PERCEIVE FAST, THINK SLOW: A COGNITIVE-INSPIRED FRAMEWORK FOR TIME SERIES ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series modeling faces persistent challenges: fixed-window tokenization misaligns with natural event boundaries, uniform computation wastes capacity on simple patterns, and static architectures cannot adapt to diverse temporal dependencies. We propose **PeCo-TS**, a cognitive-inspired framework that instantiates the principle of "perceive fast, think slow" through three key innovations: (1) *event-driven dynamic-length tokenization* that aligns tokens with semantic boundaries and reduces redundancy, (2) a *Slow–Fast dual-pathway architecture* that separates rapid perception of fine-grained variations from slower abstraction of event-level structures, and (3) *Dual-Axis Adaptive (DA^2) attention* that dynamically balances intra-series and inter-series dependencies via learnable gating. Extensive experiments across forecasting, classification, anomaly detection, and imputation demonstrate the broad applicability of PeCo-TS, yielding consistent improvements over Transformer and linear baselines, including 5.6% lower forecasting MSE, 9.3% lower imputation error, higher classification accuracy across UCR/UEA benchmarks, and a 6.7% relative F1 gain in anomaly detection. Beyond accuracy, PeCo-TS achieves favorable efficiency–performance trade-offs by leveraging event-level abstraction and complementary pathway synergy, while its learned boundaries align with real-world regime shifts, providing interpretability. These results establish PeCo-TS as a versatile backbone that unifies efficiency, adaptability, and semantic alignment for diverse time-series applications.

1 INTRODUCTION

Time series data drives critical decision-making across diverse domains including climate monitoring, energy management, financial trading, healthcare diagnostics, and industrial automation. Real-world time series exhibit rich temporal complexity: abrupt regime shifts such as market crashes or equipment failures coexist with gradual trends such as seasonal variations or long-term growth, while high-frequency noise interleaves with persistent periodic patterns such as daily cycles and weekly rhythms. To effectively support the growing spectrum of tasks, including forecasting future values, classifying temporal patterns, detecting anomalies, and imputing missing data, models must capture both transient events and long-term dependencies across multiple temporal scales.

Despite this complexity, most approaches still follow a rigid three-stage pipeline. First, they split a series into fixed-size patches and treat each patch as a token. Second, a uniform architecture (e.g., self-attentive Transformer or MLP) assigns the same amount of compute to every token. Third, task heads project hidden states to outputs (e.g., forecasting, classification, anomaly detection). While convenient, this recipe clashes with heterogeneous real-world signals and leads to three limitations: (i) *boundary misalignment*—fixed windows cut through meaningful events (e.g., crashes, daily cycles, anomaly onsets), yielding incoherent representations (Nie et al., 2023; Wu et al., 2023); (ii) *computational redundancy*—expensive attention is spent on simple trends while complex patterns remain under-modeled (Zeng et al., 2023; Chen et al., 2023); and (iii) *limited adaptivity*—static channel handling cannot balance intra-series temporal dependencies against inter-series cross-channel correlations (Zhou et al., 2023; Han et al., 2023).

Cognitive neuroscience provides a useful blueprint. Human perception operates through dual pathways: *fast perceptual streams* that capture high-frequency details for immediate responsiveness, and *slower integrative streams* that abstract low-frequency regularities into coherent events and higher-level concepts (Zacks and Swallow, 2007; Kahneman, 2011; Desimone and Duncan, 1995; Kiebel

et al., 2008). Crucially, the brain performs *adaptive event segmentation*, partitioning continuous inputs into variable-length events such as daily cycles, regime changes, or anomaly onsets, rather than rigid temporal windows (Zacks and Swallow, 2007). Higher-order processing further leverages *selective attention*, shifting focus between temporal patterns within streams and cross-modal correlations across channels (Grondin, 2010). Together, these mechanisms concentrate computation on meaningful units while maintaining efficiency through event-level abstraction.

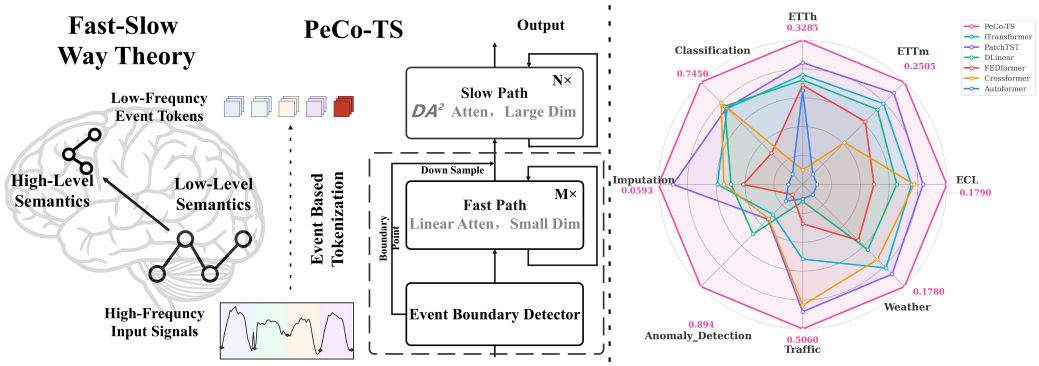

Figure 1: **Overview and highlights.** *Left:* Cognitive motivation and architecture of **PeCo-TS**, which integrates event-driven segmentation, a Fast Path for high-frequency perception, a Slow Path for event-level abstraction, and DA^2 attention for adaptive dependency modeling. *Right:* Aggregated results across four core time-series tasks show that PeCo-TS achieves consistent accuracy gains and superior performance compared to strong baselines.

Inspired by cognitive neuroscience, we propose **PeCo-TS** (Perception–Concept Transformer for Time Series), a dual-pathway framework that couples rapid perception with slower conceptual abstraction (Figure 1, left). The *Fast Path* employs point-wise embedding and linear attention to capture high-frequency details and transient events. An *event-driven tokenization* module, guided by frequency-domain boundary detection, adaptively segments sequences into variable-length tokens aligned with intrinsic dynamics. These tokens are then processed by the *Slow Path* using Dual-Axis Adaptive (DA^2) attention, which balances temporal dependencies within each series and cross-channel correlations via a learnable gating mechanism. In this way, PeCo-TS replaces rigid fixed-window patching with adaptive event segmentation, reduces computational complexity through event-level abstraction, and allocates attention more effectively while preserving both local fidelity and global coherence.

Comprehensive experiments confirm the advantages of this cognitively inspired design (Figure 1, right). Across forecasting, classification, anomaly detection, and imputation, PeCo-TS consistently outperforms state-of-the-art Transformer and linear baselines while offering superior accuracy–efficiency trade-offs. Furthermore, the learned event boundaries align well with real regime shifts and anomalies, providing intuitive interpretability and validating the semantic relevance of our adaptive segmentation. Our key contributions are threefold: (1) a novel **event-driven dynamic-length tokenization** framework that fundamentally replaces fixed-window patching with boundary-aware segmentation; (2) a **Slow–Fast dual-pathway architecture** that separates rapid perception from conceptual abstraction, mirroring the brain's perceive-fast, think-slow strategy; and (3) a **Dual-Axis Adaptive (DA^2) attention** mechanism that dynamically balances intra-series and inter-series dependencies through learnable gating, enhancing both generalization and versatility.

2 Related Work

Tokenization and Architecture Efficiency. Fixed-size patching remains the dominant paradigm in time-series Transformers. PatchTST (Nie et al., 2023) treats contiguous windows as tokens for long-horizon forecasting, while TimesNet (Wu et al., 2023) leverages 2D transformations to capture multi-period structure. Yet such rigid partitioning often cuts through semantically meaningful events, producing fragmented representations. Recent efforts aim to mitigate this: MultiResFormer (Peršak et al., 2024) constructs tokens at multiple resolutions, DeformableTST (Luo and Wang, 2024) adapts attention spans to informative time points, and token-level methods such as TOTEM (Talukder et al., 2024) or token merging (Götz et al., 2024) improve efficiency by dis-

cretization or merging. Pre-trained models like LPTM (Kamarthi and Prakash, 2024) also adopt adaptive segmentation for cross-domain learning. In parallel, lightweight alternatives (Linear (Zeng et al., 2023), TSMixer (Chen et al., 2023)) and pruning-based strategies (Wang et al., 2024a; Zhou et al., 2024) reveal redundancy in uniform Transformers. Unlike these approaches, our framework replaces fixed windows altogether with learnable, rhythm-aligned event tokens, offering an end-to-end solution where segmentation is naturally adaptive to signal dynamics, enabling semantically coherent representations and efficient capacity allocation.

Multivariate Dependencies and Channel Modeling. Handling cross-channel structure is another key challenge. While iTransformer (Zhou et al., 2023) models variables as tokens to capture inter-series relations, static designs cannot adaptively trade off intra- versus inter-series dependencies. More flexible strategies such as MCformer and MLinear (Han et al., 2023; Li et al., 2023) dynamically group channels, while MSGNet (Liu et al., 2024) incorporates frequency-aware graphs. Large-scale pre-trained models like TimesFM, Chronos, MOIRAI, Sundial, and TimesBERT (Das et al., 2024; Shchur et al., 2024; Bhatnagar et al., 2024; Liu et al., 2025; Zhang et al., 2025) further broaden the range of downstream capabilities, and domain-specific models such as PriceFM (Yu et al., 2025) tailor objectives for financial series. Our DA^2 attention differs by introducing a unified mechanism that adaptively balances intra- and inter-series correlations via learnable gating, providing a compact yet general solution across datasets.

Cognitive-Inspired Processing. Beyond engineering heuristics, cognitive neuroscience offers a principled perspective. SlowFast networks (Feichtenhofer et al., 2019) demonstrate the benefit of dual-rate pathways in video, and wavelet or multi-scale time-series methods (Wang et al., 2023; Lai et al., 2018) approximate multi-resolution patterns. However, these approaches lack explicit separation between perception and conceptual abstraction. Cognitive studies (Zacks and Swallow, 2007; Grondin, 2010) highlight the brain's dual-pathway principle of "perceive fast, think slow," where fast streams capture immediate high-frequency cues and slower pathways integrate them into higher-level abstractions. Our PeCo-TS directly operationalizes this idea, coupling event-driven segmentation with a Slow–Fast dual-pathway design, thereby moving beyond ad hoc multi-scale heuristics toward a cognitively motivated and empirically validated framework.

3 Methodology

3.1 Overview of PeCo-TS

The human brain processes continuous sensory streams through a dual-pathway system: a *fast pathway* that responds rapidly to fine-scale stimuli, and a *slow pathway* that integrates information over longer horizons to form abstract concepts. This division of labor allows cognition to capture both transient details and stable regularities. In contrast, existing Transformers for time series typically rely on a single processing pipeline with fixed patching and uniform attention, which fails to reflect the heterogeneous timescales and adaptive correlations inherent in real signals.

Inspired by this neuro-cognitive principle, we propose the **Perception–Concept Transformer for Time Series (PeCo-TS)**, a dual-pathway architecture designed to model event-driven signals with both efficiency and accuracy (see Figure 2). The framework integrates four coordinated stages: (i) *Event Boundary Detector* that identifies semantic boundaries for adaptive tokenization; (ii) *Fast Path* that captures fine-grained details through point-wise processing, followed by a segmentation-and-downsampling step that converts high-resolution features into event-level tokens; (iii) *Slow Path* with DA^2 attention that processes these event-based tokens for abstract modeling; and (iv) *Temporal Reprojection* that fuses abstract and fine-grained representations for multi-task outputs.

3.2 Event Boundary Detector

Modeling long sequences with uniform patches is not only computationally expensive but also misaligned with the event-driven nature of real signals. In practice, important transitions often occur at irregular intervals, making fixed patching prone to cutting through meaningful events. To address this, we design an **event-driven tokenization module** that detects semantic boundaries directly from the raw multivariate input $x \in \mathbb{R}^{B\times L\times C}$, ensuring that subsequent processing aligns with natural temporal structure.

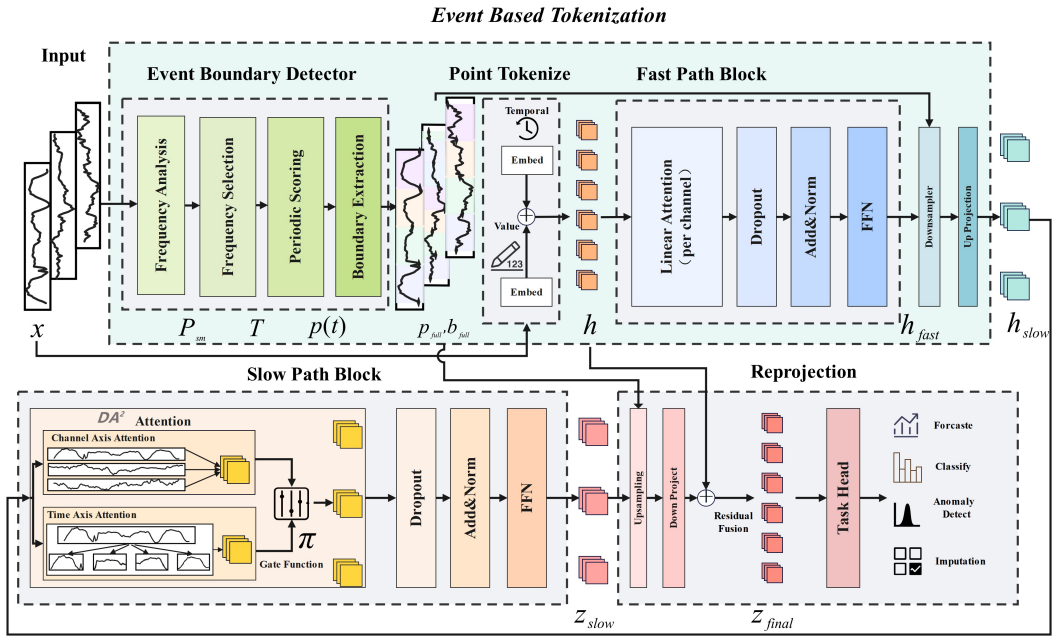

Figure 2: **Overall framework of PeCo-TS.** The framework integrates four coordinated stages: Event Boundary Detector identifies semantic boundaries for adaptive tokenization; Fast Path captures fine-grained details through point-wise processing and linear attention; Slow Path processes event-level tokens with DA^2 attention for adaptive intra- and inter-series dependencies; and Temporal Reprojection fuses abstract and fine-grained representations for multi-task outputs.

For each channel, the dominant rhythm is estimated by computing the power spectrum and applying a learnable frequency smoother g_θ:

$$P = |X|^2, \quad X = \text{FFT}(x), \quad P_{\text{sm}} = g_\theta(P). \tag{1}$$

A softmax distribution with temperature τ softly selects frequency bins to obtain an effective frequency k_{eff} and period T:

$$\alpha = \text{softmax}\Big(\frac{P_{\text{sm}}}{\tau}\Big), \quad k_{\text{eff}} = \sum_f \alpha_f \cdot f, \quad T = \frac{L}{k_{\text{eff}}}. \tag{2}$$

A differentiable cosine comb then highlights candidate boundaries:

$$p(t) = \Big(\frac{1+\cos(2\pi t/T)}{2}\Big)^\gamma, \quad \gamma > 1, \tag{3}$$

where the learnable sharpness γ adjusts boundary precision. Non-maximum suppression and thresholding convert these scores into binary boundaries $b_{\text{full}} \in \{0,1\}^{B\times L\times C}$, while soft probabilities $p_{\text{full}} \in [0,1]^{B\times L\times C}$ are retained.

This mechanism aligns tokenization with the inherent rhythm of each channel, yielding three advantages: (*i*) the number of tokens adapts to signal-specific periodicity; (*ii*) the boundaries are differentiable and trainable, enabling end-to-end optimization; and (*iii*) by operating on event tokens rather than all time steps, the complexity of later attention layers is reduced from $O(L^2)$ to $O(M^2)$ with $M \ll L$.

3.3 Fast Path: Perception of Fine Details

While boundaries guide event-level abstraction, retaining fine-grained local details remains essential for accurate modeling. Analogous to early cortical areas in human perception, the **Fast path** processes the input at its original resolution to preserve high-frequency variations and transient patterns.

Formally, each scalar observation $x_{t,c}$ is embedded into a d_f-dimensional vector via a point-wise tokenizer:

$$h \in \mathbb{R}^{B\times L\times C\times d_f}. \tag{4}$$

After reshaping and adding positional encodings, we obtain $h \in \mathbb{R}^{BC\times L\times d_f}$, which is then processed with *linear attention* to efficiently capture short-range dependencies:

$$\text{Attn}(Q,K,V) \approx \frac{\phi(Q)\left(\phi(K)^\top V\right)}{\phi(Q)\,\phi(K)^\top}, \tag{5}$$

reducing time complexity from $O(L^2)$ to $O(Ld)$.

The resulting representation h_{fast} preserves temporal precision and is projected into a higher-dimensional space $h_{\text{fast}} \in \mathbb{R}^{BC\times L\times d_h}$. Guided by the boundaries b_{full} from Section 3.2, a boundary-aware downsampler aggregates h_{fast} into variable-length event tokens:

$$h_{\text{slow}} = \text{Downsample}(h_{\text{fast}}, b_{\text{full}}) \in \mathbb{R}^{B\times C\times M\times d_h}, \quad M \ll L. \tag{6}$$

Since different channels may yield different token counts M_c, we pad sequences to $M_{\max} = \max_c M_c$ and maintain a mask $\mu \in \{0,1\}^{B\times C\times M_{\max}}$ to ensure consistent computation. This design enables the model to preserve fine details while seamlessly transitioning to event-level abstraction.

3.4 Slow Path: Conceptual Abstraction

High-level perception in the brain does not stop at detecting local events; it further integrates them into coherent concepts by linking information across time and across modalities. Following this principle, the **Slow path** in PeCo-TS takes event-level tokens as input and abstracts them into higher-order representations using a dual-axis adaptive attention mechanism.

Formally, given event tokens $h_{\text{slow}} \in \mathbb{R}^{B\times C\times M\times d_h}$ and mask μ (with M denoting $M_{\max}$), DA^2 attention decomposes modeling into two complementary axes. Along the *token axis*, attention captures temporal dependencies across events within each channel. Along the *channel axis*, attention captures correlations across channels at the same event step. Padded positions are excluded using μ (see Appendix A.4):

$$\tilde{z}_c(b,c,\cdot) = \text{Attn}_{\text{token}}\big(h_{\text{slow}}(b,c,\cdot,\cdot)\big) \in \mathbb{R}^{M\times d_h}, \tag{7}$$

$$\tilde{z}_m(b,\cdot,m) = \text{Attn}_{\text{channel}}\big(h_{\text{slow}}(b,\cdot,m,\cdot)\big) \in \mathbb{R}^{C\times d_h}. \tag{8}$$

Both outputs are reshaped to a common layout and blended by a learnable gate $\pi \in (0,1)$:

$$Y = \pi \odot \tilde{z}_m + (1-\pi)\odot \tilde{z}_c \in \mathbb{R}^{B\times C\times M\times d_h}. \tag{9}$$

Unless otherwise specified, π is a per-layer scalar broadcast as $B\times C\times M\times 1$, balancing inter-series and intra-series modeling. A finer variant allows per-position gating $\pi \in (0,1)^{B\times C\times M\times 1}$, but we use the scalar form by default for stability.

Stacking multiple DA^2 layers with residual and feedforward modules produces the abstract representation $z_{\text{slow}} \in \mathbb{R}^{B\cdot C\times M\times d_h}$, which jointly encodes long-horizon temporal dependencies and context-dependent cross-channel relations. This abstraction is particularly important for multivariate event-driven time series, where both within-series evolution and cross-series interactions carry critical semantics (see Appendix A.2).

3.5 Temporal Reprojection and Multi-task Heads

Event tokens are efficient for abstraction but not directly aligned with the fine temporal resolution required by downstream tasks. To bridge this gap, we design a **temporal reprojection layer** that upsamples event-level features back to the original scale, restoring temporal alignment while injecting high-level semantics.

Given $z_{\text{slow}} \in \mathbb{R}^{B\cdot C\times M\times d_h}$ and boundary indicators $(p_{\text{full}}, b_{\text{full}})$, the reprojection constructs convex weights $\{w_{t,i}\}_{i=1}^{M}$ for each time step t:

$$z_{\text{full}}(t) = \sum_{i=1}^{M} w_{t,i}\, z_{\text{slow}}(i), \quad \sum_{i=1}^{M} w_{t,i} = 1. \tag{10}$$

Table 1: Multivariate forecasting results with prediction lengths $S \in \{96, 192, 336, 720\}$ for all datasets and fixed lookback length $T = 96$. Results are averaged across prediction lengths. The best results are highlighted in **red** and the second best are shown in **blue**.

Dataset	PeCo-TS		iTransformer		PatchTST		TSMixer		TimesNet		Mamba		DLinear		FEDformer		Crossformer		Autoformer	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.415**	**0.427**	0.465	0.455	0.448	0.446	0.626	0.588	0.460	0.455	0.550	0.509	0.460	0.457	0.492	0.486	0.576	0.535	0.494	0.482
ETTh2	**0.378**	**0.400**	0.383	0.407	0.381	0.408	2.025	1.194	0.409	0.425	0.443	0.441	0.564	0.519	0.424	0.442	1.199	0.796	0.427	0.444
ETTm1	**0.369**	**0.388**	0.407	0.411	0.386	0.400	0.529	0.513	0.412	0.418	0.498	0.468	0.404	0.408	0.456	0.460	0.483	0.482	0.608	0.522
ETTm2	0.287	0.333	0.291	0.334	**0.285**	**0.330**	1.030	0.753	0.294	0.332	0.377	0.380	0.355	0.401	0.298	0.344	0.774	0.623	0.319	0.360
Electricity	**0.201**	**0.282**	0.225	0.308	0.210	0.297	0.233	0.340	0.297	0.376	0.209	0.311	0.225	0.319	0.279	0.375	0.498	0.534	0.306	0.395
Exchange	0.387	0.418	0.364	**0.407**	0.369	0.407	0.539	0.590	0.406	0.439	0.693	0.555	**0.339**	0.413	0.508	0.498	0.659	0.623	0.504	0.501
Traffic	**0.526**	**0.337**	0.612	0.404	**0.527**	0.339	0.606	0.407	0.903	0.523	0.679	0.381	0.672	0.419	0.713	0.446	0.564	0.394	0.712	0.448
Weather	0.260	0.283	0.267	0.287	0.260	**0.281**	**0.243**	0.309	0.262	0.288	0.295	0.315	0.265	0.317	0.308	0.353	0.281	0.319	0.338	0.378

Segments $\mathcal{S}_i = [s_i, e_i]$ are defined by consecutive boundaries in b_{full}. Within each segment, unnormalized weights are assigned as $\tilde{w}_{t,i} = \kappa(\text{dist}(t; s_i, e_i))\, \bar{p}(t)$, where $\bar{p}(t)$ is the channel-aggregated confidence from p_{full} and $\kappa(d) = \exp(-d^2/2\sigma^2)$ is a Gaussian kernel. Normalization yields

$$w_{t,i} = \frac{\tilde{w}_{t,i}}{\sum_{j=1}^{M} \tilde{w}_{t,j}}, \quad w_{t,i} = 0 \text{ if } t \notin \mathcal{S}_i. \tag{11}$$

Finally, the reprojected features are aligned with fast-path representations via a learnable output projection and residual fusion:

$$z_{\text{final}}(t) = W_{\text{out}}\, z_{\text{full}}(t) + h_{\text{fast}}(t), \quad W_{\text{out}} \in \mathbb{R}^{d_f \times d_h}. \tag{12}$$

The unified representation $z_{\text{final}} \in \mathbb{R}^{B \cdot C \times L \times d_f}$ forms a shared basis for diverse tasks—classification, imputation, anomaly detection, forecasting, and pretraining. This feedback from abstraction to detail resembles *predictive coding*, ensuring that conceptual modeling remains consistent with fine-grained temporal alignment (see Appendix A.5).

4 Experiments

We evaluate **PeCo-TS** on four fundamental time-series tasks—forecasting, classification, anomaly detection, and imputation—using widely adopted benchmarks: forecasting on ETTh1/h2, ETTm1/m2, Electricity, Exchange, Traffic, and Weather (Zhou et al., 2021; Trindade, 2015; Lai et al., 2017; Lai and contributors, 2017; Li et al., 2018; for Biogeochemistry , data origin; Wang et al., 2024b); classification on seven UCR/UEA datasets (Chen et al., 2015; Bagnall et al., 2018); anomaly detection on MSL, PSM, SMAP, SMD, and SWAT (Hundman et al., 2018; Abdulaal et al., 2021; Su et al., 2019; Goh et al., 2016); and imputation on ETTh/ETTm/Electricity/Weather. This comprehensive evaluation setting ensures coverage of both short- and long-horizon prediction, univariate and multivariate inputs, and diverse application domains.

4.1 Broad Applicability Validated by Multi-Task Results

Across all four task categories, PeCo-TS consistently outperforms strong baselines, demonstrating the versatility of its cognitive-inspired dual-pathway design. In forecasting, it achieves an average 5.6% reduction in MSE compared with Transformer-based competitors, with robustness across horizons and datasets (Table 1; Appendix, Table 2). In classification, it surpasses leading alternatives (Table 3), highlighting its ability to learn transferable representations. For anomaly detection, PeCo-TS raises the average F1 score from 0.837 to 0.893, a 6.7% relative gain (Table 4), while in imputation it reduces reconstruction error by 9.3% on average (Table 5).

Taken together, these dense and consistent improvements across heterogeneous datasets substantiate the broad applicability and robustness of PeCo-TS. Rather than relying on task-specific heuristics or bespoke tuning, the cognitively inspired separation of perception and abstraction emerges as a general modeling principle for time series, validating PeCo-TS as a versatile backbone for real-world applications.

4.2 Advantages over Fixed Patching

A key limitation of conventional Transformers for time series lies in their rigid fixed-window tokenization, which fragments signals and often cuts through natural temporal boundaries. In con-

trast, our learnable, event-driven segmentation produces variable-length tokens that adapt to intrinsic rhythms, such as daily cycles or volatility bursts, thereby aligning representation with the underlying event structure.

To validate its effectiveness, we compare our segmentation against fixed-patch baselines across two complementary dimensions: prediction horizon and input length. As shown in Figure 3a, event-driven segmentation consistently achieves lower MSE across horizons, with relative gains ranging from 4.7% on Weather to 7.3% on ETTm1. Figure 3b further confirms robustness under varying input sequence lengths: our method maintains superior performance regardless of the temporal context size. Notably, the advantage of event-driven segmentation becomes more pronounced as input or prediction length increases. Short patches tend to split coherent events into fragments and introduce redundant tokens into higher layers, leading to inefficiency, while long patches often merge multiple events into a single token, causing semantic overlap and learning difficulty. By contrast, event-driven segmentation preserves semantic integrity within tokens while maintaining computational efficiency, thereby scaling gracefully with longer horizons and context windows.

Together, these results provide strong evidence that event-driven segmentation fundamentally improves over arbitrary fixed patches. Qualitative visualizations in the Appendix (Figures 6–12) further show that learned boundaries align with key temporal events, yielding more interpretable and generalizable representations. This alignment underpins the consistent quantitative gains observed across datasets, establishing event-driven tokenization as a principled foundation for time-series modeling.

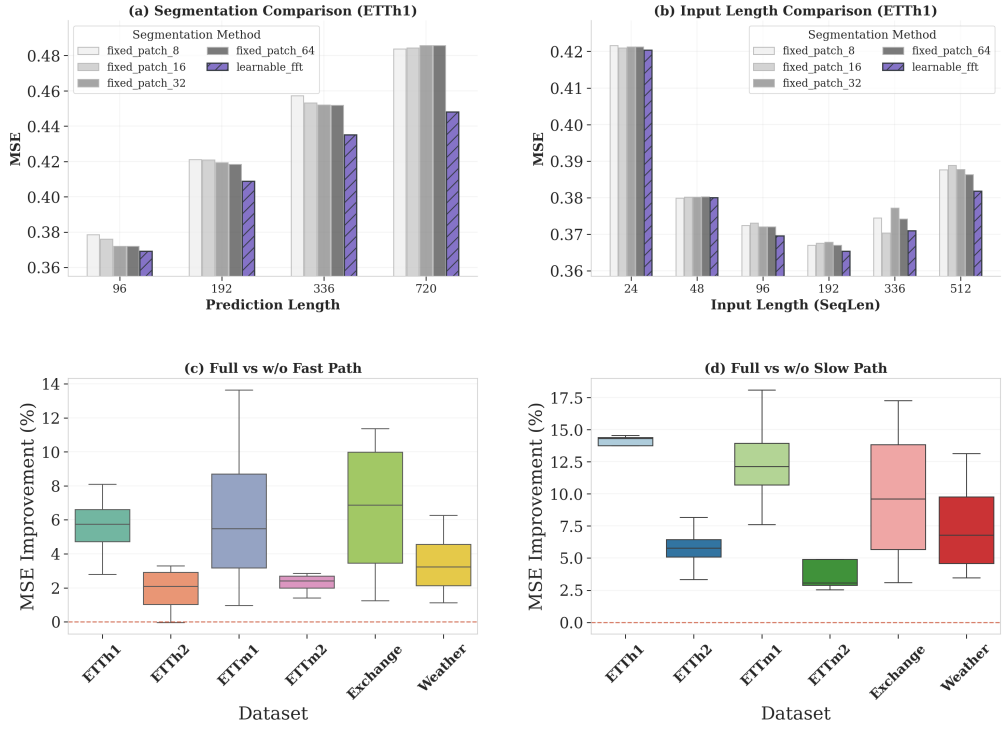

Figure 3: **Comprehensive analysis of PeCo-TS key components.** (a) **Segmentation method comparison (ETTh1):** Adaptive event-driven segmentation consistently outperforms fixed patches across prediction horizons, with learnable FFT achieving the lowest MSE. (b)**Input length sensitivity analysis (ETTh1):** Performance comparison across varying input sequence lengths and fixed 96 prediction horizon demonstrates the robustness of learnable FFT segmentation, maintaining superior performance regardless of input length variations. (c)(d) **Dual-pathway architecture validation:** Comprehensive ablation study demonstrates the complementary nature of Fast and Slow pathways. The asymmetric contributions validate our cognitive-inspired hypothesis that rapid perception and slower reasoning serve distinct but synergistic roles in time series modeling.

4.3 Synergistic Roles of Cognitive Fast and Slow Paths

Inspired by the brain's dual-pathway system, PeCo-TS explicitly separates rapid perception from slower conceptual abstraction. To assess the necessity of this design, we perform ablation studies by removing either the Fast or Slow path. As shown in Figure 3c,d, eliminating the Fast path leads to an average 10.0% drop in MSE improvement, while removing the Slow path results in a 4.6% reduction. This asymmetric degradation underscores their complementary functions: the Fast path preserves high-frequency details crucial for precise temporal alignment (e.g., anomaly detection), whereas the Slow path processes event-level tokens to capture long-range dependencies efficiently and allocate modeling capacity to complex structures.

Beyond accuracy, this division of labor also contributes to efficiency. As reported in Appendix A.12, PeCo-TS attains higher accuracy with fewer parameters and lower latency compared with strong baselines. These empirical savings align with the theoretical benefits of event-driven compression and the practical synergy of the dual pathways. Together, the results validate our cognitive-inspired hypothesis: rapid perception and slower abstraction are distinct yet synergistic mechanisms that jointly yield a more effective and efficient time-series model.

4.4 Adaptive Intra- and Inter-Series Dependency Modeling

A distinctive advantage of PeCo-TS lies in its DA^2 attention, which adaptively balances intra-series and inter-series dependencies rather than committing to fixed channel-independent or channel-dependent designs. As shown in Appendix Figure 20, DA^2 consistently outperforms both alternatives across benchmarks, with the largest margin on ETTh2 (5.2% average MSE reduction).

Beyond accuracy, DA^2 attention dynamically adjusts its gate parameter π to reflect the correlation structure of each dataset. Figure 4 illustrates that, as training progresses, the model gradually learns dataset-specific dependency patterns: on Traffic (Chen et al., 2025), where intra-series periodicity dominates, DA^2 increases its emphasis on intra-series attention ($\pi \approx 0.56$); on Weather (Chen et al., 2025), where cross-channel correlations are stronger, the model assigns greater weight to inter-series attention ($\pi \approx 0.49$). This adaptive learning process improves predictive accuracy while offering interpretable insights into dataset-specific structures.

4.5 Cognitive Pathway Behavior and Mechanistic Insights

Figure 5 visualizes how the cognitive principle of "perceive fast, think slow" is instantiated in PeCo-TS and materializes into observable modeling behavior. The boundary detector first converts continuous signals into event-aligned tokens (Figure 5a,b), ensuring semantic integrity at the token level. These tokens then flow into two complementary pathways: the Fast path applies linear attention with strong near-diagonal focus (Figure 5d), retaining fine-grained local dependencies crucial for precise temporal alignment; the Slow path employs DA^2 attention across tokens and channels (Figure 5e,f), integrating long-range structures and cross-series correlations. Temporal reprojection (Figure 5c) feeds event-level abstractions back to the native resolution, enabling consistent reconstruction of high-frequency detail.

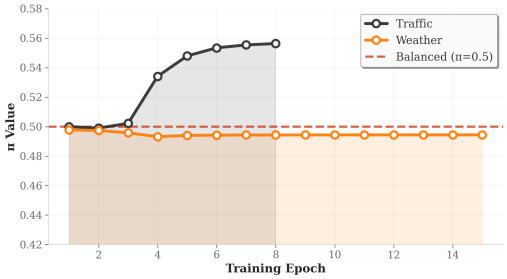

Figure 4: π evolution on Weather and Traffic datasets.

The attention heatmaps provide direct evidence for this division of labor: the Fast path concentrates on short-range patterns, while the Slow path distributes capacity over broader token and channel contexts. Moreover, the gate parameter π (Figure 21) adapts smoothly across datasets, shifting emphasis toward intra-series dependencies in periodic data (e.g., Traffic) and toward inter-series dependencies when cross-channel correlations dominate (e.g., Weather). This dynamic reallocation reflects the model's ability to specialize its reasoning strategy to the dataset at hand.

These mechanistic observations align closely with the empirical results. Event-driven segmentation explains why PeCo-TS maintains stronger margins at longer horizons (Figure 3a,b); the complementary Fast/Slow contributions account for the asymmetric error increases in ablation (Figure 3c,d); and DA^2 attention clarifies why adaptive correlation modeling consistently outperforms fixed channel-independent/dependent baselines. Even under anomaly detection and missing-data scenarios, the synergy holds: slow abstractions provide contextual guidance, while fast features anchor precise timing, yielding improved localization and robustness. Together, Figure 5 illustrates a mechanism–phenomenon–result loop: event alignment and dual-path reasoning shape interpretable attention geometry, which directly underpins the multitask gains and favorable accuracy–efficiency trade-offs observed in PeCo-TS.

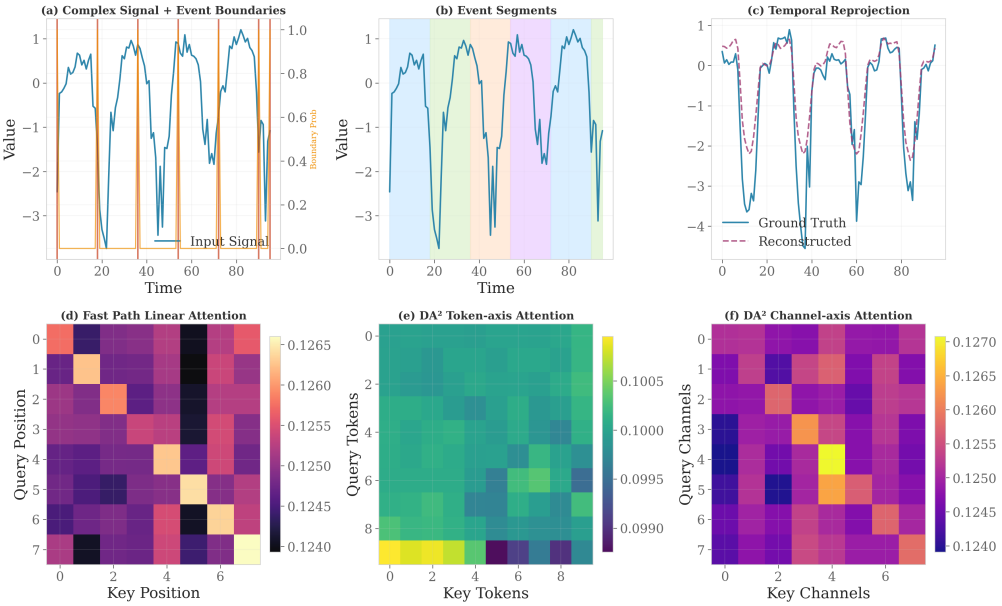

Figure 5: Cognitive architecture visualization showing the dual-pathway processing. (a) Input signal with learned event boundaries; (b) Event segments derived from boundary detection; (c) Temporal reprojection reconstructing fine-grained outputs from event-level abstractions; (d) Fast path linear attention exhibiting local temporal dependencies; (e) DA^2 token-axis attention for intra-series modeling; (f) DA^2 channel-axis attention for inter-series correlations.

5 Conclusion

This work introduced **PeCo-TS**, a cognitive-inspired framework that translates the principle of "perceive fast, think slow" into a practical architecture for time series modeling. By coupling event-driven tokenization, a dual-pathway design, and DA^2 adaptive attention, PeCo-TS directly addresses the long-standing limitations of fixed-window segmentation, uniform computation, and static channel mixing.

Extensive experiments across forecasting, classification, anomaly detection, and imputation confirm its advantages: event-driven segmentation scales gracefully with horizon and context length, the Fast and Slow paths contribute complementary precision and abstraction, and DA^2 attention adapts to dataset-specific dependency structures. Together, these mechanisms yield consistent improvements over strong baselines while reducing redundancy and enhancing efficiency.

Beyond performance gains, PeCo-TS demonstrates how cognitive processing principles can be operationalized into concrete modeling benefits. The framework offers not only a versatile backbone for diverse temporal tasks but also a blueprint for future research on event-adaptive and interpretable architectures. We believe this cognitive–computational synthesis opens promising directions for scalable pretraining, cross-domain generalization, and transparent decision-making in time series applications.

ACKNOWLEDGMENTS

We thank the reviewers for their constructive feedback and suggestions. This work was supported by research grants and computational resources that enabled comprehensive experimental validation.

ETHICS STATEMENT

This research was conducted in strict compliance with ethical standards. The datasets used in our experiments are all publicly available benchmarks or synthetically generated signals without any personally identifiable or sensitive information. No human or animal subjects were involved. All experimental protocols respect the principles of fairness, transparency, and scientific integrity. The proposed methods are intended solely for academic research purposes and do not pose foreseeable risks of harm or misuse.

REPRODUCIBILITY STATEMENT

We have taken concrete steps to ensure the reproducibility of our results.

- **Code and Models:** The full implementation of our PeCo-TS architecture, including the segmentation module, DA^2-Attention mechanism, and experimental pipelines, is available at the following anonymous repository: https://anonymous.4open.science/r/PeCO-TS-Code-102C
- **Datasets:** All datasets used are standard public benchmarks (ETT, Electricity, Exchange, Traffic, Weather, UCR/UEA, MSL, PSM, SMAP, SMD, SWAT). Detailed preprocessing instructions are included in the repository.
- **Configurations:** Hyperparameters, training schedules, and random seeds are documented in configuration files for exact replication.
- **Results:** Reported metrics are averaged over multiple runs to mitigate randomness, and raw logs/checkpoints are provided for verification.

USE OF LLMS

We acknowledge the use of large language models (LLMs) during the preparation of this work. ChatGPT (GPT-5) was employed **only** for the following purposes:

- **Writing Assistance:** Refining the clarity, conciseness, and readability of manuscript drafts, without altering the underlying technical content.
- **Formatting Support:** Generating LaTeX snippets for figures, tables, and equations.
- **Code Review:** Checking for consistency in implementation details and documenting modules (e.g., `LearnableFFTSegmenter`, `DualAxisAdaptiveAttention`).

Importantly, all core ideas, model designs, algorithm implementations, and experiments were conceived and executed by the authors. The LLM was not used to generate research hypotheses, design experiments, or produce empirical results. We take full responsibility for the originality, correctness, and integrity of this work.

REFERENCES

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023. OpenReview.net. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023. OpenReview.net. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, pages 11121–11128. AAAI Press, 2023. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

Si-An Chen, Chun-Liang Li, Sercan Ö. Arik, Nathanael C. Yoder, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023, 2023. URL https://openreview.net/forum?id=wbpxTuXgm0.

Yong Zhou, Sheng Jin, Xiaoyun Liu, and Zhongfeng Wang. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Lijun Han, Xingchen Ye, Zheng Xu, and Yufeng Chen. Mcformer: Multivariate time series forecasting with mixed-channels transformer. *arXiv preprint arXiv:2403.09223*, 2023.

Jeffrey M Zacks and Khena M Swallow. Event segmentation. *Current Directions in Psychological Science*, 16(2):80–84, 2007.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

Stefan J Kiebel, Jean Daunizeau, and Karl J Friston. A hierarchy of time-scales in the brain. *PLoS Computational Biology*, 4(11):e1000209, 2008.

Simon Grondin. Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics*, 2010.

Egon Peršak, Miguel F. Anjos, Sebastian Lautz, and Aleksandar Kolev. Multiple-resolution tokenization for time series forecasting with an application to pricing. *arXiv preprint arXiv:2407.03185*, 2024. URL https://arxiv.org/abs/2407.03185.

Donghao Luo and Xue Wang. Deformabletst: Transformer for time series forecasting without over-reliance on patching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a0b1082fc7823c4c68abcab4fa850e9c-Paper-Conference.pdf.

Sabera Talukder, Yisong Yue, and Georgia Gkioxari. Totem: Tokenized time series embeddings for general time series analysis. In *ICLR 2024 Workshop on Learning from Time Series for Health*, 2024. URL https://openreview.net/forum?id=QlTLkH6xRC.

Leon Götz, Marcel Kollovieh, Stephan Günnemann, and Leo Schwinn. Efficient time series processing for transformers and state-space models through token merging. *arXiv preprint arXiv:2405.17951*, 2024. URL https://arxiv.org/abs/2405.17951.

Harshavardhan Kamarthi and B. Aditya Prakash. Large pre-trained time-series models for cross-domain time series analysis. In *Advances in Neural Information Processing Systems (NeurIPS), Poster Track*, 2024. URL https://arxiv.org/abs/2311.11413.

Lei Wang, Ming Zhang, and Xiaoyun Liu. Efficient pruning of time series transformers. *arXiv preprint arXiv:2412.12883*, 2024a.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Foundation models for time series analysis: A tutorial and survey. *arXiv preprint arXiv:2403.14735*, 2024.

Shizhan Li, Yangdong Jin, Youfang Xuan, Xiaoqian Zhou, Wenxuan Chen, Yu-Xiang Wang, and Xifeng Yan. Mlinear: The search for linear time series forecasting. *arXiv preprint arXiv:2305.10721*, 2023.

Xiang Liu, Qingsong Hu, Yuxuan Zhang, Lixiang Nie, Shuaiqiang Yao, and Jianwei Yin. Msgnet: Multi-scale graph neural networks for multivariate time series forecasting. *AAAI*, 2024.

Ankur Das, Alejandro Ruiz, Cuong Nguyen, Samy Boodman, Xi Cheng, et al. Timesfm: Time series foundation models. *arXiv preprint arXiv:2403.07784*, 2024.

Oleksandr Shchur, Manuel Raedler, Tim Januschowski, Jan Gasthaus, and Valentin Flunkert. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Aseem Bhatnagar, Alexey Gakhov, Irene Ktena, et al. Moirai: Tiny foundation models for multivariate time series forecasting. *arXiv preprint arXiv:2402.02592*, 2024.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025.

Haoran Zhang, Yong Liu, Yunzhong Qiu, Haixuan Liu, Zhongyi Pei, Jianmin Wang, and Mingsheng Long. Timesbert: A bert-style foundation model for time series understanding. *arXiv preprint arXiv:2502.21245*, 2025.

Runyao Yu, Chenhui Gu, Jochen Stiasny, Qingsong Wen, Wasim Sarwar Dilov, Lianlian Qi, and Jochen L. Cremer. Pricefm: Foundation model for probabilistic electricity price forecasting. *arXiv preprint arXiv:2508.04875*, 2025.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *ICCV*, 2019.

Xiaoming Wang, Yuting Li, and Hao Chen. Wavelet-based time series analysis for forecasting. *IEEE Transactions on Signal Processing*, 2023.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. *SIGIR*, 2018.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianmin Li, Hanning Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 2021.

Artur Trindade. Electricityloaddiagrams20112014 [dataset]. UCI Machine Learning Repository, 2015. https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. *arXiv preprint arXiv:1703.07015*, 2017. LSTNet — used Exchange dataset; code/data at https://github.com/laiguokun/multivariate-time-series-data.

Guokun Lai and contributors. multivariate-time-series-data (exchange, electricity, traffic, ...). GitHub repository, 2017. https://github.com/laiguokun/multivariate-time-series-data (commonly used source for Exchange dataset).

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR)*, 2018.

Max Planck Institute for Biogeochemistry (data origin) and community mirrors. Jena climate (jena weather) dataset. Public meteorological observations (mirrored on Kaggle and other hosts), 2009. Kaggle mirror example: https://www.kaggle.com/datasets/harishedison/jena-weather-dataset.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*, 2024b. Time Series Library (TSLib) — code: https://github.com/thuml/Time-Series-Library.

Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, 2015. http://www.cs.ucr.edu/~eamonn/time_series_data/.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Kyle Hundman, Valerii Constantinou, Rishav Laporte, Ian Colwell, and Tim Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *KDD*, 2018.

A. Abdulaal, Z. Liu, and T. Lancewicki. Pooled server metrics (psm) dataset. GitHub / RANSynCoders, 2021. https://github.com/eBay/RANSynCoders.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

J. Goh, S. Adepu, K. N. Junejo, and A. Mathur. A dataset to support research in the design of secure water treatment systems. In *11th International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater)*, 2016. dataset available from iTrust SUTD: https://itrust.sutd.edu.sg/itrust-labs_datasets/.

Yu Chen, Nathalia Céspedes, and Payam Barnaghi. A closer look at transformers for time series forecasting: Understanding why they work and where they struggle. In *Forty-second International Conference on Machine Learning*, 2025.

A Appendix

A.1 Complexity Analysis of Event-driven Segmentation

Let $L \in \mathbb{N}$ denote the input length, $M \in \{1, \dots, L\}$ the number of event tokens after segmentation, and $d \in \mathbb{N}$ the hidden dimension.

Standard attention. Self-attention scales quadratically:

$$\mathcal{C}_{\text{full}} = O(L^2 d). \tag{13}$$

Event-driven segmentation. Our segmentation compresses L time steps into M tokens, where $M \approx L/T$ and $T > 0$ is the effective period estimated by the detector. Thus,

$$\mathcal{C}_{\text{event}} = O(M^2 d) = O\left(\frac{L^2}{T^2} d\right). \tag{14}$$

Reduction ratio. The relative savings is

$$\frac{\mathcal{C}_{\text{event}}}{\mathcal{C}_{\text{full}}} \approx \frac{1}{T^2}. \tag{15}$$

For periodicities $T \in [8, 32]$ commonly observed in climate, ECG, and engine vibration data, this is a cost ratio of $1/64$ to $1/1024$. Importantly, segmentation also improves *statistical efficiency*, where boundaries align with rhythmic units, concentrating attention capacity on semantically coherent chunks, rather than arbitrary windows. This is analogous to parsing sentences by words instead of fixed-length character spans.

Distributional view. Let T be a random effective period supported on $[T_{\min}, T_{\max}] \subset (0, \infty)$ and assume $M = \lceil L/T \rceil$. Then

$$\mathbb{E}\big[\mathcal{C}_{\text{event}}\big] \in O\Big(d\,\mathbb{E}[M^2]\Big) \subseteq O\Big(d\,\mathbb{E}[(L/T+1)^2]\Big) = O\Big(d\,L^2\,\mathbb{E}[T^{-2}] + d\,L\,\mathbb{E}[T^{-1}] + d\Big). \tag{16}$$

Therefore, whenever $\mathbb{E}[T^{-2}]$ is finite and bounded by $c/T_{\min}^2$, the expected reduction factor satisfies

$$\frac{\mathbb{E}[\mathcal{C}_{\text{event}}]}{\mathcal{C}_{\text{full}}} \le \frac{c}{T_{\min}^2} + O\Big(\tfrac{1}{L}\Big), \quad L \to \infty. \tag{17}$$

Proposition (piecewise-constant exactness). Suppose the input is piecewise constant with segment boundaries equal to the detector boundaries and the slow-path attention is applied only across segments. Then M equals the number of pieces and $O(M^2 d)$ attention achieves the same result as $O(L^2 d)$ full attention restricted to piecewise-constant hypotheses. *Proof.* On each segment the representation is constant, so aggregating to one token per segment is a sufficient statistic. Attention between segments in token space is identical to attention between any representatives in the original space. The quadratic pair count reduces from L^2 to M^2.

A.2 Theoretical Properties of DA^2 Attention

We now analyze the expressive capacity of the proposed dual-axis adaptive attention.

Formulation. Given token-level attention $\tilde{z}_c$ and channel-level attention $\tilde{z}_m$, DA^2 combines them as

$$Y(\pi) = \pi \cdot \tilde{z}_m + (1 - \pi) \cdot \tilde{z}_c, \quad \pi = \sigma(\theta) \in (0, 1). \tag{18}$$

Degenerate cases.

- $\pi = 0$: $Y(0) = \tilde{z}_c$, equivalent to independent channel-wise Transformers (no cross-channel interactions).
- $\pi = 1$: $Y(1) = \tilde{z}_m$, equivalent to fully shared cross-channel attention (ignoring per-channel dynamics).

Thus, DA^2 strictly generalizes both extremes.

Lemma (convex combination and stability). For any $\pi \in (0, 1)$,

$$\|Y(\pi)\| \leq \pi\|\tilde{z}_m\| + (1-\pi)\|\tilde{z}_c\|, \tag{19}$$

implying stability and boundedness. The output lies in the convex hull of the two attention branches, ensuring that DA^2 cannot underperform both simultaneously.

Proposition (Lipschitz inheritance). If the token- and channel-attention maps are L_c- and L_m-Lipschitz w.r.t. inputs, then for any fixed $\pi \in [0, 1]$, $Y(\pi)$ is $L(\pi)$-Lipschitz with $L(\pi) \leq (1-\pi)L_c + \pi L_m$. *Proof.* By triangle inequality and linearity of the convex mixing.

Proposition (richness via convex blending). Let $\mathcal{H}_c, \mathcal{H}_m$ be hypothesis classes realized by the two branches. Then the closure of $\mathcal{H}_{\text{DA}^2} = \{\pi h_m + (1-\pi)h_c\}$ under composition with standard MLP blocks strictly contains $\mathcal{H}_c \cup \mathcal{H}_m$ provided $\mathcal{H}_c \not\subset \mathcal{H}_m$ and $\mathcal{H}_m \not\subset \mathcal{H}_c$. *Sketch.* There exist functions realizable only by mixtures of h_c and h_m (e.g., requiring simultaneous temporal and cross-channel interactions). Post-mixing MLPs preserve separability, yielding strictly larger expressivity.

Expressivity. Consider the hypothesis class $\mathcal{H}_c$ defined by token-attention and $\mathcal{H}_m$ defined by channel-attention. Then

$$\mathcal{H}_{\text{DA}^2} = \{\pi h_m + (1-\pi)h_c : h_m \in \mathcal{H}_m,\ h_c \in \mathcal{H}_c,\ \pi \in (0, 1)\}. \tag{20}$$

This is strictly larger than $\mathcal{H}_c \cup \mathcal{H}_m$, since convex combinations allow intermediate solutions that neither pure axis can represent alone. In other words, DA^2 spans a richer functional space without increasing asymptotic complexity.

A.3 Analysis and Visualization of Event-boundary Segmentation

The event boundary detector introduces three trainable factors: (i) spectral smoothing kernel g_θ, (ii) softmax temperature τ, and (iii) sharpness γ.

Spectral smoothing. g_θ acts as a localized convolution over the frequency axis, emphasizing task-relevant bands. This is equivalent to learning a prior over plausible periodicities.

Soft frequency selection. The softmax distribution

$$\alpha_f = \frac{\exp(P_{\text{sm}}(f)/\tau)}{\sum_{f'} \exp(P_{\text{sm}}(f')/\tau)} \tag{21}$$

ensures differentiability. Lower τ sharpens α into hard frequency selection, while higher τ encourages broader distributions. During training, τ adapts to balance stability and discriminability.

Differentiable comb scoring. By raising the cosine comb to a learnable exponent γ, the segmenter interpolates between smooth sinusoidal modulation ($\gamma \approx 1$) and sharp periodic spikes ($\gamma \gg 1$). This provides a continuous control of boundary sparsity.

Visualization. The segmentation process demonstrates each stage: raw spectrum to smoothed spectrum to softmax weighting to cosine comb peaks to event boundaries. This progression highlights that segmentation is not a fixed heuristic but a differentiable, learnable module, as evidenced by the boundary alignment with natural signal dynamics shown in Figures 6 to 12.

A.4 Masked softmax with padding

Let $A \in \mathbb{R}^{N\times N}$ be attention logits and $m \in \{0, 1\}^N$ a binary keep-mask (1 for valid, 0 for padded). Define the masked logits

$$\tilde{A}_{ij} = \begin{cases} A_{ij}, & m_j = 1, \\ -\infty, & m_j = 0, \end{cases}$$

and the masked-softmax as

$$\text{softmax}_j(\tilde{A}_{ij}) = \frac{\exp(\tilde{A}_{ij})}{\sum_{k:m_k=1} \exp(\tilde{A}_{ik})}.$$

Equivalently, one can compute $\mathrm{softmax}(A + (1 - m) \cdot (-M))$ with a large $M \gg 0$. In our implementation for DA^2 attention, the per-channel per-batch mask μ provides m along the token axis for token-attention and along the channel-token pairing for channel-attention. This guarantees that padded positions neither receive nor contribute probability mass.

A.5 Information Preservation in Temporal Reprojection

The reprojection operator maps event-level embeddings $z_{\text{slow}} \in \mathbb{R}^{M \times d}$ to time-resolved outputs $z_{\text{full}}(t)$:

$$z_{\text{full}}(t) = \sum_{i=1}^{M} w_{t,i}\, z_{\text{slow}}(i), \quad \sum_i w_{t,i} = 1,\ w_{t,i} \geq 0. \tag{22}$$

Lemma (convexity and boundedness). Since $z_{\text{full}}(t)$ is a convex combination, for any norm $\|\cdot\|$,

$$\|z_{\text{full}}(t)\| \ \leq\ \sum_{i=1}^{M} w_{t,i}\, \|z_{\text{slow}}(i)\| \ \leq\ \max_i \|z_{\text{slow}}(i)\|. \tag{23}$$

Thus reprojection does not inflate magnitudes beyond the convex hull of the inputs.

Proposition (approximation error bound). Let $h_{\text{high}}(t)$ denote the high-dimensional fast representation at time t. Then

$$\|z_{\text{full}}(t) - h_{\text{high}}(t)\|_2 \ \leq\ \sum_{i=1}^{M} w_{t,i}\, \|z_{\text{slow}}(i) - h_{\text{high}}(t)\|_2 \ \leq\ \max_i \|z_{\text{slow}}(i) - h_{\text{high}}(t)\|_2. \tag{24}$$

This shows that the reprojection error is bounded by the convex combination (and hence by the maximum) of per-segment discrepancies, and does not grow with sequence length.

Theorem (exactness for piecewise-constant signals). Suppose the time axis is partitioned by the detector into M segments and $z_{\text{slow}}(i)$ equals the segment-wise mean of $h_{\text{high}}(t)$ on segment i. If $w_{t,i}$ are the standard barycentric weights induced by segment lengths (row-stochastic and segment-local), then $z_{\text{full}}(t) = h_{\text{high}}(t)$ for any piecewise-constant h_{high} aligned with the segmentation. *Proof.* On each segment the mean equals the constant value; barycentric reconstruction reproduces the constant exactly, and off-segment weights vanish.

Theoretical Analysis. Temporal reprojection can be viewed as a form of predictive coding, where abstract hypotheses $z_{\text{full}}(t)$ are continuously projected back to the temporal stream, and reconstruction errors serve as alignment signals. This guarantees both *fidelity* (preserving local detail) and *consistency* (maintaining event-level abstraction).

Properties of boundary-guided reprojection weights

We now justify the definition of $w_{t,i}$ constructed from $(p_{\text{full}}, b_{\text{full}})$.

Setup. Let $\{\mathcal{S}_i = [s_i, e_i]\}_{i=1}^{M}$ be a partition of the time axis induced by b_{full} (consecutive ones indicate boundaries). For any t, define unnormalized segment-local weights

$$\tilde{w}_{t,i} = \begin{cases} \kappa\big(\operatorname{dist}(t; s_i, e_i)\big)\, \bar{p}(t), & t \in \mathcal{S}_i, \\ 0, & \text{otherwise}, \end{cases}$$

where $\kappa : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is bounded and nonincreasing, and $\bar{p}(t) \in [0, 1]$ is a channel-aggregated soft confidence from p_{full}. Set

$$w_{t,i} = \frac{\tilde{w}_{t,i}}{\sum_{j=1}^{M} \tilde{w}_{t,j}} \quad \text{whenever } \sum_j \tilde{w}_{t,j} > 0, \quad \text{and} \quad w_{t,i} = \frac{\mathbf{1}\{t \in \mathcal{S}_i\}}{\#\{j : t \in \mathcal{S}_j\}} \text{ otherwise.}$$

Lemma (nonnegativity, locality, partition-of-unity). For every t, $w_{t,i} \geq 0$, $w_{t,i} = 0$ if $t \notin \mathcal{S}_i$, and $\sum_{i=1}^{M} w_{t,i} = 1$. *Proof.* Nonnegativity and locality follow from $\tilde{w}_{t,i} \geq 0$ and its definition. When $\sum_j \tilde{w}_{t,j} > 0$, normalization yields a convex combination with unit sum. If the denominator vanishes (measure-zero edge case only when $\bar{p}(t) = 0$ for all active segments), the fallback uniform average over active segments preserves unit sum.

Lemma (stability). If κ is bounded by K and Lipschitz with constant L_κ, and $\bar{p}$ is bounded and Lipschitz with constant L_p, then $w_{t,i}$ is bounded and piecewise-Lipschitz in t away from segment boundaries. *Sketch.* Products and sums of Lipschitz functions preserve Lipschitzness; division by a denominator bounded away from zero on each segment interior preserves regularity.

Proposition (consistency with segmentation). Suppose $z_{\text{slow}}(i)$ summarizes segment $\mathcal{S}_i$ (e.g., mean of h_{high} on $\mathcal{S}_i$). Then $z_{\text{full}}(t)$ is a segment-local convex interpolation of adjacent segment summaries and thus cannot introduce off-segment leakage. *Proof.* By locality and partition-of-unity, only indices i with $t \in \mathcal{S}_i$ contribute, and the coefficients form a convex combination.

Theorem (exactness for piecewise-constant signals). If h_{high} is piecewise constant on $\{\mathcal{S}_i\}$ and $z_{\text{slow}}(i)$ equals the segment mean, then with any segment-local $w_{t,i}$ as above that is constant on each segment (e.g., $\kappa \equiv 1$, constant $\bar{p}$ per segment), one has $z_{\text{full}}(t) = h_{\text{high}}(t)$ for all t. *Proof.* On $\mathcal{S}_i$, $h_{\text{high}}(t) \equiv c_i$ and $z_{\text{slow}}(i) = c_i$. Since $w_{t,j} = 0$ for $j \neq i$ and $\sum_j w_{t,j} = 1$, we obtain $z_{\text{full}}(t) = w_{t,i} c_i = c_i = h_{\text{high}}(t)$.

These results justify the boundary-guided construction: it yields nonnegative, local, normalized weights tied to detected events, admits smooth interpolations via κ and $\bar{p}$, and recovers exact reconstruction for signals aligned with the learned segmentation.

A.6 Event-Driven Segmentation Visualization

Enhanced spectral flux (ESF). To compare with the boundary proposal of PeCo-TS, we compute a spectral change cue that emphasizes onsets and regime shifts. Let $S_t(f)$ denote the magnitude spectrum at time t and frequency f, obtained from a short-time FFT over the original input x with a Hann window. We apply spectral whitening using a robust per-band statistic $M(f)$ (median over a local temporal window) and bandlimited smoothing h along the frequency axis:

$$\hat{S}_t(f) = \frac{S_t(f)}{M(f) + \varepsilon}, \quad \tilde{S}_t(f) = (h * \hat{S}_t)(f). \tag{25}$$

The enhanced spectral flux is the half-wave rectified frame-to-frame spectral increment and normalized to $[0, 1]$ across t, optionally with frequency weights $w(f)$:

$$\mathrm{ESF}(t) = \sum_f w(f) \left[\tilde{S}_t(f) - \tilde{S}_{t-1}(f) \right]_+, \quad [x]_+ = \max(x, 0). \tag{26}$$

Figures 6 to 12 show that the event boundary detector places boundaries at semantically meaningful transitions across datasets. On ETTh1, peaks cluster around daily and weekly regime shifts; on Traffic, boundaries concentrate at rush-hour onsets and weekend changes. The ETTm1/m2 and ETTh2 results indicate cross-resolution robustness, adapting segment lengths to mid- versus low-frequency rhythms. Weather boundaries densify near storm fronts, and Exchange boundaries align with volatility bursts and macro events. This adaptivity avoids both under- and over-segmentation, preserving coherent events while minimizing token count.

A.7 Multi-task Evaluation

We report MSE/MAE for forecasting and imputation, accuracy for classification, and precision/F1 for anomaly detection. Training uses PyTorch with Adam optimizer (lr=1e-4, batch size 32); event segmentation combines FFT, autocorrelation, and Hilbert transforms; DA^2 Attention employs eight heads with dataset-adaptive gating parameter π. All experiments run on RTX 3090 GPUs. Complete results are shown in Table 2– 5. The best results are highlighted in **red** and the second best are shown in **blue**. Among the various models, PeCo-TS exhibits superior multitask performance. To provide a clear comparison among different models, we list supplementary prediction showcases of three representative datasets in Figures 13–15.

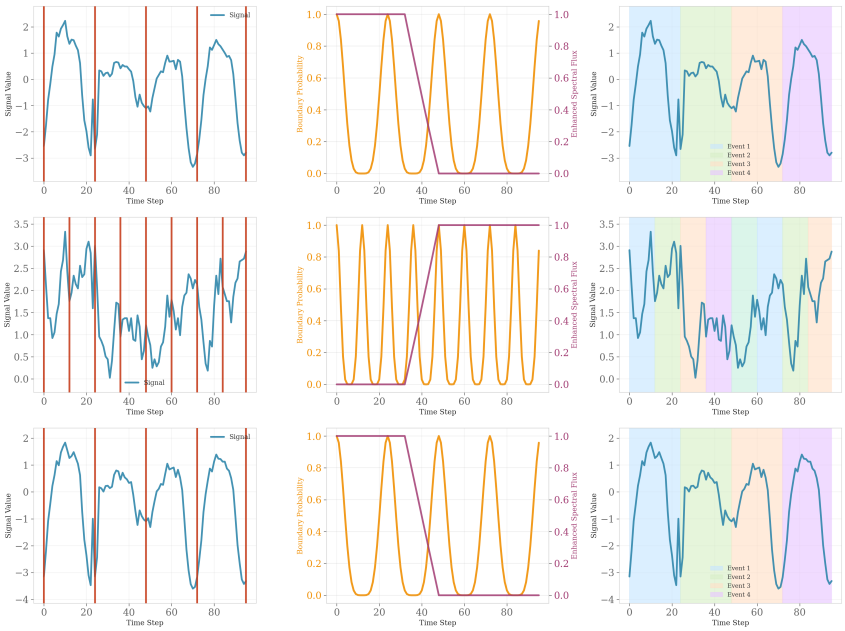

Figure 6: ETTh1 segmentation with an event boundary detector. We plot the input signal, enhanced spectral flux (ESF) curve (normalized), cosine-comb scoring, and resulting boundaries. ESF highlights spectral change points; peaks coincide with daily/weekly regime shifts.

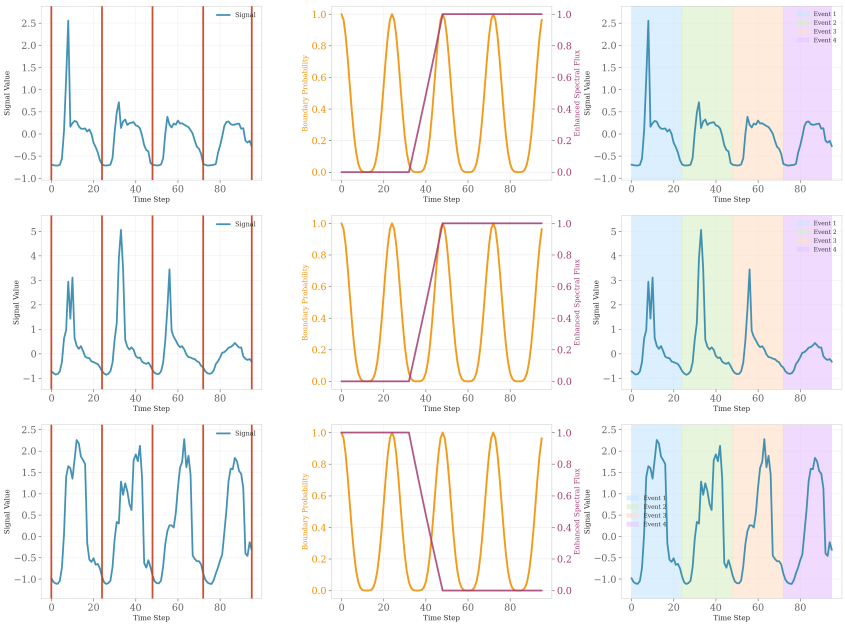

Figure 7: Traffic segmentation with an event boundary detector. ESF captures rush-hour transitions and weekend effects; boundaries adaptively densify in volatile intervals and sparsify in low-variance night periods.

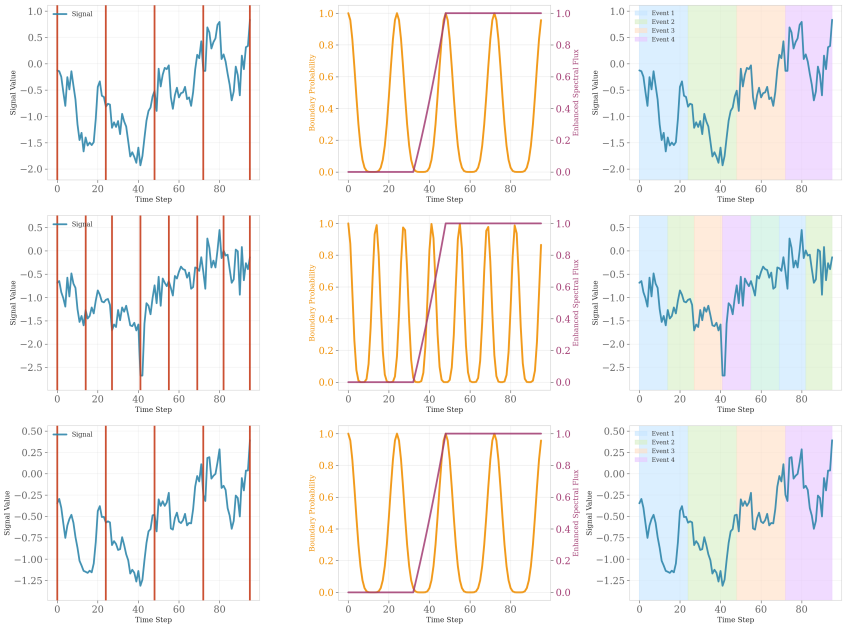

Figure 8: ETTh2 segmentation analysis. ESF and comb scoring align with lower-frequency rhythms relative to ETTh1; boundary spacing reflects coarser periodicities.

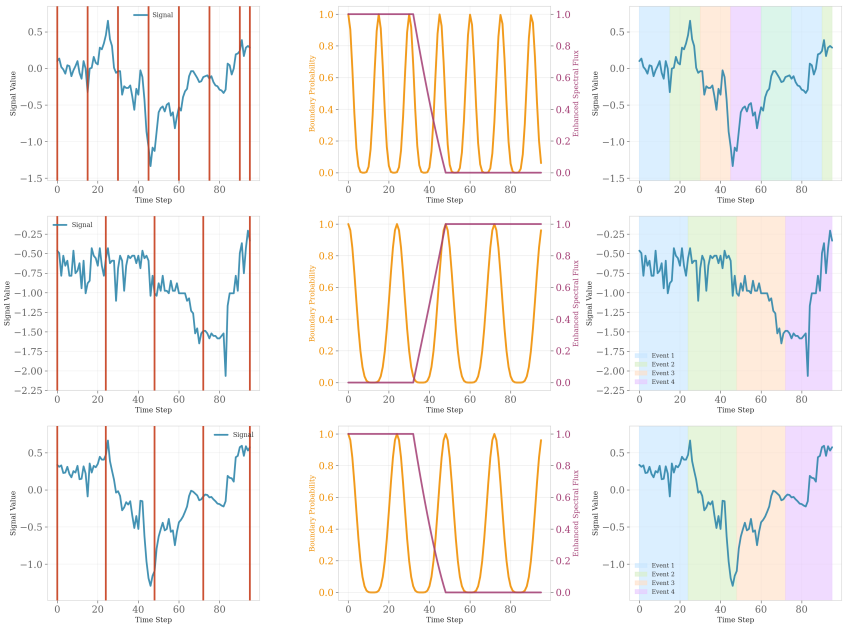

Figure 9: ETTm1 segmentation analysis. Minute-level series exhibits mid-frequency rhythms; ESF peaks are more frequent than hourly datasets, yielding finer-grained event tokens.

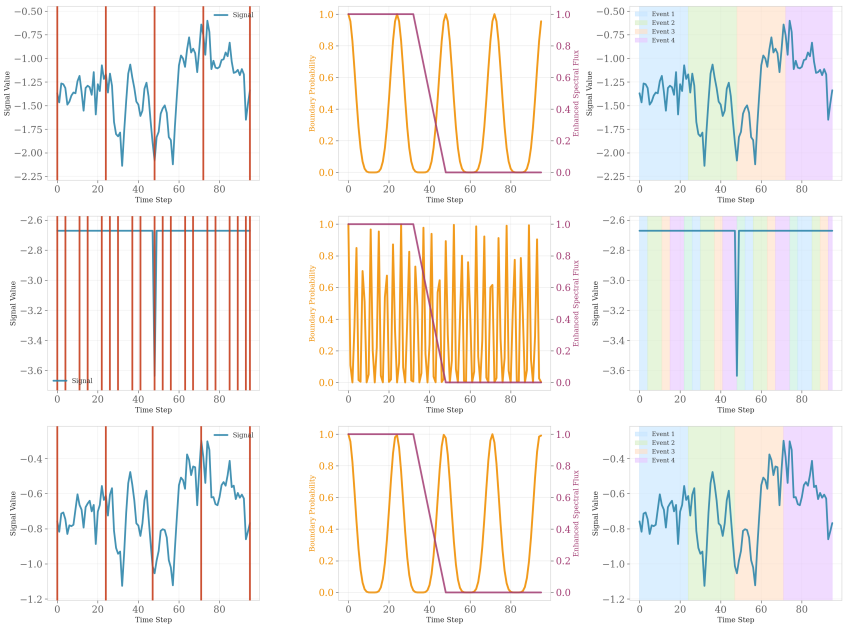

Figure 10: ETTm2 segmentation analysis. Similar to ETTm1 with dataset-specific periodicities; learnable smoothing adapts to suppress spurious high-frequency flux.

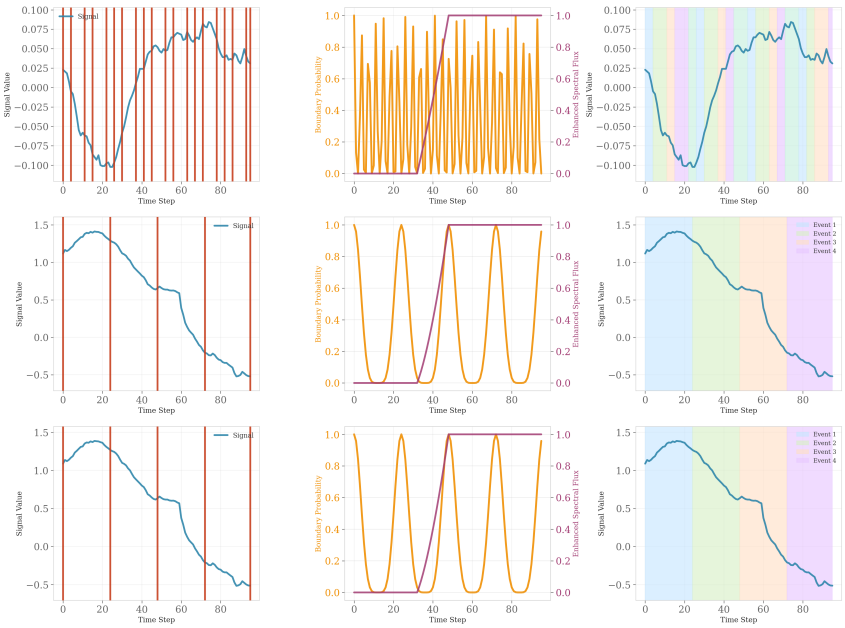

Figure 11: Weather segmentation analysis. ESF peaks densify near synoptic events (fronts/storms), indicating sensitivity to transient meteorological regimes beyond simple diurnal periodicity.

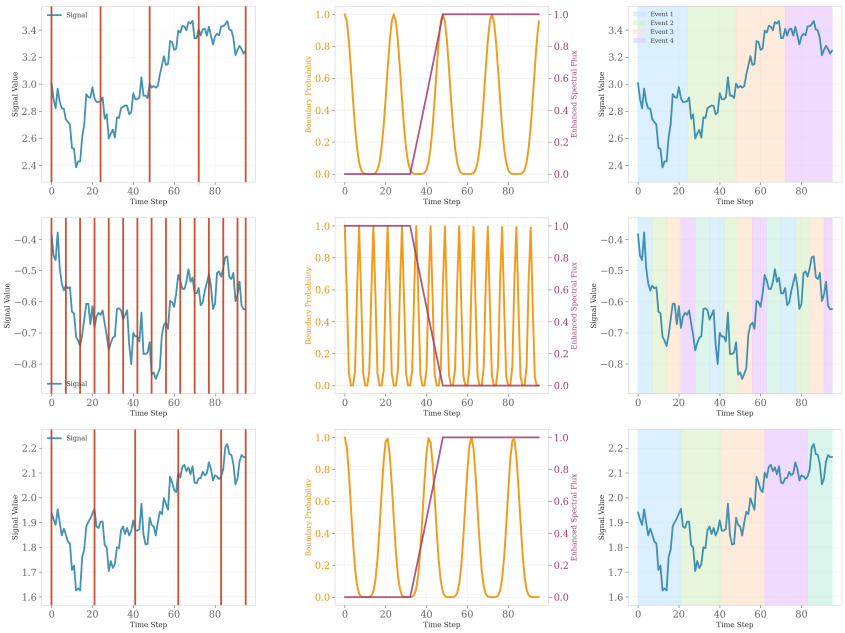

Figure 12: Exchange segmentation analysis. ESF highlights volatility bursts; boundaries concentrate around macroeconomic announcements and major market moves, while remaining sparse during stable phases.

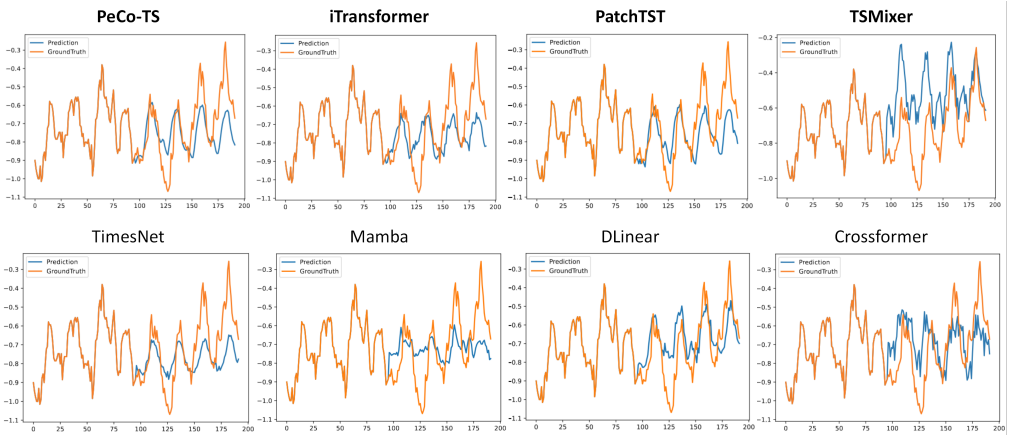

Figure 13: Visualization of input-96-predict-96 results on the ETTh1 dataset.

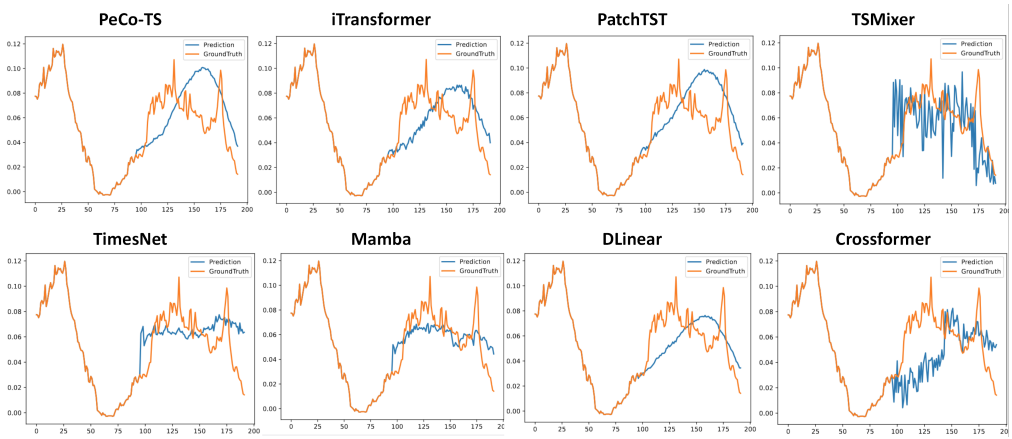

Figure 14: Visualization of input-96-predict-96 results on the Weather dataset.

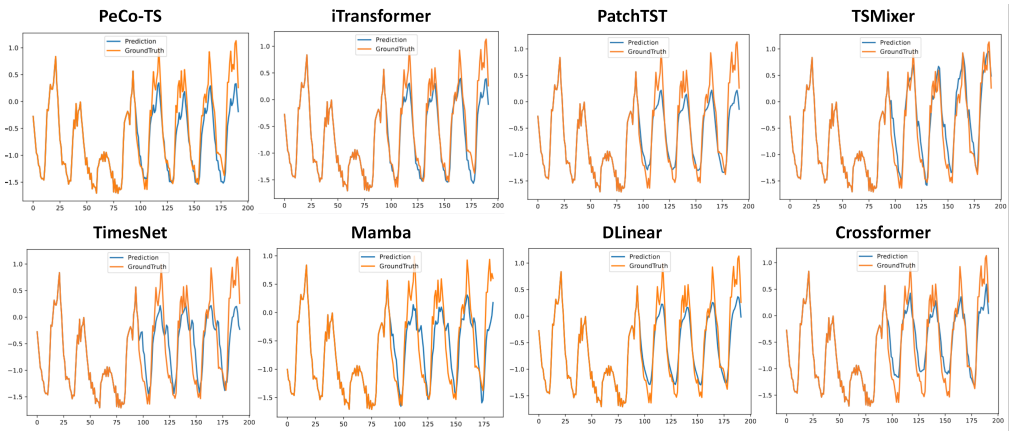

Figure 15: Visualization of input-96-predict-96 results on the ECL dataset.

Table 2: Multivariate forecasting results with prediction lengths $S \in \{96, 192, 336, 720\}$ for all datasets and fixed lookback length $T = 96$.

Models		PeCo-TS		iTransformer		PatchTST		TSMixer		TimesNet		Mamba		DLinear		FEDformer		Crossformer		Autoformer	
Metric	PredLen	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.369**	**0.396**	0.395	0.409	0.377	0.397	0.494	0.502	0.389	0.412	0.486	0.452	0.396	0.411	0.411	0.440	0.421	0.438	0.417	0.437
	192	**0.409**	**0.422**	0.449	0.441	0.425	0.428	0.581	0.557	0.439	0.442	0.555	0.506	0.445	0.440	0.451	0.462	0.466	0.465	0.505	0.476
	336	**0.440**	**0.438**	0.492	0.465	0.473	0.458	0.677	0.618	0.494	0.471	0.537	0.500	0.487	0.465	0.529	0.504	0.683	0.608	0.537	0.505
	720	**0.444**	**0.454**	0.522	0.504	0.518	0.501	0.752	0.674	0.518	0.494	0.624	0.577	0.513	0.510	0.576	0.538	0.732	0.631	0.516	0.511
ETTh2	96	**0.288**	**0.342**	0.300	0.350	0.295	0.347	1.056	0.807	0.330	0.370	0.347	0.378	0.341	0.395	0.338	0.381	0.594	0.580	0.344	0.384
	192	**0.375**	**0.396**	0.382	0.400	0.376	0.398	2.587	1.403	0.394	0.410	0.455	0.445	0.482	0.479	0.416	0.429	1.059	0.718	0.432	0.441
	336	**0.416**	**0.423**	0.424	0.432	0.421	0.433	2.407	1.347	0.471	0.468	0.429	0.443	0.593	0.542	0.460	0.470	1.394	0.882	0.459	0.466
	720	0.433	**0.437**	**0.426**	0.445	0.431	0.453	2.051	1.218	0.442	0.452	0.541	0.497	0.840	0.661	0.483	0.488	1.749	1.003	0.475	0.486
ETTm1	96	**0.320**	**0.360**	0.341	0.376	0.324	0.365	0.479	0.470	0.336	0.376	0.372	0.391	0.346	0.374	0.405	0.433	0.363	0.403	0.593	0.508
	192	**0.329**	**0.366**	0.381	0.395	0.365	0.386	0.480	0.482	0.387	0.402	0.436	0.421	0.382	0.391	0.432	0.446	0.393	0.419	0.562	0.500
	336	**0.372**	**0.389**	0.419	0.419	0.393	0.408	0.541	0.525	0.414	0.422	0.558	0.511	0.415	0.415	0.460	0.464	0.459	0.460	0.644	0.542
	720	**0.457**	**0.438**	0.486	0.456	0.460	0.443	0.616	0.574	0.513	0.472	0.625	0.548	0.473	0.451	0.527	0.496	0.716	0.644	0.634	0.538
ETTm2	96	0.181	0.264	0.184	0.267	**0.178**	**0.260**	0.250	0.366	0.188	0.268	0.196	0.275	0.193	0.293	0.194	0.283	0.305	0.377	0.218	0.300
	192	**0.246**	0.308	0.253	0.312	0.247	0.308	0.492	0.559	0.252	**0.307**	0.302	0.342	0.285	0.361	0.260	0.320	0.459	0.472	0.277	0.335
	336	0.311	0.350	0.315	0.352	**0.309**	0.347	0.833	0.735	0.317	**0.346**	0.372	0.393	0.385	0.429	0.320	0.357	0.647	0.632	0.336	0.373
	720	0.411	0.409	0.412	0.406	**0.407**	**0.403**	2.544	1.352	0.421	0.406	0.637	0.510	0.556	0.523	0.420	0.416	1.686	1.010	0.445	0.432
Electricity	96	**0.179**	**0.261**	0.196	0.281	0.189	0.277	0.200	0.305	0.276	0.358	0.188	0.290	0.210	0.301	0.228	0.339	0.198	0.299	0.290	0.387
	192	**0.184**	**0.267**	0.206	0.293	0.193	0.283	0.220	0.331	0.285	0.367	0.204	0.308	0.210	0.305	0.269	0.366	0.226	0.319	0.252	0.356
	336	**0.200**	**0.284**	0.226	0.313	0.209	0.298	0.242	0.353	0.296	0.378	0.207	0.313	0.223	0.319	0.289	0.383	0.757	0.719	0.285	0.381
	720	0.241	**0.317**	0.270	0.347	0.251	0.331	0.271	0.372	0.333	0.402	**0.237**	0.333	0.258	0.350	0.331	0.413	0.813	0.799	0.399	0.455
Exchange	96	**0.082**	**0.203**	0.087	0.207	0.084	0.203	0.232	0.388	0.111	0.238	0.127	0.258	0.094	0.227	0.162	0.293	0.215	0.342	0.156	0.287
	192	0.190	0.312	**0.180**	0.303	0.181	**0.302**	0.464	0.549	0.209	0.333	0.287	0.391	0.186	0.325	0.282	0.387	0.507	0.566	0.276	0.383
	336	0.335	0.423	0.333	**0.419**	0.337	0.421	0.754	0.720	0.374	0.448	0.651	0.603	**0.327**	0.435	0.443	0.488	0.764	0.711	0.473	0.514
	720	0.942	0.735	0.856	0.700	0.875	0.703	**0.705**	0.701	0.931	0.735	1.706	0.970	0.749	**0.664**	1.147	0.824	1.150	0.873	1.111	0.820
Traffic	96	**0.506**	**0.327**	0.574	0.386	0.509	0.331	0.578	0.388	0.868	0.499	0.679	0.383	0.696	0.429	0.653	0.420	0.516	0.367	0.701	0.447
	192	0.517	**0.332**	0.584	0.390	**0.514**	0.337	0.579	0.394	0.919	0.537	0.645	0.367	0.646	0.407	0.660	0.417	0.543	0.385	0.714	0.450
	336	**0.521**	0.338	0.613	0.405	0.522	**0.334**	0.604	0.409	0.898	0.514	0.636	0.360	0.653	0.410	0.740	0.457	0.583	0.402	0.708	0.451
	720	0.563	**0.352**	0.676	0.434	**0.558**	0.353	0.664	0.439	0.927	0.542	0.755	0.414	0.694	0.429	0.798	0.488	0.615	0.423	0.723	0.444
Weather	96	0.178	**0.220**	0.183	0.225	0.179	0.220	**0.163**	0.241	0.180	0.225	0.195	0.243	0.196	0.257	0.212	0.283	0.189	0.259	0.261	0.328
	192	0.227	0.263	0.234	0.266	0.228	**0.260**	**0.210**	0.285	0.227	0.266	0.252	0.291	0.236	0.294	0.299	0.356	0.258	0.281	0.308	0.359
	336	0.281	0.299	0.287	0.304	0.279	**0.297**	**0.267**	0.333	0.281	0.304	0.327	0.342	0.283	0.333	0.318	0.359	0.277	0.321	0.355	0.392
	720	0.355	0.349	0.362	0.352	0.355	**0.347**	**0.332**	0.379	0.360	0.355	0.406	0.385	0.347	0.384	0.405	0.412	0.398	0.415	0.428	0.431

Table 3: Time-series classification results on UCR/UEA benchmarks. Metric is Accuracy (%, higher is better). All methods follow dataset-standard train/test splits and z-score normalization.

Dataset	PeCo-TS	iTransformer	PatchTST	DLinear	FEDformer	Crossformer	Autoformer
EthanolConcentration	**0.3270**	0.2852	0.2814	0.2928	0.2776	0.3030	0.2433
FaceDetection	0.6831	0.6654	**0.6864**	0.6822	0.6751	0.6512	0.5951
JapaneseVowels	0.9676	**0.9757**	0.9595	0.9649	0.9674	0.9757	0.9649
SelfRegulationSCP1	0.9144	**0.9215**	0.8737	0.9147	0.5802	0.9147	0.5631
SelfRegulationSCP2	**0.5560**	0.5444	0.5278	0.5444	0.5278	0.5467	0.5333
SpokenArabicDigits	**0.9818**	0.9804	0.9741	0.9650	0.9782	0.9627	0.9759
UWaveGestureLibrary	0.7919	0.8594	**0.8625**	0.8219	0.5656	0.8531	0.5000

A.8 Segmentation Method Comparison

Across ETTh1, ETTm1, and Weather (Figures 16 to 18), the event boundary approach consistently produces cleaner, more stable boundaries than fixed windows or heuristic detectors. Competing methods either miss critical regime shifts (under-segmentation) or fragment coherent trends (over-segmentation), while our method achieves tighter alignment with intrinsic periodicities, which later translates into lower forecasting error and better anomaly localization.

A.9 Fast-Slow Path Comparison

The Fast path preserves high-frequency cues, improving short-horizon fidelity, while the Slow path enforces long-range consistency via event abstractions. Figure 19 shows complementary error profiles; combining both reduces both bias (trend errors) and variance (spiky mispredictions).

A.10 DA^2 Ablations

DA^2 adaptively allocates capacity between intra-series and inter-series attention. Figure 20 confirms consistent gains over fixed CI/CD strategies across datasets. Learned allocations correlate with dataset structure: higher inter-series emphasis on Electricity/Traffic (strong cross-channel coupling), and higher intra-series emphasis on ETT variants (dominant per-channel temporal patterns).

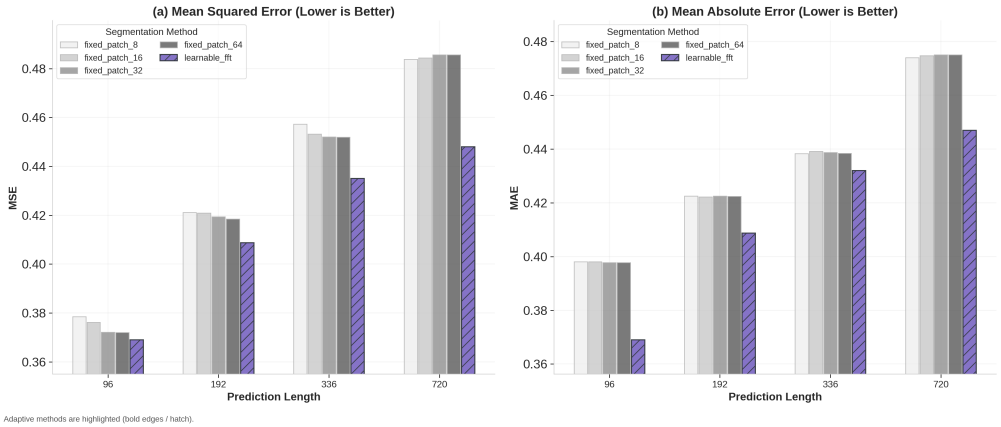

Figure 16: Segmentation method comparison on ETTh1. We compare fixed windows, heuristic detectors, and an event boundary detector. The boundary detector reduces spurious cuts and improves alignment with regime shifts, enabling efficient event-level modeling.

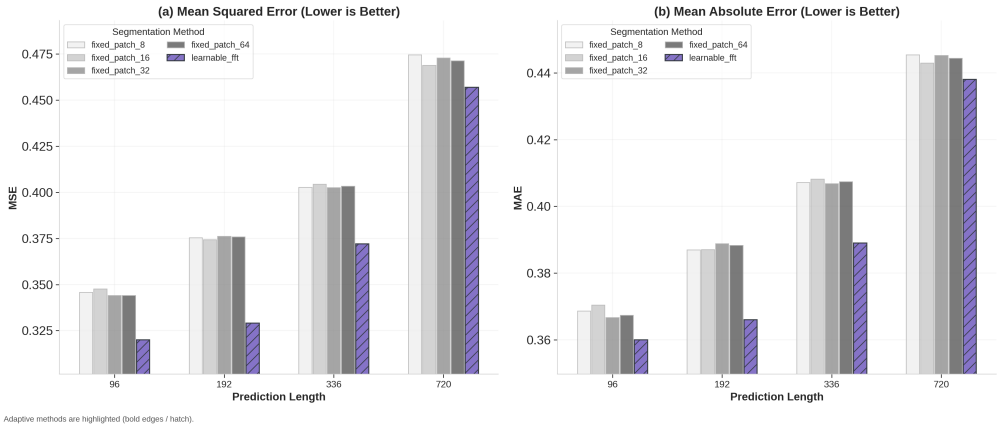

Figure 17: Segmentation method comparison on ETTm1. Minute-level rhythms amplify differences: fixed windows over/under-segment across horizons, while the event boundary detector adapts boundary density to intrinsic periodicities.

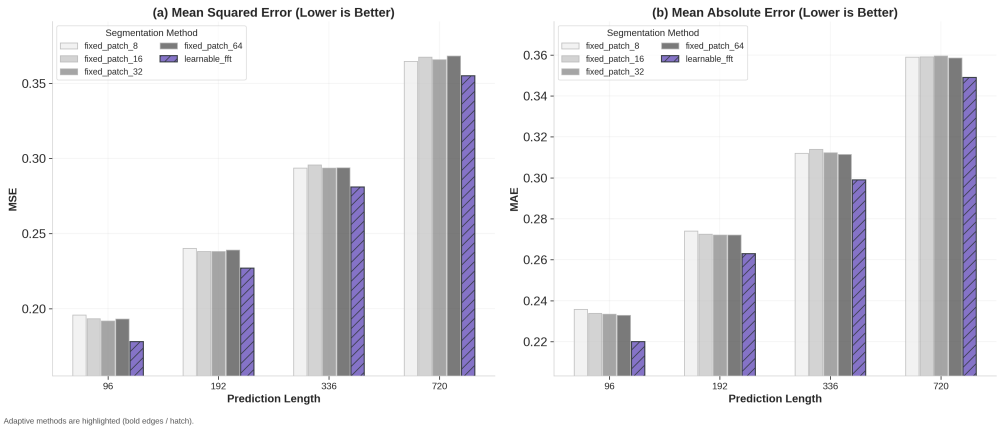

Figure 18: Segmentation method comparison on Weather. Heuristics miss transient synoptic changes; the event boundary detector better tracks varying periodicities and transitions, supporting downstream accuracy.

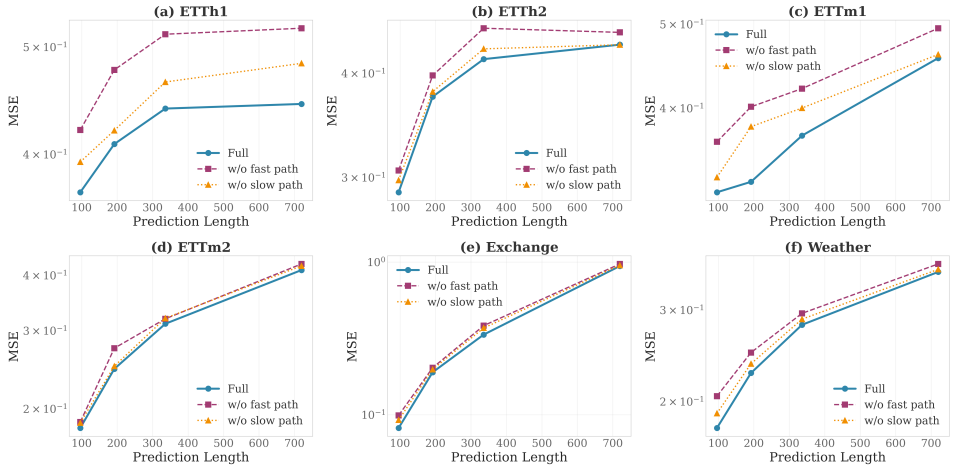

Figure 19: Fast vs. Slow path comparison. Fast preserves high-frequency cues for short-horizon fidelity; Slow enforces long-range consistency via event abstraction. Fusion reduces both bias and variance across datasets.

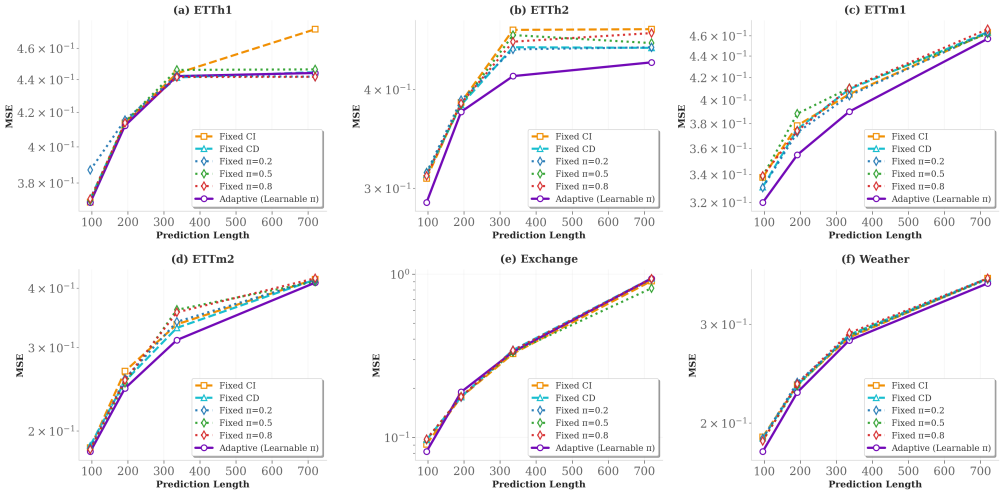

Figure 20: DA^2 vs. fixed CI/CD channel strategies. Adaptive gating π learns dataset-specific allocations, outperforming fixed extremes across multivariate benchmarks.

Table 4: Anomaly detection results on MSL, PSM, SMAP, SMD, and SWAT. We report Precision and F1 (higher is better) under the standard contiguous-window detection protocol; thresholds selected on validation splits.

Dataset	PeCo-TS		iTransformer		PatchTST		DLinear		FEDformer		Crossformer		Autoformer	
Metric	Precision	F1-Score	Precision	F1-Score	Precision	F1-Score	Precision	F1-Score	Precision	F1-Score	Precision	F1-Score	Precision	F1-Score
MSL	0.9036	**0.8331**	0.8615	0.7253	0.8860	0.7913	0.8969	0.8187	**0.9068**	0.8230	0.9031	0.8060	0.9054	0.8187
PSM	0.9864	0.9626	0.9797	0.9532	0.9910	0.9626	0.9864	**0.9661**	**0.9999**	0.9007	0.9729	0.9239	0.9999	0.8823
SMAP	**0.9330**	**0.8615**	0.9069	0.6675	0.8988	0.6726	0.8987	0.6729	0.9015	0.6863	0.8998	0.6874	0.9127	0.7411
SMD	**0.7891**	**0.8482**	0.7627	0.8057	0.7648	0.8104	0.7612	0.8007	0.7278	0.7688	0.7204	0.7758	0.7298	0.7723
SWAT	0.9810	**0.9635**	0.9221	0.9265	0.9124	0.8673	0.9227	0.9266	0.9995	0.7918	0.9782	0.9063	**0.9996**	0.7918

Table 5: Imputation results on ETTh/ETTm/Electricity/Weather. We report MSE/MAE (lower is better) under random missingness with multiple mask rates.

Dataset	Mask Ratio	PeCo-TS		iTransformer		PatchTST		DLinear		FEDformer		Crossformer		AutoFormer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Etth1	12.5%	0.1010	0.2089	0.0987	0.2218	0.0928	0.2010	0.1117	0.2319	**0.0735**	**0.1958**	0.1064	0.2283	0.0978	0.2266
	25.0%	0.1230	0.2315	0.1250	0.2504	0.1063	**0.2165**	0.1496	0.2690	**0.1055**	0.2365	0.1172	0.2427	0.1208	0.2516
	37.5%	0.1470	0.2540	0.1573	0.2818	**0.1188**	**0.2306**	0.1874	0.3004	0.1411	0.2748	0.1293	0.2572	0.1550	0.2851
	50.0%	0.1755	0.2755	0.2177	0.3325	**0.1403**	**0.2486**	0.2316	0.3326	0.2006	0.3293	0.1478	0.2764	0.2119	0.3348
Etth2	12.5%	0.0635	0.1615	0.0932	0.2080	**0.0570**	**0.1510**	0.1091	0.2229	0.1296	0.2437	0.1215	0.2310	0.1702	0.2887
	25.0%	0.0682	0.1692	0.1209	0.2395	**0.0620**	**0.1594**	0.1449	0.2593	0.1788	0.2911	0.1335	0.2435	0.2229	0.3302
	37.5%	0.0719	0.1727	0.1485	0.2650	**0.0674**	**0.1670**	0.1794	0.2895	0.2335	0.3323	0.1451	0.2565	0.2781	0.3632
	50.0%	0.0846	0.1914	0.1931	0.3026	**0.0736**	**0.1753**	0.2161	0.3186	0.3462	0.3988	0.1614	0.2710	0.3747	0.4198
Ettm1	12.5%	**0.0338**	0.1294	0.0456	0.1474	0.0396	**0.1280**	0.0556	0.1612	0.0448	0.1594	0.0436	0.1487	2.010	1.204
	25.0%	**0.0397**	**0.1260**	0.0605	0.1723	0.0420	0.1318	0.0766	0.1906	0.0531	0.1633	0.0466	0.1524	1.109	0.8591
	37.5%	**0.0450**	**0.1380**	0.0774	0.1959	0.0466	0.1390	0.0998	0.2175	0.0809	0.2013	0.0506	0.1580	0.3463	0.4382
	50.0%	**0.0517**	0.1470	0.1067	0.2316	0.0523	**0.1470**	0.1286	0.2463	0.1278	0.2545	0.0567	0.1677	0.3391	0.4195
ETTm2	12.5%	**0.0253**	**0.0911**	0.0518	0.1514	0.0254	0.0931	0.0662	0.1707	0.0601	0.1681	0.0557	0.1576	2.788	1.326
	25.0%	0.0277	0.0999	0.0707	0.1789	**0.0277**	**0.0982**	0.0893	0.2007	0.0921	0.2089	0.0741	0.1802	0.9562	0.7293
	37.5%	**0.0300**	**0.1010**	0.0915	0.2043	0.0301	0.1028	0.1117	0.2256	0.1328	0.2464	0.0796	0.1779	1.463	0.8603
	50.0%	0.0340	0.1150	0.1176	0.2327	**0.0332**	**0.1079**	0.1382	0.2514	0.2415	0.3297	0.0877	0.1861	0.6442	0.5610
ECL	12.5%	**0.0492**	**0.1413**	0.0724	0.1895	0.0526	0.1550	0.0844	0.2063	0.1808	0.3204	0.0640	0.1792	0.1875	0.3259
	25.0%	**0.0559**	**0.1521**	0.0898	0.2134	0.0623	0.1692	0.1131	0.2427	0.2020	0.3367	0.0716	0.1899	0.2123	0.3442
	37.5%	**0.0651**	**0.1654**	0.1068	0.2344	0.0726	0.1826	0.1412	0.2731	0.2205	0.3512	0.0804	0.2025	0.2289	0.3557
	50.0%	**0.0796**	**0.1853**	0.1259	0.2553	0.0874	0.2022	0.1726	0.3034	0.2425	0.3670	0.0901	0.2155	0.2600	0.3768
Weather	12.5%	**0.0285**	0.0555	0.0376	0.0858	0.0287	**0.0485**	0.0380	0.0885	0.0425	0.1033	0.2314	0.3437	0.0387	0.0947
	25.0%	0.0310	**0.0056**	0.0460	0.1054	**0.0310**	0.0531	0.0471	0.1074	0.0568	0.1305	0.1888	0.2963	0.0398	0.0973
	37.5%	**0.0330**	**0.0560**	0.0549	0.1209	0.0350	0.0588	0.0558	0.1216	0.0732	0.1575	0.1156	0.2205	0.0399	0.0967
	50.0%	**0.0360**	**0.0600**	0.0671	0.1407	0.0378	0.0626	0.0663	0.1368	0.1134	0.2095	0.1655	0.2691	0.0432	0.1017

A.11 π EVOLUTION

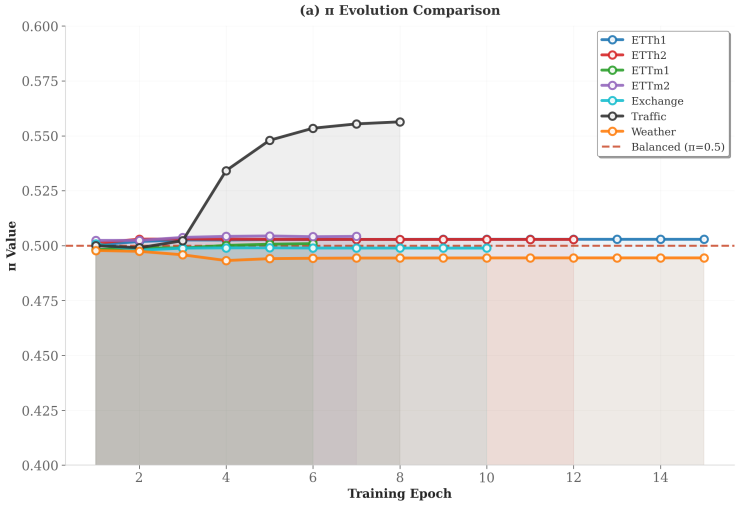

Figure 21: Training-time evolution of π. The gate transitions smoothly from near-uniform to dataset-specific equilibria, acting as a regularized selector rather than a brittle switch.

The gate π evolves smoothly during training from near-uniform to dataset-specific allocations (Figures 21). This behavior indicates a regularized selector rather than a brittle switch, stabilizing with-

out collapse. Per-dataset shifts reflect structural differences (e.g., sensor versus market data), explaining robust cross-dataset performance without architecture changes.

A.12 Model Efficiency

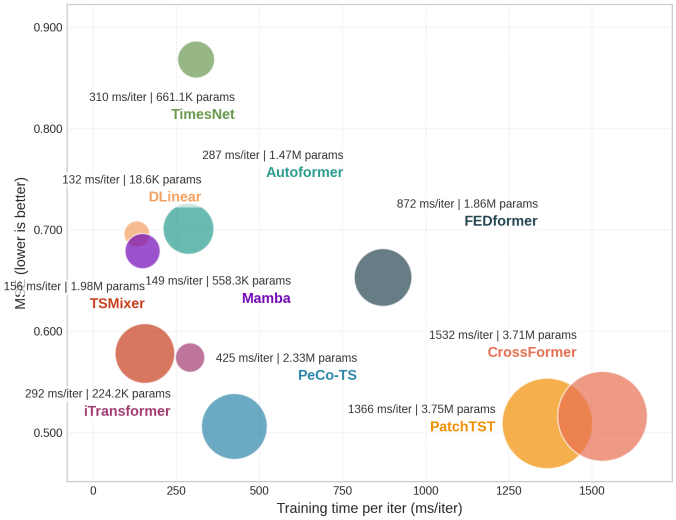

Figure 22: Model efficiency on Traffic (input-96, predict-96). PeCo-TS achieves higher accuracy with fewer parameters and lower latency than strong baselines, consistent with theoretical complexity reductions.

Under identical settings (input-96, predict-96), PeCo-TS attains higher accuracy with fewer parameters and lower latency (Figure 22). These empirical savings match the theoretical reduction from event-driven compression (Appendix A.1) and the practical ablations showing complementary contributions of Fast/Slow paths.