CONTEXT-CONDITIONED TOKEN FORMATION FOR MEDICAL TIME SERIES CLASSIFICATION FROM SHORT TO LONGER WINDOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Local discriminative patterns in medical time series often depend on neighboring waveforms and the overall physiological state, whereas conventional token formation compresses local observations before contextual modeling, potentially weakening fine-grained cues that can only be identified with sufficient context. To address this issue, we propose ReConTok, a context-conditioned token formation framework for medical time series classification. Its core module, Context-guided Token Gating and Fusion (CoTGF), jointly exploits the current local representation, neighborhood context, and global summary to conditionally gate and fuse fine-grained features. The same mechanism is applied to both temporal and channel tokens while leaving the subsequent Transformer interaction unchanged. Across five EEG/ECG datasets, ReConTok achieves the best F1 on four datasets and shows consistent advantages in Recall. Controlled tokenizer comparisons, short-versus-long window experiments, and robustness analyses further validate its effectiveness. These results demonstrate that directly incorporating context into token formation yields more discriminative and robust representations for medical time series. Code is available in the supplementary material.

1 INTRODUCTION

Medical time series, such as electroencephalograms (EEG) and electrocardiograms (ECG), record changes in physiological activity over time, while synchronous multichannel measurements provide complementary physiological observations (Rim et al., 2020). Identifying disease-related patterns and physiological states in these signals underpins computer-aided diagnosis and clinical monitoring (Hannun et al., 2019; Tatum et al., 2018; Perslev et al., 2021).

Medical time-series classification relies on discriminative local patterns whose meaning depends on their context. EEG waveforms must be interpreted alongside background activity, temporal organization, and their distribution across electrodes (Kural et al., 2020a;b); ECG diagnosis likewise depends on waveform components and relationships among leads (Smith et al., 2012; Kontos et al., 2022). Physiological states and signal characteristics also vary substantially across subjects (Ding et al., 2025; Jemal et al., 2024). Thus, the discriminative value of a local pattern changes with its neighborhood, the overall state, and the particular sample.

The range of available context also affects the discriminative information a model can access. Existing medical time-series benchmarks commonly use short windows, whereas complete recordings preserve broader temporal relationships and state changes. For example, PTB-XL diagnostic labels describe complete ECG examinations, although common protocols further divide them into short windows (Wagner et al., 2020). Extending the input range provides more complete physiological context but also introduces more local patterns and background information, placing greater demands on representation construction.

Transformers provide a foundation for modeling long-range context (Vaswani et al., 2017). Prior work improves time-series modeling through efficient attention (Kitaev et al., 2020; Zhou et al., 2021), temporal-structure modeling (Wu et al., 2021; Zhou et al., 2022; Liu et al., 2022), and token organization (Nie et al., 2023; Zhang et al., 2024b; Liu et al., 2024). For medical time series, Medformer (Wang et al., 2024a) models multigranular and cross-channel patterns, TeCh (Yu et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2026) strengthens global interactions between temporal and channel tokens, and MedSpaformer (Ye et al., 2026) reduces redundant tokens. These methods primarily improve interaction, aggregation, or selection after tokens are formed, while the information available to subsequent encoders remains limited by the quality of token formation.

Existing tokenization methods improve token representations through temporal patching (Nie et al., 2023), multiscale representations (Zhang et al., 2024a), variable-wise organization (Liu et al., 2024), and selective reassembly (Wu et al., 2025). Yet token content is still typically produced by a mapping shared across samples, without explicitly using the current sample's context to organize information within a token. In heterogeneous medical time series, important patterns may be compressed into compact tokens together with background information; this problem becomes more pronounced as the input range or patch granularity grows. Effective modeling therefore calls for the current sample's context to participate directly in token formation, alongside interactions among tokens.

We propose **ReConTok**, a context-conditioned token formation framework for medical time-series classification. Its core module, **Context-guided Token Gating and Fusion (CoTGF)**, combines the current local representation, neighborhood information, and global context to conditionally gate and fuse fine-grained features in both temporal and channel tokens. ReConTok concentrates its changes in token formation before the Transformers, leaving subsequent token interactions unchanged.

Our main contributions are:

- **A context-conditioned token formation framework.** We propose ReConTok, which lets context from the current sample participate directly in token formation and dynamically organize local information while leaving subsequent Transformer interactions unchanged.
- **A unified temporal and channel conditioning mechanism.** We develop CoTGF to conditionally reorganize fine-grained features using current, neighborhood, and global context, while accommodating the different structures of temporal and channel tokens.
- **Systematic empirical evaluation.** Across five EEG/ECG datasets, ReConTok consistently improves classification, particularly Recall, while controlled tokenizer comparisons, ablations, robustness tests, and short-/long-window analyses further support context-conditioned token formation.

2 Related Work

2.1 Token Formation in Time Series

Token formation determines how continuous observations are organized into representational units before contextual modeling, directly affecting the granularity and content available to subsequent Transformers (Dosovitskiy et al., 2021). Vision and multimodal research has explored adaptive token formation through token generation (Ryoo et al., 2021), token merging and fusion (Bolya et al., 2023; Kim et al., 2024; Wang et al., 2022), and token sampling and pruning (Rao et al., 2021; Fayyaz et al., 2022; Yin et al., 2022), with the goals of reducing redundancy and improving adaptability to the input.

For time series, PatchTST (Nie et al., 2023) and Crossformer (Zhang & Yan, 2023) construct patch tokens through local temporal aggregation. MTST (Zhang et al., 2024a) uses multiscale token representations, while iTransformer (Liu et al., 2024) organizes variables as tokens, extending token organization from the perspectives of temporal granularity and variable relationships. More recently, SRSNet (Wu et al., 2025) selects and rearranges high-information patches to construct an input-dependent representation space.

These methods have advanced token granularity, organization, and selection. However, token content is generally produced by a construction mapping shared across samples, without explicitly conditioning the aggregation of fine-grained information on the context of the current sample. For samples that differ in temporal dynamics, variable relationships, and the distribution of discriminative information, this shared construction may limit the representation of sample-specific information.

2.2 Medical Time-Series Classification

Medical time-series classification requires modeling temporal dynamics and cross-channel relationships in multichannel physiological signals such as EEG and ECG (Liu et al., 2021; Roy et al., 2019).

Early methods relied primarily on handcrafted time-domain, frequency-domain, and physiological features; convolutional neural networks (CNNs) and recurrent neural networks (RNNs) later began learning representations directly from raw signals (Lawhern et al., 2018; Hannun et al., 2019; Esteva et al., 2019). More recently, Transformers have strengthened the modeling of long-range temporal dependencies and multichannel relationships.

For medical time series, Medformer (Wang et al., 2024a) models patterns at different temporal scales through multigranular patches, TeCh (Yu et al., 2026) constructs temporal and channel tokens and strengthens their global interactions, and MedSpaformer (Ye et al., 2026) combines multiscale representations with token sparsification to reduce redundancy. These methods primarily enhance interaction, aggregation, or selection after tokens have been formed, whereas the information available to a downstream encoder remains limited by the quality of the representations produced during token formation.

Medical time series exhibit substantial sample heterogeneity: discriminative temporal patterns, channel relationships, and physiological states can differ across subjects (Ding et al., 2025; Jemal et al., 2024). Consequently, token construction shared across samples may not adequately adapt to sample-specific information distributions. ReConTok introduces current, neighborhood, and global context into token formation, allowing local content to be organized dynamically according to the current input while leaving subsequent Transformer interactions unchanged.

In addition, existing medical time-series benchmarks commonly use short-window classification, dividing continuous recordings into shorter segments for training and evaluation (Wang et al., 2024a; Yu et al., 2026). Some diagnostic labels, however, correspond to longer recordings or subject-level states (Wagner et al., 2020). Extending the input range preserves broader physiological context, but also introduces more local patterns and background information, making the retention of discriminative information during token formation more consequential. ReConTok's context-conditioned formation mechanism provides a more flexible representational basis for using such longer continuous inputs.

3 PRELIMINARIES

Recordings and derived windows from the same subject share physiological characteristics. Randomly splitting windows could therefore place the same subject in both training and test sets and overstate performance on unseen subjects (Wang et al., 2024b). Subject-independent evaluation keeps all data from each subject in one training, validation, or test split. To examine the value of longer continuous context, we compare paired PTB-XL inputs from each ECG examination: its first second (Short) and the complete 10 seconds (Long). The longer input preserves temporal relationships and state changes that may be absent from its one-second counterpart. Labels, sample counts, and subject splits are held fixed, so only the observed duration changes. Formal definitions appear in Appendix F.

4 METHOD

4.1 FRAMEWORK OVERVIEW

Given an input window $X \in \mathbb{R}^{T \times C}$, ReConTok addresses the difficulty of adapting a shared token construction rule to the sample-specific distribution of discriminative information. In medical time series, the significance of a local pattern often depends on neighboring waveforms, temporal organization, and cross-channel information (Kural et al., 2020a;b; Kontos et al., 2022). We therefore retain fine-grained information before token aggregation and use current, neighborhood, and global context to adjust its contribution, forming token representations suited to the current input.

As shown in Figure 1(a), ReConTok constructs tokens from two complementary temporal and channel perspectives. The temporal branch treats a local time segment as a token and retains its fine-grained channel representation. The channel branch treats an individual channel as a token and retains its fine-grained temporal representation. These branches can thus adjust the contributions of different channels and time positions, respectively.

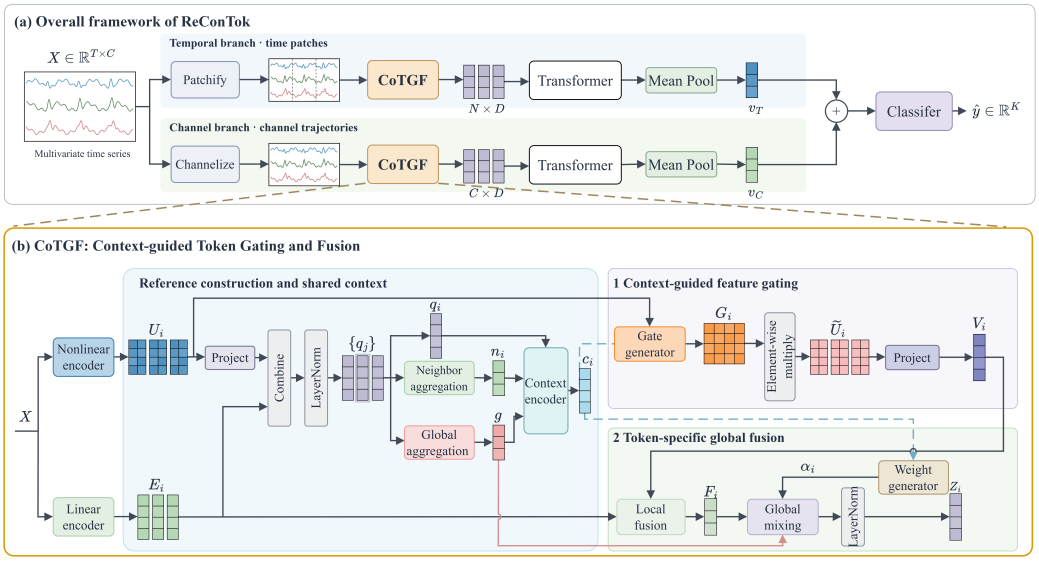


Figure 1: Overview of ReConTok and Context-guided Token Gating and Fusion (CoTGF). (a) The temporal and channel branches form context-conditioned tokens, encode them with separate Transformers, and fuse their representations for classification. (b) CoTGF builds the condition vector c_i from the reference representation q_i, neighborhood context n_i, and global summary g, then regulates token formation through the gate G_i and fusion coefficient α_i.

Both branches use **Context-guided Token Gating and Fusion (CoTGF)** for conditioned token construction, as shown in Figure 1(b). CoTGF combines a base representation and fine-grained features to form a reference representation, then uses neighborhood information and a global summary to construct a context condition. This condition modulates the contribution of fine-grained features and controls the fusion of global information. The resulting temporal and channel tokens are passed to separate standard Transformer encoders; their pooled representations are fused for classification. ReConTok thus changes token formation before the Transformers while leaving the subsequent token interaction mechanisms unchanged.

4.2 Context Condition Construction

The two branches retain fine-grained channel features within each temporal token and fine-grained temporal features within each channel token. Directly aggregating these features would form tokens mainly from information within each token, without indicating whether a local pattern is discriminative in the context of neighboring observations and the overall physiological state of the current sample. CoTGF therefore first constructs a context condition from the current input and then uses it to guide final token formation.

CoTGF begins with a reference token that has not yet been adjusted by context. Inspired by the feature-wise core aggregation in SOFTS (Han et al., 2024), we define a dynamic aggregation operator P_b that allows different feature dimensions to aggregate information from different fine-grained positions. For $A \in \mathbb{R}^{L_b \times M_b \times f}$, we compute

$$\begin{aligned} H_{i,m} &= W_2^P \,\mathrm{GELU}(W_1^P A_{i,m} + b_1^P) + b_2^P, & \omega_{i,m,k} &= \mathrm{softmax}_m(H_{i,m,k}/\tau), \\ \bar{H}_{i,k} &= \sum_{m=1}^{M_b} \omega_{i,m,k} H_{i,m,k}, & P_b(A)_i &= W_{\mathrm{out}}^P \bar{H}_i + b_{\mathrm{out}}^P \in \mathbb{R}^D. \end{aligned} \tag{1}$$

where $k = 1, \ldots, f$, and the softmax operates only over the fine-grained positions m within a token. Unlike fixed averaging or a single projection over all positions, this aggregation lets different feature dimensions attend to different fine-grained positions, providing a content-adaptive initial representation for compact token construction.

Before introducing the context condition, we obtain

$$V_i^{0,b} = P_b(U^b)_i, \qquad q_i^b = \mathrm{LN}_{\mathrm{ref}}^b\left(E_i^b + V_i^{0,b}\right), \tag{2}$$

where $q_i^b \in \mathbb{R}^D$ is the i-th reference token. It combines the base representation from the linear path with the initial aggregation of fine-grained features, describing the token before conditioning. This two-stage design lets the model first observe the current input and then use the resulting context to guide final token formation.

The reference token q_i^b alone is insufficient to determine the relative importance of local information within the current sample. We therefore introduce relational context n_i^b and a global summary g^b. In the temporal branch, n_i^{T} describes local changes among neighboring time segments. In the channel branch, n_i^{C} captures the relationship between the current channel and other channels:

$$\begin{aligned} n_i^{\mathrm{T}} &= \mathrm{Agg}\left(\{q_j^{\mathrm{T}} : j \in \mathcal{N}(i)\}\right), \\ n_i^{\mathrm{C}} &= \sum_{j \neq i} \mathrm{softmax}_{j \neq i}\left(\frac{(W_Q q_i^{\mathrm{C}})^\top (W_K q_j^{\mathrm{C}})}{\sqrt{D}}\right) q_j^{\mathrm{C}}. \end{aligned} \tag{3}$$

We aggregate all reference tokens from the current input to obtain a global summary,

$$g^b = \mathrm{Pool}_b\left(q_1^b, \ldots, q_{L_b}^b\right) \in \mathbb{R}^D, \tag{4}$$

which provides context for the overall physiological state of the input.

Finally, the current reference representation, relational context, and global summary are jointly encoded into a token-specific condition vector:

$$d_i^b = \mathrm{LN}_{3D}^b([q_i^b; n_i^b; g^b]), \qquad c_i^b = \mathrm{LN}_r^b\big(W_{c2}^b\,\mathrm{GELU}(W_{c1}^b d_i^b + b_{c1}^b) + b_{c2}^b\big) \in \mathbb{R}^r. \tag{5}$$

Thus, c_i^b reflects the token itself, its relationship to relevant time segments or physiological channels, and the overall state of the current input. It guides context-conditioned token formation in the next stage.

4.3 Context-guided Token Formation

Given the condition vector c_i^b constructed from the current representation, neighborhood information, and global summary, CoTGF introduces context directly into token formation. It first modulates the contribution of fine-grained positions within a token before aggregation, and then controls the fusion of local content with the global state.

Context-conditioned gating. For the fine-grained representation U_i^b of token i, we define

$$\begin{aligned} \ell_{i,m,:}^b &= W_o^b\,\mathrm{GELU}\left(W_u^b U_{i,m,:}^b + b_u^b + W_c^b c_i^b\right) + b_o^b, \\ \pi_{i,m,k}^b &= \mathrm{softmax}_m(\ell_{i,m,k}^b), \qquad G_{i,m,k}^b = 1 + \rho_b(M_b \pi_{i,m,k}^b - 1), \qquad \rho_b = \sigma(\theta_\rho^b). \end{aligned} \tag{6}$$

The softmax runs over the positions m within a token. It therefore redistributes fine-grained information across channels in the temporal branch and across time positions in the channel branch. Centering the gate at 1 preserves a path for the original information, while ρ_b controls the strength of modulation. The conditioned content representation is

$$V_i^b = P_b\big(G_i^b \odot U_i^b\big) \in \mathbb{R}^D, \tag{7}$$

where P_b is shared with the reference stage, so context acts before the final aggregation.

Local and global information fusion. The base representation E_i^b and conditioned content V_i^b are first fused locally. Inspired by the star-shaped series-core fusion in SOFTS (Han et al., 2024), the global summary g^b then provides sample-level context, weighted for each token by its condition vector:

$$\begin{aligned} F_i^b &= W_f^b[E_i^b; V_i^b] + b_f^b, \qquad \alpha_i^b = \sigma(W_\alpha^b c_i^b + b_\alpha^b), \\ Z_i^b &= \mathrm{LN}_{\mathrm{out}}^b\big((1-\alpha_i^b)F_i^b + \alpha_i^b g^b\big). \end{aligned} \tag{8}$$

Here, G_i^b controls the composition of fine-grained information within the token, while α_i^b controls how much of the global state is mixed with the locally fused representation. Together, they form a token representation adapted to the current input.

4.4 RECONTOK

Temporal dynamics and cross-channel relationships are complementary structures in multivariate medical time series (Qiu et al., 2025; Yu et al., 2026; Ye et al., 2026). ReConTok applies CoTGF to both temporal and channel branches. The temporal branch adjusts the contribution of different channels to each temporal token, whereas the channel branch adjusts the contribution of different time positions to each channel token. Both types of tokens are thereby adapted to the current sample before entering their encoders.

After CoTGF, the conditioned tokens enter separate Transformer encoders and are pooled into sample-level representations:

$$Z'^b = \mathrm{Enc}_b(Z^b), \qquad v_b = \mathrm{Pool}_b(Z'^b) \in \mathbb{R}^D, \qquad b \in \{\mathrm{T}, \mathrm{C}\}. \tag{9}$$

The temporal branch models dynamic dependencies among time segments and uses pooling weighted by each patch's valid coverage. The channel branch models relationships among physiological channels and uses mean pooling.

The two branch representations are added in a shared feature space, and a linear classification head produces the prediction:

$$v = \sum_{b \in \mathcal{B}} v_b, \qquad o = W_{\mathrm{cls}} v + b_{\mathrm{cls}}, \qquad \hat{y} = \mathrm{softmax}(o), \tag{10}$$

where $\mathcal{B} \subseteq \{\mathrm{T}, \mathrm{C}\}$ denotes the enabled branches and $\hat{y} \in [0,1]^K$ is the predicted class-probability vector. The model is optimized end-to-end with cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{B} \sum_{a=1}^{B} \log \hat{y}_{a,y_a}. \tag{11}$$

ReConTok concentrates its modeling changes in token formation before the Transformers; the subsequent encoding, pooling, and classification follow standard structures.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. Following the data construction, preprocessing, and split protocol of Medformer (Wang et al., 2024a), we evaluate on five medical time-series datasets. APAVA (Escudero et al., 2006; Smith et al., 2017), ADFTD (Miltiadous et al., 2023), and TDBRAIN (van Dijk et al., 2022) are EEG classification tasks; PTB (Bousseljot et al., 1995; Goldberger et al., 2000) and PTB-XL (Wagner et al., 2020) are ECG classification tasks. Table 1 summarizes the processed short-window benchmark.

Table 1: Processed short-window dataset statistics.

Dataset	Subjects	Samples	Classes	Channels	Time points
APAVA	23	5,967	2	16	256
TDBRAIN	50	6,000	2	32	256
ADFTD	88	69,752	3	19	256
PTB	198	64,356	2	15	300
PTB-XL	17,596	191,400	5	12	250

All five tasks use subject-level training, validation, and test splits. Methods evaluated on the same dataset use identical processed inputs and splits. Preprocessing and split details are given in Appendix B.

Paired short- and long-window data. For the PTB-XL (Wagner et al., 2020) short-/long-window experiment, we use the official 100 Hz release (`records100`). Each complete 10-second, 12-lead ECG examination is a Long input, while its first second is the paired Short input. Both retain the same examination-level label and subject split, and each channel is normalized separately. Sample counts and construction details are given in Appendix B.2.

Table 2: Short-window Test Macro-F1 on five MedTS datasets (%, mean ± sample SD; ↑). The best is **bold red**; the second is underlined blue. Complete six-metric results and Avg are given in Appendix A, Table 5.

Models	ADFTD (3 Classes)	APAVA (2 Classes)	TDBRAIN (2 Classes)	PTB (2 Classes)	PTB-XL (5 Classes)
Autoformer	42.59±1.85	68.06±1.94	78.64±4.24	63.69±3.84	48.85±2.27
FEDformer	43.91±1.37	73.51±3.39	70.36±4.47	67.14±4.37	47.89±8.44
Informer	45.74±1.38	69.47±5.06	80.16±1.84	70.84±3.47	60.44±0.43
iTransformer	46.79±1.13	72.30±1.79	64.79±2.49	79.06±1.06	56.20±0.19
MTST	44.31±1.74	64.01±3.16	75.30±2.72	67.38±3.71	61.43±0.58
Nonformer	46.96±1.35	69.74±3.84	78.36±2.88†	70.90±1.00	59.10±0.66
PatchTST	41.97±1.37	55.97±3.10	71.99±5.98	64.36±3.38	62.61±0.34
Reformer	47.94±0.69	75.93±1.82	75.63±1.67	69.65±3.88	60.69±0.18
Transformer	48.09±1.59	73.75±5.38	79.36±1.86	68.47±2.19	59.05±0.25
Medformer	50.65±1.51	76.31±0.71	71.58±6.95	79.18±3.31	62.02±0.37
TeCh	48.84±1.72	**86.30±1.06**	81.99±1.24	81.97±4.07	62.44±0.27
ReConTok	**53.84±0.74**	85.25±1.65	**83.40±1.05**	**83.77±1.68**	**63.18±0.20**

Baselines. We compare 11 Transformer-based methods: Transformer (Vaswani et al., 2017), Reformer (Kitaev et al., 2020), Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), Non-stationary Transformer (Liu et al., 2022), PatchTST (Nie et al., 2023), MTST (Zhang et al., 2024b), iTransformer (Liu et al., 2024), and the medical time-series methods Medformer (Wang et al., 2024a) and TeCh (Yu et al., 2026). We reproduce and tune the baselines locally on a reconstructed TDBRAIN cohort; for the other four datasets, we use reported results.

Implementation details. Our implementation follows the methodological settings of Medformer (Wang et al., 2024a) and TeCh (Yu et al., 2026). ReConTok is implemented in PyTorch with cross-entropy loss and RAdam; checkpoint selection and early stopping use validation macro-F1. Appendix C summarizes the search dimensions and candidate ranges without listing dataset-specific final settings. For the updated TDBRAIN V3.1 cohort, we rematch patients and healthy controls using demographic and acquisition-session information and tune each baseline on the validation set; the candidate ranges and cohort construction are described in Appendices C and B.3, respectively.

Evaluation metrics. We report Accuracy, Precision, Recall, F1, AUROC, and AUPRC. All but Accuracy use macro averaging; AUROC and AUPRC use one-vs-rest computation. Experiments use training seeds 42–46, and results are reported as mean ± sample standard deviation. The main table shows F1 only; all six metrics and their arithmetic mean (Avg) are in Appendix A.

5.2 Main Results

Table 2 summarizes short-window classification F1 across the five medical time-series datasets; the complete metrics appear in Appendix Table 5. ReConTok achieves the highest F1 on ADFTD, TDBRAIN, PTB, and PTB-XL, and ranks second only to TeCh on APAVA. These results suggest that context-conditioned token formation helps retain discriminative evidence across classes.

ReConTok redesigns only token representation formation before the Transformers, leaving subsequent token interactions unchanged. The results show that incorporating current, neighborhood, and global context when forming tokens from local features can improve classification across modalities and dataset sizes. They support our central claim that performance depends on what tokens represent before they interact, as well as on how they interact.

5.3 Ablation of Token Construction and Conditioning

On the five short-window tasks, we compare the original linear encoder, variants without Feature Gating or Global Mixing, and the full ReConTok. All variants share data splits, reference architecture, augmentation, and classification backend; only token formation changes. Checkpoints are selected by validation F1. Table 3 reports test F1, with other metrics in Appendix D. ReConTok achieves the highest F1 on all five datasets, improving over the original linear encoder by 2.32,

0.67, 2.41, 5.90, and 2.46 percentage points, respectively. The full model also consistently outperforms both ablated variants, indicating that Feature Gating and Global Mixing make complementary contributions.

Table 3: Token formation ablation: Test F1 (%). Bold indicates the highest score for each dataset.

Variant	ADFTD	APAVA	TDBRAIN	PTB	PTB-XL
Linear encoder	51.52±1.88	84.58±2.29	80.99±1.09	77.87±5.38	60.72±1.06
w/o Feature Gating	52.91±1.70	84.44±1.95	81.67±0.75	82.61±4.71	63.16±0.40
w/o Global Mixing	52.83±1.76	85.22±1.84	82.89±1.69	80.68±4.47	63.05±0.78
ReConTok	**53.84±0.74**	**85.25±1.65**	**83.40±1.05**	**83.77±1.68**	**63.18±0.20**

Controlled comparison with an SRS tokenizer. We compare ReConTok with an SRS tokenizer (Wu et al., 2025) on ADFTD, APAVA, TDBRAIN, and PTB-XL. The two methods share the classification backbone, Transformer interactions, data splits, and training settings; only temporal token construction differs. Checkpoints are selected by validation F1.

ReConTok yields higher Accuracy and F1 on all four datasets, with F1 gains of 3.11, 5.66, 3.44, and 2.98 percentage points, respectively. Because the downstream Transformer is identical, this controlled comparison isolates the contribution of contextual information to token construction.

Table 4: Controlled temporal-tokenizer comparison: test Accuracy and F1 (%). Bold indicates the higher score in each column.

	ADFTD		APAVA		TDBRAIN		PTB-XL	
Method	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
ReConTok	**55.48±1.14**	**53.84±0.74**	**86.16±1.47**	**85.25±1.65**	**83.44±1.03**	**83.40±1.05**	**74.01±0.43**	**63.18±0.20**
SRS tokenizer	52.13±1.14	50.73±1.32	80.39±3.65	79.59±3.42	79.98±4.01	79.95±4.03	72.48±0.46	60.21±0.56

5.4 Robustness to Local Noise

PTB-XL has 12 ECG leads. We add Gaussian noise $\beta\epsilon$ only to the final lead and gradually increase β, simulating contamination of a local observation while leaving the other leads intact. The curves in Figure 2b show test F1 at different noise intensities, and the bands show variation in the results.

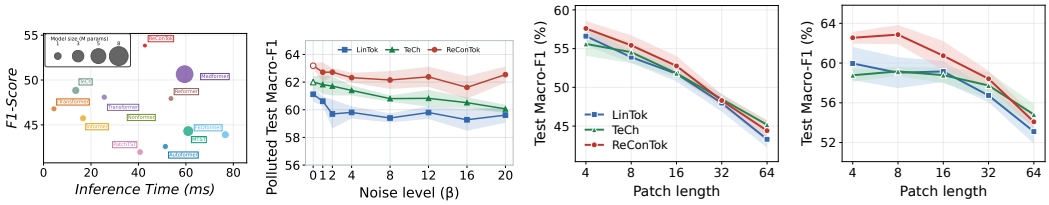
ReConTok maintains the highest performance at every noise intensity and retains a stable advantage under strong noise. When a local input is corrupted, allowing neighborhood and global context to participate in token formation can reduce the effect of encoding that noise directly into the representation, making the tokens more robust to local contamination.

5.5 Short- and Long-Window Experiments

The original PTB-XL ECG recordings last 10 seconds. From the same recordings, we construct Short inputs (the first second) and Long inputs (the complete 10 seconds). The two settings share sample counts, labels, and splits, differing only in the available temporal context. We also increase patch length from 4 to 64 to examine different token granularities. The curves in Figure 2c–d show test F1, and the bands show variation in the results.

With the same number of samples, Long inputs perform substantially better overall than Short inputs. ReConTok's best F1 rises from 57.60% to 62.86%, indicating that longer continuous context supplies diagnostic information unavailable in short windows. As patch length increases, performance generally declines, suggesting that coarser tokens compress local physiological details. This decline is markedly slower with Long inputs, indicating that broader context can compensate for information lost in local representations.

ReConTok achieves the highest performance at most patch lengths. It changes only token formation before the Transformer, without altering downstream interactions. Thus, the benefit of long-window modeling depends not only on token interactions but also on how local information is organized into tokens. By introducing neighborhood and global context during token formation, ReConTok forms representations that better align with the available context and make fuller use of longer physiological recordings.



(a) ADFTD efficiency. (b) PTB-XL robustness. (c) Short input. (d) Long input.

Figure 2: Efficiency, robustness, and long-context analyses. (a) Efficiency comparison on ADFTD: the horizontal axis shows forward-pass time per batch during second-epoch validation, and the vertical axis shows Test F1 under the full training budget. Bubble area is proportional to parameter count. (b) PTB-XL robustness to Gaussian noise of intensity β added to the final ECG channel. (c)–(d) PTB-XL results with short 1-s and long 10-s inputs. The horizontal axis shows patch length and the vertical axis shows Test F1.

5.6 Efficiency Analysis

Figure 2a compares classification performance, inference time, and parameter count on ADFTD; bubble area represents model size. ReConTok achieves the highest F1 while retaining a relatively small model and moderate inference cost, providing a favorable performance–efficiency trade-off. Its substantial gain comes from improving token formation before the Transformer, without requiring a more complex token interaction mechanism.

5.7 Patient-Level Recognition Stability and Missing-Input Robustness

We compare ReConTok with a Linear tokenizer in terms of within-patient prediction stability and robustness to missing inputs, keeping the downstream Transformer architecture fixed.

Patient-window recognition. For each patient, we calculate the proportion of correctly classified test windows and then compute the difference between ReConTok and the Linear tokenizer. Positive values in Appendix Figure 4a mean that ReConTok correctly identifies more windows from the same patient. The mean difference is positive on all five datasets, with gains of 4.94 and 2.10 percentage points on APAVA and PTB-XL, respectively. This indicates that context-conditioned token formation improves recognition stability across local windows from the same patient.

Missing-input robustness. At inference time only, we randomly set a proportion of input channels to zero, using the same missing-channel positions for both methods. Appendix Figure 4b–d shows F1 as the missing proportion increases on PTB-XL, PTB, and TDBRAIN. ReConTok performs better at every missing level. At 30% missing channels, its F1 values on PTB, PTB-XL, and TDBRAIN are 0.837, 0.591, and 0.762, exceeding the Linear tokenizer's 0.791, 0.573, and 0.608. Together, these analyses show that context-conditioned token formation improves both recognition stability across a patient's windows and representation robustness when some inputs are unavailable.

6 Conclusion

We revisit how token representations are formed in medical time-series Transformers and propose ReConTok, which conditionally reorganizes fine-grained information using current, neighborhood, and global context before tokens are finalized, while keeping subsequent Transformer interactions unchanged. Experiments on five medical time-series datasets show consistent classification gains and advantages in Recall. Ablations and a controlled comparison with an SRS tokenizer further support the effectiveness of context-conditioned token formation. Noise perturbation, missing-input, and patient-level analyses indicate more stable and robust representations. The PTB-XL short-/long-window experiment shows that, with sample counts held fixed, longer continuous context improves classification and ReConTok remains more stable across patch granularities. Overall, these findings show that medical time-series modeling depends on how context-relevant discriminative information is organized and retained before tokens enter the encoder. Incorporating context directly into token formation provides a route to more discriminative and robust representations.

AI USE STATEMENT

We used ChatGPT (OpenAI) to draft portions of the paper and to aid and polish the writing, including translation and language editing. We also used it to assist with data processing. The authors reviewed the AI-assisted text and take responsibility for the final manuscript and reported results.

REFERENCES

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token Merging: Your ViT But Faster. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das internet. *Biomedical Engineering / Biomedizinische Technik*, 40(s1):317–318, 1995. doi: 10.1515/bmte.1995.40.s1.317. URL https://www.degruyterbrill.com/document/doi/10.1515/bmte.1995.40.s1.317/html.

Cheng Ding, Tianliang Yao, Chenwei Wu, and Jianyuan Ni. Advances in Deep Learning for Personalized ECG Diagnostics: A Systematic Review Addressing Inter-Patient Variability and Generalization Constraints. *Biosensors and Bioelectronics*, 271:117073, 2025. doi: 10.1016/j.bios.2024.117073. URL https://www.sciencedirect.com/science/article/pii/S0956566324010807.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

J. Escudero, D. Abásolo, R. Hornero, P. Espino, and M. López. Analysis of electroencephalograms in Alzheimer's disease patients with multiscale entropy. *Physiological Measurement*, 27(11):1091–1106, 2006. doi: 10.1088/0967-3334/27/11/004. URL https://doi.org/10.1088/0967-3334/27/11/004.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019. doi: 10.1038/s41591-018-0316-z. URL https://doi.org/10.1038/s41591-018-0316-z.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive Token Sampling for Efficient Vision Transformers. In *Proceedings of the European Conference on Computer Vision*, 2022. doi: 10.1007/978-3-031-20083-0_24. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/4901_ECCV_2022_paper.php.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215. URL https://doi.org/10.1161/01.CIR.101.23.e215.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. SOFTS: Efficient multivariate time series forecasting with series-core fusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 64145–64175, 2024.

Awni Y. Hannun et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 2019. doi: 10.1038/s41591-018-0268-3. URL https://www.nature.com/articles/s41591-018-0268-3.

Imene Jemal, Lina Abou-Abbas, Khadidja Henni, Amar Mitiche, and Neila Mezghani. Domain Adaptation for EEG-Based, Cross-Subject Epileptic Seizure Prediction. *Frontiers in Neuroinformatics*, 18:1303380, 2024. doi: 10.3389/fninf.2024.1303380. URL https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2024.1303380/full.

Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token Fusion: Bridging the Gap Between Token Pruning and Token Merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1383–1392, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Kim_Token_Fusion_Bridging_the_Gap_Between_Token_Pruning_and_Token_WACV_2024_paper.html.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

Michael C. Kontos et al. 2022 ACC Expert Consensus Decision Pathway on the Evaluation and Disposition of Acute Chest Pain in the Emergency Department: A Report of the American College of Cardiology Solution Set Oversight Committee. *Journal of the American College of Cardiology*, 2022. doi: 10.1016/j.jacc.2022.08.750. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10691881/.

Mustafa Aykut Kural et al. Criteria for defining interictal epileptiform discharges in EEG: A clinical validation study. *Neurology*, 2020a. doi: 10.1212/WNL.0000000000009439. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7526669/.

Mustafa Aykut Kural et al. Optimized set of criteria for defining interictal epileptiform EEG discharges. *Clinical Neurophysiology*, 2020b. doi: 10.1016/j.clinph.2020.06.026. URL https://www.sciencedirect.com/science/article/pii/S1388245720304089.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c. URL https://doi.org/10.1088/1741-2552/aace8c.

Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. Deep learning in ECG diagnosis: A review. *Knowledge-Based Systems*, 227:107187, 2021. doi: 10.1016/j.knosys.2021.107187. URL https://doi.org/10.1016/j.knosys.2021.107187.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0718. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8 (6):95, 2023. doi: 10.3390/data8060095. URL https://www.mdpi.com/2306-5729/8/6/95.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/10876.

Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jorgen Jennum, and Christian Igel. U-Sleep: resilient high-frequency sleep staging. *npj Digital Medicine*, 2021. doi: 10.1038/s41746-021-00440-5. URL https://www.nature.com/articles/s41746-021-00440-5.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1185–1196, 2025. doi: 10.1145/3690624.3709325.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In *NeurIPS*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/747d3443e319a22747fbb873e8b2f9f2-Abstract.html.

Beanbonyka Rim, Nak-Jun Sung, Sedong Min, and Min Hong. Deep Learning in Physiological Signal Data: A Survey. *Sensors*, 20(4):969, 2020. doi: 10.3390/s20040969. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7071412/.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H. Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019. doi: 10.1088/1741-2552/ab260c. URL https://doi.org/10.1088/1741-2552/ab260c.

Michael S. Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive Space-Time Tokenization for Videos. In *NeurIPS*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/6a30e32e56fce5cf381895dfe6ca7b6f-Paper.pdf.

Keith Smith, Daniel Abásolo, and Javier Escudero. Accounting for the complex hierarchical topology of EEG phase-based functional connectivity in network binarisation. *PLOS ONE*, 12(10): e0186164, 2017. doi: 10.1371/journal.pone.0186164. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0186164.

Stephen W. Smith, Kenneth W. Dodd, Timothy D. Henry, David M. Dvorak, and Lesly A. Pearce. Diagnosis of ST-elevation myocardial infarction in the presence of left bundle branch block with the ST-elevation to S-wave ratio in a modified Sgarbossa rule. *Annals of Emergency Medicine*, 2012. doi: 10.1016/j.annemergmed.2012.07.119. URL https://pubmed.ncbi.nlm.nih.gov/22939607/.

William O. Tatum et al. Clinical utility of EEG in diagnosing and monitoring epilepsy in adults. *Clinical Neurophysiology*, 2018. doi: 10.1016/j.clinph.2018.01.019. URL https://pubmed.ncbi.nlm.nih.gov/29483017/.

Hanneke van Dijk, Guido van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database. *Scientific Data*, 9:333, 2022. doi: 10.1038/s41597-022-01409-z. URL https://www.nature.com/articles/s41597-022-01409-z.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7:154, 2020. doi: 10.1038/s41597-020-0495-6. URL https://www.nature.com/articles/s41597-020-0495-6.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A Multi-Granularity Patching Transformer for Medical Time-Series Classification. In *Advances in*

Neural Information Processing Systems, volume 37, 2024a. doi: 10.52202/079017-1145. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3fe2a777282299ecb4f9e7ebb531f0ab-Abstract-Conference.html.

Yihe Wang, Taida Li, Yujun Yan, Wenzhan Song, and Xiang Zhang. How to evaluate your medical time series classification?, 2024b. URL https://arxiv.org/abs/2410.03057. Preprint; formal publication not verified.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal Token Fusion for Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12186–12195, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Multimodal_Token_Fusion_for_Vision_Transformers_CVPR_2022_paper.html.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Xingjian Wu, Xiangfei Qiu, Hanyin Cheng, Zhengyu Li, Jilin Hu, Chenjuan Guo, and Bin Yang. Enhancing Time Series Forecasting through Selective Representation Spaces: A Patch Perspective. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0789. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/21beeac772f6f84d1517a2fc81e582d0-Abstract-Conference.html.

Jiexia Ye, Weiqi Zhang, Ziyue Li, Jia Li, and Fugee Tsung. MedSpaformer: A Transferable Transformer with Multi-Granularity Token Sparsification for Medical Time Series Classification. In *AAAI*, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/40001.

Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive Tokens for Efficient Vision Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10809–10818, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yin_A-ViT_Adaptive_Tokens_for_Efficient_Vision_Transformer_CVPR_2022_paper.html.

Guoqi Yu, Juncheng Wang, Chen Yang, Jing Qin, Angelica I. Aviles-Rivero, and Shujun Wang. Decentralized Attention Fails Centralized Signals: Rethinking Transformers for Medical Time Series. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/oral/10007399.

Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution Time-Series Transformer for Long-term Forecasting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4222–4230, 2024a. URL https://proceedings.mlr.press/v238/zhang24l.html.

Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution time-series transformer for long-term forecasting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4222–4230. PMLR, 2024b. URL https://proceedings.mlr.press/v238/zhang24l.html.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vSVLM2j9eie.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

A Complete Main Results

Table 5 gives all six test metrics and Avg for the main comparison in Table 2. Results are percentages. The six metrics are shown as means ± sample standard deviations over five seeds; Avg is the arithmetic mean of the six metrics and is shown without a standard deviation. Baseline results for ADFTD, APAVA, PTB, and PTB-XL come from TeCh (Yu et al., 2026) and Medformer (Wang et al., 2024a); TDBRAIN baselines were reproduced locally on the matched cohort. † indicates the complete five-seed Lite selection for TDBRAIN Nonformer; the other baselines use Extended selections.

Table 5: Complete short-window results on five MedTS datasets (%; ↑). Avg is the arithmetic mean of the six metrics. The best is **bold red**; the second is underlined blue.

Datasets	Models	Accuracy	Precision	Recall	F1 score	AUROC	AUPRC	Avg
ADFTD (3-Classes)	Autoformer	45.25±1.48	43.67±1.94	42.96±2.03	42.59±1.85	61.02±1.82	43.10±2.30	46.43
	FEDformer	46.30±0.59	46.05±0.76	44.22±1.38	43.91±1.37	62.62±1.75	46.11±1.44	48.20
	Informer	48.45±1.96	46.54±1.68	46.06±1.84	45.74±1.38	65.87±1.27	47.60±1.30	50.04
	iTransformer	52.60±1.59	46.79±1.27	47.28±1.29	46.79±1.13	67.26±1.16	49.53±1.21	51.71
	MTST	45.60±2.03	44.70±1.33	45.05±1.30	44.31±1.74	62.50±0.81	45.16±0.85	47.89
	Nonformer	49.95±1.05	47.71±0.97	47.46±1.50	46.96±1.35	66.23±1.37	47.33±1.78	50.94
	PatchTST	44.37±0.95	42.40±1.13	42.06±1.48	41.97±1.37	60.08±1.50	42.49±1.79	45.56
	Reformer	50.78±1.17	49.64±1.49	49.89±1.67	47.94±0.69	69.17±1.58	51.73±1.94	53.19
	Transformer	50.47±2.14	49.13±1.83	48.01±1.53	48.09±1.59	67.93±1.59	48.93±2.02	52.09
	Medformer	53.27±1.54	51.02±1.57	50.71±1.55	50.65±1.51	70.93±1.19	51.21±1.32	54.63
	TeCh	54.54±0.70	53.02±0.87	49.25±1.01	48.84±1.72	68.67±1.05	50.62±1.26	54.16
	ReConTok	**55.48±1.14**	**54.42±0.46**	**53.71±0.91**	**53.84±0.74**	**73.21±0.83**	**55.75±1.24**	**57.74**
APAVA (2-Classes)	Autoformer	68.64±1.82	68.48±2.10	68.77±2.27	68.06±1.94	75.94±3.61	74.38±4.05	70.71
	FEDformer	74.94±2.15	74.59±1.50	73.56±3.55	73.51±3.39	83.72±1.97	82.94±2.37	77.21
	Informer	73.11±4.40	75.17±6.06	69.17±4.56	69.47±5.06	70.46±4.91	70.75±5.27	71.36
	iTransformer	74.55±1.66	74.77±2.10	71.76±1.72	72.30±1.79	85.59±1.55	84.39±1.57	77.23
	MTST	71.14±1.59	79.30±0.97	65.27±2.28	64.01±3.16	68.87±2.34	71.06±1.60	69.94
	Nonformer	71.89±3.81	71.80±4.58	69.44±3.56	69.74±3.84	70.55±2.96	70.78±4.08	70.70
	PatchTST	67.03±1.65	78.76±1.28	59.91±2.02	55.97±3.10	65.65±0.28	67.99±0.76	65.89
	Reformer	78.70±2.00	82.50±3.95	75.00±1.61	75.93±1.82	73.94±1.40	76.04±1.14	77.02
	Transformer	76.30±4.72	77.64±5.95	73.09±5.01	73.75±5.38	72.50±6.60	73.23±7.60	74.42
	Medformer	78.74±0.64	81.11±0.84	75.40±0.66	76.31±0.71	83.20±0.91	83.66±0.92	79.74
	TeCh	**86.86±1.09**	86.85±1.29	**86.10±1.00**	**86.30±1.06**	**94.02±0.52**	**93.79±0.56**	**88.99**
	ReConTok	86.16±1.47	**86.93±1.31**	84.46±1.74	85.25±1.65	92.69±1.19	92.68±1.27	88.03
TDBRAIN (2-Classes)	Autoformer	78.88±4.15	80.16±4.10	78.88±4.15	78.64±4.24	84.47±5.18	83.25±5.34	80.71
	FEDformer	70.81±4.04	71.98±3.47	70.81±4.04	70.36±4.47	81.39±3.26	79.27±3.17	74.10
	Informer	80.25±1.78	80.76±1.56	80.25±1.78	80.16±1.84	89.03±1.08	88.46±1.60	83.15
	iTransformer	65.46±2.90	67.00±4.32	65.46±2.90	64.79±2.49	73.80±5.59	73.25±5.42	68.29
	MTST	75.40±2.70	75.80±2.73	75.40±2.70	75.30±2.72	82.79±4.14	80.68±4.63	77.56
	Nonformer†	78.44±2.83	78.79±2.66	78.44±2.83	78.36±2.88	87.50±2.03	86.45±2.27	81.33
	PatchTST	72.17±5.87	72.65±5.70	72.17±5.87	71.99±5.98	79.60±5.65	76.55±5.84	74.19
	Reformer	75.83±1.51	76.67±1.18	75.83±1.51	75.63±1.67	84.71±1.05	84.37±1.38	78.84
	Transformer	79.40±1.82	79.59±1.66	79.40±1.82	79.36±1.86	88.29±0.79	87.67±1.36	82.28
	Medformer	71.85±6.97	72.84±7.36	71.85±6.97	71.58±6.95	79.56±8.69	78.48±9.60	74.36
	TeCh	82.06±1.18	82.54±0.90	82.06±1.18	81.99±1.24	90.74±1.18	90.68±1.19	85.01
	ReConTok	**83.44±1.03**	**83.76±1.08**	**83.44±1.03**	**83.40±1.05**	**92.20±0.35**	**92.34±0.38**	**86.43**
PTB (2-Classes)	Autoformer	73.35±2.10	72.11±2.89	63.24±3.17	63.69±3.84	78.54±3.48	74.25±3.53	70.86
	FEDformer	76.05±2.54	77.58±3.61	66.10±3.55	67.14±4.37	85.93±4.31	82.59±5.42	75.90
	Informer	78.69±1.68	82.87±1.02	69.19±2.90	70.84±3.47	92.09±0.53	90.02±0.60	80.62
	iTransformer	83.89±0.71	88.25±1.18	76.39±1.01	79.06±1.06	91.18±1.16	90.93±0.98	84.95
	MTST	76.59±1.90	79.88±1.90	66.31±2.95	67.38±3.71	86.86±2.75	83.75±2.84	76.80
	Nonformer	78.66±0.49	82.77±0.86	69.12±0.87	70.90±1.00	89.37±2.51	86.67±2.38	79.58
	PatchTST	74.74±1.62	76.94±1.51	63.89±2.71	64.36±3.38	88.79±0.91	83.39±0.96	75.35
	Reformer	77.96±2.13	81.72±1.61	68.20±3.35	69.65±3.88	91.13±0.74	88.42±1.30	79.51
	Transformer	77.37±1.02	81.84±0.66	67.14±1.80	68.47±2.19	90.08±1.76	88.22±1.68	78.85
	Medformer	83.50±2.01	85.19±0.94	77.11±3.39	79.18±3.31	92.81±1.48	90.32±1.54	84.69
	TeCh	85.96±2.52	**89.92±0.74**	79.43±4.13	81.97±4.07	94.57±0.70	**94.36±0.66**	87.70
	ReConTok	**86.89±1.18**	89.13±0.90	**81.40±1.79**	**83.77±1.68**	**94.73±0.65**	94.28±0.66	**88.37**
PTB-XL (5-Classes)	Autoformer	61.68±2.72	51.60±1.64	49.10±1.52	48.85±2.27	82.04±1.44	51.93±1.71	57.53
	FEDformer	57.20±9.47	52.38±6.09	49.04±7.26	47.89±8.44	82.13±4.17	52.31±7.03	56.82
	Informer	71.43±0.32	62.64±0.60	59.12±0.47	60.44±0.43	88.65±0.09	64.76±0.17	67.84
	iTransformer	69.28±0.22	59.59±0.45	54.62±0.18	56.20±0.19	86.71±0.10	60.27±0.21	64.45
	MTST	72.14±0.27	63.84±0.72	60.01±0.81	61.43±0.58	88.97±0.33	65.83±0.51	68.70
	Nonformer	70.56±0.55	61.57±0.66	57.75±0.72	59.10±0.66	88.32±0.36	63.40±0.29	66.78
	PatchTST	73.23±0.25	65.70±0.64	60.82±0.76	62.61±0.34	89.74±0.19	67.32±0.22	69.90
	Reformer	71.72±0.43	63.12±1.02	59.20±0.75	60.69±0.18	88.80±0.24	64.72±0.47	68.04
	Transformer	70.59±0.44	61.57±0.65	57.62±0.35	59.05±0.25	88.21±0.16	63.36±0.29	66.73
	Medformer	72.87±0.23	64.14±0.42	60.60±0.46	62.02±0.37	89.66±0.13	66.39±0.22	69.28
	TeCh	73.53±0.07	**65.92±0.52**	60.61±0.59	62.44±0.27	90.03±0.12	67.19±0.25	69.95
	ReConTok	**74.01±0.43**	65.66±1.01	**61.94±0.15**	**63.18±0.20**	**90.18±0.21**	**67.65±0.49**	**70.44**

B DATASETS AND PREPROCESSING

The short-window benchmark follows Medformer's data preparation and loading protocols (Wang et al., 2024a). Table 6 reports the numbers of processed examples in each partition. Appendix B.2 describes the paired short- and long-window data. Appendix B.3 describes the construction of the TDBRAIN V3.1 matched cohort.

Table 6: Training, validation, and test partitions. Counts are input examples. Subject counts in parentheses are ordered as train/validation/test. Percentages in subject-based rules apply within each class, with integer rounding.

Dataset	Train	Validation	Test	Partition rule
APAVA	3,123	1,413	1,431	Fixed subjects (15/4/4)
TDBRAIN	4,080	960	960	Matched pairs (34/8/8)
ADFTD	40,446	14,658	14,648	60/20/20% subjects (51/18/19)
PTB	41,995	12,993	9,368	55/15/30% subjects (108/29/61)
PTB-XL	112,570	39,280	39,550	60/20/20% subjects

B.1 EEG AND ECG DATASETS

APAVA. APAVA (Smith et al., 2017) contains EEG from 12 Alzheimer's disease patients and 11 healthy controls. We use Medformer's 256-point windows with a 128-point stride and fixed subject partitions: prepared-data IDs 15, 16, 19, and 20 for validation; 1, 2, 17, and 18 for testing; and the remainder for training.

TDBRAIN. TDBRAIN (van Dijk et al., 2022) provides resting-state EEG. We construct a V3.1 subset for Parkinson's disease (PD) versus healthy control (HC) classification, with 25 participants per class and 32 signal channels (26 scalp EEG and six auxiliary channels). Appendix B.3 describes the cohort construction and processing.

ADFTD. ADFTD (Miltiadous et al., 2023) contains EEG from 36 participants with Alzheimer's disease, 23 with frontotemporal dementia, and 29 healthy controls. We use 256 Hz signals and non-overlapping one-second windows, discarding incomplete final windows. Within each class, subjects are partitioned in the prepared order using the ratios in Table 6.

PTB. We use the myocardial-infarction/healthy-control subset of PTB (Bousseljot et al., 1995; Goldberger et al., 2000). Medformer's preprocessing resamples recordings from 1,000 to 250 Hz, applies standard scaling, and extracts R-peak-aligned heartbeats padded to 300 points. Subjects are partitioned within each class in the prepared order at a 55/15/30% ratio.

PTB-XL. We use Medformer's single-label construction of PTB-XL (Wagner et al., 2020): each record's highest-confidence ECG annotation code (SCP code) is mapped to normal ECG, myocardial infarction, ST-segment or T-wave changes, conduction disturbance, or hypertrophy. Participants are retained when all their records share the same retained class. In the short-window benchmark, signals sampled at 250 Hz are split into non-overlapping one-second windows. Subjects are partitioned within each class in the prepared order, giving 10,556/3,519/3,521 subjects for training/validation/testing.

Normalization. Each EEG/ECG example is standardized per channel by subtracting its temporal mean and dividing by its temporal standard deviation, set to one for constant channels.

B.2 PAIRED SHORT- AND LONG-WINDOW DATA

The paired long-window comparison includes PTB-XL (Wagner et al., 2020) only. It uses the official downsampled 100 Hz version (`records100`): each complete 10-second, 12-lead ECG record is a long sample, and its first 100 points form the short sample. The two views share the examination-level label and subject partitions and are standardized separately per channel after cropping.

Table 7: Paired-view input lengths and partition sizes. Train, validation, and test counts apply to each view separately.

Dataset	Rate (Hz)	Short points	Long points	Train	Validation	Test
PTB-XL	100	100	1,000	11,257	3,928	3,955

B.3 TDBRAIN COHORT CONSTRUCTION AND PROCESSING

Cohort selection. We match controls to patients to reduce demographic and session differences between the groups. We remove entries with missing participant IDs or diagnoses and entries marked `REPLICATION`, standardize diagnosis labels, and deduplicate IDs by retaining their first occurrence. Requiring exactly one first-session, closed-eye resting recording (`ses-1/restEC`) per participant leaves 25 PD patients and 179 HC candidates.

Matching objective. For patient i and candidate control j, the matching cost is

$$c_{ij} = 1.5\frac{|a_i - a_j|}{10} + 2\delta(g_i, g_j) + 0.5\delta(s_i, s_j) + 1.5\delta(t_i, t_j) + |n_i - n_j|, \tag{12}$$

where a, g, s, t, and n denote age in years, recorded gender category, session season, session time, and recorded number of sessions, respectively; subscripts identify the participant. Categorical strings are trimmed and uppercased. The mismatch indicator δ equals zero for identical values (including two missing values) and one otherwise. For age and session count, the corresponding term is replaced by 5 and 1, respectively, when exactly one value is missing, or by zero when both are missing. Unparseable or non-finite numeric values are treated as missing. Session time is missing for all PD participants.

We minimize total cost by one-to-one assignment over the 25×179 matrix using `scipy.optimize.linear_sum_assignment`. The resulting pairs have a mean absolute age difference of 3.07 years and matching gender categories in all 25 pairs.

Partitions and signal processing. In patient-table order, the first 17 matched pairs form training, the next four validation, and the final four testing. Each pair stays in the same partition, and the cohort and partitions are fixed across training seeds. Recordings are linearly resampled from 500 to 256 Hz and segmented into non-overlapping 256-point windows aligned around the recording midpoint, discarding incomplete end segments. After excluding status and trigger channels, each participant contributes 120 windows with the 32 channels described in Appendix B.1, standardized per channel as above.

C IMPLEMENTATION DETAILS

ReConTok search space. We use staged discrete searches rather than a full Cartesian product. The candidate ranges cover learning rates from 10^{-5} to 10^{-3}, representation widths $D \in \{64, 128, 256, 512\}$, predefined patch-length/stride pairs within 1–16, Temporal and Channel branch depths $\{0, 1, 3, 6\}$, and training budgets from 10 to 100 epochs. Structural candidates include no/fixed/learned positional encoding; no/shared/branch-specific LayerNorm; global-summary injection before or after fusion or conditioning-only; neighborhood radius $k \in \{0, 1\}$; Mean Pool, Low-rank, or CoTAR Project; and shared or separate dynamic aggregation and final linear maps. Augmentation types and strengths are also compared, with tested strengths spanning 0.1–0.8. Each stage evaluates only its prespecified candidate combinations. Checkpoints are selected by validation macro-F1; budget and configuration selection follow the protocol reported for the corresponding experiment.

TDBRAIN baseline search space. Baseline tuning covers model capacity and architecture-specific parameters: $d_{\text{model}} \in \{64, 128, 256\}$, $d_{\text{ff}} \in \{128, 256, 512\}$, and encoder depths $\{3, 6, 9\}$. Model-specific candidates span moving-average windows of 13–49, 16–64 Fourier modes, patch lengths of 4–192, attention factors of 1–5, Reformer bucket sizes and hash counts of 2–8, and projector widths of 64–256 with depths of 1–3. Augmentation candidates include no augmentation and drop augmentation at strengths 0.25 or 0.4. Checkpoints and candidate configurations are selected

using validation macro-F1. We report the search ranges rather than dataset-specific final training settings.

D Complete Ablation Results

Table 8 reports all test metrics for the four variants in the main text across five datasets (percentages, mean ± sample standard deviation). Checkpoints are selected by validation F1. Linear encoder is the original linear tokenizer; w/o Feature Gating and w/o Global Mixing remove the corresponding CoTGF components.

Table 8: Complete token formation ablation results (Test %, mean ± sample SD).

Dataset	Variant	Accuracy	Macro-Prec.	Macro-Rec.	F1	AUROC	AUPRC
ADFTD	Linear encoder	53.28±2.11	52.23±1.79	51.59±1.35	51.52±1.88	71.12±2.31	53.27±2.87
	w/o Feature Gating	55.15±1.64	54.14±1.57	52.81±1.75	52.91±1.70	72.97±1.12	55.26±1.27
	w/o Global Mixing	55.32±1.22	54.22±1.94	52.59±1.84	52.83±1.76	72.90±1.46	55.42±1.98
	ReConTok	55.48±1.14	54.42±0.46	53.71±0.91	53.84±0.74	73.21±0.83	55.75±1.24
APAVA	Linear encoder	85.60±1.91	86.89±1.85	83.89±2.58	84.58±2.29	93.78±1.11	93.85±1.12
	w/o Feature Gating	85.58±1.56	87.00±0.76	83.49±2.11	84.44±1.95	92.37±1.52	92.43±1.50
	w/o Global Mixing	85.95±1.57	86.14±1.37	84.81±2.14	85.22±1.84	93.03±1.06	93.03±1.07
	ReConTok	86.16±1.47	86.93±1.31	84.46±1.74	85.25±1.65	92.69±1.19	92.68±1.27
TDBRAIN	Linear encoder	81.02±1.07	81.21±0.98	81.02±1.07	80.99±1.09	89.41±0.78	89.42±1.13
	w/o Feature Gating	81.73±0.76	82.13±0.88	81.73±0.76	81.67±0.75	90.66±0.55	90.74±0.71
	w/o Global Mixing	82.92±1.67	83.12±1.62	82.92±1.67	82.89±1.69	91.51±1.27	91.60±1.32
	ReConTok	83.44±1.03	83.76±1.08	83.44±1.03	83.40±1.05	92.20±0.35	92.34±0.38
PTB	Linear encoder	83.30±2.95	88.23±1.46	75.53±4.98	77.87±5.38	92.98±1.83	92.75±1.61
	w/o Feature Gating	86.29±2.90	89.39±0.93	80.25±4.74	82.61±4.71	93.94±1.14	93.67±1.03
	w/o Global Mixing	84.88±2.91	87.95±2.44	78.28±4.44	80.68±4.47	92.66±3.47	92.31±3.21
	ReConTok	86.89±1.18	89.13±0.90	81.40±1.79	83.77±1.68	94.73±0.65	94.28±0.66
PTB-XL	Linear encoder	72.80±0.32	65.58±0.59	59.05±1.15	60.72±1.06	89.34±0.13	65.69±0.26
	w/o Feature Gating	73.99±0.33	67.02±0.90	61.43±0.60	63.16±0.40	90.14±0.14	68.01±0.40
	w/o Global Mixing	74.16±0.40	66.61±0.68	61.35±0.88	63.05±0.78	90.13±0.28	67.73±0.67
	ReConTok	74.01±0.43	65.66±1.01	61.94±0.15	63.18±0.20	90.18±0.21	67.65±0.49

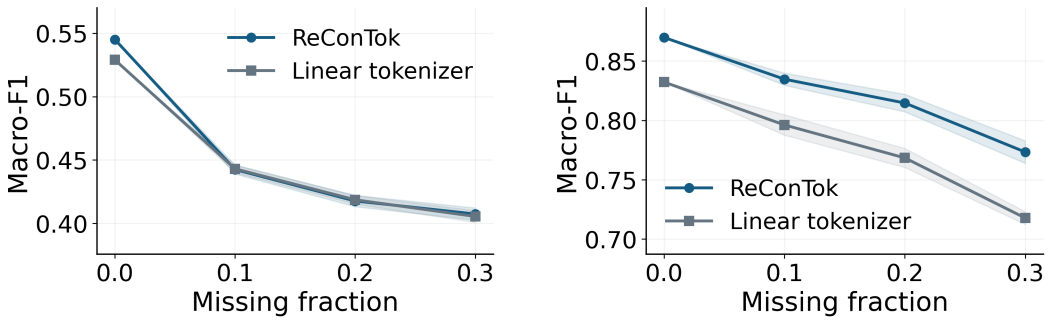
E Supplementary Protocol and Results for Patient-Window Recognition and Inference-Time Perturbations

Model sources and selection. ReConTok weights come from the frozen P0 archive. The Linear tokenizer is a separately trained native linear-tokenizer control, with weights from the Pure-L batch. Both methods use the same processed data and subject partitions; all predictions and weights were verified against source signatures and hashes. Table 9 lists the models, training budgets, and validation-selected ReConTok epochs. Checkpoints during training were selected by validation F1.

Table 9: Frozen single-model configurations used in the A analyses. Budgets are maximum training epochs; the epoch column gives the ReConTok checkpoint selected by validation F1.

Dataset	Token branch	Budget	ReConTok epoch	Training batch
ADFTD	Temporal + Channel	30	20	128
APAVA	Temporal + Channel	30	30	128
PTB	Channel	80	57	128
PTB-XL	Temporal	30	14	128
TDBRAIN	Temporal	60	48	128

Patient-window recognition statistics. We aligned ReConTok and Linear-tokenizer predictions by dataset, split, and unique window ID, and verified labels, subject identities, and class order entry by entry. For each subject with a fixed label, we first computed the fraction of windows correctly classified as that subject's true class, then averaged equally across subjects within each class and across classes. We estimated 95% intervals with 10,000 subject-level bootstrap replicates, resampling all windows from a subject as one cluster. Table 10 reports the complete comparison across five cohorts; the intervals describe subject-sampling variation conditional on the given model.



(a) ADFTD. (b) APAVA.

Figure 3: Inference-time missing-channel curves for the two cohorts not shown in the main text. Blue denotes ReConTok and gray the Linear tokenizer. Each panel uses independent axes; shading shows the standard deviation over 10 shared perturbations.

Table 10: Patient-window recognition rates for ReConTok and the Linear tokenizer (percentage points). Intervals are 95% subject-level bootstrap intervals.

Dataset	Test subjects	ReConTok	Linear tokenizer	Δ [95% CI]
ADFTD	19	54.92	53.21	$1.70\ [-1.59, 5.15]$
APAVA	4	84.27	79.33	$4.94\ [2.10, 7.78]$
PTB	61	80.90	79.14	$1.76\ [-3.33, 6.41]$
PTB-XL	3,521	60.04	57.94	$2.10\ [1.19, 3.04]$
TDBRAIN	8	84.58	81.25	$3.33\ [-1.88, 10.52]$

Fixed-weight inference-time perturbations. For missing-channel experiments, we selected $\max(1, \lfloor Cf{+}0.5 \rfloor)$ channels in processed-input space and zeroed each selected channel over the full window, with $f \in \{0.1, 0.2, 0.3\}$. Source-matched TDBRAIN windows shared the masked channels; for other datasets, masks were selected per window. The local-noise analysis covers ADFTD, APAVA, PTB-XL, and TDBRAIN: we selected approximately 10% of channels and a contiguous span of approximately 20% of time points, then added Gaussian noise scaled by each channel's training-set standard deviation, at levels $\{0.5, 1, 2\}$. Each level was repeated 10 times. Each paired test window for ReConTok and the Linear tokenizer shared the same missing region or noise locations. Perturbations were applied only at test inference; models were not retrained. Shading in the figures denotes the standard deviation across the 10 perturbation runs.

Complete curves and calculation. Normalized AUC is computed by trapezoidal integration over the clean input and three prespecified perturbation levels, divided by the maximum level. It is a weighted average of absolute F1, not retention relative to the clean-input score. Table 11 reports the complete comparison across five cohorts.

Table 11: Normalized AUC of absolute F1 under inference-time perturbations. Values are means over 10 perturbations; – indicates that the local-noise metric was not reported.

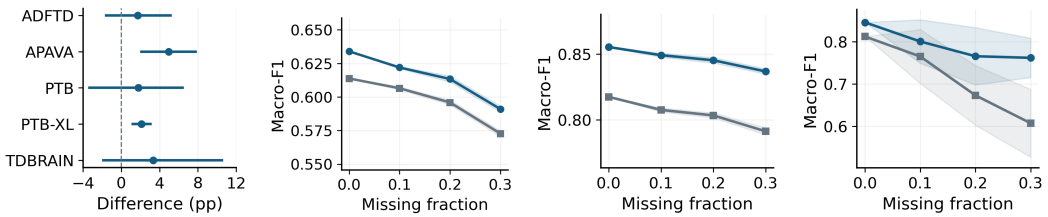
Dataset	Missing channels ReConTok	Missing channels Linear tokenizer	Local noise ReConTok	Local noise Linear tokenizer
ADFTD	0.4455	0.4430	0.5131	0.5068
APAVA	0.8236	0.7798	0.8658	0.8333
PTB	0.8469	0.8051	–	–
PTB-XL	0.6159	0.5986	0.6301	0.6128
TDBRAIN	0.7900	0.7161	0.8405	0.8062

Inference cost. We measured both models on the same physical RTX 3090 GPU in float32, using batch sizes 1 and 32. For each setting, we performed 50 warm-up and 200 timed forward passes, repeating the measurement three times. Table 12 reports the median of the three per-run medians.

System load may affect absolute latency, so these are single-device observations. Values are listed as ReConTok / Linear tokenizer.

Table 12: Inference cost of the selected ReConTok / Linear-tokenizer models. Latency is measured at batch size 1 and throughput at batch size 32.

Dataset	Parameters (M)	Latency (ms)	Throughput (samples/s)
ADFTD	0.555/0.421	14.87/8.03	1592/2756
APAVA	8.332/6.396	18.46/12.20	798/1444
PTB	1.999/1.659	4.08/1.63	6131/12448
PTB-XL	0.146/0.113	6.72/3.61	3707/6846
TDBRAIN	1.006/0.807	15.90/5.61	1235/2375



(a) Patient-window recognition difference. (b) PTB-XL missing channels. (c) PTB missing channels. (d) TDBRAIN missing channels.

Figure 4: Patient-level recognition stability and robustness to missing input. (a) Difference in patient-window recognition rates between ReConTok and the Linear tokenizer; (b)–(d) test F1 at different proportions of missing channels.

F Task Definitions and Subject-Level Evaluation

Medical time series typically have a hierarchical structure spanning subjects, recordings, and segments. To clarify the experimental settings used later, we introduce subject-independent splitting and define short- and long-window inputs.

F.1 Subject-Independent Evaluation

Multiple recordings from the same subject and the segments derived from them share individual physiological characteristics. A random segment-level split may therefore place data from the same subject in both the training and test sets, overestimating generalization to unseen subjects (Wang et al., 2024b). Subject-independent evaluation consequently uses subjects as the splitting unit. Let the training, validation, and test subject sets be $\mathcal{S}_{\mathrm{tr}}$, $\mathcal{S}_{\mathrm{val}}$, and $\mathcal{S}_{\mathrm{te}}$, respectively, with

$$\mathcal{S}_{\mathrm{tr}} \cap \mathcal{S}_{\mathrm{val}} = \mathcal{S}_{\mathrm{tr}} \cap \mathcal{S}_{\mathrm{te}} = \mathcal{S}_{\mathrm{val}} \cap \mathcal{S}_{\mathrm{te}} = \emptyset. \tag{13}$$

All data from a given subject are assigned to the same split.

F.2 Short- and Long-Window Classification

Labels in medical time series may correspond to a local event, a complete examination, or a subject-level state, whereas a model's input at each prediction typically covers only one continuous portion of the signal. Label granularity and input granularity therefore do not always match. Here, label granularity denotes the temporal or subject-level scope represented by the supervision label, while input granularity denotes the range of signal actually observed for a single prediction. Existing medical time-series classification protocols often divide continuous recordings into short segments and inherit examination- or subject-level labels (Wang et al., 2024a; Yu et al., 2026). Thus, different short windows may share a label while containing different diagnostic evidence.

PTB-XL (Wagner et al., 2020) provides complete 10-second multilead ECG examinations with examination-level diagnostic labels, allowing the same recording to be evaluated using either a short

window or the complete continuous signal. The former measures the discriminative ability of local patterns, while the latter further assesses the model's use of broader physiological context.

Formally, let the v-th recording of subject u be $S_{u,v} \in \mathbb{R}^{L_{u,v} \times C}$, with diagnostic label $\bar{y}_{u,v} \in \{1, \ldots, K\}$. Given an observation duration ℓ and sampling rate f_s, a segment starting at position t_j is extracted as

$$X^{(\ell)}_{u,v,j} = S_{u,v}[t_j : t_j + T_\ell], \qquad T_\ell = \ell f_s, \qquad y^{(\ell)}_{u,v,j} = \bar{y}_{u,v}. \tag{14}$$

When $\ell = \ell_{\mathrm{S}}$, we refer to the task as short-window classification; when $\ell = \ell_{\mathrm{L}} > \ell_{\mathrm{S}}$, we refer to it as long-window classification. Both predict diagnostic labels at the same level and differ only in the range of continuous physiological context available to the model for each input.