}^{t+1}$ can be further denoted as

$$p(\boldsymbol{\Theta}_{\mathrm{d}}^{t+1}|\boldsymbol{\Theta}_{\mathrm{d}}^{t},\tilde{\boldsymbol{Z}}_{\mathrm{d}}^{t}) = p(\boldsymbol{\Theta}_{\mathrm{d}}^{t+1}|\boldsymbol{\Theta}_{\mathrm{d}}^{t},\mathcal{L}_{\mathrm{total}}^{t})p(\mathcal{L}_{\mathrm{total}}^{t}|\tilde{\boldsymbol{Z}}_{\mathrm{d}}^{t}), \mathrm{d} \in \mathcal{I}, \quad (8)$$

Scenarios	Metrics	LaST	TST	TS2Vec	Ti-MAE	PatchTST	TSLANet	ALLMT	SimMTM	VisionTS	**PATK**
Epilepsy	Acc	79.40	80.21	89.40	89.71	78.81	90.11	78.20	95.66	95.71	**95.75**
	Pre	79.20	80.11	90.39	78.36	83.21	93.21	72.39	94.63	94.70	**94.88**
	Re	72.34	80.00	90.21	67.45	80.00	80.21	90.21	**91.43**	90.66	91.39
	F1	76.88	78.51	86.88	68.66	78.51	80.21	91.35	**92.92**	92.36	92.50
Gesture	Acc	64.17	72.17	69.01	76.88	67.37	70.51	68.51	78.33	78.23	**78.53**
	Pre	70.36	70.60	65.42	70.35	67.40	77.31	66.72	78.26	78.47	**79.34**
	Re	66.17	69.17	68.54	76.75	69.17	80.21	68.54	78.33	78.46	**78.53**
	F1	63.76	68.01	65.70	74.29	68.01	80.21	65.70	76.47	77.66	**78.51**
EMG	Acc	76.34	78.34	78.54	89.99	84.84	86.11	90.54	97.56	97.69	**98.32**
	Pre	76.37	77.11	80.40	80.65	76.41	76.41	89.60	93.33	94.26	**95.33**
	Re	73.33	80.30	67.85	85.59	80.30	80.21	89.85	94.04	94.87	**95.34**
	F1	72.75	68.89	67.66	86.83	68.89	80.21	89.66	95.14	93.47	**96.34**
ECG	Acc	64.17	72.17	69.01	76.80	53.17	54.21	54.81	78.33	76.38	**79.53**
	Pre	70.36	70.60	65.42	70.35	31.11	51.51	57.92	78.57	76.24	**79.34**
	Re	66.17	69.17	68.54	76.75	69.17	80.21	68.54	78.33	76.29	**78.53**
	F1	63.76	68.01	65.70	74.29	68.01	80.21	65.70	76.47	76.76	**78.51**
FD-A	Acc	40.00	52.33	43.67	60.00	69.82	75.41	48.67	63.11	65.49	**79.50**
	Pre	36.34	58.02	39.83	59.37	71.22	77.91	23.73	67.82	67.26	**78.87**
	Re	35.59	56.82	38.94	60.72	56.82	80.21	38.94	61.30	60.66	**81.40**
	F1	30.45	55.39	37.29	63.59	55.39	80.21	37.29	63.70	64.16	**80.00**
FD-B	Acc	40.21	56.40	47.90	60.38	79.90	87.71	78.93	69.40	70.16	**88.45**
	Pre	39.89	51.58	43.39	61.59	78.78	87.32	78.39	86.99	74.39	**88.99**
	Re	34.55	54.50	48.42	63.02	78.50	86.21	78.42	76.41	75.83	**88.13**
	F1	36.92	59.34	43.89	64.87	78.34	87.21	78.89	75.11	74.62	**88.73**

Table 1: The comparison results with different lengths.

where $p(\mathbf{\Theta}_{\mathrm{d}}^{t+1}|\mathbf{\Theta}_{\mathrm{d}}^{t}, \mathcal{L}_{\text{total}}^{t})$ denotes the parameter update for the networks, $p(\mathcal{L}_{\text{total}}^{t}|\tilde{\boldsymbol{Z}}_{\mathrm{d}}^{t})$ represents the loss $\mathcal{L}_{\text{total}}^{t}$ calculation based on the obtained tokens $\tilde{\boldsymbol{Z}}_{\mathrm{d}}^{t}$. Consequently, the PA-HMM will converge towards the desired distributions of weights, which correlate with segments that exhibit superior feature representation ability, pertaining to the semantic tokens in sequence.

By encoding these cross-domain interactions within latent state transitions, this PA-HMM enables reciprocal modulation between time-varying feature sampling and frequency-selective token prioritization, while adaptively assimilating downstream semantic objectives into emission probabilities. This tripartite co-optimization mechanism inherently balances physical interpretability grounded in time-series variations with task-driven representational efficacy, creating an emergent equilibrium where temporal-spectral saliency and semantic relevance mutually reinforce through Markovian state evolution.

Moreover, this PA-HMM can be seamlessly incorporated to mainstream Transformers (Oh, Wang, and Wiens 2018; Devlin et al. 2018) and large-scale time series models (Zhou et al. 2023; Liu et al. 2025; Jin et al. 2023; Woo et al. 2024; Shi et al. 2024), enhancing their capability to capture multi-scale semantic patterns in time-series.

Pre-training Task of PATK

We introduce a token-based contrastive learning strategy designed for our PA-HMM. In the t^{th} training iteration within a batch comprising B series, the tokens produced are input into the feature extractor network $F(\cdot;\boldsymbol{\theta}_F^t)$ with parameters $\boldsymbol{\theta}_F^t$. This network may utilize either a transformer or other established large-scale models. The process of feature extraction is described as

$$\boldsymbol{r}^{t,n} = F(\tilde{\boldsymbol{Z}}_{\mathrm{R}}^{t}, \tilde{\boldsymbol{Z}}_{\mathrm{S}}^{t}, \mathbf{\Theta}^{t}; \boldsymbol{\theta}_F^{t}), n = 1, \cdots, C \times B, \quad (9)$$

where $\mathbf{\Theta}^t = \{\mathbf{\Theta}_{\mathrm{R}}^t, \mathbf{\Theta}_{\mathrm{S}}^t\}$ and $\boldsymbol{r}^{t,n}$ denotes the extracted features. To simplify the description, let $\mathcal{D}_{\boldsymbol{r}}^t = \{\boldsymbol{r}^{t,n}\}_{n=1}^{C\times B}$ denote the set of projected embedded features, which are fed to a contrastive learning phase. In particular, we generate positive and negative pairs at the token level, where tokens originating from the same series are identified as positives, and those from different series are regarded as negatives. Then, the token contrastive loss $\mathcal{L}_{\text{TCL}}^t$ can be formulated as

$$\mathcal{L}_{\text{TCL}}^{t} = -\sum_{\boldsymbol{r}^{t,n}\in\mathcal{D}_{\boldsymbol{r}}^{t}} \frac{1}{C-1} \sum_{\boldsymbol{r}^{t,i}\in\mathcal{P}_{\boldsymbol{r}}^{+}(\boldsymbol{r}^{t,n})} \log \frac{\exp(\boldsymbol{r}^{t,n}\cdot\boldsymbol{r}^{t,i}/\rho)}{\sum_{\boldsymbol{r}^{t,a}\in\mathcal{A}_{\boldsymbol{r}}(\boldsymbol{r}^{t,n})}\exp(\boldsymbol{r}^{t,n}\cdot\boldsymbol{r}^{t,a}/\rho)}, \quad (10)$$

where $\boldsymbol{r}^{t,n}$ is the k^{th} anchor, and $\mathcal{A}_{\boldsymbol{r}}(\boldsymbol{r}^{t,n}) = \mathcal{D}_{\boldsymbol{r}}^t/\{\boldsymbol{r}^{t,n}\}$ represents the subset of $\mathcal{D}_{\boldsymbol{r}}^t$ excluding $\boldsymbol{r}^{t,n}$. $\boldsymbol{r}^{t,i}$ are the positives. $\mathcal{P}_{\boldsymbol{r}}^{+}(\boldsymbol{r}^{t,n})$ is the set of all positives. ρ is a scalar temperature parameter, and $\cdot$ denotes the dot product.

Furthermore, the physics-aware loss $\mathcal{L}_{\text{PA}}^t$ is formulated as:

$$\mathcal{L}_{\text{PA}}^{t} = -\sum_{c=1}^{C}\left[\sum_{i=\theta_{\mathrm{R}}^{t,c}}^{\theta_{\mathrm{R}}^{t,c}+h_{\mathrm{R}}^{t,c}} (\Delta \boldsymbol{z}_{\mathrm{H,R}}^{t,c})^2 \cdot \int_{\theta_{\mathrm{S}}^{t,c}}^{\theta_{\mathrm{S}}^{t,c}+h_{\mathrm{S}}^{t,c}} |F_{\boldsymbol{z}_{\mathrm{H,S}}^{t,c}}(\omega)|^2 d\omega\right], \quad (11)$$

where $\sum_{i=\theta_{\mathrm{R}}^{t,c}}^{\theta_{\mathrm{R}}^{t,c}+h_{\mathrm{R}}^{t,c}} (\Delta \boldsymbol{z}_{\mathrm{H,R}}^{t,c})^2$ is the high cumulative for token $\boldsymbol{z}_{\mathrm{H,R}}^{t,c}$, and the $\int_{\theta_{\mathrm{S}}^{t,c}}^{\theta_{\mathrm{S}}^{t,c}+h_{\mathrm{S}}^{t,c}} |F_{\boldsymbol{z}_{\mathrm{H,S}}^{t,c}}(\omega)|^2 d\omega$ is concentrated en-

Dataset	Len	LaST	TST	TS2Vec	Ti-MAE	PatchTST	TSLANet	ALLMT	SimMTM	VisionTS	**PATK**
ETTh1	96	0.3845	0.4011	0.4368	0.3997	0.3754	0.3804	0.3831	0.3725	**0.3715**	0.3792
	192	0.3867	0.5312	0.4556	0.4542	0.4032	0.3962	0.3962	0.4149	**0.3805**	0.3861
	336	0.4431	0.4744	0.6897	0.4977	0.4223	0.4193	0.4273	0.4296	0.4155	**0.4133**
	720	0.4872	0.4718	0.4891	0.5151	0.4439	0.4634	0.4478	0.4468	0.4419	**0.4402**
ETTm1	96	0.3238	0.3386	0.6793	0.3236	0.2934	**0.3021**	0.3248	0.3227	0.3125	0.3297
	192	0.3492	0.3945	0.6731	0.3705	0.3427	0.3342	**0.3313**	0.3326	0.3335	0.3331
	336	0.3921	0.4017	0.7033	0.3977	0.3677	0.3671	0.3654	0.3941	0.3576	**0.3501**
	720	0.4011	0.4346	0.7224	0.4426	0.4144	0.4041	0.4091	0.4111	0.4004	**0.3983**
ETTm2	96	0.4001	0.4011	0.4368	0.3997	0.3728	0.3742	0.3775	0.3735	**0.3715**	0.3730
	192	0.3978	0.5312	0.4556	0.4542	0.3823	0.3892	0.3847	0.4149	0.3814	**0.3761**
	336	0.4231	0.4744	0.6897	0.4977	0.4248	0.4244	0.4183	0.4293	0.4239	**0.4183**
	720	0.4593	0.4718	0.4891	0.5151	0.4638	0.4631	0.4528	0.4468	0.4525	**0.4402**
Weather	96	0.2024	0.2921	0.4338	0.2167	0.1892	0.1708	0.1794	0.1715	0.1725	**0.1692**
	192	0.2258	0.4104	0.5086	0.2542	0.2245	0.1992	0.2015	0.1991	0.1994	**0.1981**
	336	0.2542	0.4347	0.5457	0.2907	0.2453	0.2483	0.2454	0.2463	0.2471	**0.2443**
	720	0.3448	0.5392	0.5761	0.3381	0.3192	0.3203	0.3208	0.3182	0.3165	**0.3142**
Electricity	96	0.2043	0.2926	0.3223	0.2796	0.1493	0.1472	0.1418	0.1337	0.1382	**0.1346**
	192	0.2746	0.2705	0.3431	0.2855	0.1674	0.1454	0.1498	0.1476	0.1461	**0.1431**
	336	0.2894	0.3347	0.3623	0.3017	0.1832	0.1632	0.1679	0.1661	0.1673	**0.1601**
	720	0.2942	0.3446	0.3884	0.3336	0.2143	0.2093	0.2093	0.2031	0.2045	**0.1983**
Traffic	96	0.2926	0.3223	0.2802	0.3621	0.3451	0.3346	0.3337	0.2746	0.2725	**0.2649**
	192	0.2705	0.3431	0.2933	0.3678	0.2934	0.3523	0.3476	0.2831	0.2625	**0.2418**
	336	0.3347	0.3623	0.3075	0.3971	0.3932	0.3842	0.3661	0.3001	0.2825	**0.2761**
	720	0.3446	0.3884	0.3309	0.4392	0.4331	0.4042	0.4031	0.3283	0.3225	**0.3057**

Table 2: The forecasting results in MSE metrics.

ergy of token $\boldsymbol{z}_{\mathrm{H,S}}^{t,c}$. The total loss is formulated as

$$\mathcal{L}_{\text{total}}^{t} = \mathcal{L}_{\text{TCL}}^{t} + \lambda_{\text{PA}} \mathcal{L}_{\text{PA}}^{t}, \tag{12}$$

where λ_{PA} is weight parameter to balance the concentrations between physics-significance and task-specific significance.

Simulation and Analysis

Experimental Settings

Datasets. Fifteen time series datasets evaluate the PATK framework for both classification and forecasting tasks. Specifically, following the setting in (Dong et al. 2024), the SleepEEG, Epilepsy, HAR, Gesture, ECG, EMG, FD-A, and FD-B datasets are utilized for classification tasks. For the forecasting tasks, the ETT 4 subsets, Weather, Electricity, and Traffic datasets are applied for forecasting tasks.

Implementation Details. In PATK, PA-HMM uses a one-layer FCN as the network $H_{\text{d}}(\cdot)$ and $G_{\text{d}}(\cdot)$. The physical distributions $\mathcal{R}$ and $\mathcal{S}$ are uniform distribution. The temperature scale ρ is 0.5, and the optimizer in the pre-training process is applied with Adadelta (Zeiler 2012) with a learning rate $\gamma = 8 \times 10^{-3}$. The weight parameter λ_{PA} is set to 1. For classification tasks, all methods follow standard protocols (Dong et al. 2024; Li et al. 2023): pre-trained on SleepEEG (178-length), then evaluated on Epilepsy (200), HAR (206), Gesture (315), EMG/ECG (1,500), and FD-A/B (5,120) under variable-length (1.5K-5K), noise (10-40dB), and missing data (10-40%) conditions. Forecasting models use ETTh2 pre-training with multi-horizon fine-tuning (96/192/336/720). Hyperparameters (sampling count C, token length h) are optimized in Supplementary A.4.7; we adopt $C = 30$, $h = 40$ for all experiments.

Comparing Methods. We evaluated nine time series methods: LaST (Wang et al. 2022) (seasonal-trend decomposition via FNN), TS2Vec (Yue et al. 2022) (1D-CNN feature extractors), TST (Zerveas et al. 2021) and Ti-MAE (Li et al. 2023) (transformer-based reconstruction), PatchTST/TSLANet/ALLMT (Nie et al. 2022; Eldele et al. 2024; Bian et al. 2024) (tokenization approaches), SimMTM (Dong et al. 2024) (1D-CNN for classification tasks, transformer-based structure for forecasting) and VisionTS (signal-to-image) (Chen et al. 2024), along with three LLM-transferred models (OFA (Zhou et al. 2023), CALF (Liu et al. 2025), Time-LLM(Jin et al. 2023)) and two time series foundation models Time-MoE (Shi et al. 2024) and Moirai (Woo et al. 2024) to validate PATK's effectiveness.

Experimental Designs. Our experiments comprise six integral components, designed to holistically assess performance, robustness, and generalizability: (1) Performance comparison with deep learning methods, (2) Performance comparison with large-scale model methods, (3) Ablation studies, (4) Semantic weight effectiveness verification, (5) Computational complexity analysis, (6) Parameter sensitivity experiments. Due to page limitations, the main text focuses on high-impact results from the first three experimental components (1-3), utilizing representative datasets to highlight key advancements. The complete dataset results (e.g., full metrics across 15+ benchmarks) and detailed outcomes for the latter three experiments (4–6) are comprehensively documented in the Supplementary A.4.

Classification Results of Deep-Learning Methods

We benchmark PATK's robustness against SOTA models under three challenging conditions: length variations, noise

Methods	OFA		CALF		Time-LLM		Moirai		Time-MoE	
	Ori	**Ours**	Ori	**Ours**	Ori	**Ours**	Ori	**Ours**	Ori	**Ours**
Epilepsy	74.13	**84.12**	75.12	**78.34**	72.38	**80.68**	74.53	**84.78**	73.49	**83.26**
HAR	62.82	**72.31**	64.52	**68.71**	63.29	**65.46**	65.82	**74.32**	64.94	**73.12**
Gesture	48.65	**50.21**	50.11	**52.21**	49.28	**51.11**	50.23	**54.83**	49.92	**53.28**
EMG	70.74	**86.25**	73.12	**85.33**	72.34	**74.56**	74.33	**85.65**	73.37	**84.63**
ECG	68.45	**71.34**	65.58	**69.34**	69.28	**70.34**	70.43	**74.86**	69.58	**72.86**
FD-A	54.98	**68.79**	59.82	**67.45**	56.37	**66.36**	58.78	**67.38**	57.83	**65.23**
FD-B	41.43	**47.38**	42.35	**45.38**	42.35	**44.68**	44.45	**49.68**	43.58	**48.92**

Table 3: The results of large-scale models on classification tasks.

Methods	OFA		CALF		Time-LLM		Moirai		Time-MoE	
	Ori	**Ours**	Ori	**Ours**	Ori	**Ours**	Ori	**Ours**	Ori	**Ours**
ETTh1	0.4251	**0.4147**	0.4311	**0.4231**	0.4324	**0.4275**	0.4024	**0.3945**	0.4012	**0.3972**
ETTm1	0.3924	**0.3837**	0.3802	**0.3711**	0.3627	**0.3561**	0.4482	**0.4345**	0.3948	**0.3855**
ETTm2	0.3785	**0.3681**	0.3937	**0.3831**	0.3698	**0.3608**	0.3124	**0.3064**	0.3172	**0.3028**
Weather	0.2583	**0.2431**	0.2765	**0.2684**	0.2446	**0.2384**	0.2434	**0.2365**	0.2654	**0.2558**
Electricity	0.2634	**0.2539**	0.2804	**0.2733**	0.1646	**0.1589**	0.2354	**0.2265**	0.2478	**0.2365**
Traffic	0.2658	**0.2593**	0.2672	**0.2584**	0.4231	**0.4170**	0.2537	**0.2441**	0.2531	**0.2475**

Table 4: The results of large-scale models on classification tasks.

corruption, and missing data. The results of noise corruption and missing data are given in Supplementary A.4.3-A.4.4.

Comparison results of varying lengths. The analysis of time series across diverse lengths is indicated in Table 1. For standard-length series (200, 206, 315), PATK matches or exceeds SOTA methods in classification accuracy. As the length of the test series extends to 1500, the benefits of the PATK model become more pronounced. In instances of exceedingly long series (5,120), while conventional approaches experience a notable decline in efficacy and typically fail in classification, the PATK model maintains satisfactory performance and exhibits significant advantages. These findings underscore the robustness of the PATK framework to variations in series length, attributed to its elastic length generalization capabilities.

Forecasting Results of Deep-Learning Methods

Table 2 validates PATK's universal applicability in forecasting tasks. On standard-length sequences, PATK matches SOTA methods. While the sequence length extends to long-term lengths, our PATK method demonstrates a substantial enhancement over the SimMTM approach, achieving improvements of 0.0163 on the ETTh1 through multi-scale tokenization, outperforming fixed-window approaches.

Results of Large-Scale Time Series Models

We validate PA-HMM's generalizability by integrating physics-aware tokens into three SOTA time series foundation models (TSFMs). As shown in Table 3, PA-HMM-enhanced variants achieve universal classification improvements, notably accuracy gains on Epilepsy and EMG datasets. Table 4 further demonstrates forecasting enhancements: OFA reduces Electricity MSE by 0.95%, while CALF and Time-LLM improve Traffic and Weather MSE by 0.88% and 0.62% respectively, validating physics-driven tokenization across temporal semantics.

Ablation Studies for PATK

Table 5 shows the results, including the physics-awareness sampling (PA), length adaptation (LA) re-weighting, the token significance (TS) re-weighting, and physics-awareness loss (PAL). Removing PA divides the input time series into uniform segments, while removing LA and TS directly applies the feature learning across the original samples. It is clear that the four parts are crucial for our performance.

Scenarios				Epilepsy	EMG	FD-B
PA	LA	TS	PAL	Acc	Acc	Acc
-	-	-	-	79.13	76.14	55.23
✓	-	-	-	85.35	80.38	60.03
✓	✓	-	-	91.27	93.35	79.32
✓	✓	✓	-	95.25	97.87	83.55
✓	✓	✓	✓	**95.75**	**98.32**	**84.45**

Table 5: The classification results for ablation studies.

Conclusion

This paper introduces PATK, a physics-aware tokenization framework that bridges the semantic-physical gap via dual physics priors: RoV encoding multi-scale temporal volatility and SEI capturing global periodicity. The PA-HMM component dynamically aligns token boundaries with intrinsic signal dynamics through neural-parameterized elastic segmentation, enabling adaptive responses to localized volatility and spectral shifts. Evaluated across 15+ benchmarks, PATK demonstrates universal compatibility and superior performance in both transformer and large-scale models on classification and forecasting tasks, resolving temporal representation decoupling through physics-grounded tokenization.

Acknowledgments

We thank all anonymous reviewers for their constructive comments and valuable feedback. This work is supported by the National Natural Science Foundation of China under Grant 62576350, 62131020, 62376283 and 62531026.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Afrifa-Yamoah, E.; Mueller, U. A.; Taylor, S.; and Fisher, A. 2020. Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications*, 27(1): e1873.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Bian, Y.; Ju, X.; Li, J.; Xu, Z.; Cheng, D.; and Xu, Q. 2024. Multi-patch prediction: Adapting llms for time series representation learning. *arXiv preprint arXiv:2402.04852*.

Cao, D.; Ye, W.; Zhang, Y.; and Liu, Y. 2024. Timedit: General-purpose diffusion transformers for time series foundation model. *arXiv preprint arXiv:2409.02322*.

Cao, S.; Yin, Y.; Huang, L.; Liu, Y.; Zhao, X.; Zhao, D.; and Huang, K. 2023. Efficient-vqgan: Towards high-resolution image generation with efficient vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7368–7377.

Challu, C.; Olivares, K. G.; Oreshkin, B. N.; Ramirez, F. G.; Canseco, M. M.; and Dubrawski, A. 2023. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 6989–6997.

Chen, M.; Shen, L.; Li, Z.; Wang, X. J.; Sun, J.; and Liu, C. 2024. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dimri, T.; Ahmad, S.; and Sharif, M. 2020. Time series analysis of climate variables using seasonal ARIMA approach. *Journal of Earth System Science*, 129: 1–16.

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

Du, T.; Wang, Y.; and Wang, Y. 2023. On the Role of Discrete Tokenization in Visual Representation Learning. In *The Twelfth International Conference on Learning Representations*.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; and Li, X. 2024. Tslanet: Rethinking transformers for time series representation learning. *arXiv preprint arXiv:2404.08472*.

Hamilton, J. D. 2020. *Time series analysis*. Princeton university press.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

Hewamalage, H.; Bergmeir, C.; and Bandara, K. 2021. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1): 388–427.

Jiang, S.; Benge, C.; and King, W. C. 2022. BERTVision–A Parameter-Efficient Approach for Question Answering. *arXiv preprint arXiv:2202.12210*.

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

Kelkar, S.; Grigsby, L.; and Langsner, J. 2007. An extension of Parseval's theorem and its use in calculating transient energy in the frequency domain. *IEEE Transactions on Industrial Electronics*, 42–45.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Li, G.; and Jung, J. J. 2023. Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges. *Information Fusion*, 91: 93–102.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, Z.; Li, S.; and Yan, X. 2023. Time Series as Images: Vision Transformer for Irregularly Sampled Time Series. *arXiv preprint arXiv:2303.12799*.

Li, Z.; Rao, Z.; Pan, L.; Wang, P.; and Xu, Z. 2023. Ti-MAE: Self-Supervised Masked Time Series Autoencoders. *arXiv preprint arXiv:2301.08871*.

Liu, P.; Guo, H.; Dai, T.; Li, N.; Bao, J.; Ren, X.; Jiang, Y.; and Xia, S.-T. 2025. Calf: Aligning llms for time series forecasting via cross-modal fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18915–18923.

Liu, X.; Liu, J.; Woo, G.; Aksu, T.; Liang, Y.; Zimmermann, R.; Liu, C.; Savarese, S.; Xiong, C.; and Sahoo, D. 2024. Moirai-MoE: Empowering Time Series Foundation Models with Sparse Mixture of Experts. *arXiv preprint arXiv:2410.10469*.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2023. GPT understands, too. *AI Open*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 1907. RoBERTa: a robustly optimized BERT pretraining approach (2019). *arXiv preprint arXiv:1907.11692*, 364.

Ma, Q.; Li, S.; and Cottrell, G. W. 2020. Adversarial joint-learning recurrent neural network for incomplete time series

classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 1765–1776.

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

Oh, J.; Wang, J.; and Wiens, J. 2018. Learning to exploit invariances in clinical time-series data using sequence transformer networks. In *Machine learning for healthcare conference*, 332–347. PMLR.

Rizvi, S. M. H. 2022. Time series deep learning for robust steady-state load parameter estimation using 1D-CNN. *Arabian Journal for Science and Engineering*, 47(3): 2731–2744.

Sezer, O. B.; Gudelek, M. U.; and Ozbayoglu, A. M. 2020. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90: 106181.

Shi, X.; Wang, S.; Nie, Y.; Li, D.; Ye, Z.; Wen, Q.; and Jin, M. 2024. Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*.

Talukder, S.; Yue, Y.; and Gkioxari, G. 2024. Totem: Tokenized time series embeddings for general time series analysis. *arXiv preprint arXiv:2402.16412*.

Tan, M.; Merrill, M. A.; Gupta, V.; Althoff, T.; and Hartvigsen, T. 2024. Are language models actually useful for time series forecasting?, 2024. *URL https://arxiv.org/abs/2406.16964*.

Tang, B.; and Matteson, D. S. 2021. Probabilistic transformer for time series analysis. *Advances in Neural Information Processing Systems*, 34: 23592–23608.

Tang, W.; Long, G.; Liu, L.; Zhou, T.; Jiang, J.; and Blumenstein, M. 2020. Rethinking 1d-cnn for time series classification: A stronger baseline. *arXiv preprint arXiv:2002.10061*, 1–7.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, S.; Li, J.; Shi, X.; Ye, Z.; Mo, B.; Lin, W.; Ju, S.; Chu, Z.; and Jin, M. 2024. Timemixer++: A general time series pattern machine for universal predictive analysis. *arXiv preprint arXiv:2410.16032*.

Wang, Z.; Xu, X.; Zhang, W.; Trajcevski, G.; Zhong, T.; and Zhou, F. 2022. Learning latent seasonal-trend representations for time series forecasting. *Advances in Neural Information Processing Systems*, 35: 38775–38787.

Woo, G.; Liu, C.; Kumar, A.; Xiong, C.; Savarese, S.; and Sahoo, D. 2024. Unified training of universal time series forecasting transformers, 2024. *URL https://arxiv.org/abs/2402.02592*, 7.

Woo, G.; Liu, C.; Sahoo, D.; Kumar, A.; and Hoi, S. 2022. CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *arXiv preprint arXiv:2202.01575*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34: 22419–22430.

Xu, J.; Wu, H.; Wang, J.; and Long, M. 2021. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*.

Yang, Z.; Garcia, N.; Chu, C.; Otani, M.; Nakashima, Y.; and Takemura, H. 2020. Bert representations for video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1556–1565.

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

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11121–11128.

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

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8552–8562.

Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, 27268–27286. PMLR.

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