BioLT: Streaming Biosignal Modeling

Anonymous authors
Paper under double-blind review

Abstract

Clinical biosignals such as ECG, PPG, and EEG are increasingly acquired as hours-long streams from bedside monitors and wearables, yet existing analysis pipelines almost universally operate on isolated short windows of around 10 seconds and cannot process the long, irregular, frequently interrupted signals. We introduce BioLT, a streaming decoder-only framework for long-context biosignal modeling that turns any window-level feature extractor or pretrained biosignal foundation model into a streaming model. BioLT combines (i) a discontinuity-aware encoding that exposes the inter-window elapsed time as an explicit input channel, so that the decoder can learn to model these gaps as clinically meaningful information, and (ii) a fixed-size compressed memory bank supervised with cumulative summary statistics, which keeps both training and inference linear in the recording length while preserving long-range context. We evaluate on CPSC-2021, VitalDB, MC-MED, and CHB-MIT for arrhythmia, sepsis, and seizure detection and for forecasting up to 20 minutes ahead, using domain features and embeddings from six biosignal foundation models. Across 21 dataset-feature settings, BioLT achieves the highest detection AUROC in 15 and the highest 20-minute forecasting AUROC in 13 settings, and shows consistent performance gains over state-of-the-art snapshot foundation-model embeddings. These highlight the importance of modeling how streaming biosignal representations evolve over time, beyond improving window-level representations alone.

1 Introduction

Human biosignals are continuous streams of physiological information. Signals such as electrocardiography (ECG), photoplethysmography (PPG), and electroencephalography (EEG) are increasingly acquired through wearable sensors, bedside monitors, and ambulatory devices, yielding long-term streams of physiology-state information. Their "streaming" nature makes them particularly valuable

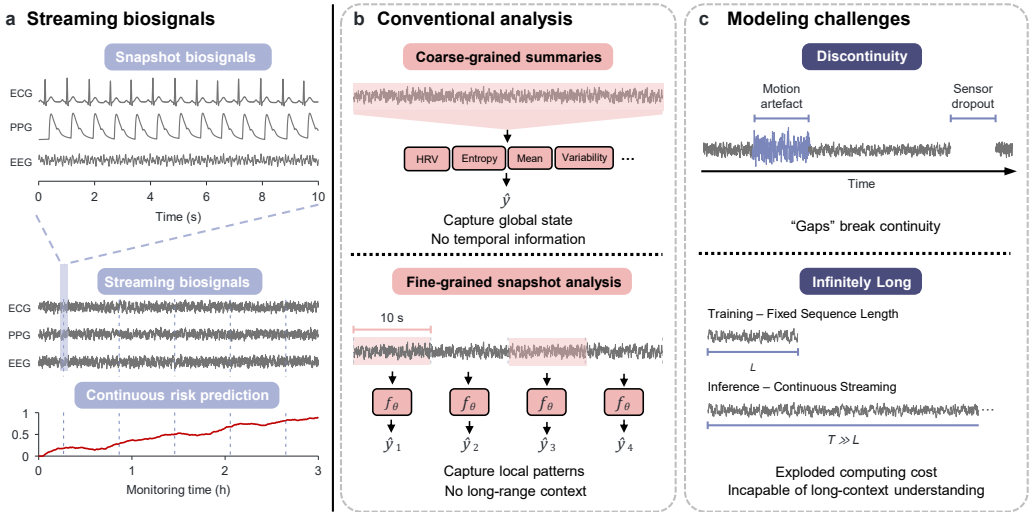


Figure 1: **Challenges of streaming biosignal modeling.** Illustration of two challenges of streaming biosignal modeling addressed by BioLT: the *discontinuity* of biosignal streams (gaps from missing or unreliable segments) and the *infinitely-long* nature of continuous monitoring (predictions must update as the trajectory unfolds), compared to coarse-grained summaries and fine-grained snapshot analysis.

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.

for clinical monitoring, where models are expected to continuously update patient-state estimates, and facilitate the early detection or forecasting of health deterioration (Blythe et al., 2022; Hannun et al., 2019; Hyland et al., 2020; Han et al., 2025).

However, existing biosignal AI methods are limited in their capacity to support such streaming analysis, as they typically operate in either of the following two extremes, as shown in Figure 1b.

(i) *Coarse-grained summaries*, where summary statistics, such as heart rate variability and entropy distribution, are calculated from long-term duration biosignals (Duong et al., 2019; Hosseini et al., 2020). These features can capture global physiological characteristics, but they discard temporal ordering and therefore obscure how patient state unfolds over time.

(ii) *Fine-grained snapshot analysis*, which operates on short biosignal segments (typically 10 seconds, i.e., windows), to derive physiologically meaningful features. These features may be manually engineered using domain knowledge (Elgendi, 2012), or more recently obtained as embeddings from foundation models (Gu et al., 2025; Li et al., 2025; Shu et al., 2025; Gu et al., 2026; Saha et al., 2025; Jiang et al., 2024; Wang et al., 2025). While effective for local pattern recognition, such methods usually treat each window as an independent sample during training or evaluation, ignoring the continuous trajectory of the underlying physiology.

Several works attempt to bridge this gap by applying sequential models over window-level "snapshot" features. They either perform pre-hoc aggregation of these window-level features, or post-hoc aggregation of window-calculated probability/logits (Supratak et al., 2017; Phan et al., 2019; Tison et al., 2018). However, they mostly need to fix a maximum sequence length during training, and could not perform properly when the biosignal is streaming over the long term - the very practical needs necessitated by continuous monitoring.

To model streaming biosignals, there are two fundamental challenges inherent to such streaming nature, which need to be addressed properly, as shown in Figure 1c.

(i) **"Discontinuity" of streaming biosignals**. One particular "sweet spot" of biosignals is their continuous nature - supporting continuous-monitoring needs. In reality, however, biosignal streams are rarely perfectly continuous. "Gaps" (no signals or unreliable segments) often arise during signal acquisition due to issues such as poor sensor contact or motion artifacts. In conventional window-level analysis, such low-quality windows may often be discarded. However, for streaming trajectory modeling, simply removing these windows breaks temporal continuity, especially for long missing intervals. Naive imputation is also insufficient, especially when gaps are long.

(ii) **"Infinitely-Long" nature of streaming biosignals**. Biosignal streams can, in principle, continue indefinitely as new observations arrive over time. This creates a challenge for models that rely on fixed-length inputs or predefined observation windows. In continuous monitoring, the model should move beyond retrospective analysis of fixed-length segments and continuously update its predictions as the physiological trajectory evolves. This requires the model to preserve local waveform information while capturing long-range temporal dependencies across an expanding stream.

To address these challenges, we propose `BioLT` (**Bio**signal **L**ong-context **T**ransformer), a streaming framework that turns snapshot-level biosignal features or foundation-model embeddings into causal patient trajectories for real-time streaming prediction. `BioLT` combines two designs: (i) a discontinuity-aware encoding that exposes inter-window elapsed time to the model (Figure 2), and (ii) a compressed streaming "memory" that preserves long-range context into a fixed number of tokens, to avoid quadratic attention costs over long streaming biosignals (Figure 4). Across ECG, PPG, and EEG tasks spanning arrhythmia, sepsis, and seizure, for both detection and forecasting, `BioLT` improves over snapshot-level and temporal aggregation baselines, supporting its ability to model biosignals as long, irregular, continuously updated streams.

2 Related Work

Biosignal modeling. Biosignal analysis is evolving in terms of both how physiological representations are learned from raw signals, and how biosignal models are guided during learning (Hong et al., 2020; Gu et al., 2025). Early approaches relied heavily on domain-driven feature engineering (such as morphological features, frequency-domain measures, and signal-quality indicators) from raw biosignals (Faust et al., 2018; Gu et al., 2025). They typically need sophisticated signal processing and a feature selection procedure, necessitating substantial expert design. Deep learning methods have shifted such analysis from handcrafted features to learn task-relevant embeddings via careful

annotation (Hannun et al., 2019; Supratak et al., 2017; Tison et al., 2018), and more recently to foundation models supporting general-purpose representations by self-supervised learning (Li et al., 2025; Shu et al., 2025; Gu et al., 2026; Saha et al., 2025; Jiang et al., 2024; Wang et al., 2025). However, most of these methods still focus on short, fixed-length windows, treating biosignals primarily as isolated snapshots rather than continuous physiological streams, leaving long-context modeling and continuous-monitoring evaluation underexplored for biosignals.

Sequence modeling. Sequential modeling has been widely studied across language, video, and time series, with recurrent networks, Transformers, and state-space models providing common tools for modeling these sequential data. A central challenge is how to extend models beyond short contexts. Existing methods such as architecture variants (Rae et al., 2020), attention variants (Press et al., 2022), and state-space models (Lahoti et al., 2026) aim to improve memory, extrapolation, or computational efficiency for long sequences. A complementary direction is to carry history in a "memory" module, which caches hidden states from previous token attentions (Dai et al., 2019). These concepts are relevant to our problem, where monitoring records may extend over long periods. However, biosignal streams introduce additional practical challenges: valid windows may be interrupted by missing or unreliable recordings, and the elapsed time between neighboring windows can itself be informative.

Real-time clinical prediction. Real-time clinical prediction aims to update risk estimates as new patient data arrive. This has been implemented using electronic health record (EHR)-based models, such as early warning scores (e.g., the Sequential Organ Failure Assessment (SOFA) or the National Early Warning Score (NEWS)) based on intermittently observed vital-sign and laboratory data (Reyna et al., 2020; Hyland et al., 2020). In contrast, biosignal-based systems, due to their "continuity", are better suited for generating real-time risks of deterioration or adverse events (Hannun et al., 2019; Tison et al., 2018). Nevertheless, in both settings, models are commonly evaluated at the window level or within a fixed-length temporal context. This leaves open the question of how to model and evaluate risk trajectories over long, irregular, continuously updated biosignal streams.

3 METHOD

Given streaming biosignals, we propose `BioLT` as a framework for modeling continuous physiological trajectories and supporting real-time risk prediction. `BioLT` builds on snapshot-level biosignal representations. Given a raw biosignal recording, we first divide it into short valid non-overlapping windows (in practice, 10 seconds) and extract one representation for each window, using either hand-crafted physiological features or embeddings from a pretrained biosignal foundation model. `BioLT` then models the resulting sequence as a causal physiological trajectory.

Let $\mathbf{z}_t \in \mathbb{R}^D$ be the representation of the t-th valid window, where D denotes the representation dimension, and let s_t be its timestamp on the clinical timeline. A recording is represented as

$$\mathcal{X} = \{(\mathbf{z}_1, s_1), \ldots, (\mathbf{z}_T, s_T)\}. \tag{1}$$

At each valid window t, `BioLT` produces $\hat{y}_t = f_\theta(\mathbf{z}_{\le t}, s_{\le t})$, where $\hat{y}_t$ denotes the predicted risk or class probability, and y_t denotes the corresponding task-specific label for detection (whether an adverse event happens at the current timepoint) or forecasting (whether adverse event will happen in the future). The prediction is causal: $\hat{y}_t$ depends only on observations available up to timestamp s_t.

`BioLT` uses a decoder-only temporal backbone to update the patient representation under a causal attention mask (Section 3.1). To address the practical challenges of streaming biosignals, we introduce two components: a discontinuity-aware token encoding that exposes inter-window elapsed time to the model (Section 3.2), and a compressed memory bank that preserves long-range trajectory context while keeping inference efficient (Section 3.3).

3.1 BASIC MODELING: DECODER-ONLY NEXT-TOKEN PREDICTION

For real-time monitoring, predictions must be made using only past and current information. This aligns the standard causal sequence-modeling setting, which has been widely used in language (Brown et al., 2020) and video (Zhao et al., 2023; Girdhar & Grauman, 2021) modeling. `BioLT` therefore uses a decoder-only temporal backbone that processes the window sequence from left to right. Given the input tokens $\mathbf{z}_{1:T}$, the decoder produces hidden states $\mathbf{h}_{1:T} = F_\theta(\mathbf{z}_{1:T})$, where a causal attention mask ensures that $\mathbf{h}_t$ depends only on $\mathbf{z}_{\le t}$. $\mathbf{h}_t$ is then used for prediction at timestamp s_t.

To learn from biosignal streams, we pretrain `BioLT` with a next-window prediction (i.e., autoregressive) task. For two neighboring windows, let $\Delta\mathbf{z}_{t+1} = \mathbf{z}_{t+1} - \mathbf{z}_t$ denote the change in their snapshot representations. `BioLT` predicts this change from current summary vector: $\widehat{\Delta\mathbf{z}}_{t+1} = \mathrm{Pred}_{\mathrm{AR}}(\mathbf{h}_t)$,

where $\mathrm{Pred}_{\mathrm{AR}}$ is a lightweight prediction head. The pretraining loss is mean squared error between the predicted and true changes, averaged over neighboring valid window pairs:

$$\mathcal{L}_{\mathrm{AR}} = \mathrm{mean}_t \left\| \widehat{\Delta \mathbf{z}}_{t+1} - \Delta \mathbf{z}_{t+1} \right\|_2^2. \tag{2}$$

By predicting changes between windows, BioLT learns temporal physiological dynamics (refer to Appendix F.4) that are useful for downstream detection and forecasting.

3.2 Module 1: The "discontinuity" of streaming biosignals.

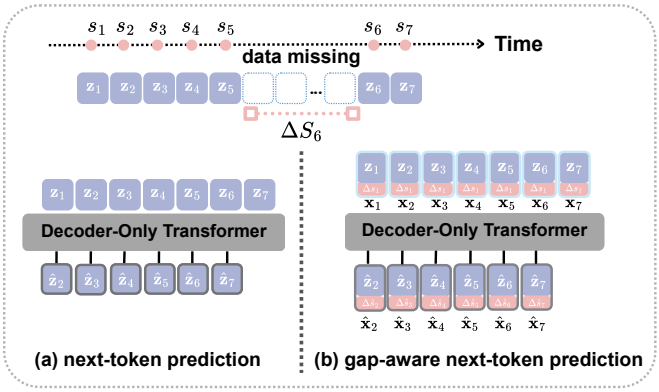


Figure 2: **Discontinuity-aware token encoding.** Neighboring valid biosignal windows may be separated by irregular elapsed time due to missing or low-quality signal segments. BioLT exposes this gap as Δs_t in the input token.

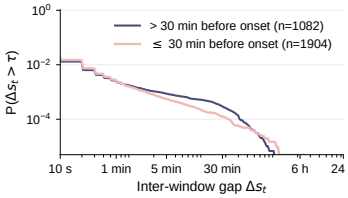


Figure 3: **Inter-window gaps are "informative" to adverse events.** Pooled CCDF of Δs_t on MC-MED ECG, comparing windows within 30 minutes before sepsis onset with those outside the forecast horizon. The different gap distributions suggest that elapsed time provides useful context for streaming biosignal prediction.

The basic causal decoder provides the streaming backbone of BioLT. The next question is how to form input tokens for biosignal streams, which are often not perfectly continuous in practice, as mentioned before (due to factors such as noise and sensor drop-out). As a result, two neighboring valid windows may be separated by a long elapsed time $(s_t - s_{t-1})$. Direct imputation is often unreliable, especially when gaps are long. Conversely, treating these windows as equally spaced can make the model confuse missing time with continuous physiological change.

We take inspirations from sequential modeling of EHRs, where "irregularity" is a persistent issue: valid clinical records may be observed at different and unevenly spaced time points. In recent Transformer-based EHR models (Renc et al., 2024), such gap information is often treated not merely as missingness to be imputed, but as a clinically meaningful signal that may reflect patient state, clinical workflow, or monitoring intensity. Instead of forcing all observations onto a regular time grid, these models encode elapsed time between observations as part of the input.

Similar "informativeness" of gap is also noted in biosignal gaps, as demonstrated in Figure 3. Therefore, we adopt the same principle for biosignal streams: rather than filling long recording gaps, we make the elapsed time between neighboring valid windows visible to BioLT.

To encode this information, we define the inter-window elapsed time as

$$\Delta s_t = s_t - s_{t-1}, \qquad \Delta s_1 = 0. \tag{3}$$

The model input token for the t-th valid window is then $\mathbf{x}_t = [\mathbf{z}_t; \Delta s_t]$. Please refer to Appendix C.2 for the detailed setup of Δs_t. The augmented token $\mathbf{x}_t$ is then used throughout BioLT, including in the next-window prediction objective of Section 3.1, as shown in Figure 2.

3.3 Module 2: The "infinite" streaming of biosignals.

Apart from the "gap" issue, the other challenge is "infinite length" incurred by real-world streaming biosignals. Such streaming may continue for hours or days, producing an increasing number of valid windows. For instance, in our datasets, individual recordings reach up to 5,907 valid windows ($\sim$16 h) on VitalDB (Appendix Table A.1). An ideal model should therefore update continuously, while not storing or attending to every previous window, to avoid prohibitive quadratic attention cost. BioLT addresses this with a compressed memory, as illustrated in Figure 4.

Chunk based streaming. Rather than attending to the whole sequence, we first divide the input into consecutive chunks comprising multiple windows (empirically set as 180 windows; refer to sensitivity analysis in Appendix F.1). Within the current chunk, BioLT applies exact causal attention for temporal modeling.

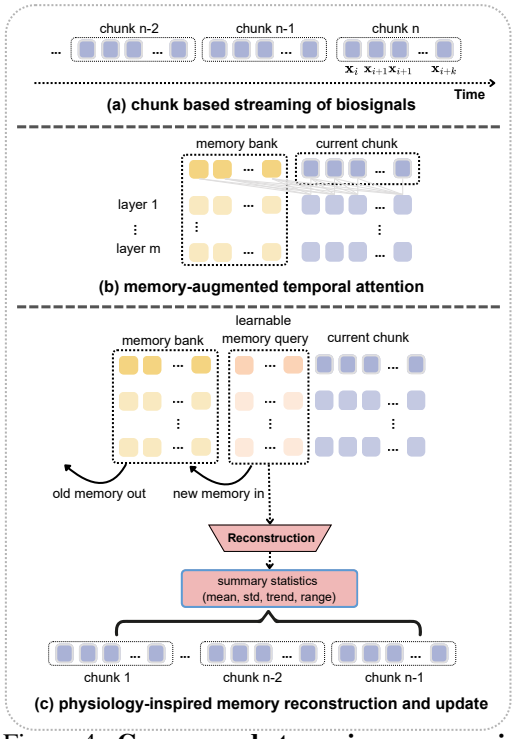


Figure 4: **Compressed streaming memory in `BioLT`. (a)** A long biosignal stream is divided into consecutive chunks. **(b)** Each temporal layer attends to both the current chunk and a fixed-size compressed memory bank. **(c)** Learnable memory queries produce new memory slots, by compressing the attention between old memory bank and last chunk; these slots are supervised by physiology-inspired trajectory summaries, and used to update the memory bank, in a first-in-first-out manner.

Memory-augmented temporal attention. To carry information across chunks, each temporal layer also attends to a fixed-size compressed "memory bank" M^{cmp} (Figure 4b), which stores compact summaries of previous chunks, and keeps updating as new chunks arrive, in a recurrent manner. Predictions therefore use both causal context within the current chunk and compressed information from earlier chunks.

Physiology-inspired memory learning. To make sure the "memory bank" preserves long-context historic information, we apply a learnable query module to compress the attentions operated between current chunk and old-"memory bank" into a small set of memory slots using learnable queries, and append these slots to the "memory bank" in a FIFO (first-in first-out) manner.

(i) Learnable query of current chunk. The memory compression is based on a learnable query mode. As shown in Figure 4c, after processing the current chunk, learnable memory query module compresses its hidden states into K ordered memory slots, $\mathbf{c}_{k,1}, \ldots, \mathbf{c}_{k,K}$. These slots are added to M^{cmp}, while the oldest slots are removed when the bank exceeds its capacity.

(ii) Learning objectives. To make sure the compressed memory preserves clinically meaningful information about the whole past trajectory, we regularize the learning using cumulative summary targets inspired by conventional long-term biosignal analysis. In conventional analysis, long recordings are often described by summary statistics such as average, variability, trend, and amplitude range. Hence, we define $\mathbf{q} = [\mu, \sigma, b, r]$, where μ, σ, b, and r denote the mean, standard deviation, linear trend, and dynamic range of the hidden states observed up to that point. A lightweight reconstruction head predicts these statistics from the compressed slot, giving $\mathcal{L}_{\text{mem}} = \text{mean SmoothL1}(\widehat{\mathbf{q}}, \mathbf{q})$, where $\widehat{\mathbf{q}}$ is the predicted summary target. The full pretraining objective is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{AR}} + \lambda_{\text{mem}} \mathcal{L}_{\text{mem}}. \tag{4}$$

`BioLT` is trained in two stages, and in both stages, the memory bank is updated sequentially using truncated backpropagation through time (TBPTT), as in previous sequential modeling work (Mikolov et al., 2010). Each prediction conditions on long-range context while backpropagation stays local to the current chunk. During the pretraining stage, `BioLT` is optimized by Equation 4 on the whole, unlabeled recordings. After pretraining, `BioLT` is fine-tuned for downstream dense prediction by applying the task loss to the output at each valid window of labeled segments of at most 500 windows.

During inference, chunks are processed sequentially, the memory bank is updated after each chunk, and predictions are produced causally over the full recording. The detailed hyperparameter settings are provided in Appendix D.4.

4 Experimental Setup

4.1 Datasets

Biosignal datasets. We evaluate `BioLT` across ECG, PPG, and EEG recordings with diverse sampling rates, label structures, and clinical time scales. (i) **CPSC-2021** (Wang et al., 2021) contains dynamic ECG recordings for atrial fibrillation (AF) analysis; each window is labeled as non-AF, paroxysmal AF, or persistent AF. (ii) **VitalDB** (Eun et al., 2026) provides intraop-

erative ECG and PPG waveforms with rhythm annotations, indicating AF or not. (iii) **MC-MED** (Kansal et al., 2025) contains visit-level ECG and PPG waveforms for sepsis onset modeling, evaluated as binary event detection and forecasting. (iv) **CHB-MIT** (Shoeb, 2009) contains pediatric scalp EEG recordings with seizure onset annotations. These datasets contain dense annotations of those specific adverse events, with details summarized in Table 1 & Appendix A.2.

Preprocessing. Each dataset is converted into per-recording sequences of non-overlapping 10-second windows, where each window has a timestamp s_t and snapshot representation $\mathbf{z}_t$ (by either domain feature engineering or foundation model, as detailed in Section 4.2). Signal cleaning and feature extraction are performed before temporal modeling, with details provided in Appendix A.1.

Table 1: **Dataset summary.** Evaluation covers 3 modalities and 4 datasets; records and 10-second windows are reported after preprocessing.

Dataset	Mod.	Task	Labels	Rec.	Win.
CPSC-2021	ECG	AF	3-class	1,424	170k
VitalDB	ECG	AF	Binary	431	571k
VitalDB	PPG	AF	Binary	382	470k
MC-MED	ECG	Sepsis	Binary	2,510	1.65M
MC-MED	PPG	Sepsis	Binary	2,715	2.12M
CHB-MIT	EEG	Seizure	Binary	655	342k

Split and Task Setup. All experiments use patient-level 60/20/20 train/validation/test splits, identical across all experiments, to ensure fair comparison and avoid cross-recording leakage. For **detection**, y_t denotes the clinical state at timestamp s_t; for **forecasting**, $y_t^{(H)}$ indicates whether the target event occurs within $[s_t, s_t + H]$. We evaluate $H \in \{1, 5, 10, 20\}$ minutes, with $H = 0$ treated as detection. Dataset-specific label construction is described in Appendix A.2.

4.2 Compared Methods

Feature-wise. We consider both hand-crafted domain features and foundation-model-derived embeddings to represent each 10-second "snapshot" window. (i) For **ECG**, domain features cover morphology, heart-rate-variability, and signal-quality features (64 features); foundation-model embeddings are extracted with CSFM (Gu et al., 2026), ECGFounder (Li et al., 2025), and CLEF (Shu et al., 2025). (ii) For **PPG**, domain features cover morphology, and HRV indices derived from systolic-peak intervals, and signal-quality features (71 features); embeddings are extracted with CSFM (Gu et al., 2026) and PulsePPG (Saha et al., 2025). (iii) For **EEG**, per-channel domain features cover time-domain statistics, frequency-band powers, wavelet time-frequency descriptors, and complexity/entropy measures (31 features $\times$ 22 channels); foundation-model embeddings are extracted with LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025). Detailed feature definitions are listed in Appendix B.1.

Temporal-modeling-wise. We compare with several temporal aggregation strategies. **Snapshot** applies an MLP classifier independently to each window and has no temporal context. **MeanPool** and **MaxPool** classify cumulative feature averages or maxima up to the current time. **LSTM** uses recurrent hidden states to carry history across windows. **TCN** applies causal dilated temporal convolutions. **Attention** is a single-head causal self-attention layer that learns a weighted sum of past projected snapshots, subsequently classified by an MLP head. **Mamba-3** (Lahoti et al., 2026) is a recent state-space sequence model whose selective scan provides linear-time recurrent context aggregation. **Transformer-XL** (Dai et al., 2019) is architecturally identical to `BioLT`, whereas the compressed memory bank is replaced by an uncompressed 180-window FIFO segment memory, isolating the effect of compression itself. It is trained with the same input tokens, autoregressive pretraining recipe, downstream training procedures, as `BioLT`.

In practice, to mimic realistic "streaming" settings, all compared methods are trained on sequences cropped to a fixed maximum length of 500 windows during downstream training and tested on the full recordings. Full architectural specifications and the shared training protocol are provided in Appendix E. Results are reported as mean$\pm$standard deviation over three random seeds.

4.3 Performance Metrics

Window-based metrics. At each 10-second window t, each model outputs a probability p_t under the causal setting. We evaluate window-level predictions using AUROC and F1. AUROC is independent of any threshold choice. For multi-class tasks, we report macro-averaged AUROC and F1, obtained by class-argmax without any operating threshold; for binary tasks, we report standard binary AUROC and F1 (95% specificity), where the operating threshold is calibrated per dataset/model/seed to attain 95% specificity on the *validation* split and then frozen before the test stream is scored.

Alarm-based metric. Apart from window-level metrics to assess individual predictions, continuous monitoring also requires evaluating whether clinical events are detected at an acceptable false-alarm burden. Following discussions with clinicians, we therefore add one event-level metric for the binary

tasks. An *event* is a ground-truth occurrence of the target condition, i.e., a consecutive sequence of windows whose label is positive ($y_t = 1$); a model raises an *alarm* at window t when its predicted probability p_t exceeds a threshold θ. An alarm is *false* if no event begins within the following 30 minutes; an event is *detected* if at least one alarm falls inside it. Sens@1FA/h is the fraction of test events detected at a threshold (θ). We choose (θ) on the validation set to yield at most one false alarm per hour and keep it fixed for testing:

$$\text{Sens@1FA/h} = \frac{\#\{\text{test events detected at } \theta^*\}}{\#\{\text{test events}\}}, \qquad \theta^* = \min\{\theta : \text{FA}_{\text{val}}(\theta) \le 1 \text{ per hour}\}. \tag{5}$$

5 Results and Discussion

5.1 Main Detection and Forecasting Results

BioLT improves detection and forecasting across biosignal tasks, datasets, modalities, and feature types. Figure 5 provides a compact overview of the main results. The left panel summarizes detection performance at $H = 0$, using AUROC and F1 across ECG, PPG, and EEG tasks. The right panel summarizes forecasting performance averaged across $H \in \{1, 5, 10, 20\}$ minutes. AUROC and F1 axes are normalized separately by the best observed value for each metric, respectively. A larger radius therefore indicates stronger relative performance. Overall, BioLT achieves the largest score on 8 of the 12 detection axes and on 8 of the 12 forecasting axes in Figure 5, including every forecasting AUROC axis. It is noteworthy that performance margins of forecasting is generally smaller, compared to detection, suggesting greater inherent difficulty of forecasting tasks.

BioLT is effective across both domain features and foundation-model embeddings. We further break down the results by dataset, compared method, and feature type in Table 2 for detection and Table 3 for 20-minute forecasting. Overall, BioLT achieves the best detection AUROC in 15 of 21 dataset-feature settings (Table 2) and the best 20-minute forecasting AUROC in 13 of 21 settings (Table 3). This demonstrates BioLT as an input-agnostic temporal backbone, compatible with both hand-crafted domain features as well as embeddings from pretrained foundation models.

BioLT performance gains reflect the value of temporal context. The advantage of BioLT is most evident in settings where temporal context provides information beyond an isolated 10-second window. For instance, at detection (Table 2), BioLT performs particularly strongly on CHB-MIT seizure detection and MC-MED sepsis detection, while gains are smaller on AF tasks where local rhythm morphology is already informative. For forecasting, margins over competing methods are generally smaller. At 20-min forecasting (Table 3), however, BioLT remains strongest in most settings, particularly for MC-MED sepsis forecasting across both ECG and PPG, in line with prior evidence that dynamic physiology changes carry prognostic information in sepsis (Zhang et al., 2026).

Effective aggregation of domain features outperforms snapshot-based foundation model features. Further into detection results (Table 2), as shown in the rightmost column, BioLT on domain features outperforms the snapshot-based method on all FM embeddings of the same datasets. Moreover, once modeled by BioLT, domain features and FM embeddings achieve broadly competitive performance across tasks, as in second-to-last column. This is particularly evident for tasks where a

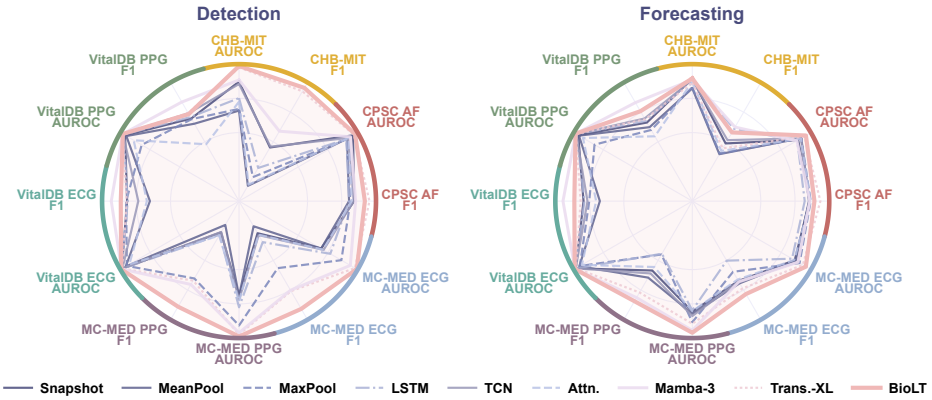


Figure 5: **Hierarchical radar comparison of detection and forecasting performance (averaged across 1,5,10,20 min) across biosignal datasets.** Each dataset is represented by two adjacent axes (AUROC and F1). The performance is averaged across all compatible features for each dataset.

Table 2: **Detection results.** Test AUROC (%, mean±std over 3 seeds). **Best mean** per row in **bold**, second best per row underlined. Rightmost column presents the performance gains using BioLT on domain features (Dom. BioLT) vs using snapshot on respective FM embeddings (Snap. FM).

Datasets	Features	Snapshot	MeanPool	MaxPool	LSTM	TCN	Attn.	Mamba-3	Trans.-XL	BioLT	Dom. BioLT vs. Snap. FM
CHB-MIT EEG seizure	Domain	$75.85_{\pm 3.60}$	$61.55_{\pm 7.70}$	$54.82_{\pm 7.95}$	$70.00_{\pm 4.19}$	$69.92_{\pm 3.32}$	$62.25_{\pm 7.52}$	$76.15_{\pm 4.45}$	$79.96_{\pm 9.71}$	**81.97**$_{\pm 7.89}$	↘
	CBraMod	$69.89_{\pm 5.36}$	$56.57_{\pm 14.18}$	$53.24_{\pm 1.94}$	$63.67_{\pm 11.32}$	$69.50_{\pm 9.16}$	$57.52_{\pm 7.04}$	$74.40_{\pm 6.23}$	**80.87**$_{\pm 10.88}$	$80.03_{\pm 9.40}$	+12.08
	LaBraM	$70.77_{\pm 3.86}$	$45.82_{\pm 4.98}$	$58.99_{\pm 4.38}$	$52.87_{\pm 14.51}$	$72.65_{\pm 5.99}$	$60.30_{\pm 2.70}$	$68.88_{\pm 6.08}$	$81.34_{\pm 3.41}$	**82.70**$_{\pm 5.74}$	+11.20
CPSC-2021 ECG AF	Domain	$92.81_{\pm 2.58}$	$87.24_{\pm 2.39}$	$86.02_{\pm 3.41}$	$86.93_{\pm 2.05}$	$88.28_{\pm 4.21}$	$88.33_{\pm 2.95}$	$91.58_{\pm 2.50}$	$96.37_{\pm 1.90}$	**96.83**$_{\pm 1.69}$	↘
	CSFM	$92.85_{\pm 3.69}$	$88.02_{\pm 6.26}$	$91.04_{\pm 5.91}$	$89.85_{\pm 4.06}$	$90.91_{\pm 6.95}$	$88.87_{\pm 8.05}$	$91.92_{\pm 4.53}$	$93.32_{\pm 5.30}$	**95.57**$_{\pm 2.37}$	+3.98
	CLEF	**92.06**$_{\pm 2.65}$	$88.01_{\pm 3.32}$	$82.78_{\pm 7.16}$	$85.56_{\pm 4.41}$	$86.23_{\pm 6.18}$	$89.55_{\pm 5.70}$	$88.06_{\pm 4.54}$	$89.74_{\pm 8.83}$	$90.23_{\pm 4.57}$	+4.77
	ECGFounder	$94.53_{\pm 4.20}$	$89.43_{\pm 4.94}$	$90.91_{\pm 5.62}$	$88.20_{\pm 7.35}$	$92.27_{\pm 6.77}$	$91.56_{\pm 6.35}$	$92.23_{\pm 6.96}$	$95.73_{\pm 3.36}$	**96.35**$_{\pm 3.08}$	+2.30
VitalDB ECG AF	Domain	$92.84_{\pm 1.36}$	$88.35_{\pm 3.20}$	$84.79_{\pm 1.98}$	$94.12_{\pm 0.56}$	$91.81_{\pm 1.64}$	$84.63_{\pm 8.96}$	$94.47_{\pm 0.86}$	$93.72_{\pm 1.17}$	**94.91**$_{\pm 1.58}$	↘
	CSFM	$94.41_{\pm 0.67}$	$91.35_{\pm 2.15}$	$91.80_{\pm 1.98}$	$94.56_{\pm 0.17}$	$93.95_{\pm 0.79}$	$91.11_{\pm 3.16}$	$94.15_{\pm 1.09}$	$93.92_{\pm 0.67}$	**94.70**$_{\pm 0.56}$	+0.50
	CLEF	$93.96_{\pm 0.32}$	$91.16_{\pm 0.91}$	$90.94_{\pm 0.58}$	$93.40_{\pm 0.23}$	$93.50_{\pm 0.13}$	$86.71_{\pm 3.63}$	$93.32_{\pm 1.01}$	$93.44_{\pm 0.89}$	**94.36**$_{\pm 0.46}$	+0.95
	ECGFounder	$94.00_{\pm 0.72}$	$91.92_{\pm 1.99}$	$91.57_{\pm 0.83}$	**94.43**$_{\pm 0.72}$	$94.08_{\pm 0.69}$	$88.83_{\pm 1.58}$	$94.30_{\pm 1.43}$	$93.01_{\pm 2.33}$	$93.48_{\pm 0.52}$	+0.91
VitalDB PPG AF	Domain	$88.90_{\pm 0.48}$	$88.55_{\pm 0.52}$	$79.07_{\pm 1.92}$	$90.73_{\pm 0.37}$	$90.54_{\pm 0.47}$	$83.77_{\pm 3.29}$	**91.98**$_{\pm 0.47}$	$90.83_{\pm 0.58}$	$90.88_{\pm 0.30}$	↘
	CSFM	$89.84_{\pm 0.45}$	$88.28_{\pm 0.85}$	$90.51_{\pm 1.03}$	$90.97_{\pm 0.51}$	$90.94_{\pm 0.75}$	$87.27_{\pm 2.08}$	$89.92_{\pm 0.44}$	**91.13**$_{\pm 0.30}$	$90.65_{\pm 0.73}$	+1.04
	PulsePPG	$86.37_{\pm 0.15}$	$85.30_{\pm 0.41}$	$56.92_{\pm 3.76}$	$86.69_{\pm 1.38}$	$88.63_{\pm 0.32}$	$74.08_{\pm 5.07}$	$88.61_{\pm 0.65}$	**89.19**$_{\pm 0.19}$	$88.96_{\pm 0.58}$	+4.51
MC-MED ECG sepsis	Domain	$54.23_{\pm 0.96}$	$52.95_{\pm 2.87}$	$62.76_{\pm 2.87}$	$61.05_{\pm 2.26}$	$57.31_{\pm 2.54}$	$53.18_{\pm 2.49}$	$74.37_{\pm 2.10}$	$78.32_{\pm 1.41}$	**78.78**$_{\pm 0.63}$	↘
	CSFM	$56.38_{\pm 0.97}$	$57.29_{\pm 1.72}$	$70.02_{\pm 1.63}$	$62.49_{\pm 0.92}$	$58.54_{\pm 0.71}$	$58.50_{\pm 3.11}$	$75.16_{\pm 0.66}$	$77.71_{\pm 0.88}$	**78.70**$_{\pm 1.33}$	+22.40
	CLEF	$54.54_{\pm 2.58}$	$55.13_{\pm 2.77}$	$69.52_{\pm 1.05}$	$63.63_{\pm 3.25}$	$56.31_{\pm 1.45}$	$57.67_{\pm 3.60}$	$75.68_{\pm 1.38}$	$78.89_{\pm 1.12}$	**79.40**$_{\pm 0.89}$	+24.24
	ECGFounder	$54.58_{\pm 1.50}$	$54.61_{\pm 1.12}$	$71.67_{\pm 0.64}$	$57.22_{\pm 4.24}$	$54.23_{\pm 1.27}$	$55.75_{\pm 1.94}$	$72.85_{\pm 0.11}$	$75.44_{\pm 0.49}$	**78.41**$_{\pm 0.73}$	+24.20
MC-MED PPG sepsis	Domain	$55.54_{\pm 0.53}$	$54.09_{\pm 0.94}$	$73.78_{\pm 0.90}$	$59.72_{\pm 2.23}$	$58.51_{\pm 1.12}$	$57.05_{\pm 2.51}$	$78.42_{\pm 0.58}$	$79.34_{\pm 0.53}$	**80.96**$_{\pm 1.10}$	↘
	CSFM	$56.93_{\pm 2.29}$	$56.78_{\pm 3.53}$	$75.23_{\pm 1.88}$	$66.33_{\pm 2.51}$	$58.30_{\pm 2.83}$	$60.44_{\pm 2.30}$	$76.54_{\pm 1.04}$	$78.82_{\pm 0.64}$	**78.96**$_{\pm 1.18}$	+24.03
	PulsePPG	$55.92_{\pm 0.81}$	$57.64_{\pm 2.31}$	$71.55_{\pm 1.85}$	$62.25_{\pm 1.26}$	$58.09_{\pm 1.07}$	$60.06_{\pm 0.87}$	$78.63_{\pm 0.64}$	$78.35_{\pm 1.26}$	**80.22**$_{\pm 1.09}$	+25.04

Table 3: **Forecasting results at horizon H=20 minutes.** Test AUROC (%, mean±std over 3 seeds). **Best mean** per row in **bold**, second best per row underlined. Rightmost column presents the performance gains using BioLT on domain features (Dom. BioLT) vs using snapshot on respective FM embeddings (Snap. FM).

Datasets	Features	Snapshot	MeanPool	MaxPool	LSTM	TCN	Attn.	Mamba-3	Trans.-XL	BioLT	Dom. BioLT vs. Snap. FM
CHB-MIT EEG seizure	Domain	**60.04**$_{\pm 8.08}$	$57.49_{\pm 10.33}$	$52.34_{\pm 4.18}$	$58.89_{\pm 6.90}$	$56.93_{\pm 9.50}$	$56.86_{\pm 16.89}$	$52.80_{\pm 3.94}$	$52.43_{\pm 9.90}$	$51.10_{\pm 9.53}$	↘
	CBraMod	$57.71_{\pm 10.42}$	$58.18_{\pm 16.99}$	$52.95_{\pm 10.75}$	**60.51**$_{\pm 13.95}$	$58.88_{\pm 13.07}$	$59.01_{\pm 14.34}$	$56.24_{\pm 9.36}$	$54.24_{\pm 9.48}$	$56.62_{\pm 13.16}$	−6.61
	LaBraM	$47.83_{\pm 5.86}$	$45.44_{\pm 7.17}$	$53.90_{\pm 12.04}$	$52.79_{\pm 9.53}$	$54.45_{\pm 12.07}$	$54.72_{\pm 9.52}$	$52.86_{\pm 6.85}$	$48.32_{\pm 7.49}$	**56.69**$_{\pm 13.83}$	+3.27
CPSC-2021 ECG AF	Domain	$75.36_{\pm 6.18}$	$78.53_{\pm 9.37}$	$71.41_{\pm 9.60}$	$77.53_{\pm 5.01}$	$78.18_{\pm 5.98}$	$81.94_{\pm 3.39}$	$76.03_{\pm 6.39}$	**87.11**$_{\pm 2.76}$	$86.43_{\pm 1.63}$	↘
	CSFM	$81.10_{\pm 4.59}$	$83.47_{\pm 1.90}$	$79.14_{\pm 9.86}$	$82.77_{\pm 4.97}$	$84.11_{\pm 1.55}$	$81.40_{\pm 3.09}$	$81.39_{\pm 4.25}$	$82.66_{\pm 2.09}$	**87.45**$_{\pm 2.08}$	+5.33
	CLEF	$76.41_{\pm 5.22}$	$80.30_{\pm 4.80}$	$79.07_{\pm 5.71}$	$77.91_{\pm 6.04}$	$80.58_{\pm 6.38}$	**82.02**$_{\pm 7.21}$	$79.14_{\pm 5.97}$	$75.46_{\pm 12.61}$	$79.67_{\pm 9.31}$	+10.02
	ECGFounder	$80.41_{\pm 6.24}$	$79.33_{\pm 4.43}$	$82.85_{\pm 4.74}$	$75.27_{\pm 5.08}$	$81.63_{\pm 2.08}$	$79.99_{\pm 2.40}$	$73.03_{\pm 0.42}$	$83.64_{\pm 2.29}$	**85.40**$_{\pm 4.38}$	+6.02
VitalDB ECG AF	Domain	$88.75_{\pm 2.12}$	$84.34_{\pm 3.69}$	$82.02_{\pm 3.68}$	$90.28_{\pm 1.46}$	$88.97_{\pm 2.08}$	$80.07_{\pm 8.56}$	$90.67_{\pm 1.28}$	$88.99_{\pm 1.81}$	**91.34**$_{\pm 1.53}$	↘
	CSFM	$89.40_{\pm 1.79}$	$87.46_{\pm 1.87}$	$87.81_{\pm 0.56}$	$89.28_{\pm 1.41}$	$89.74_{\pm 1.44}$	$87.90_{\pm 2.17}$	$90.09_{\pm 1.51}$	$90.17_{\pm 1.99}$	**90.61**$_{\pm 1.51}$	+1.94
	CLEF	**89.35**$_{\pm 1.75}$	$86.17_{\pm 1.09}$	$86.42_{\pm 1.42}$	$88.55_{\pm 2.30}$	$89.19_{\pm 1.77}$	$86.13_{\pm 2.77}$	$89.14_{\pm 2.05}$	$88.24_{\pm 2.78}$	$89.27_{\pm 1.40}$	+1.99
	ECGFounder	$89.57_{\pm 1.52}$	$86.40_{\pm 3.76}$	$87.78_{\pm 1.48}$	$88.87_{\pm 2.48}$	$90.12_{\pm 1.68}$	$85.09_{\pm 2.66}$	**90.32**$_{\pm 1.06}$	$89.02_{\pm 2.80}$	$90.18_{\pm 1.98}$	+1.77
VitalDB PPG AF	Domain	$87.49_{\pm 1.23}$	$88.36_{\pm 1.54}$	$79.14_{\pm 0.62}$	$88.81_{\pm 1.26}$	$88.58_{\pm 1.45}$	$87.49_{\pm 2.42}$	$90.38_{\pm 1.37}$	$89.19_{\pm 1.55}$	**90.43**$_{\pm 1.20}$	↘
	CSFM	$88.73_{\pm 0.73}$	$88.68_{\pm 0.64}$	$89.32_{\pm 1.85}$	$88.94_{\pm 0.56}$	$89.48_{\pm 0.44}$	$87.89_{\pm 0.28}$	$89.03_{\pm 0.94}$	**89.81**$_{\pm 0.96}$	**89.81**$_{\pm 1.08}$	+1.70
	PulsePPG	$85.05_{\pm 1.52}$	$84.59_{\pm 2.44}$	$51.36_{\pm 9.28}$	$87.38_{\pm 2.63}$	**88.38**$_{\pm 1.09}$	$81.64_{\pm 9.41}$	$88.07_{\pm 1.09}$	$87.48_{\pm 1.22}$	$88.22_{\pm 1.42}$	+5.38
MC-MED ECG sepsis	Domain	$52.47_{\pm 1.76}$	$53.26_{\pm 2.44}$	$57.25_{\pm 0.93}$	$52.84_{\pm 1.01}$	$54.04_{\pm 1.71}$	$54.18_{\pm 1.37}$	$55.19_{\pm 0.59}$	**63.24**$_{\pm 2.16}$	$62.20_{\pm 1.10}$	↘
	CSFM	$54.37_{\pm 1.87}$	$53.94_{\pm 0.93}$	$60.30_{\pm 1.21}$	$53.42_{\pm 3.18}$	$54.70_{\pm 1.71}$	$51.71_{\pm 3.41}$	$53.73_{\pm 1.29}$	$55.06_{\pm 0.99}$	**61.01**$_{\pm 2.28}$	+7.83
	CLEF	$54.03_{\pm 2.38}$	$55.10_{\pm 3.96}$	$59.20_{\pm 2.25}$	$54.26_{\pm 1.00}$	$54.64_{\pm 2.31}$	$55.48_{\pm 3.42}$	$56.86_{\pm 1.06}$	$54.72_{\pm 2.08}$	**62.24**$_{\pm 1.67}$	+8.17
	ECGFounder	$52.52_{\pm 2.30}$	$52.19_{\pm 1.87}$	$55.44_{\pm 1.51}$	$52.00_{\pm 1.69}$	$53.38_{\pm 0.92}$	$52.81_{\pm 1.91}$	$51.46_{\pm 2.28}$	$55.01_{\pm 2.34}$	**58.05**$_{\pm 1.70}$	+9.68
MC-MED PPG sepsis	Domain	$52.27_{\pm 0.73}$	$53.82_{\pm 2.23}$	$60.20_{\pm 0.82}$	$52.04_{\pm 1.52}$	$54.27_{\pm 0.79}$	$51.64_{\pm 1.17}$	$62.03_{\pm 0.56}$	$64.73_{\pm 0.79}$	**64.92**$_{\pm 0.71}$	↘
	CSFM	$52.34_{\pm 1.56}$	$51.93_{\pm 2.93}$	$60.81_{\pm 1.75}$	$52.76_{\pm 1.06}$	$53.56_{\pm 2.15}$	$51.50_{\pm 3.46}$	$58.93_{\pm 2.40}$	$52.65_{\pm 0.95}$	**63.15**$_{\pm 0.58}$	+12.58
	PulsePPG	$51.19_{\pm 0.48}$	$51.42_{\pm 1.17}$	$60.14_{\pm 0.64}$	$51.52_{\pm 1.32}$	$52.77_{\pm 0.46}$	$50.67_{\pm 1.22}$	$61.66_{\pm 1.47}$	$53.17_{\pm 0.52}$	**63.88**$_{\pm 0.80}$	+13.73

snapshot window may not be enough for detection, such as sepsis (Zhang et al., 2026) and seizure (Tatum et al., 2022) detection. This suggests that how longitudinal window-level representations are modeled can be as important as, or even more important than, the window-level feature extractor itself, which is the main motivation of BioLT.

Alarm-based evaluation. Figure 6 reports Sens@1FA/h at detection (Section 4.3). BioLT recovers the largest fraction of events (0.314), ahead of Snapshot (0.287) and Transformer-XL (0.267); MeanPool, MaxPool, and Attention recover fewer than 8%. Further discussion is available in Appendix G. In terms of all other detailed results, please refer to Appendix F.

5.2 Performance at Varied Forecasting Settings

BioLT remains competitive across forecasting horizons. We next examine how performance changes as the forecasting horizon increases. Figure 7 reports VitalDB ECG binary AF forecasting with CSFM features from detection to 20-minute forecasting. As expected, performance generally decreases as the horizon becomes longer, since the model is asked to predict further into the future from less directly informative observations. Across these horizons, BioLT remains consistently competitive with the strongest temporal baselines and often achieves the best or near-best AUROC. This suggests that BioLT does not only improve current-state detection but can also support continuously updated prediction over clinically meaningful future horizons.

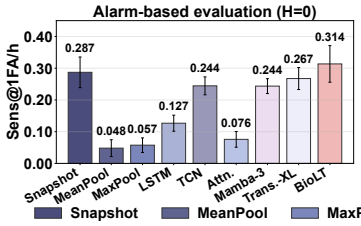


Figure 6: **Alarm-based evaluation at detection (H=0).** Average Sens@1FA/h over 10 binary dataset settings over 3 seeds.

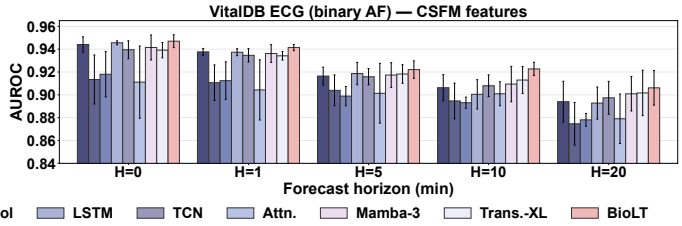


Figure 7: **Forecasting across horizons.** Forecasting performance on VitalDB ECG binary AF using CSFM features across horizons from $H = 0$ to 20 min. Error bars show the standard deviation over three seeds. Significance tests reported in Appendix F.5.

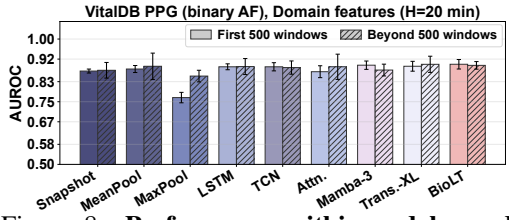


Figure 8: **Performance within and beyond the fixed-length training context.** Models are trained on at most 500 windows and tested on full recordings; bars show test AUROC within the first 500 windows and beyond, separately.

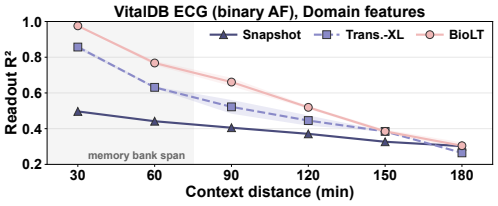


Figure 9: **Reconstruction of past context from memory alone.** Readout R^2 (mean±std over 3 seeds) of the mean domain features of the chunk located a given distance back. Shaded area: length of the `BioLT` memory bank.

`BioLT` generalizes beyond the fixed-length training context. We further evaluate whether models trained with a maximum context of 500 windows (Section 4.2) can maintain performance when applied to continuously streaming parts of longer recordings. Figure 8 separates test windows into those within the first 500 windows and those beyond this training-context length. Several baselines show variable performance after the 500-window boundary, indicating that extrapolating beyond the training context is non-trivial. In contrast, `BioLT` achieves the strongest performance both before and after the boundary, with no evident degradation beyond the training length. This supports the role of the compressed memory bank in enabling streaming inference over longer recordings.

5.3 Ablation Results

Validation of compressed memory. To evaluate the "memory" bank actually "remembers" historic information, we keep the backbone frozen, and fit on linear regression layer on top of the captured memory state to reconstruct the mean domain features of chunks located a chunks back. This is compared to the memory learned in Transformer-XL and to a Snapshot baseline that sees only the current window. As shown in Figure 9, on VitalDB ECG AF, `BioLT` retains the most information at every distance: it reconstructs the right previous chunk almost perfectly (R^2=0.98, against 0.86 for Transformer-XL and 0.50 for Snapshot) and stays above both up to 120 min, twice the nominal 60-min span of its memory bank.

Table 4: **Ablation on MC-MED sepsis forecasting (H=10).** Mean test AUROC with domain features.

Configuration	ECG	PPG
`BioLT`	**60.29±0.64**	**63.19±0.48**
w/o Δs	58.02±1.71	60.61±1.43
w/o mem. bank	56.68±0.23	57.83±0.34
w/o both	53.82±0.17	53.74±0.95

Ablation of core components. Furthermore, we ablate two key components of `BioLT`: Δs_t and memory bank. Full results shown in Table 4 demonstrate the effectiveness of both modules.

6 Conclusion

This work studies a practical yet underexplored setting in biosignal learning: how to turn short-window biosignal representations into continuously updated, long-context physiological trajectories. We introduce `BioLT`, a streaming temporal backbone that bridges snapshot representations and continuous clinical monitoring under imperfect, gapped data streams. Across overall 21 dataset-feature settings, covering arrhythmia, sepsis, and seizure tasks, and varied window-level features types, `BioLT` achieves the highest detection AUROC in 15 settings and the highest 20-minute forecasting AUROC in 13. This provides a strong basis on shifting from improving snapshot-level biosignal representations alone toward modelling longitudinal information, which is becoming increasingly accessible through advances in sensing hardware and digital health infrastructure.

AI USE STATEMENT

We used generative AI tools for code review, refine the wording and clarity of manuscript text, and assist with LaTeX formatting. We did not use generative AI tools to generate research data or experimental results. All edits were reviewed and verified by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

This work uses four publicly available, de-identified human-subject datasets (CPSC-2021, VitalDB, MC-MED, and CHB-MIT), accessed under their respective data-use terms, and involves secondary analysis only; no new data were collected.

REPRODUCIBILITY STATEMENT

Dataset preprocessing, window construction, task definitions, and data splits are described in Appendix A.1-A.3; feature and embedding extraction in Appendix B; the `BioLT` architecture in Appendix C; and the training protocol, hyperparameters, and compute in Appendix D.

REFERENCES

Robin Blythe, Rex Parsons, Nicole M White, David Cook, and Steven McPhail. A scoping review of real-time automated clinical deterioration alerts and evidence of impacts on hospitalised patient outcomes. *BMJ Quality & Safety*, 31(10):725–734, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019.

Ha Thi Hai Duong, Girmaw Abebe Tadesse, Phung Tran Huy Nhat, Nguyen Van Hao, John Prince, Tran Duc Duong, Trinh Trung Kien, Le Thanh Hoang Nhat, Le Van Tan, Chris Pugh, et al. Heart rate variability as an indicator of autonomic nervous system disturbance in tetanus. *The American journal of tropical medicine and hygiene*, 102(2):403, 2019.

Mohamed Elgendi. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews*, 8(1):14–25, 2012.

Da-In Eun, Kayoung Shim, Hyunsoo Lee, Yeji Lim, Hanbyeol Lim, Hyeonhoon Lee, Jiwon Lee, and Hyung-Chul Lee. Vitaldb arrhythmia database: An anesthesiologist-validated large-scale intraoperative arrhythmia dataset with beat and rhythm labels. *Scientific Data*, 2026.

Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, 161:1–13, 2018.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13505–13515, 2021.

Xiao Gu, Yuxuan Shu, Jinpei Han, Yuxuan Liu, Zhangdaihong Liu, James Anibal, Veer Sangha, Edward Phillips, Bradley Segal, Hang Yuan, et al. Foundation models for biosignals: A survey. *Authorea Preprints*, 2025.

Xiao Gu, Wei Tang, Jinpei Han, Veer Sangha, Fenglin Liu, Shreyank N Gowda, Antonio H Ribeiro, Patrick Schwab, Kim Branson, Lei Clifton, et al. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, 8(2):220–233, 2026.

Jinpei Han, Yinxuan Li, Xiao Gu, and A Aldo Faisal. An eeg conformer model for error feedback during human-robot interaction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11256–11262. IEEE, 2025.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, 2019.

Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in biology and medicine*, 122:103801, 2020.

Mohammad-Parsa Hosseini, Amin Hosseini, and Kiarash Ahi. A review on machine learning for eeg signal processing in bioengineering. *IEEE reviews in biomedical engineering*, 14:204–218, 2020.

Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3): 364–373, 2020.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A Kim. Mc-med, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12(1):1094, 2025.

Aakash Lahoti, Kevin Y Li, Berlin Chen, Caitlin Wang, Aviv Bick, J Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. *arXiv preprint arXiv:2603.15569*, 2026.

Jun Li, Aaron D Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *Nejm ai*, 2(7):AIoa2401033, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pp. 1045–1048. Makuhari, 2010.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations (ICLR)*, 2020.

Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ digital medicine*, 7 (1):256, 2024.

Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: The PhysioNet/Computing in cardiology challenge 2019. *Critical Care Medicine*, 48 (2):210–217, 2020.

Mithun Saha, Maxwell A Xu, Wanting Mao, Sameer Neupane, James M Rehg, and Santosh Kumar. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–35, 2025.

Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Yuxuan Shu, Peter H. Charlton, Fahim Kawsar, Jussi Hernesniemi, and Mohammad Malekzadeh. CLEF: Clinically-guided contrastive learning for electrocardiogram foundation models. *arXiv preprint arXiv:2512.02180*, 2025.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

William O Tatum, Jayanti Mani, Kazutaka Jin, Jonathan J Halford, David Gloss, Firas Fahoum, Louis Maillard, Ian Mothersill, and Sandor Beniczky. Minimum standards for inpatient long-term video-eeg monitoring: A clinical practice guideline of the international league against epilepsy and international federation of clinical neurophysiology. *Clinical Neurophysiology*, 134:111–128, 2022.

Geoffrey H. Tison, José M. Sanchez, Brandon Ballinger, Avesh Singh, Jeffrey E. Olgin, Mark J. Pletcher, Eric Vittinghoff, Emily S. Lee, Shannon M. Fan, Rachel A. Gladstone, Carlos Mikell, Nimit Sohoni, Johnson Hsieh, and Gregory M. Marcus. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA Cardiology*, 3(5):409–416, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations (ICLR)*, 2025.

Xingyao Wang, Caiyun Ma, Xiangyu Zhang, Hongxiang Gao, Gari D Clifford, and Chengyu Liu. Paroxysmal atrial fibrillation events detection from dynamic ecg recordings: The 4th china physiological signal challenge 2021. *Proc. PhysioNet*, pp. 1–83, 2021.

Rui Zhang, Fang Long, Zhanqi Zhao, Jingyi Wu, Ruoming Tan, Wen Xu, Lei Li, Yun Long, and Hongping Qu. Machine learning predicts sepsis deterioration trajectories. *npj Digital Medicine*, 2026.

Yucheng Zhao, Chong Luo, Chuanxin Tang, Dongdong Chen, Noel Codella, and Zheng-Jun Zha. Streaming video model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14602–14612, 2023.

A Dataset Details and Task Definitions

A.1 Dataset preprocessing and window construction

All datasets are represented as one sequence of windows per recording before temporal modeling. Each recording is segmented into valid, non-overlapping 10-second windows. For each valid window, we store its timestamp s_t, task label y_t, window index, and snapshot representation $\mathbf{z}_t$.

Signal cleaning and window construction are performed at the native sampling rate of each dataset. Resampling is applied only later, when preparing model-specific inputs for pretrained foundation models. This design keeps the clinical timeline shared across domain features, foundation-model embeddings and temporal aggregation models.

ECG. For ECG recordings, we use the clinically standard Lead II channel when available. VitalDB and MC-MED ECG signals are processed at 500 Hz, while CPSC-2021 ECG signals are processed at 200 Hz. ECG signals are cleaned with bandpass filtering in the 0.5-40 Hz range, and low-quality or near-flat windows are removed before feature extraction. Each remaining 10-second ECG window is z-score normalized at the window level.

PPG. For PPG recordings, we use the PLETH or PPG channel from VitalDB and MC-MED. VitalDB PPG is processed at 500 Hz and MC-MED PPG at 125 Hz. PPG signals are filtered with a 0.5-8 Hz bandpass filter to retain pulse-wave morphology while removing low-frequency drift and high-frequency noise. Near-flat windows are excluded, and valid 10-second windows are z-score normalized.

EEG. For EEG recordings, we use CHB-MIT with the major 22-channel bipolar montage. Recordings that do not contain the required channel set are excluded to maintain a fixed input shape across subjects. Signals are processed at the native 256 Hz sampling rate using MNE-based bandpass filtering from 0.1–75 Hz and a 60 Hz notch filter. Each 10-second window contains 22 channels and 2560 samples per channel. Valid windows are z-score normalized per channel.

The gaps between valid windows are preserved through the timestamp sequence and later encoded as the elapsed-time feature $\Delta s_t = s_t - s_{t-1}$, so that temporal models observe the true clinical timeline rather than an artificially regularized sequence.

Table A.1: **Per-dataset cohort statistics and per-recording length distribution.** For each (dataset, modality) pair, we report the number of recordings retained after preprocessing (Appendix A.1), the total number of valid 10-second windows, and the per-recording window-count median and maximum. Window counts convert linearly to wall-clock duration at 10 s per window.

Dataset	Mod.	Task	Labels	Rec.	Total win.	Win. / rec.	
						Median	Max
CPSC-2021	ECG	AF	3-class	1,424	170,017	54	2,465
VitalDB	ECG	AF	Binary	431	570,672	1,151	5,907
VitalDB	PPG	AF	Binary	382	470,264	1,127	3,797
MC-MED	ECG	Sepsis	Binary	2,510	1,645,535	490	4,084
MC-MED	PPG	Sepsis	Binary	2,715	2,122,840	641	4,160
CHB-MIT	EEG	Seizure	Binary	655	342,449	360	1,442

A.2 Task definitions and label construction

This section documents how the per-window labels of each detection / forecasting task in Section 5 are derived from the original dataset annotations.

CPSC-2021 (ECG, paroxysmal AF). The CPSC-2021 release contains per-recording WFDB annotations together with a recording-level AF subtype recorded in the `.hea` comment line, which is one of *non atrial fibrillation*, *paroxysmal atrial fibrillation*, or *persistent atrial fibrillation* (Wang et al., 2021). AF onset/offset events use the WFDB `'+'` symbol with auxiliary notes `'(AFIB'` (onset) and `'(N'` / `'(NORMAL'` / `'(NSR'` (return to sinus). For each 10-second window, we then assign a 3-class label combining the recording subtype with whether AF overlaps the window:

- Class 0 – *no_af*: any window from a non-AF recording, or a non-AF window inside a paroxysmal recording.
- Class 1 – *paroxysmal_af*: AF window inside a paroxysmal recording.
- Class 2 – *persistent_af*: AF window inside a persistent recording.

Windows are gap-aware: no window is allowed to span two disconnected signal segments. The CPSC-2021 rows in Table 2 reports Macro AUROC over these three classes.

VitalDB (ECG/PPG, binary AF). Rhythm labels come from the VitalDB Arrhythmia Database (Eun et al., 2026), which provides anesthesiologist-validated per-beat `rhythm_label` and `bad_signal_quality` annotations on 482 surgical cases. For each 10-second window, the dominant beat-level rhythm label is taken as the window label, mapped to integer class indices:

Index	Rhythm category
0	Normal sinus rhythm (N)
1	Atrial fibrillation or flutter (AFIB/AFL)
2	Supraventricular tachyarrhythmia (SVTA)
3	Ventricular tachycardia (VT)
4	Patterned ventricular ectopy
5	Patterned atrial ectopy
6	Sinus node dysfunction (SND)
7	Wandering atrial pacemaker / multifocal atrial tachycardia (WAP/MAT)
8	Atrioventricular block (AVB)

Beats labeled `Noise` or `Unclassifiable` (or whose flagged window overlaps a low-quality interval) are mapped to -1 and excluded from training and evaluation. The binary AF variant collapses the rhythm classes to AFIB/AFL ($=1$) versus all other valid rhythms ($=0$). The same labels are reused for VitalDB PPG, since ECG II and PLETH share the surgical timeline.

MC-MED (ECG/PPG, sepsis). The MC-MED sepsis cohort follows the CDC eSOFA (Adult Sepsis Event) surveillance criteria computed at visit (CSN) level from the public clinical tables. A visit is positive when (i) there is a *suspected infection* - a blood-culture order paired with at least one IV antibiotic order within ± 48 hours of each other and within 48 hours of ED arrival – and (ii) at least one *organ-dysfunction trigger* fires within the same window:

- Cardiovascular: any IV vasopressor (norepinephrine, dopamine, epinephrine, phenylephrine, vasopressin).
- Respiratory: initiation of mechanical / non-invasive positive-pressure ventilation.
- Renal: serum creatinine doubling from the visit minimum, or eGFR falling below half its visit maximum.
- Hepatic: total bilirubin > 2 mg/dL *and* more than $2\times$ the visit minimum.
- Hematologic: platelet count $< 100 \times 10^9$/L *and* less than $\frac{1}{2}$ of the visit maximum (with baseline > 100).
- Metabolic: lactate > 2 mmol/L.

The visit-level binary label is True if the cohort criteria are met; the trigger time (sepsis onset) is the time of the earliest organ-dysfunction event that satisfies the case definition. Detection is then framed *within* sepsis-positive visits as a pre-onset / post-onset classification: only eSOFA-positive visits are included in the MC-MED cohort, and each 10-second window is labeled 0 (pre-sepsis) if it ends strictly before the trigger time, or 1 (post-sepsis) otherwise. At forecasting horizons $H \in \{1, 5, 10, 20\}$ minutes the same trigger time defines a single $0 \to 1$ transition per visit, and a window is positive if the trigger falls within $[s_t, s_t + H]$.

CHB-MIT (EEG, seizure). CHB-MIT provides per-recording summary text files listing seizure intervals as "`Seizure` k `Start Time:` t_0 `seconds`" / "`Seizure` k `End Time:` t_1 `seconds`". We parse these intervals and assign a binary per-window label: 1 if the window overlaps any annotated seizure interval, 0 otherwise. Recordings that contain no seizures contribute only negative windows.

A.3 DATA SPLITS

All experiments use 60/20/20 train/validation/test splits at the patient level. No windows from the same identity are shared across splits. No patient contributes windows to more than one split, so temporal models are evaluated on unseen patients rather than on unseen windows of seen recordings.

Feature normalization statistics are estimated from the training split only and applied unchanged to validation and test data. All experiments are repeated over three random seeds (42, 43, 44). The random seeds determines both the patient split and the model initialization, and all methods use the same split for a given seed. Results are reported as mean and standard deviation over the three seeds.

B FEATURE AND EMBEDDING EXTRACTION

B.1 DOMAIN FEATURES

Domain features are organized by modality-specific physiological groups. Tables A.2 -A.5 provide the full grouped index used in our experiments.

Table A.2: **Domain feature atlas overview.** Features are grouped by physiological meaning and stored in a fixed column order for each modality.

Modality	Dim.	Main groups	Datasets
ECG	64	R-peak timing; wave timing; wave amplitude; HRV; SQI	VitalDB, MC-MED, CPSC-2021
PPG	71	Pulse morphology; width/amplitude/AUC; HRV; SQI	VitalDB, MC-MED
EEG	31×22	Time; frequency; time–frequency; complexity	CHB-MIT

Table A.3: **ECG feature atlas.** Feature names are grouped by physiological category.

Group	Index	Feature names
R-peak timing	0–7	a_R, RR0, RR1, RR2, RRm, RR_0_1, RR_2_1, RR_m_1
Wave timing	8–17	t_PR, t_QR, t_RS, t_RT, t_PQ, t_PS, t_PT, t_QS, t_QT, t_ST
Wave amplitude	18–27	a_PQ, a_QR, a_RS, a_ST, a_PS, a_PT, a_QS, a_QT, a_ST_QS, a_RS_QR
HRV time-domain	28–52	MeanNN, SDNN, SDANN1, SDNNI1, SDANN2, SDNNI2, SDANN5, SDNNI5, RMSSD, SDSD, CVNN, CVSD, MedianNN, MadNN, MCVNN, IQRNN, SDRMSSD, Prc20NN, Prc80NN, pNN50, pNN20, MinNN, MaxNN, HTI, TINN
HRV frequency-domain	53–62	ULF, VLF, LF, HF, VHF, TP, LFHF, LFn, HFn, LnHF
Signal quality	63	sqi

Table A.4: **PPG feature atlas.** PPG features include pulse morphology, width, amplitude, area, signal quality, and HRV descriptors.

Group	Index	Feature names
Timing morphology	0–6	Tpi, Tpp, Tsys, Tdia, Tsp, Tdp, deltaT
Systolic width	7–13	Tsw10, Tsw25, Tsw33, Tsw50, Tsw66, Tsw75, Tsw90
Diastolic width	14–20	Tdw10, Tdw25, Tdw33, Tdw50, Tdw66, Tdw75, Tdw90
Pulse width	21–27	Tpw10, Tpw25, Tpw33, Tpw50, Tpw66, Tpw75, Tpw90
Amplitude and area	28–34	Asp, Adn, Adp, Aoff, AUCpi, AUCsys, AUCdia
Signal quality	35	SQI
HRV descriptors	36–70	MeanNN, SDNN, SDANN1, SDNNI1, SDANN2, SDNNI2, SDANN5, SDNNI5, RMSSD, SDSD, CVNN, CVSD, MedianNN, MadNN, MCVNN, IQRNN, SDRMSSD, Prc20NN, Prc80NN, pNN50, pNN20, MinNN, MaxNN, HTI, TINN, ULF, VLF, LF, HF, VHF, TP, LFHF, LFn, HFn, LnHF

Table A.5: **EEG feature atlas.** The same 31 features are computed for each of the 22 EEG channels. For channel i, columns $31 \times i$ to $31 \times i + 30$ hold these features in this order.

Offset	Group	Feature names
0–6	Time domain	time_mean, time_variance, time_skewness, hjorth_activity, hjorth_mobility, hjorth_complexity, rms
7–17	Frequency domain	delta_power, theta_power, alpha_power, beta_power, gamma_power, theta_alpha_ratio, peak_frequency, spectral_centroid, spectral_bandwidth, spectral_edge, total_power
18–25	Time–frequency	wavelet_energy_d1, wavelet_energy_d2, wavelet_energy_d3, wavelet_energy_d4, wavelet_energy_d5, wavelet_energy_a5, wavelet_entropy, wavelet_energy_ratio
26–30	Complexity	sample_entropy, approx_entropy, perm_entropy, higuchi_fd, petrosian_fd

Table A.6: **Foundation-model embedding streams.** Each pretrained model produces one embedding per valid 10-second source window. Embeddings are aligned with the corresponding timestamps, labels, and domain-feature windows.

Modality	Foundation model	Embedding dim.	Datasets
ECG	CSFM	768	CPSC-2021, VitalDB, MC-MED
ECG	CLEF	1024	CPSC-2021, VitalDB, MC-MED
ECG	ECGFounder	1024	CPSC-2021, VitalDB, MC-MED
PPG	CSFM	768	VitalDB, MC-MED
PPG	PulsePPG	512	VitalDB, MC-MED
EEG	LaBraM	200	CHB-MIT
EEG	CBraMod	200	CHB-MIT

B.2 Foundation-model embeddings

In addition to hand-crafted domain features, we extract snapshot-level representations from pretrained biosignal foundation models. Model-specific resampling and reshaping are performed only at the foundation-model input stage, after windowing. Each embedding is attached to the same window, timestamp s_t, and label y_t as the corresponding domain-feature vector. The resulting embedding streams are summarized in Table A.6.

CSFM (Gu et al., 2026) is a cardiac multimodal foundation model pretrained on data from 1.7 M individuals, with separate ECG and PPG encoders that share a common representation space. We feed each 10-second window resampled to 250 Hz (2500 samples) and read the encoder's pre-classifier embedding (d=768). For single-lead ECG datasets, the input is treated as Lead II, following CSFM's configuration. The PPG encoder receives the 10-second window at 250Hz as a single channel.

ECGFounder (Li et al., 2025) is an ECG foundation model trained on more than 10 M recordings, available as 1-lead and 12-lead variants. We applied its 1-lead version to extract the encoder's pre-classifier embedding (d=1024).

CLEF (Shu et al., 2025) is a clinically-guided contrastive ECG foundation model. The released model weights is compatible with single-lead input, so we apply CLEF to Lead II only. Each 10-second window is resampled to 500 Hz (5000 samples) and passed through the encoder, yielding a per-window embedding of dimension 1024.

PulsePPG (Saha et al., 2025) is an open-source PPG foundation model trained on field-collected wearable data. Inputs are single-channel PPG at 50 Hz (500 samples per 10-second window). The model uses internal temporal pooling, so we read its global pooled embedding (d=512) as the snapshot-level representation.

LaBraM (Jiang et al., 2024) is an EEG foundation model pretrained with vector-quantized masked-patch reconstruction over a large multi-source EEG corpus. Each 10-second EEG window is resampled to 200 Hz and reshaped into patches of length 200 samples (1 s), giving an input tensor of shape $(C, P, 200)$ per window. CHB-MIT bipolar derivations are mapped to LaBraM's 128-channel positional vocabulary by aligning the first pole of each bipolar pair to the closest pretraining channel; the same channel mapping is used for every window. We use its 200-dimensional pre-classifier embedding.

CBraMod (Wang et al., 2025) is an EEG foundation model that interleaves spatial and temporal attention. Each 10-second window is resampled to 200 Hz and split into exactly ten 200-sample temporal patches per channel, matching the input format of the released model; we reuse LaBraM's 128-channel positional mapping for CHB-MIT bipolar derivations. Per-window embeddings are 200-dimensional.

C BioLT Architecture Details

C.1 Decoder backbone

BioLT uses a 4-layer decoder-only Transformer with hidden dimension 128, 4 attention heads, feed-forward dimension 512, GELU activations, and dropout rate 0.1. Each input token is the concatenation of the snapshot representation and elapsed-time channel, $\mathbf{x}_t = [\mathbf{z}_t; \Delta s_t]$, projected to

the model dimension by a linear layer. The decoder uses causal self-attention, so each hidden state $\mathbf{h}_t$ only depends on current and past windows.

For long recordings, attention is computed over the current chunk together with the compressed memory bank M^{cmp}. Specifically, each layer attends to the concatenation of memory slots and current-chunk tokens, $[M^{\mathrm{cmp}}; \mathrm{KV}_{\mathrm{chunk}}]$, while preserving causal masking within the current chunk. We use ALiBi positional biases (Press et al., 2022), which is able to generalize to sequences longer than those seen during training. The resulting hidden states are used for both autoregressive pretraining and downstream dense prediction. During pretraining, a lightweight linear head predicts the next-window representation change from each hidden state.

C.2 Discontinuity-aware elapsed-time encoding

Clipping. Following Section 3.2, we append the elapsed time Δs_t to each token, clipped at τ=300 s:

$$\Delta s_t \;=\; \min\big(\max(s_t - s_{t-1},\, 0),\, \tau\big), \qquad \Delta s_1 = 0.$$

We clip (Δs_t) to limit the influence of very long gaps while retaining the relative duration of shorter gaps. This threshold covers 96.03% of observed discontinuities; the remaining 3.97% (3,486 gaps) are truncated. Per-dataset gap distributions are reported in Table A.7.

Table A.7: **Inter-window gap distribution per dataset.** Steps between adjacent 10-s windows are not counted. "≤300 s" is the fraction of gaps within the clip τ; ">1 h" counts gaps longer than 1 hour.

Corpus	Gaps	Median	p90	p95	p99	Max	≤300 s	>1 h
CPSC-2021 ECG	647	20 s	80 s	130 s	265 s	420 s	99.85%	0
VitalDB ECG	9,386	20 s	60 s	99 s	290 s	15,876 s	99.07%	4
VitalDB PPG	389	20 s	41 s	52 s	1,005 s	6,461 s	98.46%	1
CHB-MIT EEG	0	—	—	—	—	—	—	0
MC-MED ECG	21,815	20 s	157 s	956 s	3,772 s	29,945 s	92.45%	235
MC-MED PPG	55,511	28 s	70 s	138 s	1,871 s	23,725 s	96.86%	190
All datasets	**87,748**	**27 s**	**81 s**	**190 s**	**2,375 s**	**29,945 s**	**96.03%**	**430**

Normalization and concatenation. Similar to other features in $\mathbf{z}_t$, the clipped Δs_t is z-scored using statistics from valid training windows and then concatenated with the snapshot representation. The same normalization parameters are used for validation, testing, and all downstream tasks within each dataset-modality pair.

C.3 Compressed streaming memory

Memory write. When chunk k finishes, `BioLT` compresses its hidden states $\mathbf{H}_k \in \mathbb{R}^{C\times d_{\mathrm{model}}}$ into K memory slots with a single cross-attention layer whose K queries $\mathbf{Q}_{\mathrm{W}}$ are learned:

$$\mathbf{c}_{k,1:K} = \mathrm{softmax}\left(\frac{\mathbf{Q}_{\mathrm{W}}\mathbf{H}_k^{\top}}{\sqrt{d_{\mathrm{model}}}}\right)\mathbf{H}_k\, W_{\mathrm{W}},$$

where $W_{\mathrm{W}} \in \mathbb{R}^{d_{\mathrm{model}}\times d_{\mathrm{model}}}$ is an output projection applied to each slot. We use chunk size $C = 180$ and $K = 16$ slots per chunk. The K slots are ordered in time, and please refer to Appendix C.4 on how they are learned and stored.

Memory bank update. The new slots are appended to the compressed memory bank M^{cmp}, which contains at most 32 slots. Once the bank is full, the oldest slots are removed first (FIFO). Since $\mathbf{H}_k$ is computed with access to the existing memory bank, newly written slots can also encode information from earlier chunks. In terms of positional embeddings of the memory bank, we apply cyclic learnable embeddings composed of 256 learned vectors.

Attention over memory. For each new chunk, every decoder layer attends to the current chunk and the compressed memory bank M^{cmp}. Current windows are kept at full resolution, while earlier context is represented by the memory slots. The memory and computation cost of inference is therefore fixed per chunk.

C.4 MEMORY RECONSTRUCTION TARGET

Cumulative target. To guide the compressed memory bank to preserve useful trajectory information, we supervise each memory slot with cumulative summary statistics of the past. For slot j in chunk k, the target is

$$\mathbf{q}_{k,j} = [\mu, \sigma, b, r],$$

where μ, σ, b, and r denote the feature-wise mean, standard deviation, linear trend, and dynamic range of all valid hidden states observed from the start of the recording up to that slot. Chunk k is divided into K equal contiguous sub-blocks of C/K windows, and the target of slot j is computed from all valid hidden states from the start of the recording up to the end of sub-block j, each slot therefore summarizes a slightly longer history than the previous one. The linear trend b is the feature-wise least-squares slope of the hidden states against the index i of valid windows counted from the start of the recording, scaled by the number of windows so that it equals the fitted change over these windows:

$$b = (n-1)\,\frac{\sum_i (i-\bar{i})(\mathbf{h}_i - \bar{\mathbf{h}})}{\sum_i (i-\bar{i})^2},$$

where n is the number of valid windows summarized and $\bar{i}$, $\bar{\mathbf{h}}$ are their means. These targets are computed only from past information and detached from the gradient.

Reconstruction head and loss. A lightweight reconstruction head predicts the cumulative target from each compressed slot. The memory loss is

$$\mathcal{L}_{\text{mem}} = \frac{1}{|\mathcal{S}|} \sum_{(k,j)\in\mathcal{S}} \text{SmoothL1}\left(\widehat{\mathbf{q}}_{k,j}, \mathbf{q}_{k,j}\right),$$

where $\mathcal{S}$ denotes the valid compressed slots. The full pretraining objective is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{AR}} + \lambda_{\text{mem}} \mathcal{L}_{\text{mem}},$$

with $\lambda_{\text{mem}} = 0.005$.

C.5 TRAINING VERSUS STREAMING INFERENCE

Pretraining. During pretraining, each recording is split into chunks of $C = 180$ windows. `BioLT` is optimized with the pretraining objective $\mathcal{L}_{\text{pre}}$, which combines autoregressive next-window prediction and memory reconstruction. To keep trainable memory fixed regardless of recording length, we use truncated backpropagation through time: memory slots are written at chunk boundaries, appended to M^{cmp}, and detached before processing subsequent chunks.

Downstream fine-tuning. After pretraining, `BioLT` is fine-tuned for dense prediction, with the supervised task loss applied at every valid window. Multiclass tasks use cross-entropy, while binary tasks use the same binary classification loss as the main experiments. During fine-tuning, the compressed memory bank is still updated sequentially across chunks, and it is detached from the computation graph after every $S = 2$ consecutive chunks (Table A.8). This allows each prediction to condition on long-range history while limiting backpropagation to the current TBPTT segment of two chunks.

Streaming inference. At inference time, each recording is processed sequentially in chunks. After each chunk, the memory writer updates M^{cmp}, and FIFO eviction keeps the memory bank at a fixed capacity of 32 slots. Predictions are emitted at every valid window, so the output trajectory has the same temporal resolution as the input stream. Because the memory bank keeps updating with fixed length, `BioLT` can be applied to arbitrarily long recordings with fixed per-chunk computation and memory cost.

D TRAINING AND EVALUATION PROTOCOL

D.1 PRETRAINING OBJECTIVE

Pretraining optimization uses AdamW with learning rate 10^{-4}, weight decay 10^{-4}, gradient clipping at 0.5, batch size 64, for up to 100 epochs. The validation pretraining loss drives a

`ReduceLROnPlateau` schedule (factor 0.5, patience 5, floor 10^{-8}) and an early-stopping criterion (patience 10). We save the model weights at the lowest validation $\mathcal{L}_{\text{pre}}$ and use them as the initialization for downstream fine-tuning.

D.2 Downstream fine-tuning

For downstream tasks, the pretrained `BioLT` backbone is fine-tuned with a per-window MLP head $\text{Linear}(d_{\text{model}}, 64) \to \text{ReLU} \to \text{Dropout}(0.3) \to \text{Linear}(64, O)$, where O matches the number of classes of each task. Supervision is dense: every valid window contributes a gradient, with padding and partially-covered forecasting positions masked out. We use class-weighted cross-entropy for multiclass tasks (CPSC-2021 3-class AF), the binary focal loss (Lin et al., 2017) (γ=2.0, per-dataset α) for binary tasks; class weights and α are derived from training-split label counts and held fixed across seeds.

We fine-tune with AdamW using learning rate 10^{-4}, weight decay 10^{-4}, gradient clipping at 0.5, batch size 32, and at most 100 epochs. Model selection is based on validation AUROC, and final test metrics are reported from the model weights selected on validation.

D.3 Sequence length, chunk size, and truncation

`BioLT` is pretrained on whole recordings and fine-tuned with the task loss on at most 500 windows. Inference is performed over the full recording. Each sequence is processed in causal chunks of $C = 180$ windows. After each chunk, the memory writer produces $K = 16$ compressed slots, which are appended to M^{cmp}. The memory bank is capped at 32 slots with FIFO eviction.

During training, older memory states are detached from the computation graph, so the model carries historical information forward without backpropagating through the entire recording. During inference, chunks are processed sequentially, allowing `BioLT` to operate beyond the training context with fixed per-chunk memory and computation.

D.4 Hyperparameters

Table A.8 presents the hyperparameters used throughout the experiments; values are identical across datasets, horizons, and seeds unless noted.

D.5 Compute resources

All experiments were run on 2 NVIDIA GeForce RTX 4090 GPUs with PyTorch `2.11.0` (CUDA `13.0`) and Python `3.10`.

E Baseline Details

All baselines use the same input as `BioLT`: a causal sequence of per-window snapshot representations $\{\mathbf{z}_t\}_{t=1}^{T}$, where $\mathbf{z}_t \in \mathbb{R}^D$ is either a domain feature vector or a foundation-model embedding. All methods produce logits at every valid window and are trained with the same dense per-window supervision, patient-level splits, and optimization protocol as `BioLT`. The baselines differ only in how they aggregate the history $\mathbf{z}_{\leq t}$ before prediction.

E.1 Snapshot and pooling baselines

Snapshot. The Snapshot baseline classifies each window independently. The context vector at time t is simply

$$\mathbf{u}_t = \mathbf{z}_t.$$

The same MLP prediction head is applied to each timestep. This baseline has no access to temporal history.

Table A.8: **Hyperparameter summary of `BioLT`.**

Group	Hyperparameter	Value
Architecture	Decoder layers L	4
	Hidden dim d_{model}	128
	Heads H (per-head dim d_{head})	4 (32)
	FFN dim d_{ff}	512
	Dropout (attn / resid / FFN)	0.1 / 0.1 / 0.1
	Norm / activation / position bias	Pre-LN / GELU / ALiBi
Memory	Chunk size C (windows)	180
	Slots per chunk K	16
	Bank capacity (slots)	32
	Reconstruction weight λ_{mem}	0.005
Input encoding	Δs_t clip τ	300 s
	Δs_t normalization	z-score (train-only)
Pretraining	Optimizer	AdamW
	Learning rate / weight decay	10^{-4} / 10^{-4}
	Gradient clip / batch size / epochs	0.5 / 64 / 100
	TBPTT segments S	1
	Context window (tokens)	full recording
	LR schedule (monitor / factor / patience)	ReduceLROnPlateau (val $\mathcal{L}_{\text{pre}}$ / 0.5 / 5)
Fine-tuning	Optimizer	AdamW
	Learning rate / weight decay	10^{-4} / 10^{-4}
	Gradient clip / batch size / max epochs	0.5 / 32 / 100
	TBPTT segments S	2
	Context window (tokens)	500
	Early stopping (monitor / patience)	val Macro-AUROC / 10
Evaluation	Patient split (per dataset, per seed)	stratified 60/20/20
	Seeds	42, 43, 44

MeanPool / MaxPool. MeanPool and MaxPool are non-parametric causal aggregation baselines. For each valid window t, they compute the running mean or element-wise maximum over all previous valid windows:

$$\mathbf{u}_t^{\text{mean}} = \frac{1}{|\mathcal{V}_t|} \sum_{s \in \mathcal{V}_t} \mathbf{z}_s, \qquad \mathbf{u}_t^{\max} = \max_{s \in \mathcal{V}_t} \mathbf{z}_s,$$

where $\mathcal{V}_t \subseteq \{1, \ldots, t\}$ excludes padded or invalid positions. The aggregated vector is then passed to the shared MLP prediction head.

E.2 Sequence model baselines

LSTM. The LSTM baseline uses a unidirectional LSTM with hidden dimension 128. The snapshot sequence is processed causally, and the hidden state at each timestep is used as the context vector for per-window prediction.

TCN. The TCN baseline uses a causal temporal convolutional network with hidden dimension 128. Convolutions use left-only padding so that the output at time t depends only on $\mathbf{z}_{\leq t}$. The per-timestep output is used as the context vector for prediction.

Attention. The Attention baseline first projects each snapshot representation to a 128-dimensional working space and then applies a single causal self-attention layer. A lower-triangular attention mask prevents access to future windows. The attention output at each timestep is passed to the prediction head.

Mamba-3. Mamba-3 (Lahoti et al., 2026) is implemented as a causal state-space sequence model with hidden dimension 128. It processes the snapshot stream in order, and the per-step output is used as the context vector for dense prediction.

Transformer-XL. Transformer-XL (Dai et al., 2019) uses relative-position attention and a FIFO memory that stores the hidden states of the previous 180 windows without compression. We used the same chunk size, input tokens, autoregressive pretraining, and fine-tuning pipelines following BioLT for a fair comparison. The main difference between the two models is uncompressed versus compressed cross-chunk memory.

Common training protocol. All baselines are trained on fixed-length 500-window recordings and evaluated on full test recordings. Baselines without a cross-chunk memory see only the 500-window recording during training, whereas Transformer-XL and BioLT have their memory initialized from the preceding chunks without gradients. They use the same patient-level $60/20/20$ splits, three random seeds, train-split feature normalization, validation-based model selection, and downstream optimization settings as BioLT, so that performance differences mainly reflect the temporal aggregation mechanism.

F Additional Results

F.1 Chunk size and TBPTT hyperparameter

We investigate the chunked-attention chunk size (C) and the number of truncated-BPTT segments (S) retained for gradient flow. Test AUROC (%) and F1 (%) are reported as mean±std over 3 seeds.

Table A.9: **Effect of chunk size and TBPTT segments on BioLT; CPSC-2021 with domain features at forecasting horizon H=10 min.** AUROC is Macro AUROC for the 3-class task.

Chunk	TBPTT	AUROC (%)	F1 (%)
45	1	84.33±2.40	59.61±5.05
45	2	84.44±2.88	59.49±4.80
45	4	83.92±2.32	59.01±5.93
60	1	84.61±2.88	60.11±3.52
60	2	84.54±2.27	59.16±4.44
60	4	84.37±1.55	59.79±5.00
90	1	83.18±2.00	59.98±3.81
90	2	81.96±2.30	60.16±4.36
90	4	82.82±2.32	60.05±4.87
120	1	83.73±1.39	61.27±2.08
120	2	84.56±1.30	60.08±4.22
120	4	84.59±1.67	59.75±4.43
180	1	83.95±2.45	**63.33±0.33**
180	2	**84.66±1.17**	62.51±1.46
180	4	83.88±2.17	61.93±1.98

F.2 Memory capacity and slots per chunk

We investigate the compressed-memory bank capacity $|M^{\mathrm{cmp}}|$ and the number of slots K written per completed chunk on CPSC-2021 at forecasting horizon H=10 min, holding the chunk size fixed at the default C=180 (the chunk-size axis is studied separately in Table A.9). The grid is $|M^{\mathrm{cmp}}| \in \{16, 32, 64\}$ slots and $K \in \{4, 8, 16\}$, with each setting averaged over 3 seeds (42,43,44). The default BioLT configuration ($|M^{\mathrm{cmp}}|$=32, K=16) attains the best Macro AUROC (86.93±1.36 %).

F.3 Extended memory-capacity analysis

Table A.11 extends the capacity of the memory banks to up to 2,048 slots, enough to cover 64 h before any slots are explicitly evicted. This is longer than the longest recording in our datasets (16.4 h on VitalDB ECG). A 64-fold increase in capacity changes AUROC by $+0.50$ (ECG) and $+0.26$ (PPG) points, both inside the seed std. The performance clearly saturates beyond 512. Thus a bank capacity of 32 is enough for these tasks. Larger banks do not improve accuracy beyond seed-to-seed variation. A likely reason is the recurrent update of the bank (Appendix C.3): each new slot is written from hidden states that have already attended to the bank, so information older than the bank's nominal span can persist in later slots.

Table A.10: **Effect of compressed-memory capacity $|M^{\mathrm{cmp}}|$ and slots-per-chunk K on `BioLT`; CPSC-2021 with domain features at forecasting horizon H=10 min, chunk size fixed at C=180.** AUROC is Macro AUROC for the 3-class task. Mean±std over 3 seeds (42/43/44). Default configuration ($|M^{\mathrm{cmp}}|$=32, K=16) is highlighted; best per column in bold.

$\lvert M^{\mathrm{cmp}}\rvert$	K	**AUROC (%)**
16	4	83.26±2.42
16	8	84.39±2.25
16	16	85.60±1.26
32	4	83.57±2.56
32	8	84.28±2.19
32	16	**86.93±1.36**
64	4	83.57±2.30
64	8	84.34±2.07
64	16	86.21±0.93

Table A.11: **Memory-bank capacity analysis.** VitalDB ECG and PPG binary AF forecasting at H=5 min with domain features; test AUROC (%, mean±std over 3 seeds). "History span" is the length of recording covered by the memory bank's slots at 16 slots per 30-minute chunk.

Bank capacity	**History span**	**VitalDB ECG**	**VitalDB PPG**
32 (default)	1 h	92.49±1.77	90.26±0.27
128	4 h	93.26±1.20	90.50±0.62
512	16 h	92.99±0.85	90.52±0.56
1024	32 h	93.02±0.86	90.52±0.56
2048	64 h	92.99±0.84	90.52±0.57

F.4 Validation of the pretraining objective

The autoregressive pretraining objective predicts the change $\Delta\mathbf{z}_{t+1} = \mathbf{z}_{t+1} - \mathbf{z}_t$ of the frozen window embedding under mean-squared error.

To verify that the trained predictor captures real dynamics, we compare its validation MSE against three predictors that does not require learning: *always-zero* (copy the previous window), the *mean predictor* (predict the empirical mean of the target), and *persistence* (predict the previous change $\Delta\mathbf{z}_t$). Table A.12 shows the trained model beats always-zero and the mean predictor in all dataset combinations, by 12.7-36.9%.

We additionally pretrain a contrastive predictive coding (CPC) (van den Oord et al., 2018) variant under matched hyperparameters (K=12 prediction steps, cross-recording negatives). CPC trains properly rather than collapsing (validation InfoNCE 1.51-3.12 against a chance floor of $\log 65 = 4.17$) but loses all 6 matched downstream comparisons, by a mean of 2.69 AUROC points (Table A.13).

Table A.12: **Trivial-predictor controls for the autoregressive objective.** Validation MSE on the $\Delta\mathbf{z}$ target (lower is better); mean over 3 seeds.

Predictor	**CPSC-2021**	**MC-MED ECG**	**VitalDB ECG AF**
Trained AR (ours)	**0.6633**	**0.5572**	**0.4211**
Always-zero (copy previous window)	0.7979	0.7019	0.5583
Mean predictor	0.7979	0.7019	0.5583
Persistence	2.3765	2.0838	1.6396

F.5 Statistical significance

Table A.14 reports a two-sided paired Wilcoxon signed-rank test of `BioLT` against each baseline at every horizon. The unit of analysis is the per-dataset three-seed mean AUROC over the 21 dataset-feature settings of Table 2 (n=21 pairs per test). `BioLT` is significantly ahead of the six conventional aggregation baselines at every horizon and of Mamba-3 at every horizon except H=1. The same test against Transformer-XL yields significance at H=0, H=5, H=10, and H=20.

Table A.13: **Pretraining objective: autoregressive (AR) against CPC.** Downstream test AUROC (%, mean±std over 3 seeds) with domain features.

Objective	Horizon	CPSC-2021	MC-MED ECG	VitalDB ECG AF
AR (ours)	H=0	**95.75±3.30**	**79.94±1.16**	**94.82±1.57**
CPC	H=0	93.71±4.87	77.06±2.17	94.14±1.23
AR (ours)	H=20	**87.66±1.71**	**62.87±1.47**	**91.06±2.13**
CPC	H=20	81.96±4.60	59.94±0.04	89.14±3.26

Table A.14: **Paired Wilcoxon signed-rank test of `BioLT` against each baseline.** The median AUROC difference (`BioLT` − baseline, points) across the 21 dataset-feature settings is reported; * $p<0.05$, ** $p<0.01$.

Baseline	H=0	H=1	H=5	H=10	H=20
Snapshot	+4.02**	+2.43**	+2.53**	+2.64**	+4.99**
MeanPool	+9.58**	+3.46**	+3.53**	+3.39**	+5.86**
MaxPool	+8.67**	+2.95**	+3.47**	+3.06**	+2.86**
LSTM	+9.89**	+6.11**	+0.59**	+4.61**	+3.91**
TCN	+8.54**	+2.60**	+1.34**	+1.83**	+2.37**
Attention	+14.89**	+6.74**	+5.17**	+5.95**	+5.23**
Mamba-3	+2.54**	+0.53	+1.15*	+0.91*	+2.22**
Transformer-XL	+0.62**	+0.27	+0.39*	+1.52**	+1.76**

F.6 Computational cost

Table A.15 benchmarks `BioLT` streaming inference on a single RTX 4090 at batch size 1. The memory bank is kept between chunks and each computational cost measurement only covers the processing of one new chunk. Per-update cost is constant in recording length: across a 64-fold increase in history (0.5 h to 32 h), latency varies by under 2%, and each update allocates about 3.6 MB in addition to the memory bank. `BioLT` has 1,280,515 parameters; for comparison, Snapshot has 16.8 K, Mamba-3 294 K, and Transformer-XL 874 K.

Table A.15: **Streaming inference cost of `BioLT`** (RTX 4090, batch size 1, median over repeated runs). Cost is constant in recording length.

History	Latency	Throughput	Incremental GPU	FLOPs/update
0.5 h (180 windows)	36.6 ms	4,913 win/s	3.47 MB	386 M
2 h (720 windows)	37.3 ms	4,822 win/s	3.62 MB	398 M
8 h (2,880 windows)	36.8 ms	4,895 win/s	3.62 MB	398 M
32 h (11,520 windows)	37.1 ms	4,856 win/s	3.62 MB	398 M

F.7 Qualitative examples of the alarm-based evaluation

Figures A.1 and A.2 show the alarm probability of different domain feature-based methods on two CHB-MIT test patients. Each panel shows one model's predicted probability p_t over time, its threshold θ chosen on the validation split at no more than one false alarm per hour. The ground truth seizures are shown as vertical yellow shades. Every predicted alarm that exceeded threshold θ has a tick at the bottom, green when it falls inside a seizure, blue when it precedes a seizure onset by at most 30 minutes, and red for false alarms. An event counts as detected when at least one alarm falls inside it.

For the first patient (Figure A.1, a six-hour window with ten seizures), `BioLT` detects eight of the ten seizures, whereas the baselines either remain below their thresholds (Snapshot, MeanPool, TCN) or exceed them in long stretches away from the seizures (LSTM, Transformer-XL), which produces alarms in the windows preceding onsets but none inside the short seizures.

For the second patient (Figure A.2, a four-hour window with four seizures), `BioLT` detects three of the four seizures with six false alarms, and Transformer-XL detects two events with three false alarms. The aggregate results over all binary settings are in Figure 6.

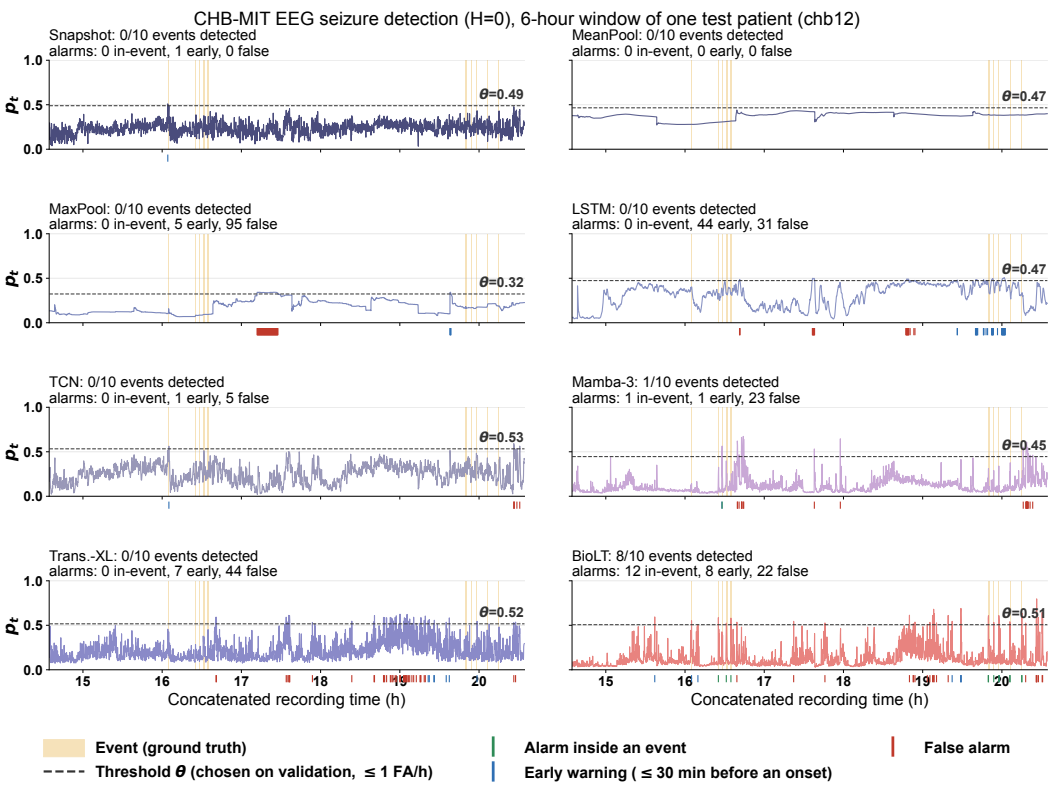


Figure A.1: **Alarm-based evaluation on one CHB-MIT test patient (chb12; six-hour window, ten seizures).** The patient's recordings are concatenated in order, the memory bank is reset at each file boundary. Each panel: predicted probability p_t of one model, its validation-chosen threshold θ (dashed), seizures (shaded), and alarms as ticks (green: inside a seizure; blue: within 30 minutes before an onset; red: false alarm). Panel titles give the number of seizures detected and alarm counts.

G LIMITATIONS

`BioLT` has several limitations. First, the evaluation is retrospective, and event timestamps come from annotations that may not match exact adverse event onset. For instance, in MC-MED, sepsis onset is defined by the first organ-dysfunction trigger under a case definition that also requires suspected infection, and may therefore occur later than the underlying deterioration onset; the cohort also includes only sepsis-positive visits.

Second, the datasets differ in annotation granularity, recording quality, and clinical setting, limiting direct comparison across tasks.

Another limitation, related to deployment, is the alarm sensitivity-false alarm trade-off. At a budget of one false alarm per hour, all evaluated methods, including `BioLT`, detect at most 31% of events. This level of sensitivity is unlikely to be sufficient for standalone clinical alerting, where missed events may be clinically consequential. The current results should therefore be interpreted as evidence for improved event detection under a fixed alarm burden, rather than readiness for autonomous deployment.

Future work should evaluate calibration, clinical utility, and prospective deployment in continuous monitoring workflows.

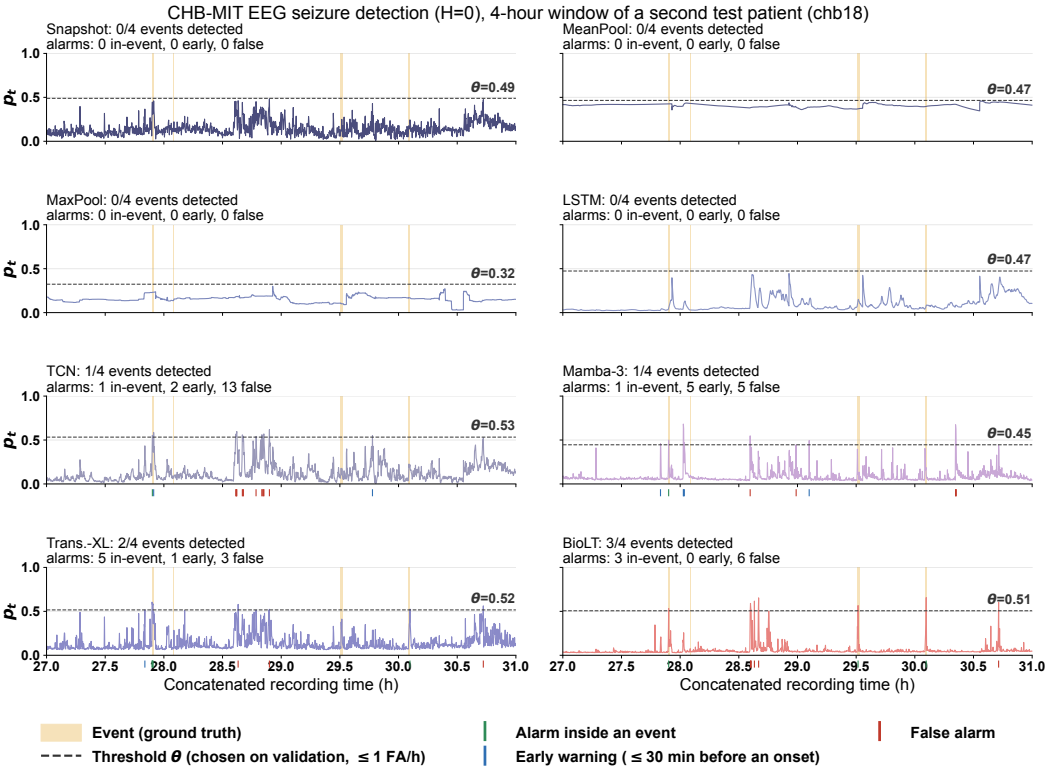


Figure A.2: **Alarm-based evaluation on a second CHB-MIT test patient (chb18; four-hour window, four seizures).** Same layout as Figure A.1.

Table A.16: **Forecasting results at horizon $H=1$ minute.** Test AUROC (%, mean±std over 3 seeds). **Best** per row highlighted in **bold**, second best per row underlined.

Datasets	Features	Snapshot	MeanPool	MaxPool	LSTM	TCN	Attention	Mamba-3	Trans.-XL	BioLT
CHB-MIT EEG seizure	Domain	67.27±8.12	59.36±8.52	54.24±8.77	64.88±7.56	63.81±8.91	60.13±10.07	66.47±4.36	68.08±10.79	**69.03±10.85**
	CBraMod	61.73±6.80	56.82±13.18	52.85±6.71	58.73±12.71	55.32±8.93	59.08±6.41	67.67±11.03	65.36±15.15	**68.51±15.14**
	LaBraM	60.16±2.96	44.28±4.20	58.58±7.03	47.17±11.36	59.81±6.99	53.08±2.00	56.76±6.44	68.41±3.96	**68.68±6.64**
CPSC-2021 ECG AF	Domain	84.75±0.70	83.63±1.06	84.86±2.44	82.04±2.40	82.26±1.52	85.98±3.35	83.25±1.03	**93.39±1.37**	93.03±0.30
	CSFM	84.64±2.13	85.72±4.32	88.84±6.46	85.04±1.28	86.97±4.66	84.02±2.73	85.36±2.88	90.80±1.50	**91.64±1.38**
	CLEF	80.17±3.58	**82.86±3.39**	79.91±7.41	73.10±4.99	78.54±5.55	82.27±6.36	78.90±7.44	81.38±9.94	80.48±9.18
	ECGFounder	85.08±4.03	87.19±4.59	89.16±6.25	82.53±7.12	87.31±5.37	83.06±7.67	83.32±5.53	88.89±3.48	**89.91±4.98**
VitalDB ECG AF	Domain	92.11±0.90	87.96±3.15	84.04±2.11	93.44±0.48	91.31±1.52	84.54±8.84	94.01±0.66	93.06±1.03	**94.16±1.14**
	CSFM	93.76±0.30	91.06±1.56	91.24±1.65	93.73±0.31	93.46±0.59	90.43±2.64	93.62±0.78	93.42±0.37	**94.14±0.27**
	CLEF	93.33±0.32	90.53±0.58	90.63±0.77	92.61±0.57	93.08±0.50	87.10±3.95	92.67±1.34	92.60±0.66	**93.58±0.37**
	ECGFounder	93.29±0.39	91.41±1.79	90.88±0.87	**93.93±0.25**	93.54±0.65	88.21±1.55	93.70±1.13	92.73±2.16	92.69±0.88
VitalDB PPG AF	Domain	88.70±0.38	88.34±0.43	79.20±2.21	90.72±0.38	90.32±0.57	84.03±3.18	**91.89±0.47**	90.55±0.54	90.89±0.33
	CSFM	89.73±0.42	88.25±0.75	90.51±1.20	90.85±0.44	90.42±0.51	87.27±1.84	89.83±0.44	**90.99±0.25**	90.96±0.11
	PulsePPG	86.14±0.36	85.10±0.58	56.07±8.32	87.07±1.41	88.55±0.45	73.88±4.88	88.55±0.60	**89.13±0.63**	88.57±0.43
MC-MED ECG sepsis	Domain	50.86±0.41	51.21±1.08	49.86±0.88	51.49±2.62	52.58±3.91	50.88±1.36	53.67±2.62	57.57±1.38	**57.62±0.30**
	CSFM	50.71±1.28	51.25±1.79	52.82±0.82	48.85±1.75	53.80±0.27	49.77±2.11	51.83±1.50	**54.09±1.11**	51.64±2.91
	CLEF	50.56±1.35	51.38±2.36	52.49±3.12	46.63±0.32	54.41±2.06	48.95±0.68	**56.18±2.57**	55.38±2.99	54.16±4.43
	ECGFounder	50.03±1.39	49.97±0.91	51.49±2.89	47.49±1.62	53.34±1.15	50.64±1.42	51.34±1.46	**53.42±2.72**	51.85±2.83
MC-MED PPG sepsis	Domain	50.95±0.09	52.81±0.84	51.80±1.25	48.62±1.82	53.51±1.68	52.02±1.31	61.03±0.84	**62.56±0.42**	60.79±1.60
	CSFM	50.87±1.72	51.32±2.49	54.08±2.87	49.95±0.90	53.34±0.99	49.52±0.99	**57.56±1.36**	53.30±1.62	56.16±1.02
	PulsePPG	51.14±1.01	51.97±1.54	54.04±1.20	49.90±0.38	52.48±1.62	50.67±1.14	**60.14±0.26**	51.62±1.05	57.28±1.37

Table A.17: **Forecasting results at horizon H=5 minutes.** Test AUROC (%, mean±std over 3 seeds). **Best** per row highlighted in **bold**, second best per row underlined.

Datasets	Features	Snapshot	MeanPool	MaxPool	LSTM	TCN	Attention	Mamba-3	Trans.-XL	BioLT
CHB-MIT EEG seizure	Domain	63.33±9.41	60.41±6.32	54.84±6.30	**66.07±3.45**	60.50±8.52	61.47±10.86	58.11±7.02	62.54±5.91	60.35±11.86
	CBraMod	58.66±9.53	56.95±13.90	50.67±8.73	60.81±12.12	56.04±9.03	59.98±9.56	**62.82±11.49**	53.22±8.73	60.48±13.07
	LaBraM	51.66±4.97	43.63±4.75	**57.38±9.77**	45.79±5.21	55.14±10.92	45.98±7.30	49.28±4.57	43.99±7.35	46.30±4.36
CPSC-2021 ECG AF	Domain	77.98±3.53	81.62±4.61	84.68±2.21	78.46±1.64	80.21±2.72	83.90±4.78	78.03±4.04	**90.35±1.69**	88.94±0.41
	CSFM	82.02±1.47	85.40±3.81	85.73±6.09	83.51±1.54	85.07±2.58	82.02±2.33	82.59±5.04	85.40±4.32	**89.20±1.78**
	CLEF	77.59±3.57	80.97±3.57	79.20±5.00	72.55±5.60	77.58±1.60	**81.70±5.05**	76.92±2.88	79.39±10.87	78.92±10.18
	ECGFounder	82.05±3.11	85.67±4.56	**87.03±6.50**	86.41±4.56	86.41±5.72	81.06±3.56	78.73±3.79	86.00±6.77	86.11±6.11
VitalDB ECG AF	Domain	90.01±0.74	86.73±3.52	83.12±2.93	91.81±0.50	89.61±2.58	81.65±7.60	**92.40±0.09**	91.09±0.82	91.96±1.01
	CSFM	91.64±0.78	90.39±1.35	89.89±0.84	91.86±0.98	91.58±0.71	90.14±2.62	91.73±1.09	91.82±0.82	**92.21±0.76**
	CLEF	91.22±1.11	88.81±0.49	88.94±1.82	91.44±1.09	91.25±1.40	85.98±3.51	91.09±1.30	90.73±1.39	**91.70±1.03**
	ECGFounder	91.36±1.06	88.61±3.70	89.98±1.07	91.54±1.11	91.00±1.87	87.41±1.82	**91.83±0.59**	91.05±2.70	90.78±2.08
VitalDB PPG AF	Domain	88.00±0.24	87.93±0.51	79.29±2.16	90.03±0.46	89.74±0.52	85.18±3.24	**91.26±0.51**	89.97±0.54	90.35±0.37
	CSFM	89.44±0.21	88.14±0.28	90.16±0.92	90.40±0.36	89.97±0.44	87.10±0.50	89.60±0.17	90.48±0.48	**90.76±0.21**
	PulsePPG	85.64±0.39	84.22±0.38	57.44±2.33	87.58±1.66	**88.63±0.28**	77.59±8.39	88.40±0.39	88.21±0.52	88.17±0.67
MC-MED ECG sepsis	Domain	50.94±0.82	53.00±1.97	53.32±1.50	51.41±1.70	53.17±2.93	51.25±1.77	56.25±1.35	59.61±1.78	**59.69±2.19**
	CSFM	51.36±1.55	52.17±1.39	53.59±0.59	48.64±1.13	53.39±0.90	50.72±3.10	51.37±1.76	**56.62±0.47**	55.30±0.91
	CLEF	52.67±2.00	53.28±2.18	54.56±3.25	51.70±3.17	56.45±1.42	52.35±1.98	56.73±0.75	55.21±2.06	**57.76±1.22**
	ECGFounder	51.05±2.35	51.41±1.56	51.14±0.36	49.11±1.58	52.82±0.54	50.58±2.33	50.24±1.65	54.70±1.05	**55.89±2.00**
MC-MED PPG sepsis	Domain	50.78±0.74	53.14±1.17	53.22±0.53	50.66±2.91	53.64±0.34	51.90±1.90	61.15±0.18	60.81±3.52	**62.72±1.19**
	CSFM	51.14±1.75	51.44±2.45	55.73±1.07	50.15±1.08	53.73±1.69	50.29±1.00	57.44±1.97	54.48±2.40	**58.90±0.43**
	PulsePPG	50.53±0.79	52.22±1.49	54.94±0.67	51.13±1.88	53.14±0.54	50.40±0.19	**60.14±0.33**	56.50±1.98	59.72±0.52

Table A.18: **Forecasting results at horizon H=10 minutes.** Test AUROC (%, mean±std over 3 seeds). **Best** per row highlighted in **bold**, second best per row underlined.

Datasets	Features	Snapshot	MeanPool	MaxPool	LSTM	TCN	Attention	Mamba-3	Trans.-XL	BioLT
CHB-MIT EEG seizure	Domain	63.11±9.21	60.00±7.28	56.22±6.08	**65.62±5.12**	60.01±9.34	61.78±11.76	60.55±4.70	58.27±6.30	57.56±12.61
	CBraMod	58.68±10.41	56.83±14.11	51.69±9.76	**62.72±13.62**	58.39±10.92	59.02±11.76	61.12±10.27	54.94±7.13	59.89±14.31
	LaBraM	50.66±5.21	44.33±5.17	**55.45±12.11**	50.79±6.52	53.93±10.72	50.37±6.79	50.48±5.11	48.86±8.97	47.63±6.31
CPSC-2021 ECG AF	Domain	77.66±4.32	79.83±7.26	81.66±3.42	78.28±1.73	80.24±3.62	83.91±4.72	78.43±5.65	**87.00±1.96**	86.93±1.36
	CSFM	82.68±2.12	85.90±2.89	85.44±3.86	83.51±2.77	86.30±1.43	81.68±2.34	81.30±5.05	87.44±3.37	**88.13±1.91**
	CLEF	76.79±4.49	79.84±3.93	78.08±4.85	71.17±2.52	76.77±2.81	**81.64±5.95**	77.48±3.81	76.30±9.92	77.82±10.18
	ECGFounder	80.02±1.04	83.09±4.23	85.86±5.12	83.37±4.63	86.55±2.27	80.90±2.64	75.29±1.68	85.57±4.98	**88.67±4.51**
VitalDB ECG AF	Domain	89.35±1.19	85.82±3.75	83.91±3.27	91.18±0.53	89.48±2.78	81.59±9.46	**91.58±0.40**	89.91±1.36	91.55±1.85
	CSFM	90.63±1.14	89.47±1.56	89.31±0.49	90.05±1.28	90.80±0.95	90.09±1.05	90.94±1.55	91.29±1.18	**92.27±0.58**
	CLEF	90.24±1.25	87.79±0.67	87.67±1.42	90.36±1.32	90.57±1.49	85.85±2.13	90.00±1.88	88.60±2.74	**90.62±1.09**
	ECGFounder	90.60±1.06	87.79±3.78	89.45±1.33	90.96±0.99	91.15±1.00	86.01±2.68	**91.19±0.75**	89.26±3.24	91.04±1.70
VitalDB PPG AF	Domain	87.67±0.73	87.97±0.77	79.56±1.79	89.05±0.63	88.74±1.49	85.86±3.03	**90.84±0.87**	89.60±0.92	90.15±0.22
	CSFM	89.29±0.40	88.45±0.22	90.10±1.09	89.13±1.24	89.74±0.14	87.40±0.28	89.38±0.29	89.83±0.68	**90.23±0.70**
	PulsePPG	85.25±0.99	84.50±1.13	58.97±2.46	87.25±1.73	**88.46±0.73**	81.86±9.76	87.91±0.66	87.68±0.93	87.89±0.15
MC-MED ECG sepsis	Domain	52.18±1.39	53.29±2.07	55.13±1.13	50.45±1.54	53.36±1.32	53.18±0.59	55.05±0.79	**60.46±3.38**	60.29±0.64
	CSFM	52.98±1.40	53.53±0.78	56.28±1.35	51.39±2.49	52.57±0.78	51.34±3.36	52.15±1.72	54.75±0.77	**58.42±1.30**
	CLEF	53.19±1.89	53.94±2.75	56.16±2.47	51.25±1.48	54.05±2.54	53.15±2.42	56.40±1.68	54.41±2.95	**59.79±1.72**
	ECGFounder	52.16±2.48	51.94±1.88	53.17±0.75	49.67±1.81	50.99±1.83	51.42±2.88	50.23±1.93	53.61±0.89	**57.37±1.43**
MC-MED PPG sepsis	Domain	51.23±0.22	53.33±1.34	55.90±0.43	49.71±2.10	53.35±0.20	51.40±1.30	61.48±0.07	**63.91±0.39**	63.19±0.48
	CSFM	51.64±1.54	51.11±2.51	57.71±0.76	50.84±1.34	52.61±2.33	51.74±2.74	57.40±1.35	52.51±0.55	**60.77±0.43**
	PulsePPG	50.75±0.50	52.06±1.09	55.83±1.56	50.69±0.84	53.09±0.19	51.00±1.41	60.61±0.83	56.30±2.82	**61.52±0.78**