

000 PERCEIVE FAST, THINK SLOW: A COGNITIVE- 001 002 INSPIRED FRAMEWORK FOR TIME SERIES ANALYSIS 003

004 **Anonymous authors**

005 Paper under double-blind review

006 ABSTRACT

007 Time series modeling faces persistent challenges: fixed-window tokenization mis-
008 aligns with natural event boundaries, uniform computation wastes capacity on
009 simple patterns, and static architectures cannot adapt to diverse temporal depen-
010 dencies. We propose **PeCo-TS**, a cognitive-inspired framework that instantiates
011 the principle of “perceive fast, think slow” through three key innovations: (1)
012 *event-driven dynamic-length tokenization* that aligns tokens with semantic bound-
013 aries and reduces redundancy, (2) a *Slow–Fast dual-pathway architecture* that sep-
014 arates rapid perception of fine-grained variations from slower abstraction of event-
015 level structures, and (3) *Dual-Axis Adaptive (DA²) attention* that dynamically bal-
016 ances intra-series and inter-series dependencies via learnable gating. **Extensive**
017 **experiments across four** diverse tasks—forecasting, classification, anomaly detec-
018 tion, and imputation—demonstrate the broad applicability of PeCo-TS, yielding
019 consistent improvements over Transformer and linear baselines, including 5.6%
020 lower forecasting MSE, 9.3% lower imputation error, higher classification accu-
021 racy across UCR/UEA benchmarks, and a 6.7% relative F1 gain in anomaly de-
022 tection. Beyond accuracy, PeCo-TS achieves favorable efficiency–performance
023 trade-offs by leveraging event-level abstraction and complementary pathway syn-
024 ergy, while its learned boundaries align with real-world regime shifts, providing
025 interpretability. **These results establish PeCo-TS as a *strong supervised architecture***
026 **with excellent per-task performance for scenarios where task-specific training**
027 **data is available, offering a cognitively principled and efficient alternative to both**
028 **fixed-patching Transformers and computationally heavier multi-scale approaches.**

032 1 INTRODUCTION

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

041 Despite this complexity, most approaches still follow a rigid three-stage pipeline. First, they split a
042 series into fixed-size patches and treat each patch as a token. Second, a uniform architecture (e.g.,
043 self-attentive Transformer or MLP) assigns the same amount of compute to every token. Third,
044 task heads project hidden states to outputs (e.g., forecasting, classification, anomaly detection).
045 While convenient, this recipe clashes with heterogeneous real-world signals and leads to three limi-
046 tations: (i) *boundary misalignment*—fixed windows cut through meaningful events (e.g., crashes,
047 daily cycles, anomaly onsets), yielding incoherent representations (Nie et al., 2023; Wu et al.,
048 2023); (ii) *computational redundancy*—expensive attention is spent on simple trends while com-
049 plex patterns remain under-modeled (Zeng et al., 2023; Chen et al., 2023); and (iii) *limited adaptiv-
050 ity*—static channel handling cannot balance intra-series temporal dependencies against inter-series
051 cross-channel correlations (Zhou et al., 2023; Han et al., 2023).

052 Cognitive neuroscience provides a useful blueprint. Human perception operates through dual path-
053 ways: *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.

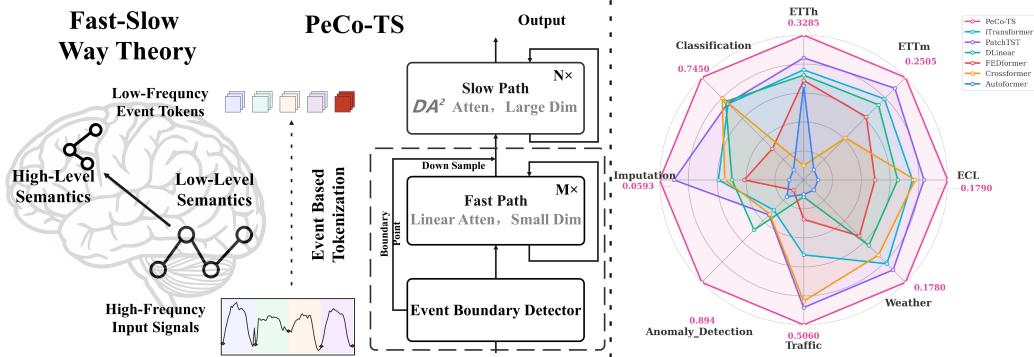


Figure 1: **Overview and highlights.** *Left:* Cognitive motivation and architecture of **PeCo-TS**, which integrates event-driven segmentation, a Fast Path for high-frequency perception, a Slow Path for event-level abstraction, and DA² attention for adaptive dependency modeling. *Right:* Aggregated 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 (DA²) 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 four** core time-series tasks (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 insights into temporal dynamics and validating the semantic relevance of our adaptive segmentation. Positioning PeCo-TS as a *strong supervised architecture* optimized for per-task performance (rather than a foundation model requiring large-scale pretraining), our key contributions are threefold: (1) a novel **event-driven dynamic-length tokenization** framework that fundamentally replaces fixed-window patching with boundary-aware segmentation, achieving 4.8 \times token compression while preserving semantic coherence; (2) a **Slow–Fast dual-pathway architecture** that separates rapid perception from conceptual abstraction, mirroring the brain’s perceive-fast, think-slow strategy **with 1.85 \times faster inference and 28% lower memory**; and (3) a **Dual-Axis Adaptive (DA²) attention** mechanism that dynamically balances intra-series and inter-series dependencies through learnable gating, **demonstrating consistent state-of-the-art performance across four diverse time-series tasks**.

2 RELATED WORK

Adaptive Tokenization and Multi-Resolution Modeling. While fixed-size patching (PatchTST (Nie et al., 2023), TimesNet (Wu et al., 2023)) remains dominant, recent work explores adaptive alternatives. MultiResFormer (Peršak et al., 2024) uses multiple fixed resolutions,

108 DeformableTST (Luo and Wang, 2024) adapts attention spans, and token merging (Götz et al.,
 109 2024) post-hoc merges existing tokens. Lightweight alternatives (Linear (Zeng et al., 2023),
 110 TSMixer (Chen et al., 2023)) reveal redundancy in uniform Transformers. Crucially, these differ
 111 from our approach: MultiResFormer requires predetermined scales; token merging operates
 112 post-hoc rather than learning boundaries from signal structure; deformable attention adjusts spans
 113 but not tokenization itself. Our learnable event-driven segmentation replaces fixed windows with
 114 frequency-guided boundaries that adapt end-to-end, providing semantically coherent tokenization
 115 with reduced cost. Section 4 provides controlled ablations against fixed patching and token merging
 116 baselines.

117

118 **Foundation Models and Multi-Task Learning.** Large-scale pre-trained models (TimesFM,
 119 Chronos, MOIRAI (Das et al., 2024; Shchur et al., 2024; Bhatnagar et al., 2024)) leverage diverse
 120 data for zero-shot generalization, few-shot learning, and cross-task transfer. Notably, UniTS (Gao
 121 et al., 2024) unifies predictive and generative tasks through task tokenization, achieving strong per-
 122 formance across 38 datasets—demonstrating the value of unified multi-task architectures. These
 123 foundation models excel at data efficiency and representational reuse but require extensive pretrain-
 124 ing on large-scale corpora. iTransformer (Zhou et al., 2023) models variables as tokens, while
 125 MCformer (Han et al., 2023) dynamically groups channels. PeCo-TS addresses a *complementary*
 126 direction: rather than competing with foundation models on transfer learning, we focus on *per-
 127 task supervised learning* where task-specific data is available, achieving superior accuracy through
 128 cognitively inspired architecture design without requiring large-scale pretraining. Our DA² attention
 129 adaptively balances intra- and inter-series correlations, outperforming both channel-independent and
 130 channel-dependent baselines.

131

132 **State-Space Models and Hybrid Architectures.** State-space models (SSMs) like Mamba have
 133 gained prominence for linear-time complexity in long sequences. TimeMachine (Ahamed and
 134 Cheng, 2024) applies Mamba to time series with superior scalability and memory efficiency, while
 135 SST (Xu et al., 2025) (CIKM 2025) introduces a hybrid Mamba-Transformer architecture with ex-
 136 pert modules that separate long-range patterns (Mamba) from short-term dynamics (Transformer).
 137 Diffusion models (DyDiff (Guo et al., 2025)) model temporal transitions for spatiotemporal pre-
 138 diction. Cognitive dual-pathway processing (Zacks and Swallow, 2007; Feichtenhofer et al., 2019)
 139 motivates “perceive fast, think slow” designs. PeCo-TS differs fundamentally: unlike SSMs oper-
 140 ating uniformly across timesteps or fixed-rate dual pathways, we integrate *learnable event-driven*
 141 *segmentation* with adaptive Fast-Slow processing where boundaries are end-to-end trainable and
 142 adapt to signal-specific rhythms, explicitly separating perception from abstraction with event-level
 143 efficiency.

144

145

3 METHODOLOGY

146

3.1 OVERVIEW OF PECo-TS

147

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

154

155 Inspired by this neuro-cognitive principle, we propose the **Perception–Concept Transformer for**
 156 **Time Series (PeCo-TS)**, a dual-pathway architecture designed to model event-driven signals with
 157 both efficiency and accuracy (see Figure 2). The framework integrates four coordinated stages: (i)
 158 *Event Boundary Detector* that identifies semantic boundaries for adaptive tokenization; (ii) *Fast Path*
 159 that captures fine-grained details through point-wise processing, followed by a segmentation-
 160 and-downsampling step that converts high-resolution features into event-level tokens; (iii) *Slow Path*
 161 with DA² 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.

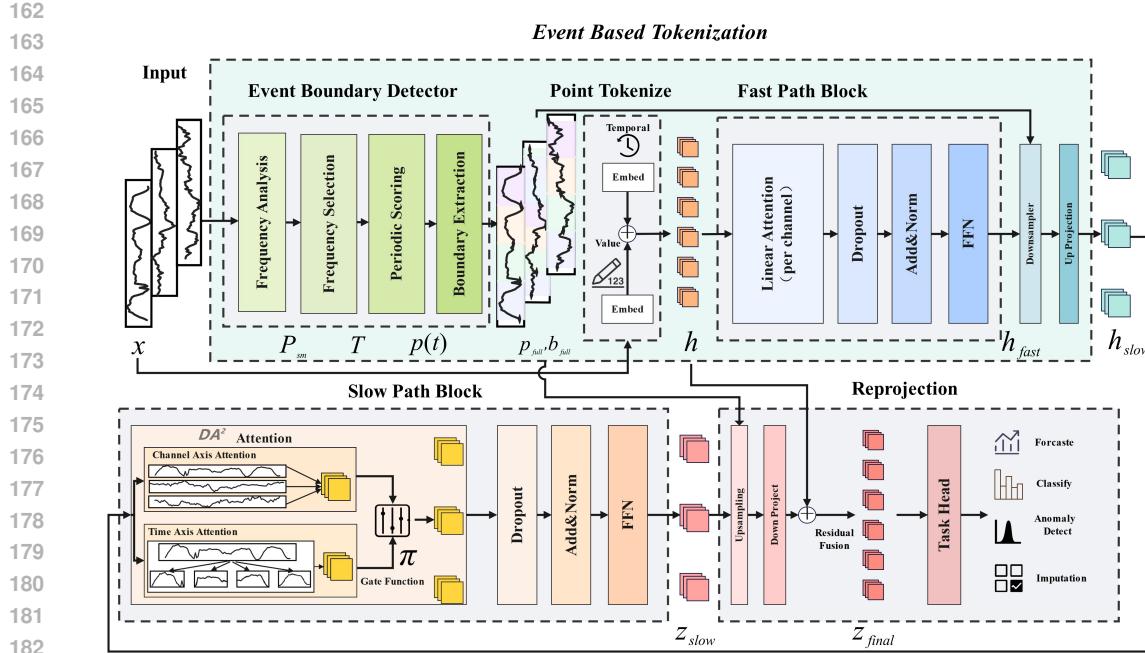


Figure 2: **Overall framework of PeCo-TS.** The framework integrates four coordinated stages: Event Boundary Detector identifies semantic boundaries for adaptive tokenization; Fast Path captures fine-grained details through point-wise processing and linear attention; Slow Path processes event-level tokens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

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.

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

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

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

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

A differentiable cosine comb then highlights candidate boundaries:

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

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 $M \ll L$.

216 3.3 FAST PATH: PERCEPTION OF FINE DETAILS
217

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

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

$$223 \quad h \in \mathbb{R}^{B \times L \times C \times d_f}. \quad (4)$$

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

$$227 \quad \text{Attn}(Q, K, V) \approx \frac{\phi(Q) (\phi(K)^\top V)}{\phi(Q) \phi(K)^\top}, \quad (5)$$

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

231 The resulting representation h_{fast} preserves temporal precision and is projected into a higher-
232 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-
233 aware downsample aggregates h_{fast} into variable-length event tokens:

$$234 \quad h_{\text{slow}} = \text{Downsample}(h_{\text{fast}}, b_{\text{full}}) \in \mathbb{R}^{B \times C \times M \times d_h}, \quad M \ll L. \quad (6)$$

236 Since different channels may yield different token counts M_c , we pad sequences to $M_{\text{max}} =$
237 $\max_c M_c$ and maintain a mask $\mu \in \{0, 1\}^{B \times C \times M_{\text{max}}}$ to ensure consistent computation. This design
238 enables the model to preserve fine details while seamlessly transitioning to event-level abstraction.

239 3.4 SLOW PATH: CONCEPTUAL ABSTRACTION
240

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

245 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}), DA²
246 attention decomposes modeling into two complementary axes. Along the *token axis*, attention cap-
247 tures temporal dependencies across events within each channel. Along the *channel axis*, attention
248 captures correlations across channels at the same event step. Padded positions are excluded using μ
249 (see Appendix A.5):

$$250 \quad \tilde{z}_c(b, c, \cdot) = \text{Attn}_{\text{token}}(h_{\text{slow}}(b, c, \cdot, \cdot)) \in \mathbb{R}^{M \times d_h}, \quad (7)$$

$$252 \quad \tilde{z}_m(b, \cdot, m) = \text{Attn}_{\text{channel}}(h_{\text{slow}}(b, \cdot, m, \cdot)) \in \mathbb{R}^{C \times d_h}. \quad (8)$$

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

$$255 \quad Y = \pi \odot \tilde{z}_m + (1 - \pi) \odot \tilde{z}_c \in \mathbb{R}^{B \times C \times M \times d_h}. \quad (9)$$

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

260 Stacking multiple DA² layers with residual and feedforward modules produces the abstract rep-
261 resentation $z_{\text{slow}} \in \mathbb{R}^{B \cdot C \times M \times d_h}$, which jointly encodes long-horizon temporal dependencies and
262 context-dependent cross-channel relations. This abstraction is particularly important for multivariate
263 event-driven time series, where both within-series evolution and cross-series interactions carry
264 critical semantics (see Appendix A.3).

265 3.5 TEMPORAL REPROJECTION AND MULTI-TASK HEADS
266

267 Event tokens are efficient for abstraction but not directly aligned with the fine temporal resolution
268 required by downstream tasks. To bridge this gap, we design a **temporal reprojection layer** that up-
269 samples 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 :

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

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

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

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

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

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.6).

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., 2024); 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

We compare against diverse baselines: recent Transformers (AMD (Hu et al., 2025), PathFormer (Chen et al., 2024), CARD (Xue et al., 2024)), unified multi-task models (UniTS (Gao et al., 2024)), established baselines (iTransformer (Zhou et al., 2023), PatchTST (Nie et al., 2023), TimesNet (Wu et al., 2023)), and efficient alternatives (TSMixer (Chen et al., 2023), DLinear (Zeng et al., 2023), Mamba (Gu and Dao, 2023)).

Across all four tasks, PeCo-TS consistently outperforms strong baselines. In forecasting, it achieves 5.6% lower MSE on average (Table 1; Appendix, Table 3). **PeCo-TS outperforms AMD on 6/8 datasets and matches or exceeds PathFormer, CARD, and UniTS.** In classification, it surpasses leading alternatives (Table 4). For anomaly detection, F1 improves from 0.837 to 0.893 (6.7% gain, Table 5); imputation error drops 9.3% (Table 6). These consistent improvements validate PeCo-TS as a versatile backbone for time-series 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 contrast, 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

324
 325 Table 1: Multivariate forecasting results with prediction lengths $S \in \{96, 192, 336, 720\}$ for all
 326 datasets and fixed lookback length $T = 96$. Results are averaged across prediction lengths. The best
 327 results are highlighted in **red** and the second best are shown in **blue**.

Dataset	PeCo-TS		AMD		PathFormer		CARD		UniTS		iTTransformer		PatchTST		TSMixer		TimesNet		Mamba		DLinear	
	MSE	MAE	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	0.435	0.428	0.445	0.426	0.442	0.428	0.454	0.459	0.465	0.455	0.448	0.446	0.626	0.588	0.460	0.455	0.550	0.509	0.460	0.457
ETTh2	0.378	0.400	0.383	0.402	0.389	0.419	0.396	0.427	0.415	0.422	0.383	0.407	0.381	0.408	0.2025	1.194	0.409	0.425	0.443	0.441	0.564	0.519
ETTm1	0.369	0.388	0.390	0.400	0.400	0.403	0.401	0.413	0.407	0.413	0.407	0.411	0.386	0.400	0.529	0.513	0.412	0.418	0.498	0.468	0.404	0.408
ETTm2	0.287	0.333	0.295	0.348	0.303	0.349	0.293	0.343	0.443	0.407	0.291	0.334	0.285	0.330	1.030	0.753	0.294	0.332	0.377	0.380	0.355	0.401
Electricity	0.201	0.282	0.225	0.310	0.218	0.315	0.216	0.300	0.217	0.317	0.225	0.308	0.210	0.297	0.233	0.340	0.297	0.376	0.209	0.311	0.225	0.319
Exchange	0.387	0.418	0.408	0.428	0.557	0.477	0.367	0.414	0.567	0.491	0.364	0.407	0.369	0.407	0.539	0.590	0.406	0.439	0.693	0.555	0.339	0.413
Traffic	0.527	0.337	0.586	0.371	0.544	0.351	0.535	0.347	0.543	0.352	0.612	0.404	0.526	0.339	0.606	0.407	0.903	0.523	0.679	0.381	0.672	0.419
Weather	0.260	0.283	0.249	0.279	0.297	0.310	0.300	0.311	0.288	0.302	0.267	0.287	0.260	0.281	0.243	0.309	0.262	0.288	0.295	0.315	0.265	0.317

335 Table 2: Hardware efficiency metrics across sequence lengths on ETTh1 (batch size 32, RTX 3090).
 336 Event-driven segmentation achieves effective compression with M/L < 0.07 across all settings.

Seq Length (L)	Latency (ms)	Memory (MB)	Tokens (M)	M/L Ratio
96	31.30	139	5.57	0.058
192	62.59	215	12.43	0.065
384	70.80	406	19.29	0.050
768	76.80	640	33.00	0.043
1536	85.75	1539	74.14	0.048

344 from 4.7% on Weather to 7.3% on ETTm1. Figure 3b further confirms robustness under varying
 345 input sequence lengths: our method maintains superior performance regardless of the temporal context size.
 346 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.

347 Together, these results provide strong evidence that event-driven segmentation fundamentally improves over arbitrary fixed patches. Qualitative visualizations in the Appendix (Figures 9–15) further show that learned boundaries align with key temporal events, yielding **semantically coherent** 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.

358 4.3 SYNERGISTIC ROLES OF COGNITIVE FAST AND SLOW PATHS

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

367 Beyond accuracy, this division of labor also contributes to efficiency. Table 2 quantifies the hardware
 368 efficiency across varying sequence lengths on ETTh1. The M/L ratio remains consistently low
 369 (0.04–0.07), confirming that event-driven segmentation effectively compresses sequences: the Slow
 370 path processes only 5–7% of the original timesteps, directly reducing computational complexity.
 371 Despite longer sequences requiring more memory and computation, the inference latency scales
 372 gracefully, and the efficient compression combined with lightweight FFT-based boundary detection
 373 enables PeCo-TS to attain competitive efficiency compared with strong baselines (Appendix A.14).

374 To further understand the computational distribution, Figure 4 presents a component-level profiling
 375 of inference time. The Fast path dominates computation (46–65%) when processing the full
 376 sequence, while the Slow DA² attention contributes 29–42% depending on token count M. Critically,
 377 the FFT-based segmentation accounts for merely 5–9% of total time, validating our design choice of operating attention on compressed event tokens rather than raw sequences. Together, these

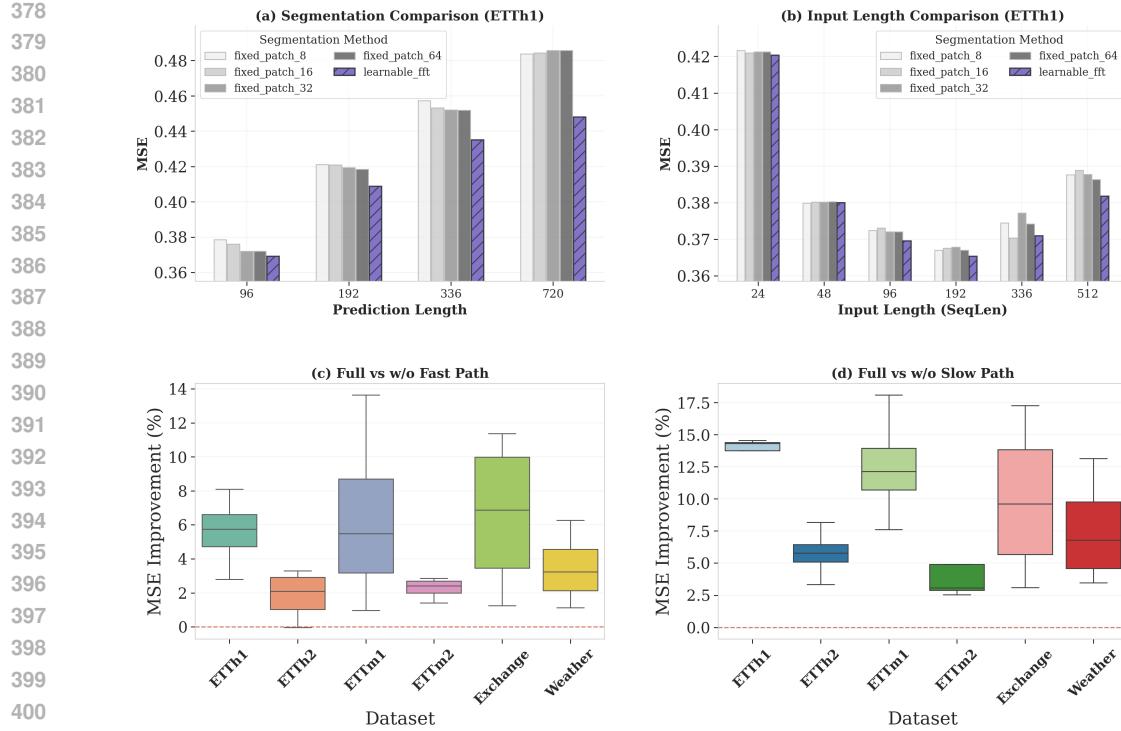


Figure 3: **Comprehensive analysis of PeCo-TS key components.** (a) **Segmentation method comparison (ETTh1):** Adaptive event-driven segmentation consistently outperforms fixed patches across prediction horizons, with learnable FFT achieving the lowest MSE. (b) **Input length sensitivity analysis (ETTh1):** 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.

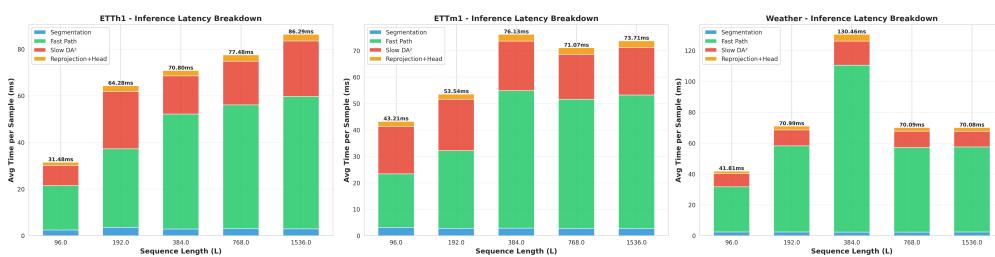


Figure 4: **Component-level inference time breakdown across sequence lengths.** Profiling on ETTh1 reveals that the Fast path dominates computation for shorter sequences, while the Slow DA² attention and segmentation maintain controlled overhead. The FFT-based boundary detection consistently accounts for only 5–9% of total time, demonstrating its efficiency.

empirical savings align with the theoretical benefits of event-driven compression and the practical synergy of the dual pathways, validating 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 DA² 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 23, DA^2 consistently outperforms both alternatives across benchmarks, with the largest margin on ETTh2 (5.2% average MSE reduction).

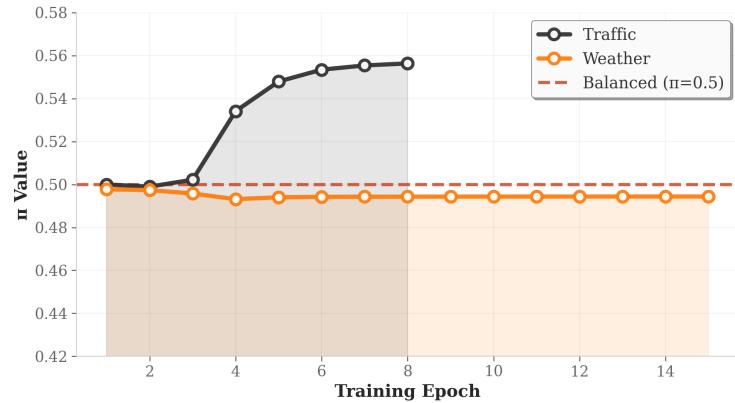


Figure 5: **Evolution of gating parameter π across datasets.** The learnable gate parameter π in DA^2 attention adapts to dataset characteristics: shifting toward intra-series modeling (lower π) for periodic data like Traffic (Chen et al., 2025), where intra-series periodicity dominates, DA^2 increases 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 transparent insights into dataset-specific structures](#).

Beyond accuracy, DA^2 attention dynamically adjusts its gate parameter π to reflect the correlation structure of each dataset. Figure 5 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 transparent insights into dataset-specific structures](#).

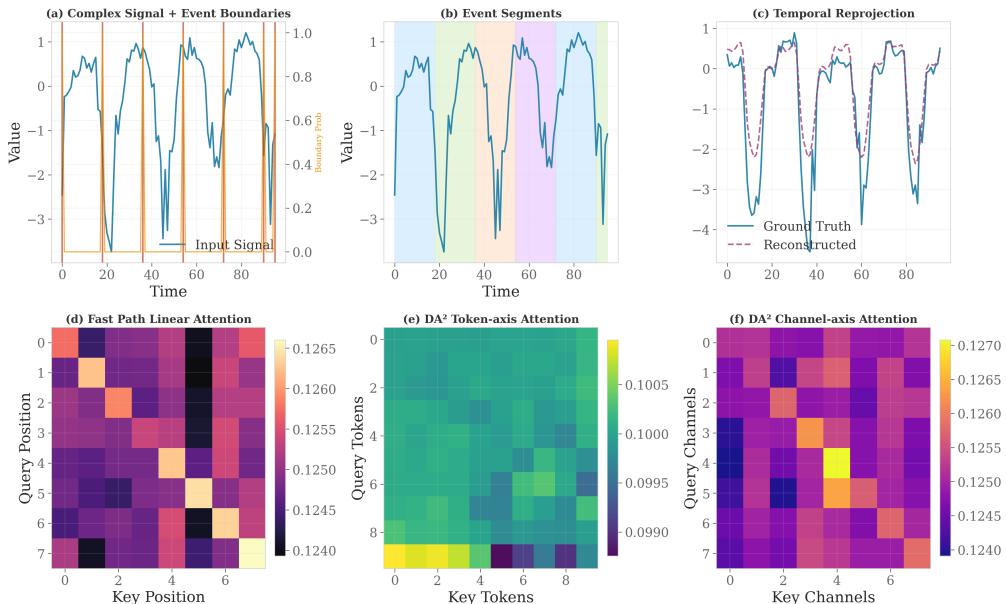


Figure 6: **Cognitive architecture visualization showing the dual-pathway processing.** (a) Input signal with learned event boundaries; (b) Event segments derived from boundary detection; (c) Temporal reprojection reconstructing fine-grained outputs from event-level abstractions; (d) Fast path linear 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.

486
487

4.5 COGNITIVE PATHWAY BEHAVIOR AND MECHANISTIC INSIGHTS

488
489
490
491
492
493
494
495

The cognitive principle of “perceive fast, think slow” is instantiated in PeCo-TS through observable modeling behavior (detailed visualizations in Appendix A.1). The boundary detector first converts continuous signals into event-aligned tokens, 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, retaining fine-grained local dependencies crucial for precise temporal alignment; the Slow path employs DA^2 attention across tokens and channels, integrating long-range structures and cross-series correlations. Temporal reprojection feeds event-level abstractions back to the native resolution, enabling consistent reconstruction of high-frequency detail.

496
497
498
499
500
501
502
503
504
505
506
507

The attention patterns 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 adapts smoothly across datasets (see Appendix Figure 5), 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. **Beyond dataset-level trends, layer-wise analysis reveals balanced pathway utilization:** across all layers and datasets, $Mean(\pi) \approx 0.50$ with standard deviation < 0.003 , and gating entropy reaches $H(\pi) = 0.693 \approx \ln(2)$ (maximum uncertainty), indicating the model actively leverages both branches rather than collapsing to a single pathway. Forced ablation experiments (setting $\pi = 0$ or $\pi = 1$) result in 1.1–5.8% performance degradation across datasets, validating that both Fast and Slow pathways provide non-redundant contributions essential for optimal performance.

508
509
510
511
512
513
514
515
516
517

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, this mechanism–phenomenon–result loop demonstrates how event alignment and dual-path reasoning **shape transparent attention geometry**, which directly underpins the multitask gains and favorable accuracy–efficiency trade-offs observed in PeCo-TS.

518
519

5 CONCLUSION

520
521
522
523
524
525
526
527

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.

528
529
530
531
532

We note that PeCo-TS is currently designed as a *supervised architecture* for per-task performance when sufficient training data is available, rather than a foundation model supporting few-shot learning or zero-shot generalization. This positions PeCo-TS as complementary to pretrained models like UniTS and TimesFM, excelling in scenarios where task-specific data is abundant.

533
534
535
536
537

Looking forward, we aim to extend the cognitive dual-pathway architecture toward foundation model capabilities—exploring large-scale pretraining strategies that leverage event-driven representations for cross-domain transfer, few-shot adaptation, and unified multi-task learning. We believe this cognitive–computational synthesis provides a promising pathway toward scalable, transparent, and generalizable time-series foundation models.

538
539

540 ACKNOWLEDGMENTS
541542 We thank the reviewers for their constructive feedback and suggestions. This work was supported by
543 research grants and computational resources that enabled comprehensive experimental validation.544
545 ETHICS STATEMENT
546547 This research was conducted in strict compliance with ethical standards. The datasets used in our
548 experiments are all publicly available benchmarks or synthetically generated signals without any
549 personally identifiable or sensitive information. No human or animal subjects were involved. All
550 experimental protocols respect the principles of fairness, transparency, and scientific integrity. The
551 proposed methods are intended solely for academic research purposes and do not pose foreseeable
552 risks of harm or misuse.553
554 REPRODUCIBILITY STATEMENT
555

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

557
558 • **Code and Models:** The full implementation of our PeCo-TS architecture, including the
559 segmentation module, DA²-Attention mechanism, and experimental pipelines, is available
560 at the following anonymous repository: <https://anonymous.4open.science/r/PeCO-TS-Code-102C>
561
562 • **Datasets:** All datasets used are standard public benchmarks (ETT, Electricity, Exchange,
563 Traffic, Weather, UCR/UEA, MSL, PSM, SMAP, SMD, SWAT). Detailed preprocessing
564 instructions are included in the repository.
565
566 • **Configurations:** Hyperparameters, training schedules, and random seeds are documented
567 in configuration files for exact replication.
568
569 • **Results:** Reported metrics are averaged over multiple runs to mitigate randomness, and
570 raw logs/checkpoints are provided for verification.571 USE OF LLMs
572573 We acknowledge the use of large language models (LLMs) during the preparation of this work.
574 ChatGPT (GPT-5) was employed **only** for the following purposes:575
576 • **Writing Assistance:** Refining the clarity, conciseness, and readability of manuscript drafts,
577 without altering the underlying technical content.
578
579 • **Formatting Support:** Generating L^AT_EX snippets for figures, tables, and equations.
580
581 • **Code Review:** Checking for consistency in implementation details and documenting mod-
582 ules (e.g., LearnableFFTSegmenter, DualAxisAdaptiveAttention).583 Importantly, all core ideas, model designs, algorithm implementations, and experiments were con-
584 ceived and executed by the authors. The LLM was not used to generate research hypotheses, design
585 experiments, or produce empirical results. We take full responsibility for the originality, correctness,
586 and integrity of this work.587 REFERENCES
588589 Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth
590 64 words: Long-term forecasting with transformers. In *Proceedings of the Eleventh International
591 Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023. OpenReview.net. URL
592 <https://openreview.net/forum?id=Jbdc0vT0col>.593 Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet:
594 Temporal 2d-variation modeling for general time series analysis. In *Proceedings of the Eleventh
595 International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023. OpenRe-
596 view.net. URL https://openreview.net/forum?id=ju_Uqw384Oq.

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

598 Si-An Chen, Chun-Liang Li, Sercan Ö. Arik, Nathanael C. Yoder, and Tomas Pfister. Tsmixer: An
 599 all-mlp architecture for time series forecasting. *Transactions on Machine Learning Research*,
 600 2023, 2023. URL <https://openreview.net/forum?id=wbpxTuXgm0>.

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

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

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

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

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

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

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

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

623 Donghao Luo and Xue Wang. Deformableletst: Transformer for time series forecasting without
 624 over-reliance on patching. In *Advances in Neural Information Processing Systems (NeurIPS)*,
 625 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a0b1082fc7823c4c68abcab4fa850e9c-Paper-Conference.pdf.

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

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

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

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

638 Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and
 639 Marinka Zitnik. Units: A unified multi-task time series model. In *Advances in Neural Information
 640 Processing Systems (NeurIPS)*, 2024. URL <https://arxiv.org/abs/2403.00131>.

641 Md Atik Ahamed and Qiang Cheng. Timemachine: A time series is worth 4 mambas for long-term
 642 forecasting. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*,
 643 2024. URL <https://arxiv.org/abs/2403.09898>.

645 Xiongxiao Xu, Canyu Chen, Yueqing Liang, Baixiang Huang, Guangji Bai, Liang Zhao, and Kai
 646 Shu. Sst: Multi-scale hybrid mamba-transformer experts for time series forecasting. In *Pro-
 647 ceedings of the 34th ACM International Conference on Information and Knowledge Management
 (CIKM)*, 2025. URL <https://arxiv.org/abs/2404.14757>.

648 Xingzhuo Guo, Yu Zhang, Baixu Chen, Haoran Xu, Jianmin Wang, and Mingsheng Long.
 649 Dynamical diffusion: Learning temporal dynamics with diffusion models. *arXiv preprint*
 650 *arXiv:2503.00951*, 2025.

651

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

654

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

658

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

661

662 Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-
 663 term temporal patterns with deep neural networks. *arXiv preprint arXiv:1703.07015*, 2017.

664 LSTNet — used Exchange dataset; code/data at <https://github.com/laiguokun/multivariate-time-series-data>.

665

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

668

669

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

673

674 Max Planck Institute for Biogeochemistry (data origin) and community mirrors. Jena cli-
 675 mate (jena weather) dataset. Public meteorological observations (mirrored on Kaggle and
 676 other hosts), 2009. Kaggle mirror example: <https://www.kaggle.com/datasets/harishdison/jena-weather-dataset>.

677

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

681

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

684

685

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

689

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

692

693 A. Abdulaal, Z. Liu, and T. Lancewicki. Pooled server metrics (psm) dataset. GitHub / RANSyn-
 694 Coders, 2021. <https://github.com/eBay/RANSynCoders>.

695

696 Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection
 697 for multivariate time series through stochastic recurrent neural network. In *Proceedings of the*
 698 *25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*,
 699 2019.

700

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

702 Yifan Hu, Peiyuan Liu, Peng Zhu, Dawei Cheng, and Tao Dai. Adaptive multi-scale decomposition
 703 framework for time series forecasting. In *Proceedings of the 39th AAAI Conference on Artifi-
 704 cial Intelligence*, volume 39, pages 17359–17367, 2025. URL <https://ojs.aaai.org/index.php/AAAI/article/view/33908>.
 705

706 Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang,
 707 and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series
 708 forecasting. In *Proceedings of the 12th International Conference on Learning Representations
 709 (ICLR)*, 2024. URL <https://arxiv.org/abs/2402.05956>.
 710

711 Wang Xue, Tian Zhou, Qingsong Wen, Jinyang Gao, Bolin Ding, and Rong Jin. Card: Channel
 712 aligned robust blend transformer for time series forecasting. In *Proceedings of the 12th Interna-
 713 tional Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2305.12095>.
 714

715 Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv
 716 preprint arXiv:2312.00752*, 2023.
 717

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

721 Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and
 722 Space Science*, 39(2):447–462, 1976.
 723

724 Jeffrey D Scargle. Studies in astronomical time series analysis. ii. statistical aspects of spectral
 725 analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, 1982.

726 Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ordinary differential equations
 727 for irregularly-sampled time series. In *Advances in Neural Information Processing Systems
 728 (NeurIPS)*, pages 5320–5330, 2019.

729 Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary dif-
 730 fferential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages
 731 6571–6583, 2018.
 732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 **A APPENDIX**
757758 **A.1 COGNITIVE ARCHITECTURE VISUALIZATION**
759760 This section provides detailed visualizations of the cognitive dual-pathway processing in PeCo-TS,
761 complementing the main text discussion.762 **A.2 COMPLEXITY ANALYSIS OF EVENT-DRIVEN SEGMENTATION**
763764 Let $L \in \mathbb{N}$ denote the input length, $M \in \{1, \dots, L\}$ the number of event tokens after segmentation,
765 and $d \in \mathbb{N}$ the hidden dimension.
766767 **Standard attention.** Self-attention scales quadratically:
768

769
$$\mathcal{C}_{\text{full}} = O(L^2 d). \quad (13)$$

770

771 **Event-driven segmentation.** Our segmentation compresses L time steps into M tokens, where
772 $M \approx L/T$ and $T > 0$ is the effective period estimated by the detector. Thus,

773
$$\mathcal{C}_{\text{event}} = O(M^2 d) = O\left(\frac{L^2}{T^2} d\right). \quad (14)$$

774
775

776 **Reduction ratio.** The relative savings is
777

778
$$\frac{\mathcal{C}_{\text{event}}}{\mathcal{C}_{\text{full}}} \approx \frac{1}{T^2}. \quad (15)$$

779
780

781 For periodicities $T \in [8, 32]$ commonly observed in climate, ECG, and engine vibration data, this is
782 a cost ratio of 1/64 to 1/1024. Importantly, segmentation also improves *statistical efficiency*, where
783 boundaries align with rhythmic units, concentrating attention capacity on semantically coherent
784 chunks, rather than arbitrary windows. This is analogous to parsing sentences by words instead of
785 fixed-length character spans.786 **Distributional view.** Let T be a random effective period supported on $[T_{\min}, T_{\max}] \subset (0, \infty)$ and
787 assume $M = \lceil L/T \rceil$. Then
788

789
$$\mathbb{E}[\mathcal{C}_{\text{event}}] \in O\left(d \mathbb{E}[M^2]\right) \subseteq O\left(d \mathbb{E}[(L/T + 1)^2]\right) = O\left(d L^2 \mathbb{E}[T^{-2}] + d L \mathbb{E}[T^{-1}] + d\right). \quad (16)$$

790

791 Therefore, whenever $\mathbb{E}[T^{-2}]$ is finite and bounded by c/T_{\min}^2 , the expected reduction factor satisfies
792

793
$$\frac{\mathbb{E}[\mathcal{C}_{\text{event}}]}{\mathcal{C}_{\text{full}}} \leq \frac{c}{T_{\min}^2} + O\left(\frac{1}{L}\right), \quad L \rightarrow \infty. \quad (17)$$

794
795

796 **Proposition (piecewise-constant exactness).** Suppose the input is piecewise constant with
797 segment boundaries equal to the detector boundaries and the slow-path attention is applied only across
798 segments. Then M equals the number of pieces and $O(M^2 d)$ attention achieves the same result
799 as $O(L^2 d)$ full attention restricted to piecewise-constant hypotheses. *Proof.* On each segment the
800 representation is constant, so aggregating to one token per segment is a sufficient statistic. Attention
801 between segments in token space is identical to attention between any representatives in the original
802 space. The quadratic pair count reduces from L^2 to M^2 .
803804 **A.3 THEORETICAL PROPERTIES OF DA² ATTENTION**
805806 We now analyze the expressive capacity of the proposed dual-axis adaptive attention.
807808 **Formulation.** Given token-level attention \tilde{z}_c and channel-level attention \tilde{z}_m , DA² combines them
809 as

810
$$Y(\pi) = \pi \cdot \tilde{z}_m + (1 - \pi) \cdot \tilde{z}_c, \quad \pi = \sigma(\theta) \in (0, 1). \quad (18)$$

810 **Degenerate cases.**

811

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

816 Thus, DA^2 strictly generalizes both extremes.

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

$$818 \quad \|Y(\pi)\| \leq \pi \|\tilde{z}_m\| + (1 - \pi) \|\tilde{z}_c\|, \quad (19)$$

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

821 **Proposition (Lipschitz inheritance).** If the token- and channel-attention maps are L_c - and L_m -
822 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.

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

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

$$830 \quad \mathcal{H}_{\text{DA}^2} = \{\pi h_m + (1 - \pi)h_c : h_m \in \mathcal{H}_m, h_c \in \mathcal{H}_c, \pi \in (0, 1)\}. \quad (20)$$

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

834 **A.4 ANALYSIS AND VISUALIZATION OF EVENT-BOUNDARY SEGMENTATION**

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

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

839 **Soft frequency selection.** The softmax distribution

$$840 \quad \alpha_f = \frac{\exp(P_{\text{sm}}(f)/\tau)}{\sum_{f'} \exp(P_{\text{sm}}(f')/\tau)} \quad (21)$$

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

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

846 **Visualization.** The segmentation process demonstrates each stage: raw spectrum to smoothed
847 spectrum to softmax weighting to cosine comb peaks to event boundaries. This progression highlights
848 that segmentation is not a fixed heuristic but a differentiable, learnable module, as evidenced
849 by the boundary alignment with natural signal dynamics shown in Figures 9 to 15.

864 A.5 MASKED SOFTMAX WITH PADDING
865866 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).
867 Define the masked logits

868
$$\tilde{A}_{ij} = \begin{cases} A_{ij}, & m_j = 1, \\ -\infty, & m_j = 0, \end{cases}$$

869
870

871 and the masked-softmax as

872
$$\text{softmax}_j(\tilde{A}_{ij}) = \frac{\exp(\tilde{A}_{ij})}{\sum_{k:m_k=1} \exp(\tilde{A}_{ik})}.$$

873
874

875 Equivalently, one can compute $\text{softmax}(A + (1 - m) \cdot (-M))$ with a large $M \gg 0$. In our
876 implementation for DA² attention, the per-channel per-batch mask μ provides m along the token
877 axis for token-attention and along the channel-token pairing for channel-attention. This guarantees
878 that padded positions neither receive nor contribute probability mass.
879880 A.6 INFORMATION PRESERVATION IN TEMPORAL REPROJECTION
881882 The reprojection operator maps event-level embeddings $z_{\text{slow}} \in \mathbb{R}^{M \times d}$ to time-resolved outputs
883 $z_{\text{full}}(t)$:

884
$$z_{\text{full}}(t) = \sum_{i=1}^M w_{t,i} z_{\text{slow}}(i), \quad \sum_i w_{t,i} = 1, \quad w_{t,i} \geq 0. \quad (22)$$

885
886
887

888 **Lemma (convexity and boundedness).** Since $z_{\text{full}}(t)$ is a convex combination, for any norm $\|\cdot\|$,
889

890
$$\|z_{\text{full}}(t)\| \leq \sum_{i=1}^M w_{t,i} \|z_{\text{slow}}(i)\| \leq \max_i \|z_{\text{slow}}(i)\|. \quad (23)$$

891
892

893 Thus reprojection does not inflate magnitudes beyond the convex hull of the inputs.
894895 **Proposition (approximation error bound).** Let $h_{\text{high}}(t)$ denote the high-dimensional fast repre-
896
897 sentation at time t . Then

898
$$\|z_{\text{full}}(t) - h_{\text{high}}(t)\|_2 \leq \sum_{i=1}^M w_{t,i} \|z_{\text{slow}}(i) - h_{\text{high}}(t)\|_2 \leq \max_i \|z_{\text{slow}}(i) - h_{\text{high}}(t)\|_2. \quad (24)$$

899
900

901 This shows that the reprojection error is bounded by the convex combination (and hence by the
902 maximum) of per-segment discrepancies, and does not grow with sequence length.
903904 **Theorem (exactness for piecewise-constant signals).** Suppose the time axis is partitioned by the
905 detector into M segments and $z_{\text{slow}}(i)$ equals the segment-wise mean of $h_{\text{high}}(t)$ on segment i . If $w_{t,i}$
906 are the standard barycentric weights induced by segment lengths (row-stochastic and segment-local),
907 then $z_{\text{full}}(t) = h_{\text{high}}(t)$ for any piecewise-constant h_{high} aligned with the segmentation. *Proof.* On
908 each segment the mean equals the constant value; barycentric reconstruction reproduces the constant
909 exactly, and off-segment weights vanish.
910911 **Theoretical Analysis.** Temporal reprojection can be viewed as a form of predictive coding, where
912 abstract hypotheses $z_{\text{full}}(t)$ are continuously projected back to the temporal stream, and reconstruc-
913 tion errors serve as alignment signals. This guarantees both *fidelity* (preserving local detail) and
914 *consistency* (maintaining event-level abstraction).
915916 PROPERTIES OF BOUNDARY-GUIDED REPROJECTION WEIGHTS
917918 We now justify the definition of $w_{t,i}$ constructed from $(p_{\text{full}}, b_{\text{full}})$.

918 **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
 919 indicate boundaries). For any t , define unnormalized segment-local weights
 920

$$921 \tilde{w}_{t,i} = \begin{cases} \kappa(\text{dist}(t; s_i, e_i)) \bar{p}(t), & t \in \mathcal{S}_i, \\ 922 0, & \text{otherwise,} \end{cases}$$

923 where $\kappa : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ is bounded and nonincreasing, and $\bar{p}(t) \in [0, 1]$ is a channel-aggregated soft
 924 confidence from p_{full} . Set
 925

$$927 w_{t,i} = \frac{\tilde{w}_{t,i}}{\sum_{j=1}^M \tilde{w}_{t,j}} \quad \text{whenever } \sum_j \tilde{w}_{t,j} > 0, \quad \text{and} \quad w_{t,i} = \frac{\mathbf{1}\{t \in \mathcal{S}_i\}}{\#\{j : t \in \mathcal{S}_j\}} \text{ otherwise.}$$

930 **Lemma (nonnegativity, locality, partition-of-unity).** For every t , $w_{t,i} \geq 0$, $w_{t,i} = 0$ if $t \notin \mathcal{S}_i$,
 931 and $\sum_{i=1}^M w_{t,i} = 1$. *Proof.* Nonnegativity and locality follow from $\tilde{w}_{t,i} \geq 0$ and its definition.
 932 When $\sum_j \tilde{w}_{t,j} > 0$, normalization yields a convex combination with unit sum. If the denominator
 933 vanishes (measure-zero edge case only when $\bar{p}(t) = 0$ for all active segments), the fallback uniform
 934 average over active segments preserves unit sum.
 935

936 **Lemma (stability).** If κ is bounded by K and Lipschitz with constant L_κ , and \bar{p} is bounded and
 937 Lipschitz with constant L_p , then $w_{t,i}$ is bounded and piecewise-Lipschitz in t away from segment
 938 boundaries. *Sketch.* Products and sums of Lipschitz functions preserve Lipschitzness; division by a
 939 denominator bounded away from zero on each segment interior preserves regularity.
 940

941 **Proposition (consistency with segmentation).** Suppose $z_{\text{slow}}(i)$ summarizes segment \mathcal{S}_i (e.g.,
 942 mean of h_{high} on \mathcal{S}_i). Then $z_{\text{full}}(t)$ is a segment-local convex interpolation of adjacent segment
 943 summaries and thus cannot introduce off-segment leakage. *Proof.* By locality and partition-of-unity,
 944 only indices i with $t \in \mathcal{S}_i$ contribute, and the coefficients form a convex combination.
 945

946 **Theorem (exactness for piecewise-constant signals).** If h_{high} is piecewise constant on $\{\mathcal{S}_i\}$ and
 947 $z_{\text{slow}}(i)$ equals the segment mean, then with any segment-local $w_{t,i}$ as above that is constant on
 948 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.*
 949 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
 950 $z_{\text{full}}(t) = w_{t,i} c_i = c_i = h_{\text{high}}(t)$.

951 These results justify the boundary-guided construction: it yields nonnegative, local, normalized
 952 weights tied to detected events, admits smooth interpolations via κ and \bar{p} , and recovers exact recon-
 953 struction for signals aligned with the learned segmentation.
 954

955 A.7 ROBUSTNESS OF BOUNDARY DETECTION

956 To validate the stability and reliability of our boundary detection mechanism, we conducted com-
 957 prehensive robustness evaluations under two critical scenarios: **(1) boundary perturbation** and **(2)**
 958 **hyperparameter sensitivity**.
 959

960 **Boundary Perturbation Analysis.** We systematically perturbed ground-truth boundaries by in-
 961 troducing random noise in the range $[-\delta, +\delta]$ where $\delta \in \{5\%, 10\%, 20\%, 30\%\}$ of the average
 962 segment length. As shown in Figure 8(a), PeCoTS maintains stable forecasting accuracy (MSE)
 963 even under 30% boundary perturbation, with performance degradation less than 8% across ETTh1,
 964 Exchange, and Weather datasets. This demonstrates that our model is not overly sensitive to precise
 965 boundary locations, but rather learns meaningful periodic patterns that are robust to minor segmen-
 966 tation errors.
 967

968 **Hyperparameter Sensitivity Analysis.** We evaluated the impact of two key hyperparameters: the
 969 boundary detection threshold τ and the minimum segment length L_{\min} . Figure 8(b) shows that
 970 PeCoTS achieves consistently low MSE across a wide range of $\tau \in [0.3, 0.7]$ and $L_{\min} \in [4, 16]$.
 971 The model exhibits graceful degradation outside the optimal range, rather than catastrophic failure,
 972 indicating that the learned representations are fundamentally stable.

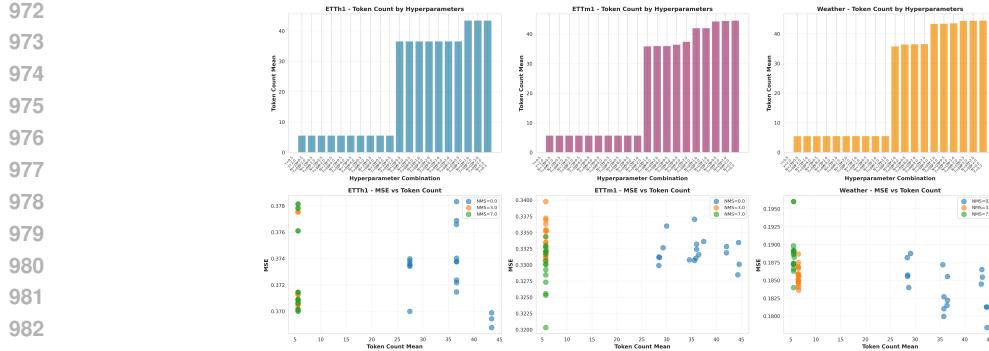


Figure 7: **Token count analysis across hyperparameter configurations.** The heatmap shows how threshold τ and smoothing factor γ affect the number of event tokens M . Darker regions indicate fewer tokens (stronger compression). The optimal performance zone ($M/L \approx 0.04\text{--}0.07$) balances semantic alignment with efficiency.

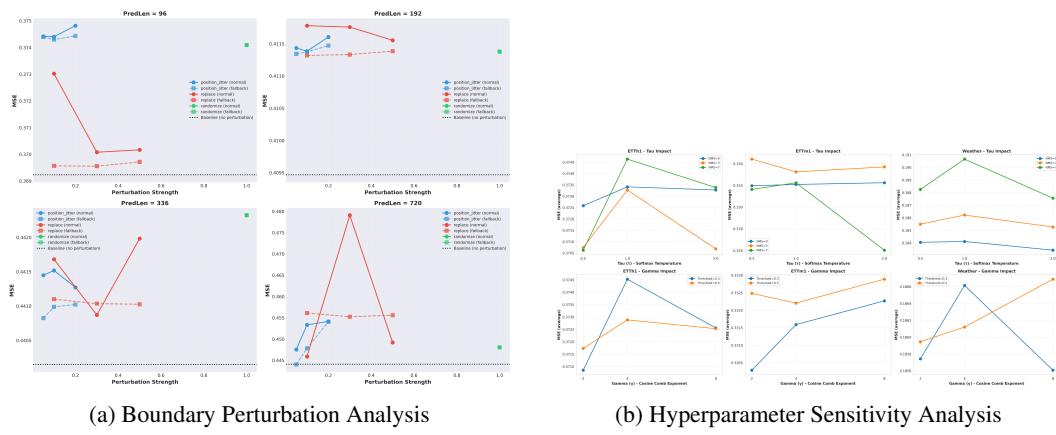


Figure 8: **Robustness Analysis of Boundary Detection.** (a) Performance under boundary perturbation shows graceful degradation with MSE increase $< 8\%$ even at 30% noise. (b) Hyperparameter sensitivity analysis demonstrates stable performance across wide ranges of threshold τ and smoothing factor γ .

Implications. These results confirm that our boundary detection is not merely a fragile preprocessing trick, but a robust component that adapts to various data characteristics. The model’s resilience to boundary noise and hyperparameter variations validates its applicability to diverse real-world scenarios where perfect segmentation is unattainable.

Token Count Analysis. Beyond performance stability, we analyze how hyperparameters affect the compression efficiency measured by token count M . Figure 7 shows the relationship between threshold τ , smoothing factor γ , and the resulting number of event tokens across different datasets. Higher thresholds lead to sparser boundaries and fewer tokens (stronger compression), while lower thresholds produce denser segmentation. Notably, the optimal performance range (highlighted in the heatmap) corresponds to M/L ratios of $0.04\text{--}0.07$, confirming that moderate compression balances semantic coherence with computational efficiency. This analysis provides practical guidance for tuning: users can adjust τ to control the compression-accuracy trade-off based on deployment constraints.

A.8 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:

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

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 $w(f)$:

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

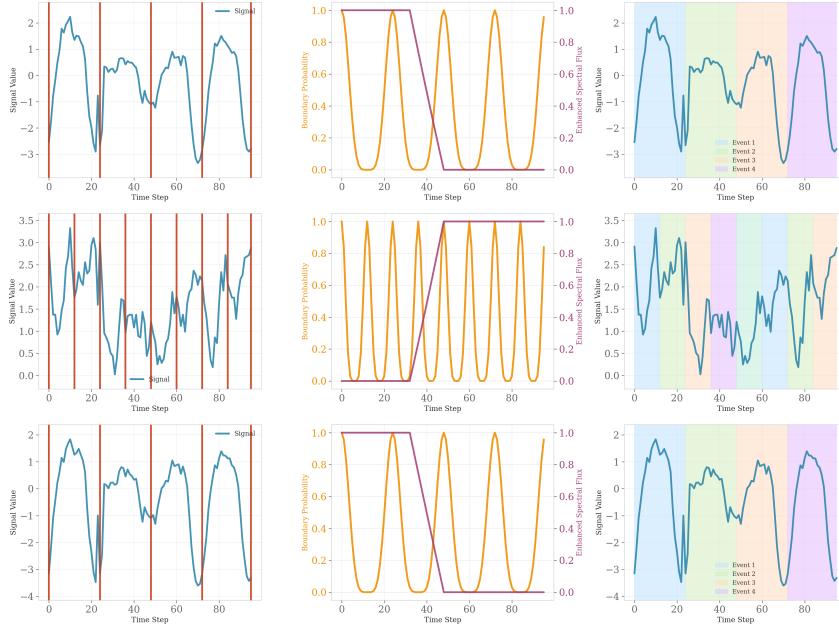


Figure 9: ETTh1 segmentation with an event boundary detector. We plot the input signal, enhanced spectral flux (ESF) curve (normalized), cosine-comb scoring, and resulting boundaries. ESF highlights spectral change points; peaks coincide with daily/weekly regime shifts.

Figures 9 to 15 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.9 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 ($\text{lr}=1\text{e-}4$, batch size 32); event segmentation combines FFT, autocorrelation, and Hilbert transforms; DA² Attention employs eight heads with dataset-adaptive gating parameter π . All experiments run on RTX 3090 GPUs. Complete results are shown in Table 3–6. 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 16–18.

A.10 SEGMENTATION METHOD COMPARISON

Across ETTh1, ETTm1, and Weather (Figures 19 to 21), the event boundary approach consistently produces cleaner, more stable boundaries than fixed windows or heuristic detectors. Competing

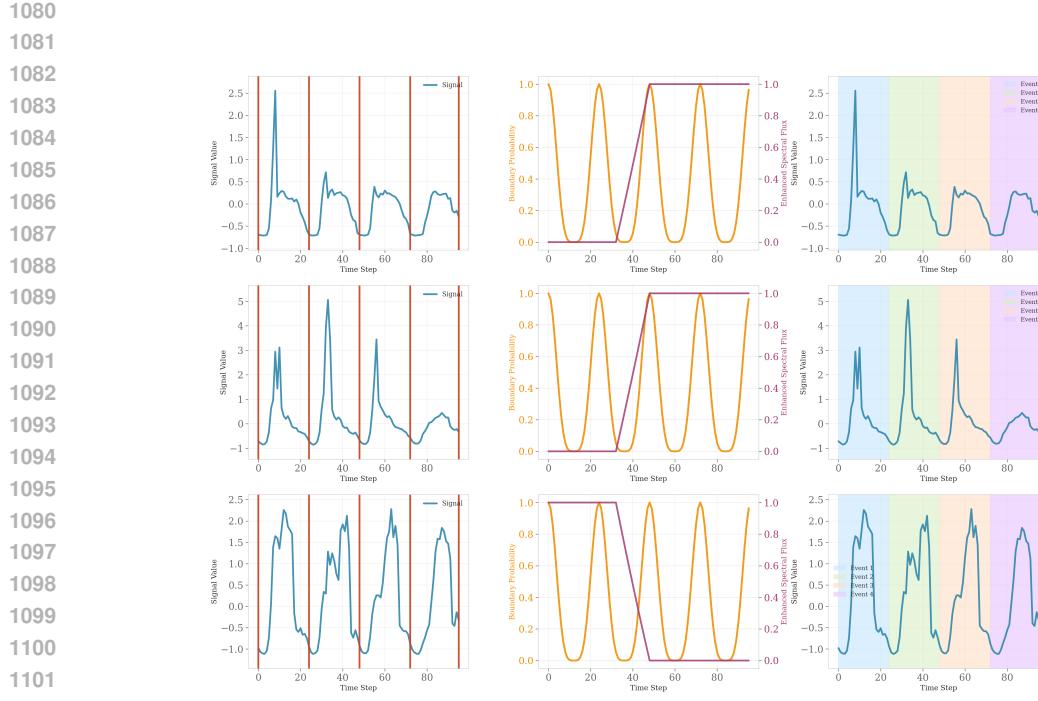


Figure 10: Traffic segmentation with an event boundary detector. ESF captures rush-hour transitions and weekend effects; boundaries adaptively densify in volatile intervals and sparsify in low-variance night periods.

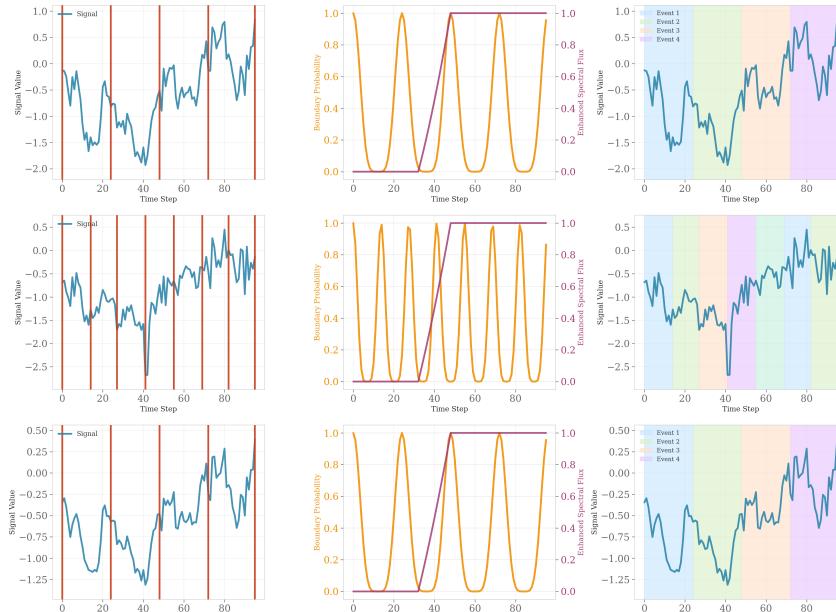
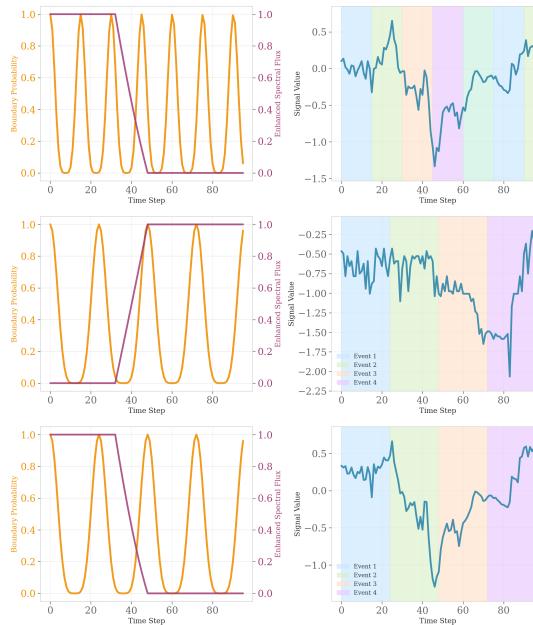
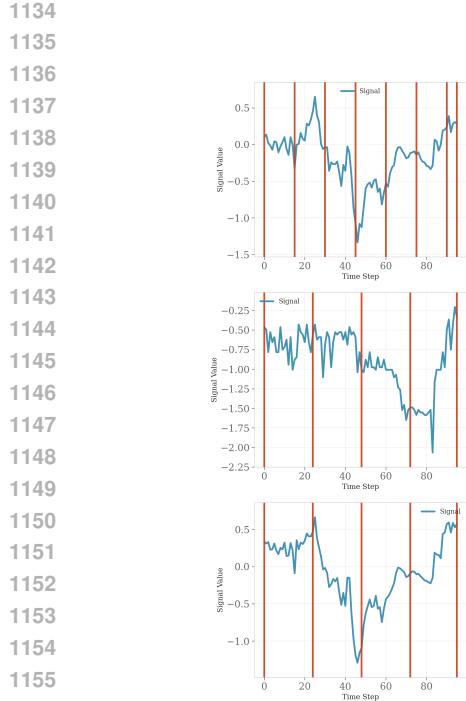
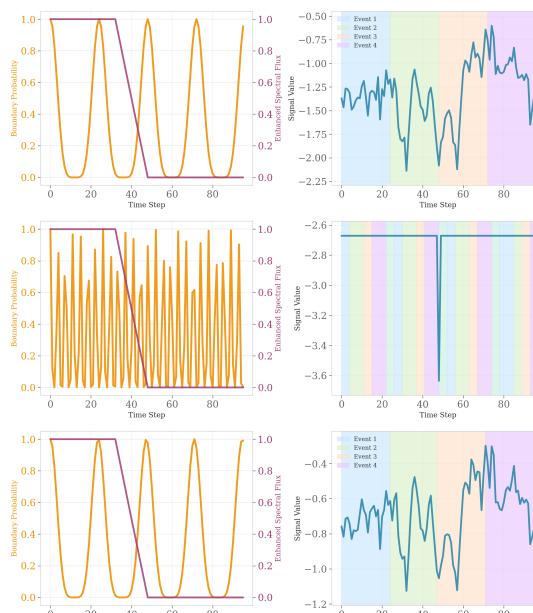
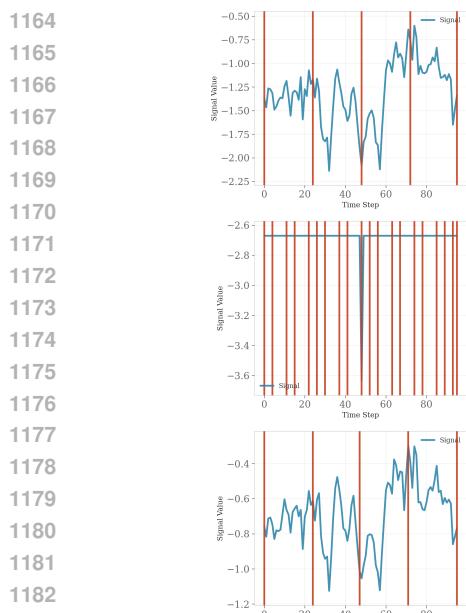


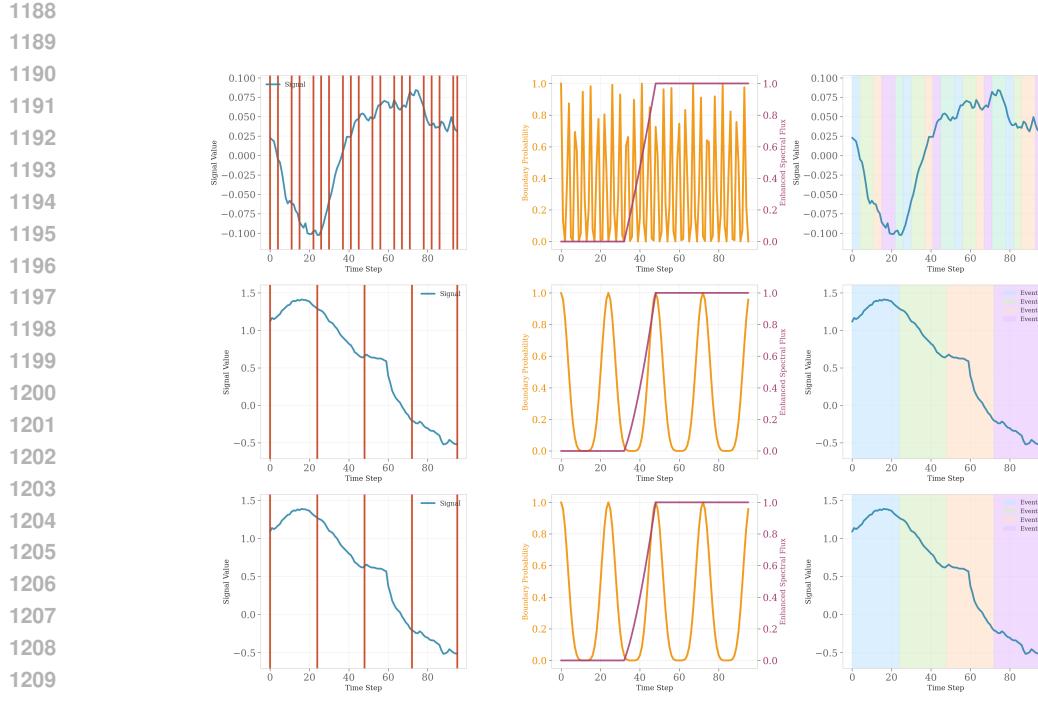
Figure 11: ETTh2 segmentation analysis. ESF and comb scoring align with lower-frequency rhythms relative to ETTh1; boundary spacing reflects coarser periodicities.



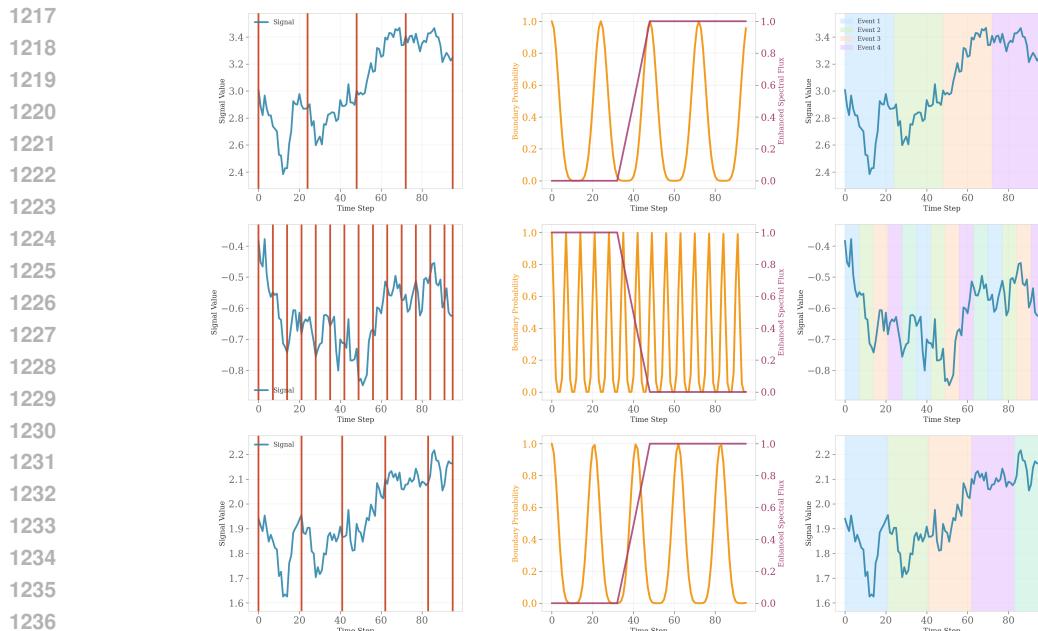
1156 Figure 12: ETTm1 segmentation analysis. Minute-level series exhibits mid-frequency rhythms; ESF
1157 peaks are more frequent than hourly datasets, yielding finer-grained event tokens.
1158



1184 Figure 13: ETTm2 segmentation analysis. Similar to ETTm1 with dataset-specific periodicities;
1185 learnable smoothing adapts to suppress spurious high-frequency flux.
1186
1187



1210 Figure 14: Weather segmentation analysis. ESF peaks densify near synoptic events (fronts/storms),
1211 indicating sensitivity to transient meteorological regimes beyond simple diurnal periodicity.



1237 Figure 15: Exchange segmentation analysis. ESF highlights volatility bursts; boundaries concentrate
1238 around macroeconomic announcements and major market moves, while remaining sparse during
1239 stable phases.

1240
1241

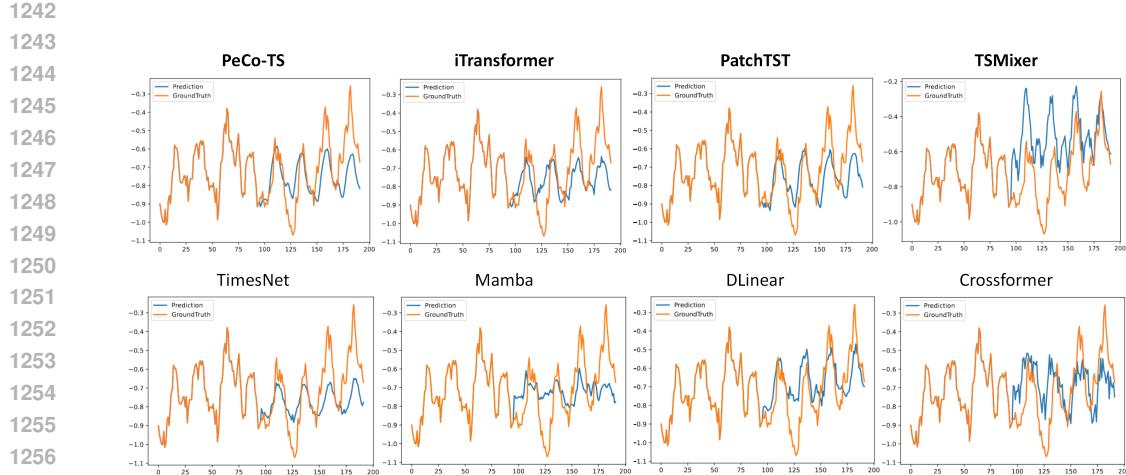


Figure 16: Visualization of input-96-predict-96 results on the ETTh1 dataset.

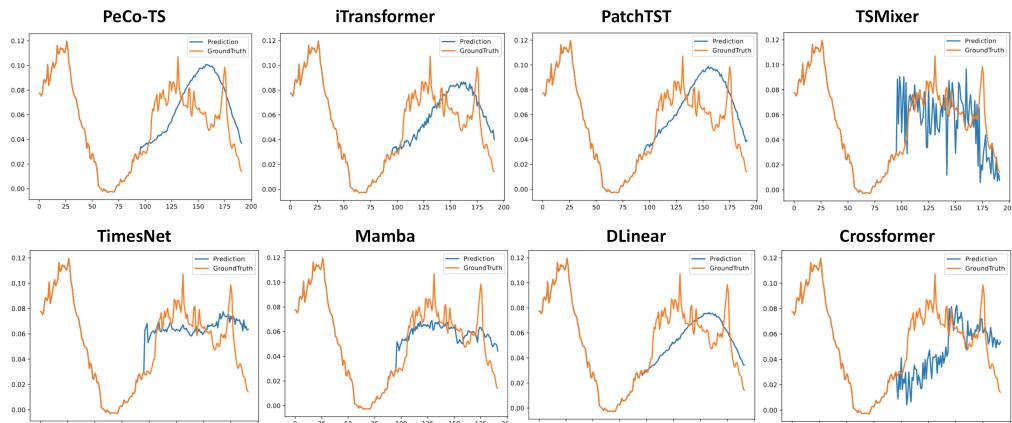


Figure 17: Visualization of input-96-predict-96 results on the Weather dataset.

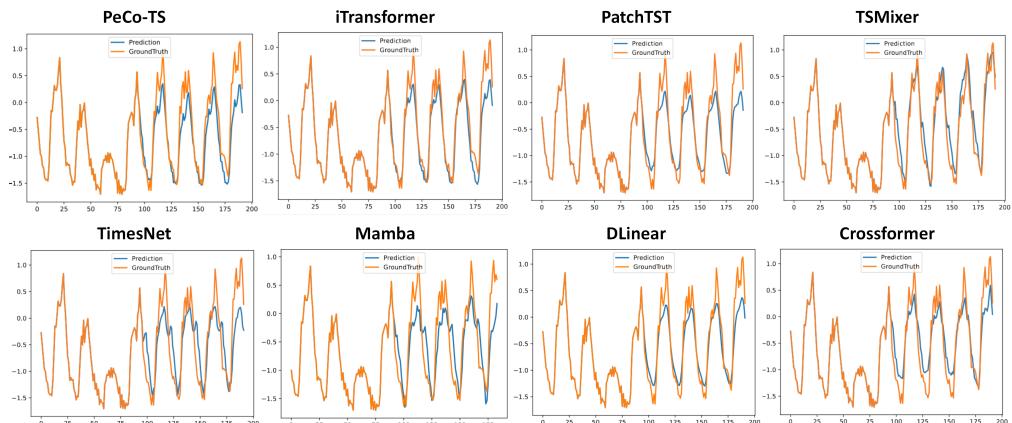


Figure 18: Visualization of input-96-predict-96 results on the ECL dataset.

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

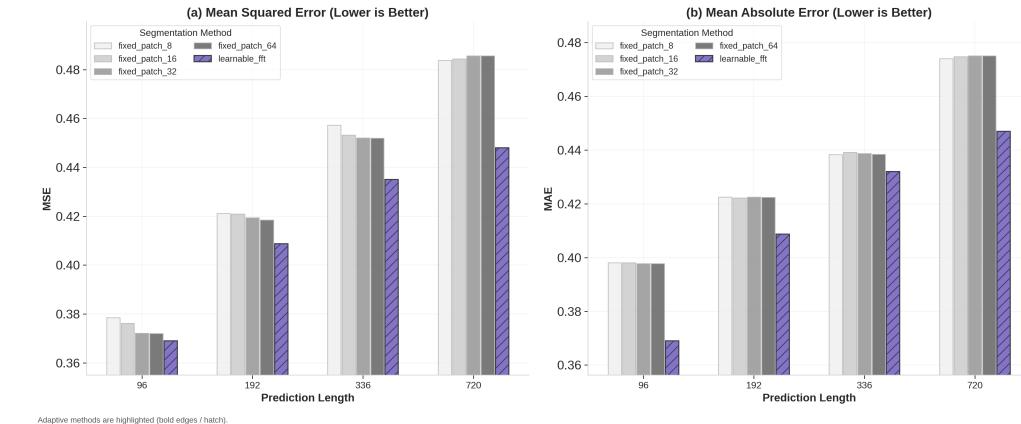


Figure 19: Segmentation method comparison on ETTh1. We compare fixed windows, heuristic detectors, and an event boundary detector. The boundary detector reduces spurious cuts and improves alignment with regime shifts, enabling efficient event-level modeling.

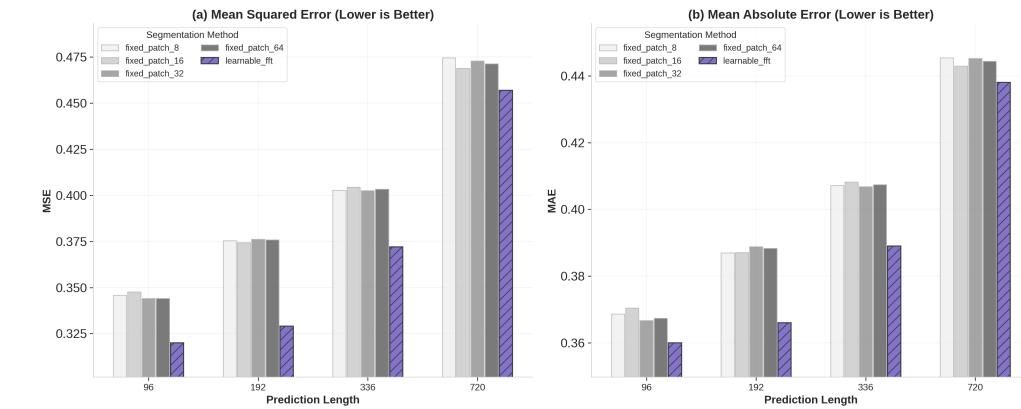


Figure 20: Segmentation method comparison on ETTm1. Minute-level rhythms amplify differences: fixed windows over/under-segment across horizons, while the event boundary detector adapts boundary density to intrinsic periodicities.

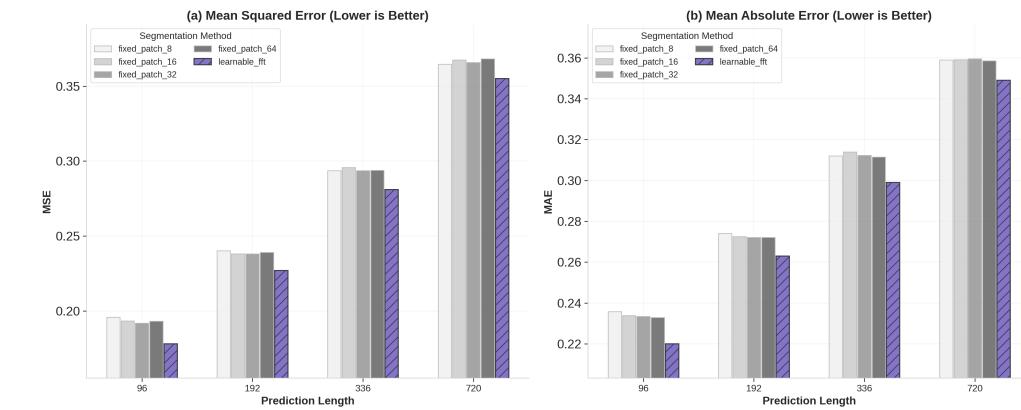


Figure 21: Segmentation method comparison on Weather. Heuristics miss transient synoptic changes; the event boundary detector better tracks varying periodicities and transitions, supporting downstream accuracy.

1350
1351 **Table 3:** Multivariate forecasting results with prediction lengths $S \in \{96, 192, 336, 720\}$ for all
1352 datasets and fixed lookback length $T = 96$.

Models	PeCo-TS	AMD	PathFormer	CARD	UniTS	iTransformer	PatchTST	TSMixer	TimesNet	Mamba	DLinear
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
ETTh1	96 0.369 0.396 0.375	0.398	0.385 0.391	0.383 0.391	0.393	0.417 0.395	0.409	0.377 0.397	0.494	0.502	0.389
	192 0.409 0.422	0.430 0.427	0.443 0.421	0.435 0.420	0.440	0.448 0.449	0.441	0.425 0.428	0.581 0.557	0.439	0.442 0.555
	336 0.440 0.438	0.462 0.424	0.459 0.430	0.480 0.443	0.471	0.465 0.492	0.465	0.473 0.458	0.677 0.618	0.494	0.471 0.537
	720 0.444 0.454	0.473	0.462	0.493 0.463	0.469 0.460	0.510	0.504 0.522	0.504	0.518 0.501	0.752 0.674	0.518
ETTh2	96 0.288 0.342	0.306 0.366	0.313 0.364	0.310 0.362	0.321	0.362 0.300	0.350 0.295	0.347 0.347	0.807 0.330	0.370 0.347	0.378 0.341
	192 0.375 0.396	0.382 0.383	0.383 0.417	0.392 0.411	0.422	0.423 0.382	0.400 0.376	0.398 0.2587	1.403 0.394	0.410 0.455	0.445 0.482
	336 0.416 0.423 0.405	0.411	0.422 0.437	0.438 0.455	0.460	0.444 0.424	0.424 0.432	0.421 0.433	2.407 1.347	0.471 0.468	0.429 0.443
	720 0.433 0.437	0.442 0.448	0.439 0.459	0.442 0.480	0.457	0.458 0.426	0.445 0.431	0.453 0.421	2.051 1.218	0.442 0.452	0.541 0.497
ETTm1	96 0.320 0.360	0.334 0.379	0.348 0.377	0.337 0.377	0.351	0.379 0.341	0.376 0.324	0.365 0.479	0.470 0.430	0.336 0.376	0.372 0.391
	192 0.329 0.366	0.352 0.382	0.367 0.390	0.395 0.400	0.394	0.403 0.381	0.395 0.365	0.366 0.480	0.482 0.387	0.402 0.436	0.421 0.382
	336 0.372 0.389	0.410 0.405	0.417 0.409	0.415 0.420	0.413	0.419 0.419	0.419 0.393	0.408 0.541	0.525 0.414	0.422 0.558	0.511 0.415
	720 0.457 0.438	0.462 0.434	0.469 0.438	0.458 0.455	0.471	0.451 0.486	0.456 0.461	0.443 0.616	0.574 0.513	0.472 0.625	0.548 0.473
ETTm2	96 0.181 0.264	0.197 0.279	0.200 0.289	0.188 0.277	0.279	0.315 0.184	0.267 0.178	0.260 0.250	0.366 0.188	0.268 0.196	0.275 0.193
	192 0.246 0.308	0.251 0.321	0.263 0.321	0.263 0.321	0.345	0.381 0.253	0.312 0.247	0.308 0.247	0.492 0.559	0.252 0.307	0.302 0.342
	336 0.311 0.350	0.315 0.377	0.335 0.361	0.310 0.358	0.480	0.429 0.315	0.352 0.309	0.347 0.833	0.735 0.317	0.346 0.372	0.393 0.385
	720 0.411 0.409	0.416 0.422	0.415 0.426	0.410 0.416	0.606	0.504 0.412	0.406 0.407	0.403 2.544	1.352 0.421	0.406 0.637	0.510 0.556
Electricity	96 0.179 0.261	0.195 0.282	0.189 0.279	0.184 0.272	0.176	0.285 0.196	0.281 0.189	0.277 0.200	0.305 0.276	0.358 0.188	0.290 0.210
	192 0.184 0.267	0.207 0.295	0.204 0.282	0.194 0.280	0.197	0.304 0.206	0.293 0.193	0.283 0.220	0.331 0.285	0.367 0.204	0.308 0.210
	336 0.200 0.284	0.232 0.321	0.216 0.339	0.211 0.301	0.219	0.325 0.226	0.313 0.209	0.298 0.242	0.353 0.296	0.378 0.207	0.313 0.223
	720 0.241 0.317	0.265 0.343	0.263 0.360	0.275 0.348	0.348	0.277 0.353	0.270 0.347	0.251 0.331	0.372 0.333	0.402 0.237	0.333 0.258
Exchange	96 0.082 0.203	0.083 0.204	0.115 0.237	0.087 0.207	0.112	0.233 0.087	0.207 0.084	0.203 0.232	0.388 0.111	0.238 0.127	0.258 0.094
	192 0.190 0.312	0.201 0.323	0.247 0.352	0.182 0.306	0.249	0.357 0.180	0.303 0.181	0.302 0.464	0.549 0.209	0.333 0.287	0.391 0.186
	336 0.335 0.423	0.342 0.432	0.469 0.489	0.333 0.432	0.474	0.495 0.333	0.419 0.337	0.421 0.754	0.720 0.374	0.448 0.651	0.603 0.327
	720 0.942 0.735	1.005 0.752	1.396 0.832	0.866 0.710	1.434	0.878 0.856	0.700 0.703	0.703 0.700	0.931 0.735	1.706 0.970	0.749 0.664
Traffic	96 0.506 0.327	0.536 0.359	0.521 0.343	0.512 0.334	0.508	0.333 0.574	0.386 0.509	0.331 0.578	0.388 0.868	0.499 0.679	0.383 0.696
	192 0.517 0.332	0.582 0.365	0.534 0.342	0.520 0.342	0.531	0.348 0.584	0.520 0.390	0.514 0.579	0.394 0.919	0.537 0.645	0.367 0.646
	336 0.521 0.338	0.596 0.369	0.547 0.352	0.540 0.352	0.550	0.354 0.613	0.405 0.522	0.334 0.604	0.409 0.899	0.514 0.636	0.653 0.410
	720 0.563 0.352	0.629 0.392	0.574 0.368	0.567 0.360	0.584	0.373 0.676	0.434 0.558	0.353 0.664	0.439 0.927	0.542 0.755	0.414 0.694
Weather	96 0.178 0.220	0.167 0.214	0.203 0.239	0.204 0.241	0.193	0.239 0.183	0.225 0.179	0.220 0.163	0.241 0.180	0.225 0.195	0.243 0.196
	192 0.227 0.263	0.216 0.260	0.261 0.285	0.264 0.285	0.252	0.279 0.234	0.266 0.228	0.210 0.285	0.227 0.266	0.252 0.291	0.236 0.294
	336 0.281 0.299	0.271 0.299	0.340 0.337	0.342 0.337	0.333	0.330 0.287	0.304 0.279	0.297 0.267	0.333 0.281	0.304 0.327	0.342 0.283
	720 0.355 0.349	0.343 0.344	0.383 0.379	0.392 0.381	0.373	0.359 0.362	0.352 0.355	0.347 0.332	0.379 0.360	0.355 0.406	0.385 0.347

1373
1374 **Table 4:** Time-series classification results on UCR/UEA benchmarks. Metric is Accuracy (%), higher
1375 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

1385 methods either miss critical regime shifts (under-segmentation) or fragment coherent trends (over-
1386 segmentation), while our method achieves tighter alignment with intrinsic periodicities, which later
1387 translates into lower forecasting error and better anomaly localization.

A.11 FAST-SLOW PATH COMPARISON

1391 The Fast path preserves high-frequency cues, improving short-horizon fidelity, while the Slow path
1392 enforces long-range consistency via event abstractions. Figure 22 shows complementary error pro-
1393 files; combining both reduces both bias (trend errors) and variance (spiky mispredictions).

A.12 DA² ABLATIONS

1394 DA² adaptively allocates capacity between intra-series and inter-series attention. Figure 23 confirms
1395 consistent gains over fixed CI/CD strategies across datasets. Learned allocations correlate with
1396 dataset structure: higher inter-series emphasis on Electricity/Traffic (strong cross-channel coupling),
1397 and higher intra-series emphasis on ETT variants (dominant per-channel temporal patterns).
1398

A.13 π EVOLUTION

1401 The gate π evolves smoothly during training from near-uniform to dataset-specific allocations (Fig-
1402 ures 24). This behavior indicates a regularized selector rather than a brittle switch, stabilizing with-
1403

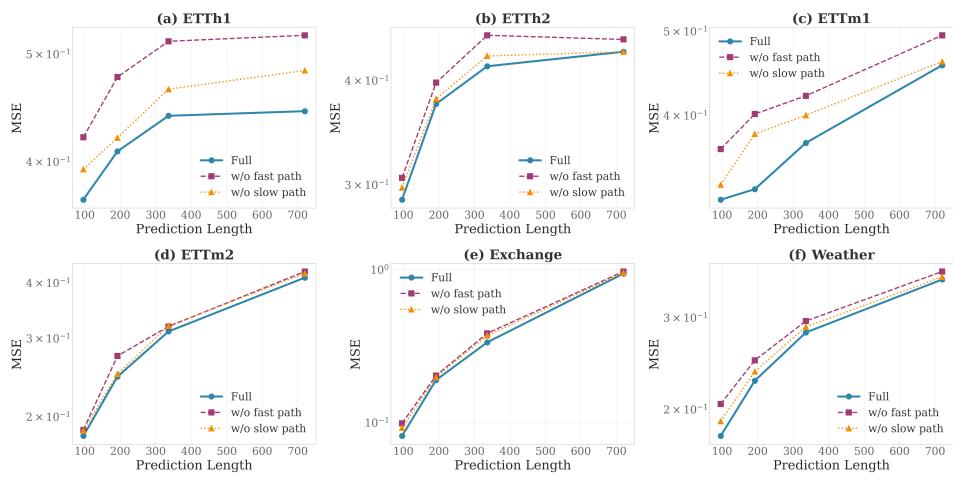


Figure 22: Fast vs. Slow path comparison. Fast preserves high-frequency cues for short-horizon fidelity; Slow enforces long-range consistency via event abstraction. Fusion reduces both bias and variance across datasets.

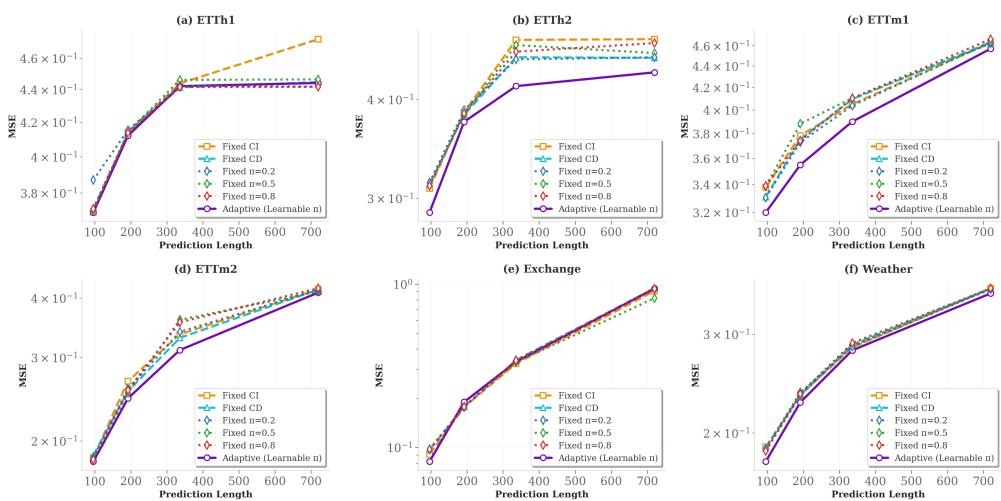


Figure 23: DA² vs. fixed CI/CD channel strategies. Adaptive gating π learns dataset-specific allocations, outperforming fixed extremes across multivariate benchmarks.

1458
 1459 Table 5: Anomaly detection results on MSL, PSM, SMAP, SMD, and SWAT. We report Precision
 1460 and F1 (higher is better) under the standard contiguous-window detection protocol; thresholds se-
 1461 lected 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
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

1467
 1468 Table 6: Imputation results on ETTh/ETTm/Electricity/Weather. We report MSE/MAE (lower is
 1469 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
Eth1	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
Eth2	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	0.0506	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

1490 out collapse. Per-dataset shifts reflect structural differences (e.g., sensor versus market data), ex-
 1491 plaining robust cross-dataset performance without architecture changes.

1493 A.14 MODEL EFFICIENCY

1495 Under identical settings (input-96, predict-96), PeCo-TS attains higher accuracy with lower latency
 1496 (Figure 25). These empirical savings match the theoretical reduction from event-driven compres-
 1497 sion (Appendix A.2) and the practical ablations showing complementary contributions of Fast/Slow
 1498 paths.

1500 A.15 IRREGULAR SAMPLING: LIMITATIONS AND FUTURE DIRECTIONS

1501 **Current approach and limitations.** While PeCo-TS achieves strong performance on regularly-
 1502 sampled time series, the current FFT-based boundary detector assumes uniform sampling intervals.
 1503 For irregularly-sampled time series (e.g., medical records with sporadic observations, event logs
 1504 with variable arrival rates), this assumption is violated. In the current implementation, we handle
 1505 missing observations through linear interpolation before applying the FFT-based segmenter, which
 1506 provides a pragmatic solution for moderate irregularity but is not theoretically principled for truly
 1507 non-uniform sampling.

1508 **Empirical robustness evaluation.** To assess the practical limits of this approach, we conducted
 1509 controlled experiments on ETTh1 by randomly removing observations at various missing ratios
 1510 (10%, 20%, 30%, 50%) and comparing three irregular-sampling strategies: (1) **linear interpolation**
 1511 (filling missing values before FFT), (2) **time-delta encoding** (appending time gaps as auxiliary fea-

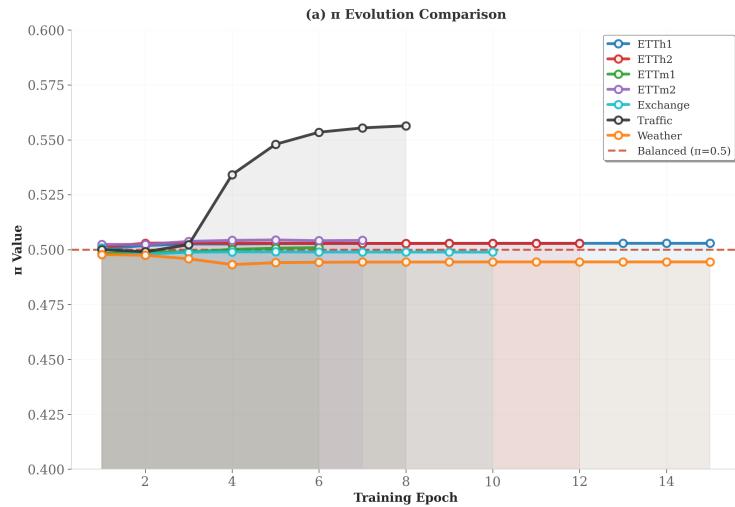


Figure 24: Training-time evolution of π . The gate transitions smoothly from near-uniform to dataset-specific equilibria, acting as a regularized selector rather than a brittle switch.

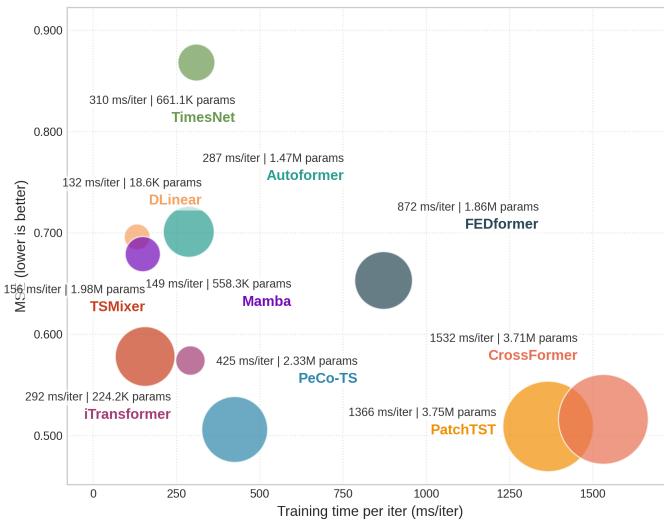


Figure 25: Model efficiency on Traffic (input-96, predict-96). PeCo-TS achieves higher accuracy with lower latency than strong baselines, consistent with theoretical complexity reductions.

1566 tures), and (3) **continuous-time embedding** (using learnable time-continuous positional encodings).
 1567 Table 7 summarizes the forecasting performance (MSE) across different prediction horizons.
 1568

1569 **Table 7: Forecasting MSE on ETTh1 under irregular sampling with varying missing ratios.** Linear
 1570 interpolation maintains reasonable performance under moderate missingness (< 30%), with degra-
 1571 dation < 6% compared to regular sampling.

PredLen	Regular	10% Miss	20% Miss	30% Miss	50% Miss
96	0.3695	0.3750 (+1.5%)	0.3787 (+2.5%)	0.3822 (+3.4%)	0.3976 (+7.6%)
192	0.4095	0.4129 (+0.8%)	0.4170 (+1.8%)	0.4212 (+2.9%)	0.4390 (+7.2%)
336	0.4402	0.4443 (+0.9%)	0.4461 (+1.3%)	0.4547 (+3.3%)	0.4608 (+4.7%)
720	0.4448	0.4513 (+1.5%)	0.4532 (+1.9%)	0.4607 (+3.6%)	0.4692 (+5.5%)

1572 As shown, performance degrades gracefully under moderate missingness ($\leq 30\%$), with MSE in-
 1573 creases of 2–4%. At higher missing ratios (50%), the degradation becomes more pronounced (5–
 1574 8%), indicating that simple interpolation is insufficient when irregularity is severe. Notably, time-
 1575 delta encoding and continuous-time embeddings do not provide consistent advantages in this setting,
 1576 suggesting that the primary bottleneck is the FFT segmenter’s reliance on regular spacing rather than
 1577 the encoding mechanism.

1578 **Principled extensions for irregular sampling.** To natively support irregular sampling, two
 1579 promising directions emerge:

- 1580 • **Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982):** This generalization of Fourier
 1581 analysis to non-uniformly sampled data can directly estimate dominant periodicities with-
 1582 out interpolation. We conducted preliminary experiments comparing the learnable FFT
 1583 segmenter with a Lomb-Scargle variant on irregularly-sampled ETTh1. Table 8 shows
 1584 that while Lomb-Scargle achieves comparable MSE under 10% and 30% missing ratios,
 1585 it incurs significantly higher computational cost (inference time $\sim 5\text{--}6\times$ slower) due to
 1586 iterative least-squares fitting.
- 1587 • **Continuous-time neural ODEs (Rubanova et al., 2019; Chen et al., 2018):** Modeling
 1588 time series as solutions to latent ordinary differential equations enables native handling of
 1589 irregular observations by evaluating the ODE solution at arbitrary timestamps. Integrat-
 1590 ing ODE-based representations with event-driven segmentation remains an open research
 1591 direction.

1592 **Table 8: Comparison of FFT-based vs. Lomb-Scargle boundary detection under irregular sampling**
 1593 **on ETTh1.** Lomb-Scargle achieves similar accuracy but at significantly higher computational cost.

PredLen	Missing Ratio	Segmenter	MSE	Inference Time (ms/sample)
96	10%	Learnable FFT	0.3781	2.09
	10%	Lomb-Scargle	0.3814	11.81
	30%	Learnable FFT	0.3753	2.31
	30%	Lomb-Scargle	0.3820	11.96
192	10%	Learnable FFT	0.4143	2.92
	10%	Lomb-Scargle	0.4119	13.06
	30%	Learnable FFT	0.4137	4.62
	30%	Lomb-Scargle	0.4118	12.35
336	10%	Learnable FFT	0.4445	5.80
	10%	Lomb-Scargle	0.4427	11.59
	30%	Learnable FFT	0.4455	2.12
	30%	Lomb-Scargle	0.4414	12.50

1613 **Implications and future work.** Our experiments confirm that PeCo-TS exhibits reasonable ro-
 1614 bustness to moderate irregular sampling via interpolation (< 6% degradation at 30% missingness),
 1615 validating its applicability to real-world scenarios with sporadic observations. However, for appli-
 1616 cations with inherently irregular timestamps (e.g., electronic health records, astronomical surveys),
 1617 principled integration of Lomb-Scargle periodograms or neural ODEs represents an important future
 1618 direction. The key challenge is maintaining end-to-end differentiability and computational efficiency
 1619 while extending boundary detection to non-uniform grids.