Extending Pretrained 10-Second ECG Foundation Models to Longer Horizons

Wei Tang[1,2] **Jinpei Han**[1] **Kangning Cui**[3,4] **Mattia Carletti**[1] **Fredrik K. Gustafsson**[1]
Shreyank N Gowda[5] **Patitapaban Palo**[1] **Anshul Thakur**[1] **Lei Clifton**[1]
Jean-michel Morel[6] **Raymond H. Chan**[6] **David A. Clifton**[1] **Xiao Gu**[1†]
[1]University of Oxford
[2]Oxford Suzhou Centre for Advanced Research
[3]City University of Hong Kong (Dongguan)
[4]City University of Hong Kong
[5]University of Nottingham
[6]Lingnan University

Abstract

Electrocardiogram (ECG) foundation models pretrained on typical diagnostic 10-second ECG segments, have demonstrated strong transferability across a range of clinical applications. However, many real-world applications produce recordings that are typically longer, and are varied in duration during inference time. These 10-second models have no built-in way to combine information across time. Extending them to longer horizons introduces two challenges: *structural* incompatibilities arising from input-length disparities, and *semantic* challenges that limit meaningful temporal aggregation. We propose a parameter-efficient framework that extends pretrained ECG foundation models to longer and variable-length ECGs without retraining the backbone. Guided by a frozen pretrained 10-second model, we introduce a lightweight plug-in module that extends the model in two complementary ways: (i) structurally compatible long-sequence processing and (ii) semantically informed temporal modeling. Experiments on multiple long-horizon ECG tasks, datasets, and foundation model backbones demonstrate that our method enables robust longer-horizon extension from pretrained snapshot models, consistently outperforming sliding-window and pooling-based baselines with strong parameter efficiency.

1 Introduction

The electrocardiogram (ECG) remains the gold standard for non-invasive cardiovascular diagnosis, providing a direct view of cardiac electrical activity. Advances in machine learning, particularly deep neural networks, have substantially improved automated ECG analysis, with models achieving expert-level performance on several diagnostic tasks [1]. Early deep learning approaches, however, primarily relied on supervised training for narrow, task-specific objectives. While effective in controlled settings, these models were constrained by limited labeled data and often failed to generalize across datasets, institutions, and patient populations [2].

To improve robustness and transferability, recent work has increasingly shifted toward ECG foundation models pretrained with self-supervised objectives on large-scale ECG datasets [3–5]. Through contrastive or generative self-supervised learning objectives, these architectures learn generic feature extractors that are robust to signal variability and improve transfer across cohorts and institutions. As

†Corresponding Author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

a result, a single pretrained foundation model backbone paired with lightweight task-specific heads can be adapted to diverse downstream applications such as diagnosis, phenotyping, and outcome prediction [6, 7].

Despite their progress, these ECG foundation models present a fundamental limitation: they are trained and used as fixed-length *snapshot* processors. They typically operate on short inputs (often 10 s) during both pretraining and inference [3, 4], which biases the learned representations toward local waveform morphology and short-term rhythm. As a result, they lack an explicit mechanism for aggregating information across extended recordings, modeling long-range temporal dependencies, or supporting coherent prediction across multiple horizons – when recording duration varies across monitoring scenarios.

This limitation is consequential for many clinically relevant settings. In contrast to short 12-lead diagnostic recordings, continuous monitoring is intended to capture sparse or intermittent events over longer horizons, often under reduced-lead configurations [8–10]. In these contexts, clinically actionable signals may be distributed unevenly over time, as in paroxysmal atrial fibrillation surveillance, sleep, or sepsis early-warning. A common workaround is sliding-window inference followed by simple aggregation (*e.g.*, mean pooling) of window-level outputs (Figure 1a). More structured variants aggregate window representations using attention or lightweight sequential modules (Figure 1b). While practical, such strategies can be computationally costly and still model cross-window dependencies only indirectly, motivating adaptation methods that explicitly extend the model temporal horizon beyond the original snapshot setting.

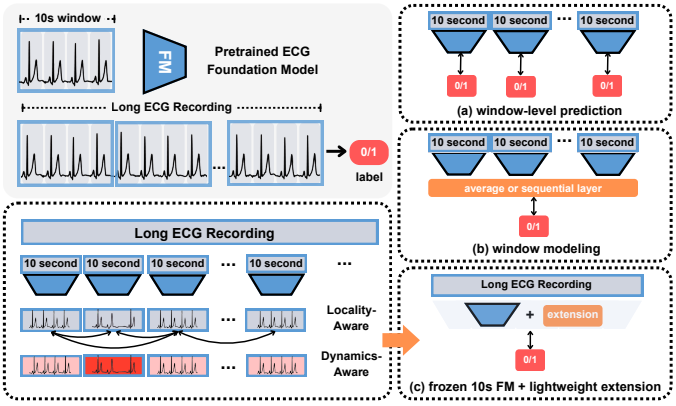


Figure 1: **Overview of the problem setting and extension strategies.** ECG foundation models are typically pretrained on short, fixed-length recordings (*e.g.*, 10 s), which makes direct use on long or variable-length recordings non-trivial. Naive extension strategies either run independent window-level predictions and aggregate the outputs as panel (a), or apply simple aggregation or sequential layers over window representations as panel (b). We propose a lightweight long-horizon extension for frozen 10-second ECG foundation models that supports variable-length longer recordings, by leveraging locality-aware inter-window similarity and dynamics-aware long-term learning, as illustrated in panel (c).

In this work, we show that pretrained snapshot ECG foundation models can be adapted to longer-duration and variable-length inputs without computationally expensive retraining. Our key observation is that these models can be repurposed for extended-context ECG analysis along three complementary dimensions, as illustrated in Figure 1 (left bottom panel).

General ECG interpretation. Pretrained ECG foundation models already acquire substantial domain knowledge for interpreting cardiac signals through large-scale pretraining. This enables knowledge to be preserved and effectively reused when extending the models to longer temporal horizons, without relearning low-level ECG semantics.

Locality-aware representation learning. Pretrained ECG foundation models are already well suited to encode short 10-second ECG windows: each segment can be meaningfully represented on its own, thus can be used as a local reference when extending the model to longer recordings. We exploit this property to preserve local semantics and align snapshot-level representations when processing long recordings.

Dynamics-aware long-term learning. The representations produced by pretrained models naturally reflect relative importance across segments, offering a basis for identifying informative events within long recordings. This property allows selective reuse and robust aggregation of local representations, which is essential for modeling sparse and transient dynamics over extended time horizons.

Based on these insights, we propose a parameter-efficient framework that extends frozen snapshot ECG foundation model to longer-horizon and variable-length settings. Rather than modifying or retraining the pretrained backbone, we reuse its embedded ECG knowledge and locality prior, while introducing a lightweight module to explicitly model long-term temporal dynamics. Across diverse datasets, tasks, and models, our method consistently improves long-horizon performance over conventional sliding-window and pooling strategies, demonstrating a practical and scalable path that goes beyond 10-second ECG understanding.

2 Related Work

ECG foundation models. Foundation models have recently been extended from vision and language to physiological signals, with ECG serving as a prominent testbed. Most existing ECG foundation models are pretrained on diagnostic ECG recordings, which are typically short in duration ($\approx$10 s) and acquired using standard 12-lead configurations [3, 4]. To date, most progress in this field has been driven along two complementary dimensions: scaling the volume and diversity of pretraining data [6, 7, 11, 12], and refining pretraining strategies to better exploit available supervision (self-supervised [13], hybrid [14], or knowledge-informed objectives [15, 16]). On the other hand, these models, pretrained in a diagnostic 10-second 12-lead regime, have been shown to generalize seamlessly to reduced-lead or single-lead settings [6, 7], which are widely used in continuous monitoring scenarios. However, they remain fundamentally limited in their ability to operate directly over longer temporal horizons, in an end-to-end manner, beyond simplistic sliding-window processing of 10 s segments.

Long-term physiology modeling. Many clinically relevant decisions depend on physiological dynamics over long horizons, which are not well captured by short snapshot windows, such as ambulatory Holter monitoring and intensive care monitoring. Recent learning-based systems often apply window slicing, which partitions a long recording into short segments, produces segment-level predictions or embeddings, and then aggregates these across time. The aggregation strategy varies from simple pooling to the more recent attention pooling or a lightweight sequential layer like LSTM to model the temporal relationships [17, 18]. However, such aggregation introduces a structural bottleneck: As most tasks are formulated as mapping a single long sequence to a single outcome, temporal dependencies within the sequence are often weakly modeled, making it difficult to localize when clinically meaningful risk signals occur. Moreover, performance can degrade when predictions are made from observations that are shorter than the horizon assumed by the aggregation, limiting utility in time-sensitive clinical decision-making.

Adaptation of biosignal foundation models. Alongside the emergence of ECG foundation models, or broader biosignal foundation models, increasing attention has been paid to adaptation under distribution shifts [19] and deployment constraints [20]. They are mostly enabled by parameter-efficient fine-tuning methods such as LoRA, prefix tuning, and adapter modules, which update only a small subset of parameters while largely preserving the backbone [21]. Despite this progress, existing biosignal adaptation work primarily targets task transfer, domain shift, and personalization under a fixed input scale. It assumes that the downstream signal can be processed within the same short window length used in pretraining. Due to the nature of most existing ECG datasets, which were typically collected in standard diagnostic settings (often 10 s), the models pretrained on these datasets fall short when applied to longer temporal horizons required in several clinical settings. A systematic approach to extend the temporal horizon of pre-trained biosignal foundation models during adaptation remains unexplored.

3 Method

We focus on ECG foundation models based on transformer architectures and denote by f_θ the backbone foundation model, parameterized by θ, whose parameters are kept frozen. The input to the backbone model is an ECG signal, processed as a sequence of fixed-length patch-level embeddings (*e.g.*, $0.1\ s$), which are augmented with positional embeddings to encode temporal order before being passed to the transformer. Specifically, the ECG input embedding is denoted as $\mathbf{X} \in \mathbb{R}^{N_{[10s]} \times D}$, where $N_{[10s]}$ is the number of patches and D is the embedding dimension. The corresponding positional embeddings $\mathbf{E}_{[10s]} \in \mathbb{R}^{N_{[10s]} \times D}$, are added to $\mathbf{X}$ and passed to the backbone model f_θ to produce a global representation of the signal, $f_\theta(\mathbf{X} + \mathbf{E}_{[10s]})$. Existing ECG foundation models are typically pretrained on short, fixed-length segments (*e.g.*, 10-second windows) and achieve strong

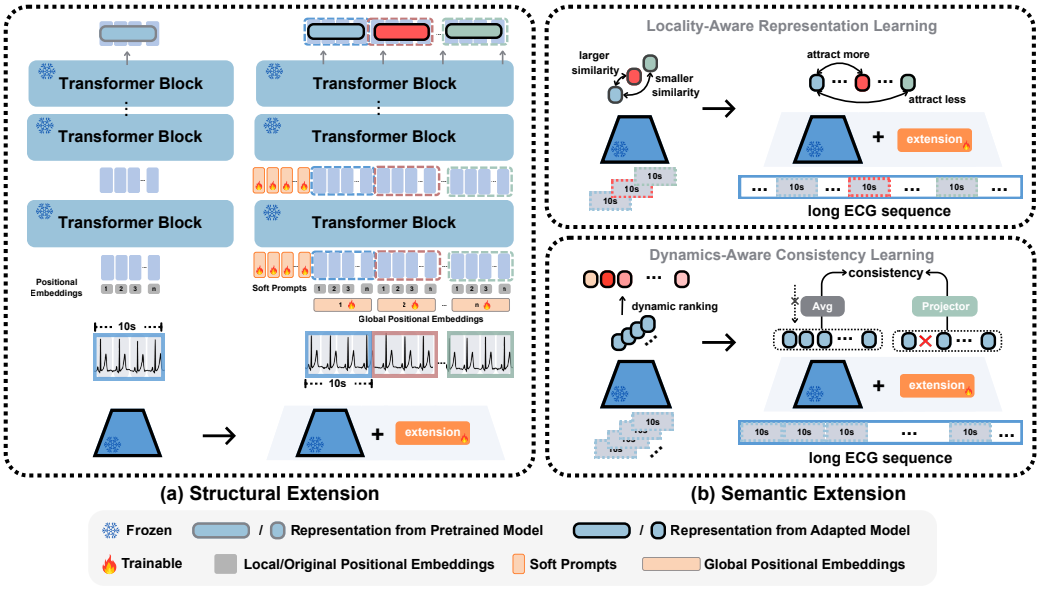


Figure 2: **Overview of our extension framework.** The extension is divided into two complementary parts: a structural extension that enables compatibility with long-horizon recordings, and a semantic extension that supports coherent representation learning over extended temporal horizons. (a) Structural extension (Section 3.1) introduces additional learnable tokens, soft prompts and global positional embeddings, by freezing the original backbone model, to make it compatible with longer recordings. (b) Semantic extension (Section 3.2) leverages snapshot-level (10-second) representations derived from the frozen backbone as supervision, enforcing semantic consistency across segments and supporting coherent long-horizon understanding.

performance by learning generalizable representations of local waveform morphology and short-term rhythm. However, applying these models directly to longer horizons poses two fundamental challenges.

- **Structural extension.** At a structural level, snapshot ECG models are not directly compatible with longer recordings. Their positional encodings, tokenization schemes, and input assumptions are optimized for short segments.
- **Semantic extension.** Beyond input compatibility, long-horizon ECG analysis requires meaningful aggregation of information across time. Clinically relevant events may be sparse, transient, or distributed unevenly over long recordings. Naively processing each 10-second segment independently and pooling predictions fails to capture long-range dependencies.

These two challenges motivate a framework that not only enables snapshot ECG foundation models to operate on long recordings, but also extends their semantic understanding to support coherent long-term modeling. To this end, we propose a framework, illustrated in Figure 2, that performs a single adaptation to long-horizon ECG signals ($\mathbf{X} \in \mathbb{R}^{N_{[L]} \times D}$, with $L > 10\,\text{s}$) and supports inference across variable signal lengths up to L without further modification. The empirical value for L considered in the main experiments is 180 s, and we further discuss behavior beyond the 3-minute setting in Appendix B.6.

3.1 Structural Extension - Enabling Long-Horizon ECG Processing

A critical barrier to adapting snapshot models to extended recordings lies in the incompatibility of positional embeddings. These embeddings $\mathbf{E}_{[10\text{s}]} \in \mathbb{R}^{N_{[10s]} \times D}$ are learned for a fixed input length and therefore encode temporal relationships that are specific to a 10-second window.

In the image domain, a common practice for extending vision transformers to different image scales [22–24], is interpolating positional embeddings. However, this strategy does not naturally transfer to time series, especially biosignals, since interpolation has fundamentally different meanings between

these two domains. The former (image) is related to resolutions, whilst the latter (biosignal) is related to the sampling rate, which is significantly entangled with their physiological information.

To address this, rather than interpolating from $N_{[10s]}$ to $N_{[L]}$, we introduce a two-level positional encoding scheme, comprising local (fine-grained, the original $\mathbf{E}_{[10s]}$) and global (coarse-grained, $\mathbf{E}_{\text{global}}$) components. The former reuses the original $\mathbf{E}_{[10s]}$ to preserve within–10-second temporal structure, whereas the latter models the relationship between consecutive 10-second segments. For a sequence of $N_{[L]}$ patches, we construct the final positional embedding $\mathbf{E}_{\text{[L]}} \in \mathbb{R}^{N_{[L]} \times D}$ as:

$$\mathbf{E}_{\text{[L]}}[t] = \mathbf{E}_{\text{[10s]}}\left[t \bmod N_{\text{[10s]}}\right] + \mathbf{E}_{\text{global}}[\lfloor t/N_{[10s]} \rfloor], \tag{1}$$

where t indexes the patch and $\lfloor t/N_{[10s]} \rfloor$ indexes its associated 10-second segment for global temporal encoding.

Local component ($\mathbf{E}_{\text{[10s]}}$). This component inherits the pre-trained positional weights, preserving sensitivity to high-frequency intra-segment structures (*e.g.*, P-QRS-T complexes). For sequences exceeding the pretraining length, we repeat these weights rather than interpolating them, preserving the temporal scale at which local waveform morphology was learned during pretraining.

Global component ($\mathbf{E}_{\text{global}}$). To anchor local windows within the broader recording context, we introduce a learnable embedding matrix $\mathbf{E}_{\text{global}} \in \mathbb{R}^{M \times D}$, where $M = N_{[L]}/N_{[10s]}$ denotes the number of 10-second segments in the signal. For simplicity and without loss of generality, we assume that $N_{[L]}$ is an integer multiple of $N_{[10s]}$. Each embedding in $\mathbf{E}_{\text{global}}$ is then broadcast across a contiguous block of $N_{[10s]}$ patches by Equation 1.

This hierarchical design enables modeling over the full recording duration while maintaining the frequency characteristics and morphological semantics learned within local segments during pretraining.

Soft prompt tuning. To facilitate the learning of global positional embeddings and enable parameter-efficient adaptation to capture long-horizon temporal relationships, we adopt soft prompt tuning by introducing additional learnable tokens that are prepended to the input sequence and inserted at intermediate layers [25].

3.2 Semantic Extension - Extending Long-Horizon ECG Understanding

A purely structural extension to longer signals does not, by itself, work directly on longer signals, as the additional components introduced by the extension are randomly initialized. Directly and solely optimizing learnable components in an end-to-end manner, by mapping the long-sequence to a single label, may results in suboptimal extension and may be insufficient for effective long-range modeling.

To address this, we take the hypothesis that the backbone foundation model f_θ itself already learns representations that are sufficiently informative at any 10-second scale locally, and when applied to consecutive 10-second segments, can be leveraged to capture dynamics over longer horizons. In this sense, we adopt a teacher-student framework, where the representation derived from each 10-second segment serves as the teacher to guide the representation learning both locally and globally over extended temporal horizons.

Locality-aware representation learning. We first enforce segment-level semantic alignment between the pretrained and extended models. The pretrained foundation model is trained on short (10-second) ECG segments and provides robust local representations of morphology, rhythm, and physiological state. When adapting the model to longer ECG sequences, these local representations may be changed by the long-horizon training objective. To reduce this risk, we constrain the student model to remain aligned with the pretrained teacher at the segment level.

We implement this using teacher-guided soft contrastive learning. Prior approaches define contrastive pairs based on temporal adjacency [26] or patient identity [27]. In contrast, we compute soft similarity targets from teacher-derived representations within each batch, thereby preserving the teacher-induced similarity structure without imposing explicit temporal or patient-level pairing assumptions. In our experiments, Appendix C.2 shows that intra-subject or inter-subject pairing constraints do not improve performance over this simple batch-level strategy.

We use pairwise similarities between teacher representations as soft supervision to guide the student via a contrastive objective. Let $\mathbf{S}^T_{i,j}$ denote the similarity (*e.g.*, negative L_2 norm) between teacher representations of local 10-s segments i and j. For an anchor segment i, if $\mathbf{S}^T_{i,j} > \mathbf{S}^T_{i,k}$, the student

is encouraged to preserve the same ordering, *i.e.*, $\mathbf{S}^S_{i,j} > \mathbf{S}^S_{i,k}$. This is implemented using a rank-based soft contrastive loss [28], which preserves the relative ranking of inter-segment similarities. Figure 2 (right upper panel) illustrates this similarity-based alignment, with the local contrastive loss formulated as below:

$$\mathcal{L}_{\text{local}} = -\frac{1}{K}\sum_{\substack{j=1 \\ j\neq i}}^{K} \log \frac{\exp(\mathbf{S}^S_{i,j}/\tau)}{\sum_{k\in\Omega_{i,j}} \exp(\mathbf{S}^S_{i,k}/\tau) + \exp(\mathbf{S}^S_{i,j}/\tau)}, \tag{2}$$

where K denotes the number of segments in the batch, and τ is a temperature parameter. The set $\Omega_{i,j}$ is defined based on the teacher similarity as $\Omega_{i,j} = \{k \mid \mathbf{S}^T_{i,k} < \mathbf{S}^T_{i,j}\}$, *i.e.*, all segments that are ranked as less similar to anchor i than segment j according to the teacher. In practice, we add a projector $g_{\text{local}}(\cdot)$ on top of the latent representation to implement this loss, following standard practice in contrastive learning [29].

Dynamics-aware consistency learning. Physiological signals such as ECG are inherently non-stationary, exhibiting pronounced temporal variability over long recordings. Rather than assuming that all 10-second segments should contribute in the same way, we use the frozen teacher representation to partition segments according to their similarity structure. This partition is used as an operational mechanism for organizing heterogeneous segments (such as rare abnormal events versus stationary patterns, or clean versus less clean segments), and empirically, we found that using both groups consistently outperforms using either group alone (Appendix D.1), which indicates that both high-variation and lower-variation parts contain useful information. The qualitative case studies are shown in Appendix D.2. We therefore encourage each group to retain information from the whole recording, to capture global dynamics.

To implement this idea, we identify a subset of 10-second segments $\mathcal{I}_{bg}$ that are most consistent with the rest of the recording, using an affinity matrix computed by the frozen teacher encoder. Intuitively, the affinity matrix encodes the pairwise similarity between all segments, and high-affinity (ha) segments $\mathcal{I}_{ha}$ are selected by ranking across the affinity matrix (refer to Appendix A.2). We then encourage these high-affinity segments to remain aligned to the global representations, remaining as consistent as possible even when the more dynamic segments are excluded. Specifically, the consistency is achieved as below:

$$\mathcal{L}_{\text{dyn}} = \frac{1}{|\mathcal{I}_{ha}|}\sum_{i\in\mathcal{I}_{ha}} \mathcal{D}_{\cos}\left(g_{\text{dyn}}(\mathbf{z}^{(i)}_S), \text{stopgrad}(\bar{\mathbf{z}}_S)\right), \tag{3}$$

where $\mathbf{z}^{(i)}_S$ denotes the student feature for 10-second segment i, $\bar{\mathbf{z}}_S$ is the global feature by averaging latent features over the entire long-horizon sequence, and $\mathcal{D}_{\cos}$ is the cosine distance. In practice, we apply a one-sided $g_{\text{dyn}}(\cdot)$ projection head to the student representations, and a stop-gradient $\text{stopgrad}(\cdot)$ operator through global representation, to prevent trivial collapse to constant representations [30].

3.3 Model Training

We train the framework end-to-end by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1\mathcal{L}_{\text{local}} + \lambda_2\mathcal{L}_{\text{dyn}}, \tag{4}$$

where $\mathcal{L}_{\text{task}}$ is the standard classification loss, λ_1 and λ_2 are scalar weighting coefficients that balance the contribution of each auxiliary loss. Additional details on the learnable components and training hyperparameters are provided in Appendix A.2 & A.3, respectively.

4 Experiments

We evaluate the proposed framework on four downstream task settings and four foundation model backbones. This setup allows us to assess long-horizon adaptation across diverse clinical scenarios, model scales, and pretraining paradigms.

Evaluation setup. Unless otherwise specified, all models are trained on 3-minute ECG recordings and evaluated on 3-minute, 2-minute, 1-minute, 30-second, and 10-second inputs without retraining. This protocol reflects realistic deployment settings in which the available monitoring duration may

Table 1: **Overview of the datasets and tasks used in our experiments.** We summarize the dataset split sizes and classification objectives for all four downstream tasks.

Task	Dataset	Train	Val	Test	Task Description
ICU monitoring	VTaC [8, 34]	3,736	460	444	Binary classification of true versus false VT alarms
ED disposition	MC-MED [10, 35]	14,152	1,791	1,778	Binary classification of *Discharge* versus {*Inpatient*, *Observation*, *ICU*}
ED triage acuity	MC-MED [10, 35]	14,068	1,780	1,764	Five-class classification of triage acuity levels: *Resuscitation*, *Emergent*, *Urgent*, *Semi-Urgent*, and *Non-Urgent*
AF detection	CPSC2021 [36]	5,277	1,709	1,723	Binary classification of AF versus non-AF

vary [31], while keeping the focus on practical extension beyond the original 10-second duration. Additional results in Appendix B.6 show that this extension remains effective at longer horizons, subject to available computational budget.

Backbone models. We use four backbone configurations from three representative ECG foundation-model families: CSFM [6] (Tiny and Base), MERL [32], and ECG-JEPA [33]. These backbones vary in model scale and pretraining objective, providing a diverse test bed for evaluating the robustness and generality of the proposed framework. Additional architectural and pretraining details are provided in Appendix A.1.

4.1 Datasets & Tasks

We consider three datasets covering four downstream task settings in total.

VTaC [8, 34] is a critical care benchmark for reducing false ventricular tachycardia (VT) alarms. We use the final 3 minutes preceding the alarm onset and formulate the task as binary classification of true versus false VT alarms.

MC-MED [10, 35] contains emergency department (ED) records and is used for two prediction tasks: *ED disposition* prediction, formulated as binary classification of *Discharge* versus *Admission*, and *triage acuity* prediction, formulated as five-class classification. For both tasks, we use the lead-II ECG signal and extract the final 3 minutes of the selected segment.

CPSC2021 [36] is a long-term Holter benchmark targeting paroxysmal atrial fibrillation (AF) events. We formulate it as a binary AF detection task using the lead-II signal and segment long recordings into non-overlapping 3-minute windows.

A complete summary of dataset statistics, task definitions, and data splits is provided in Table 1. To ensure consistent signal quality across datasets and modalities, all signals are preprocessed using a standardized pipeline before training and evaluation. Specifically, ECG and PPG signals are cleaned using functions from NeuroKit2 [37]. Signals containing extended flat regions are removed, and missing values are imputed using the mean of the corresponding channel. To match the input requirements of the backbone models, all signals are resampled to 250 Hz for CSFM-Tiny, CSFM-Base, and ECG-JEPA, and to 500 Hz for MERL.

4.2 Baseline Methods

We compare the proposed method with two groups of baseline strategies.

Snapshot Aggregation. This group includes Token Pooling and Logit Pooling. These methods partition each long ECG into non-overlapping 10-second windows and aggregate either frozen backbone features or window-level logits.

Fine-tuning-based Adaptation. This group includes Full Fine-tuning, Linear Probing, Partial Tuning, Bias Tuning, Adapter, Visual Prompt Tuning (VPT) [25, 38, 39], and an LSTM-based baseline. These methods adapt either the pretrained model parameters or the resulting sequence of snapshot features for longer recordings. For baselines that require longer inputs, positional embeddings are extended by interpolation.

Table 2: **Main adaptation results across ECG foundation models and downstream tasks.** For each task, we report performance at the longest evaluation horizon (180s) together with the average across five input lengths in the format **Avg. (Std.)**. **Best** results are shown in bold and second-best results are underlined. Full per-length results are provided in Appendix B.1.

Method	VTaC		MC-MED (ED)		MC-MED (Acuity)		CPSC2021		Overall Avg.
	180s	Avg. (Std.)	180s	Avg. (Std.)	180s	Avg. (Std.)	180s	Avg. (Std.)	
CSFM-Tiny									
Token Pooling	74.05	76.24 (2.28)	70.80	68.72 (2.69)	61.00	60.00 (1.03)	96.81	96.59 (0.39)	75.39
Logit Pooling	73.83	75.83 (1.96)	71.15	69.28 (2.26)	60.13	59.96 (0.48)	97.23	97.26 (0.16)	75.58
Full Fine-tuning	79.01	70.54 (6.89)	70.81	68.09 (3.05)	64.80	63.07 (1.18)	94.24	93.23 (1.91)	73.73
Linear Probing	69.50	71.10 (1.25)	69.38	66.89 (2.67)	60.66	60.08 (0.99)	96.46	96.91 (0.57)	73.75
Partial Tuning	86.49	73.57 (9.59)	70.91	68.47 (3.79)	63.43	63.06 (1.50)	95.48	94.87 (1.57)	74.99
Bias Tuning	64.43	66.43 (1.94)	68.13	66.28 (1.93)	56.71	54.49 (2.42)	98.46	97.62 (1.33)	71.21
Adapter	79.33	72.83 (5.08)	71.24	68.60 (3.01)	61.55	60.53 (0.94)	98.01	96.24 (2.99)	74.55
VPT	88.18	74.33 (10.78)	71.19	67.29 (5.16)	64.50	61.22 (3.17)	98.67	96.37 (3.94)	74.80
LSTM Tuning	84.97	83.01 (2.39)	71.31	65.18 (7.55)	61.97	58.80 (4.55)	98.75	98.09 (0.63)	76.27
Ours	**89.50**	**88.13** (2.32)	**72.02**	**69.79** (3.18)	**65.18**	**63.35** (2.04)	**98.86**	**98.63** (0.79)	**79.98**
CSFM-Base									
Token Pooling	77.94	79.84 (2.09)	71.23	69.30 (2.51)	66.82	64.76 (2.36)	97.02	97.33 (0.40)	77.81
Logit Pooling	77.83	79.64 (1.98)	71.59	69.90 (2.46)	67.07	64.85 (2.46)	97.54	97.52 (0.23)	77.98
Full Fine-tuning	84.23	72.46 (9.58)	72.30	69.83 (3.13)	65.09	63.47 (1.46)	95.66	95.66 (1.13)	75.35
Linear Probing	74.35	77.50 (2.81)	70.44	67.86 (2.64)	66.53	64.26 (2.32)	96.89	96.86 (0.25)	76.62
Partial Tuning	88.81	79.69 (8.23)	71.47	69.29 (3.41)	67.25	65.06 (2.11)	96.21	95.54 (1.00)	77.40
Bias Tuning	70.35	73.35 (2.02)	71.36	68.87 (3.16)	60.66	58.90 (1.77)	98.39	97.23 (1.39)	74.59
Adapter	88.43	78.95 (7.96)	71.90	69.22 (3.45)	65.91	63.72 (2.00)	98.41	96.85 (2.87)	77.19
VPT	90.24	80.72 (7.93)	72.39	68.76 (5.25)	66.00	63.78 (3.03)	98.57	96.57 (2.89)	77.46
LSTM Tuning	86.93	85.22 (1.33)	72.03	67.77 (6.33)	68.14	65.18 (2.95)	98.74	98.31 (0.47)	79.12
Ours	**91.23**	**89.62** (1.53)	**72.89**	**70.36** (3.43)	**68.89**	**66.00** (2.65)	**99.56**	**98.46** (1.08)	**81.11**
ECG-JEPA									
Token Pooling	76.76	78.54 (1.56)	69.75	67.37 (2.74)	61.38	60.33 (1.30)	90.75	90.68 (0.55)	74.23
Logit Pooling	76.23	78.10 (1.65)	69.23	67.67 (2.46)	61.23	60.11 (1.28)	90.14	90.47 (0.57)	74.09
Full Fine-tuning	88.32	82.96 (5.44)	67.96	65.44 (2.27)	62.83	60.65 (1.56)	88.67	89.59 (0.83)	74.66
Linear Probing	77.21	79.00 (1.45)	70.32	66.49 (3.04)	63.71	61.90 (1.46)	88.83	89.33 (0.34)	74.18
Partial Tuning	88.89	84.32 (5.27)	70.19	67.79 (2.92)	64.59	62.54 (2.20)	88.33	89.05 (0.76)	75.93
Bias Tuning	86.83	83.68 (5.53)	72.14	69.53 (2.77)	65.50	64.06 (1.62)	85.41	84.86 (2.20)	75.53
Adapter	70.57	67.91 (2.44)	63.06	60.68 (2.34)	60.77	60.17 (0.60)	75.88	73.17 (4.10)	65.48
VPT	89.87	82.44 (7.97)	**72.48**	67.29 (6.11)	66.18	63.89 (3.53)	93.65	84.85 (10.15)	74.62
LSTM Tuning	85.17	82.66 (2.67)	70.72	68.39 (3.15)	65.18	62.93 (2.35)	91.26	90.90 (0.49)	76.22
Ours	**91.02**	**85.65** (4.91)	**72.48**	**70.69** (3.35)	**67.53**	**64.89** (2.63)	**94.66**	**90.92** (4.25)	**78.04**
MERL									
Token Pooling	71.56	69.47 (3.97)	68.79	65.49 (3.85)	66.82	64.76 (2.36)	88.31	86.59 (3.38)	71.58
Logit Pooling	71.37	69.32 (3.88)	68.81	65.47 (3.78)	67.07	64.85 (2.46)	86.56	85.79 (2.59)	71.36
Full Fine-tuning	**88.31**	**82.08** (7.39)	67.30	65.03 (2.81)	65.09	63.47 (1.46)	**95.72**	**91.50** (7.41)	**75.52**
Linear Probing	74.91	71.94 (3.98)	68.07	64.64 (5.25)	66.53	64.26 (2.32)	84.52	82.86 (2.40)	70.93
Partial Tuning	79.06	71.59 (6.79)	68.41	64.98 (4.60)	67.25	65.06 (2.11)	88.71	87.32 (2.78)	72.24
Bias Tuning	80.63	75.62 (6.72)	69.12	65.57 (4.52)	60.66	58.90 (1.77)	94.10	88.26 (6.51)	72.09
Adapter	74.72	71.16 (4.38)	68.36	65.39 (4.77)	65.91	63.72 (2.00)	89.39	88.52 (2.24)	72.20
VPT	80.74	71.94 (8.67)	69.74	64.29 (5.69)	66.00	63.78 (3.03)	91.75	85.59 (8.69)	71.40
LSTM Tuning	83.05	78.55 (5.44)	67.74	63.86 (3.97)	68.14	65.18 (2.95)	93.55	89.01 (5.96)	74.15
Ours	82.60	76.76 (9.16)	**69.92**	**65.76** (5.32)	**68.89**	**66.00** (2.65)	93.49	89.73 (5.89)	74.56

5 Results

Main Results on Long-Horizon ECG Tasks. Table 2 summarizes the adaptation results across four backbone configurations and four downstream task settings. For each task, we report both performance at 3-minute and the average across all evaluated inference durations (3-minute, 2-minute, 1-minute, 30-second, and 10-secon). Overall, the proposed method is consistently among the strongest approaches across tasks and backbones, and in most cases outperforms both snapshot aggregation baselines and lightweight adaptation methods.

Table 3: **Guidance source comparison for MERL on VTaC.** Results are reported as AUC (%) across different inference durations. **Best** results are shown in bold.

Guidance	10s	30s	60s	120s	180s
MERL Self-guided	61.23	76.03	80.27	83.68	82.60
CSFM-Tiny guided	62.67	79.64	83.46	84.64	85.19
CSFM-Base guided	**63.16**	**81.33**	**84.51**	84.99	**86.07**
ECG-JEPA guided	63.07	80.61	84.17	**85.03**	85.78



Figure 3: **t-SNE visualizations of feature space on VTaC.** Blue and purple denote different classes. The dashed boxes highlight regions with clearer class separation, indicating stronger teacher model.

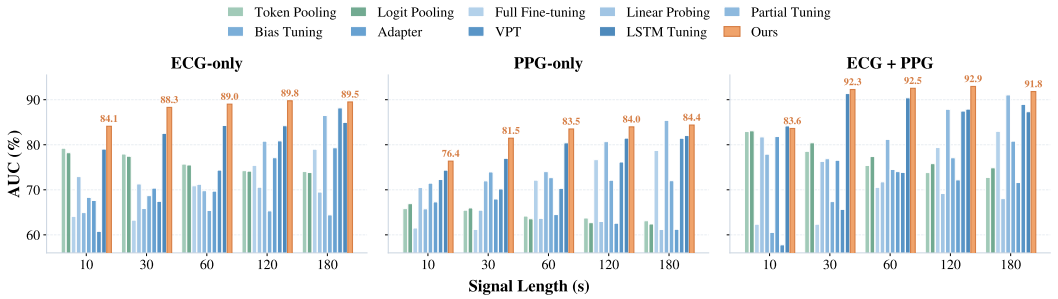


Figure 4: **Generalization across signal modalities on VTaC.** We evaluate the proposed framework using CSFM-Tiny under ECG-only, PPG-only, and joint ECG+PPG settings. Across all three modality configurations, the proposed method consistently outperforms the baseline adaptation strategies, indicating that the long-horizon extension is not specific to ECG alone.

A clear pattern also emerges across methods. Snapshot aggregation strategies such as logit pooling and token pooling provide competitive baselines in several settings, but they are typically outperformed in longer temporal context across tasks and backbones. This suggests that simple aggregation over short-window predictions or features is often insufficient for robust long-horizon adaptation.

The backbone-specific trends are also informative. On the CSFM family and ECG-JEPA, the proposed method achieves the strongest overall performance in most settings, particularly at longer input durations. On MERL, full fine-tuning remains the strongest baseline overall, whereas the proposed method remains highly competitive and often achieves the second-best performance while keeping the pretrained backbone frozen. Taken together, these results indicate that the proposed framework provides a reliable way to extend snapshot ECG foundation models across tasks, backbone scales, and pretraining paradigms.

Guidance Source Analysis. The MERL results motivate a more focused question: how much does semantic extension depend on the quality of the teacher representation? On VTaC, MERL self-guidance is noticeably weaker than full fine-tuning, and snapshot aggregation baselines are also relatively poor. This suggests that the snapshot-level MERL representations may provide a less effective guidance signal for long-horizon extension in this setting.

To examine this effect, we replace MERL self-guidance with external guidance from stronger backbone models while keeping the MERL student unchanged. As shown in Table 3, all three external guidance sources consistently improve over MERL self-guidance across inference durations, with the gains becoming more pronounced beyond 30 s. Figure 3 further provides qualitative evidence that the different backbones induce different feature-space structures on VTaC, which is consistent with their different effectiveness as guidance sources. Together, these results suggest that long-horizon semantic extension depends, at least in part, on the informativeness of the underlying snapshot-level teacher representation.

Modality-Agnostic Analysis. The long-horizon extension problem is not unique to ECG. Other physiological signals are also commonly modeled using short snapshot windows despite containing clinically meaningful information over longer durations. To test whether the proposed framework generalizes beyond ECG-only inputs, we evaluate CSFM-Tiny on VTaC using ECG-only, PPG-only, and joint ECG+PPG settings. Figure 4 shows that our method consistently outperforms the baseline adaptation strategies across all three modality configurations. This result suggests that the proposed framework is not tied to a single physiological modality, but instead addresses a broader limitation of snapshot-pretrained physiological models.

Ablation Study. We next examine how the individual components of the proposed framework contribute to the performance gains. Table 4 isolates the respective effects of the positional encoding strategy, locality-aware learning, and dynamics-aware consistency learning. Replacing repeated or interpolated positional encodings with the proposed global positional encoding already yields a clear improvement, indicating that ensuring structural compatibility between the pretrained model and longer input recordings is an important component of the overall long-horizon extension process. Adding either the locality-aware objective or the dynamics-aware objective provides further gains, while combining both produces the strongest overall performance across input durations.

Affinity-guided dropping. We further analyze the behavior of the consistency module through a comparison between affinity-guided and random segment dropping, shown in Table 5. Compared with random dropping, affinity-guided dropping is generally more effective, supporting the role of similarity-based segment structure in the dynamics-aware objective. Taken together, the results of these ablation experiments show that the observed performance gains of the proposed framework arise not from simple input resizing or generic tuning alone, but from the combination of structural extension and semantically informed temporal adaptation.

Table 4: **Ablation of adaptation components** for CSFM-Tiny on VTaC (AUC %). The best result at each duration is highlighted in **bold**.

PE Strategy	$\mathcal{L}_{\text{local}}$	$\mathcal{L}_{\text{dyn}}$	10s	30s	60s	120s	180s
Repeat	✗	✗	70.43	73.09	70.60	72.43	71.45
Interpolate	✗	✗	60.78	67.44	74.37	80.86	88.18
Global	✗	✗	78.13	83.96	87.45	89.08	88.92
Global	✓	✗	83.62	87.13	88.42	89.30	89.00
Global	✗	✓	80.01	85.90	87.80	88.86	88.90
Global	✓	✓	**84.10**	**88.29**	**89.01**	**89.76**	**89.50**

Table 5: **Comparison of affinity-guided and random dropping on VTaC.** Results are reported as AUC (%). For each model, the better result at each duration is highlighted in **bold**.

Model	Strategy	10s	30s	60s	120s	180s
CSFM-Tiny	Affinity Drop	**84.10**	**88.29**	**89.01**	89.76	**89.50**
	Random Drop	84.00	87.19	88.35	**90.03**	89.01
CSFM-Base	Affinity Drop	**87.38**	**89.15**	**89.50**	**90.85**	**91.23**
	Random Drop	86.73	88.01	88.93	90.16	90.83
ECG-JEPA	Affinity Drop	**78.86**	**82.93**	**86.09**	**89.36**	**91.02**
	Random Drop	75.92	80.79	84.87	88.92	90.01
MERL	Affinity Drop	**61.23**	**76.03**	**80.27**	**83.68**	**82.60**
	Random Drop	58.20	72.54	78.67	80.23	81.20

6 Conclusion

In this work, we studied how to extend snapshot ECG foundation models beyond their original fixed-length input setting. Without requiring full retraining, we introduced a lightweight plug-in adaptation module that enables one-time extension to longer and variable-length recordings. Across multiple downstream tasks, datasets, and backbone model families, the proposed method achieves consistently strong performance, improving over snapshot aggregation baselines and remaining competitive across diverse adaptation strategies.

A key feature of the proposed method is that the extension module is trained at the longest available temporal horizon and subsequently evaluated across a range of different input lengths. The performance obtained under this evaluation setting indicates that the adapted models can generalize effectively across variable-duration inputs. However, extending snapshot foundation models to arbitrary or substantially longer horizons under practical memory and computational constraints remains an open challenge. We therefore view the development of more inference-efficient mechanisms for longer-horizon adaptation as an important direction for future work.

Acknowledgement

This work is supported in part by the Oxford Suzhou Centre for Advanced Research. DAC was funded by an NIHR Research Professorship (NIHR302440); a Royal Academy of Engineering Research Chair; and the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE); and was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and the Pandemic Sciences Institute at the University of Oxford. XG was supported by funding from the Engineering and Physical Sciences Research Council IAA (EP/X525777/1), Medical Research Council (MR/X50273X/1), and Smart Data Research UK, a UKRI investment [grant number UKRI4004].

References

[1] Konstantinos C Siontis, Peter A Noseworthy, Zachi I Attia, and Paul A Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, 2021.

[2] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

[3] Jun Li, Aaron Aguirre, Junior Moura, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. *arXiv preprint arXiv:2410.04133*, 2024.

[4] Xiao Gu, Yuxuan Shu, Jinpei Han, Yuxuan Liu, Zhangdaihong Liu, James Anibal, Veer Sangha, Edward Phillips, Bradley Segal, Yuxuan Liu, Hang Yuan, Fenglin Liu, Kim Branson, Patrick Schwab, Danielle Belgrave, Lei Clifton, Dimitris Spathis, Vasileios Lampos, A. Aldo Faisal, and David A. Clifton. Foundation models for biosignals: A survey. 2025.

[5] Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *JAMIA open*, 8(5):ooaf122, 2025.

[6] 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.

[7] 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.

[8] Li-wei Lehman, Benjamin Moody, Harsh Deep, Feng Wu, Hasan Saeed, Lucas McCullum, Diane Perry, Tristan Struja, Qiao Li, Gari Clifford, and Roger Mark. VTaC: A benchmark dataset of ventricular tachycardia alarms from ICU monitors. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 38827–38843, 2023.

[9] Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III Waveform Database Matched Subset. *PhysioNet*, April 2020. doi: 10.13026/c2294b. URL `https://doi.org/10.13026/c2294b`. Version 1.0.

[10] 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.

[11] Zhijiang Wan, Qianhao Yu, Jia Mao, Wenfeng Duan, and Cheng Ding. Openecg: Benchmarking ecg foundation models with public 1.2 million records. *arXiv preprint arXiv:2503.00711*, 2025.

[12] Jonathan B Moody, Alexis Poitrasson-Rivière, Jennifer M Renaud, Tomoe Hagio, Fares Alahdab, Mouaz H Al-Mallah, Michael D Vanderver, Sascha N Goonewardena, Edward P Ficaro, and Venkatesh L Murthy. A foundation transformer model with self-supervised learning for ecg-based assessment of cardiac and coronary function. *NEJM AI*, 2(12):AIoa2500164, 2025.

[13] Junho Song, Jong-Hwan Jang, DongGyun Hong, Joon myoung Kwon, and Yong-Yeon Jo. Crema: A contrastive regularized masked autoencoder for robust ecg diagnostics across clinical domains, 2025. URL `https://arxiv.org/abs/2407.07110`.

[14] Shaoting Zhang, Yishan Du, Wenji Wang, Xianying He, Fangfang Cui, Liang Zhao, Bei Wang, Zhiqiang Hu, Ziqiang Wang, Qing Xia, et al. Ecgfm: A foundation model for ecg analysis trained on a multi-center million-ecg dataset. *Information Fusion*, page 103363, 2025.

[15] Han Yu, Peikun Guo, and Akane Sano. Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text. *Transactions on Machine Learning Research (TMLR)*, 2024.

[16] Yuanyuan Tian, Zhiyuan Li, Yanrui Jin, Mengxiao Wang, Xiaoyang Wei, Liqun Zhao, Yunqing Liu, Jinlei Liu, and Chengliang Liu. Foundation model of ecg diagnosis: Diagnostics and explanations of any form and rhythm on ecg. *Cell Reports Medicine*, 5(12), 2024.

[17] Wei-Long Zheng, Edilberto Amorim, Jin Jing, Wendong Ge, Shenda Hong, Ona Wu, Mohammad Ghassemi, Jong Woo Lee, Adithya Sivaraju, Trudy Pang, et al. Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. *Resuscitation*, 169:86–94, 2021.

[18] Peng Zhang, Fan Lin, Fei Ma, Yuting Chen, Siyi Fang, Haiyan Zheng, Zuwen Xiang, Xiaoyun Yang, and Qiang Li. Automatic screening of patients with atrial fibrillation from 24-h holter recording using deep learning. *European Heart Journal-Digital Health*, 4(3):216–224, 2023.

[19] Suli Wang, Yangshen Deng, Zhenghua Bao, Xinyu Zhan, and Yiqun Duan. Neurottt: Bridging pretraining-downstream task misalignment in eeg foundation models via test-time training. *arXiv preprint arXiv:2509.26301*, 2025.

[20] Rushuang Zhou, Yuanting Zhang, and Yining Dong. H-tuning: Toward low-cost and efficient ECG-based cardiovascular disease detection with pre-trained models. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=RLu1QIPiVr`.

[21] Chenrui Wu, Haishuai Wang, Xiang Zhang, Chengqi Zhang, and Jiajun Bu. Efficient personalized adaptation for physiological signal foundation model. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=55ysNwbOTI`.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

[23] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=p-BhZSz59o4`.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[25] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pages 709–727, 2022.

[26] Seunghan Lee, Taeyoung Park, and Kibok Lee. Soft contrastive learning for time series. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[27] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

[28] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36:17882–17903, 2023.

[29] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[30] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[31] Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

[32] Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *International Conference on Machine Learning (ICML)*, pages 31949–31963, 2024.

[33] Sehun Kim. Learning general representation of 12-lead electrocardiogram with a joint-embedding predictive architecture, 2026. URL https://arxiv.org/abs/2410.08559.

[34] Li-wei Lehman, Benjamin Moody, Lucas McCullum, Hasan Saeed, Harsh Deep, Diane Perry, Tristan Struja, Qiao Li, Gari Clifford, and Roger Mark. VTaC: A benchmark dataset of ventricular tachycardia alarms from ICU monitors. *PhysioNet*, October 2024. doi: 10.13026/8td2-g363. URL https://doi.org/10.13026/8td2-g363. Version 1.0.

[35] Aman Kansal, Emma Chen, Tom Jin, Pranav Rajpurkar, and David Kim. Multimodal Clinical Monitoring in the Emergency Department (MC-MED). *PhysioNet*, September 2025. doi: 10.13026/wvyw-g663. URL https://doi.org/10.13026/wvyw-g663. Version 1.0.1.

[36] 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. *PhysioNet*, June 2021. doi: 10.13026/ksya-qw89. URL https://doi.org/10.13026/ksya-qw89. Version 1.0.0.

[37] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, feb 2021. doi: 10.3758/s13428-020-01516-y. URL https://doi.org/10.3758%2Fs13428-020-01516-y.

[38] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:16664–16678, 2022.

[39] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics*, pages 487–503, 2021.

[40] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[41] Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, September 2023. doi: 10.13026/4nqg-sb35. URL https://doi.org/10.13026/4nqg-sb35. Version 1.0.

[42] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, 2020.

[43] Lei Lu, Tingting Zhu, Antonio H Ribeiro, Lei Clifton, Erying Zhao, Jiandong Zhou, Antonio Luiz P Ribeiro, Yuan-Ting Zhang, and David A Clifton. Decoding 2.3 million ECGs: interpretable deep learning for advancing cardiovascular diagnosis and mortality risk stratification. *European Heart Journal - Digital Health*, 5(3):247–259, 2024.

[44] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

[45] Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7, 02 2020. doi: 10.1038/s41597-020-0386-x.

[46] Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, Guohua Fu, Hai Yao, Dongbo Li, Hangyuan Guo, and Cyril Rakovski. Optimal multi-stage arrhythmia classification approach. *Scientific Reports*, 10, 02 2020. doi: 10.1038/s41598-020-59821-7.

[47] Yu-Jhen Chen, Chien-Liang Liu, Vincent S. Tseng, Yu-Feng Hu, and Shih-Ann Chen. Large-scale classification of 12-lead ecg with deep learning. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4, 2019. doi: 10.1109/BHI.2019.8834468.

[48] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91, October 2021.

A Implementation Details

A.1 Foundation Model Backbones

We evaluate the proposed framework using four backbone configurations from three representative ECG foundation-model families: CSFM-Tiny, CSFM-Base, ECG-JEPA, and MERL. These backbones are pretrained on different datasets and with different learning objectives, allowing us to assess the proposed method under diverse representation spaces, model scales, and inductive biases rather than tailoring it to a single architecture.

CSFM [6] is a multimodal physiological foundation model pretrained on large-scale ECG and PPG data, including more than 2 million 12-lead ECGs and 0.25 million paired PPG and single-lead ECG recordings from 1.7 million subjects. Its pretraining corpus includes MIMIC-III-WDB [9, 40], MIMIC-IV-ECG [41], and CODE-Full [42, 43]. We use two CSFM variants in our experiments: **CSFM-Tiny** and **CSFM-Base**. Both variants use a hidden dimension of 768, but differ in model scale: CSFM-Tiny uses 6 transformer layers, 8 attention heads, and an MLP hidden dimension of 1024, while CSFM-Base uses 12 transformer layers, 12 attention heads, and an MLP hidden dimension of 3072.

MERL [32] is an ECG-language foundation model trained using self-supervised and contrastive learning objectives to produce semantically meaningful ECG representations. MERL is pretrained on the MIMIC-IV-ECG dataset [41], which contains more than 0.75 million paired ECG-text samples from approximately 0.15 million subjects. We use the ViT-Tiny variant of MERL, which produces 192-dimensional features.

ECG-JEPA [33] is an ECG-specific self-supervised representation model based on the Joint-Embedding Predictive Architecture (JEPA) paradigm [44]. It is designed for multi-lead ECG representation learning using synchronized temporal masking, a lead-wise predictor, and Cross-Pattern Attention (CroPA). Following the original setup, ECG-JEPA is pretrained on the Chapman [45], Ningbo [46], and CODE-15 [47] datasets. After excluding recordings with missing values, the pretraining corpus contains 43,240 ECGs from Chapman and Ningbo and 130,900 ECGs from CODE-15. In our experiments, we use the multi-block masking variant of ECG-JEPA. The encoder adopts 12-layer transformer architectures with 16 attention heads and hidden dimension 768.

Together, these backbones cover complementary design paradigms, including multimodal physiological modeling, ECG-language representation learning, and predictive self-supervised pretraining, and therefore provide a diverse test bed for evaluating the robustness of the proposed framework.

A.2 Trainable Components

In the proposed framework, only a subset of parameters is updated during training. Specifically, the learnable components include the global positional embedding matrix $\mathbf{E}_{\mathrm{global}}$, the prompt tokens, the projection heads g_{local} and g_{dyn}, and the task-specific linear classification head. All remaining parameters, in particular the backbone encoder weights θ of the pretrained foundation model f_θ, are kept frozen throughout training.

Unless otherwise specified, we use a shared default parameter configuration across most backbone models. The number of prompt tokens is set to 100 for CSFM-Tiny, CSFM-Base, and MERL, and to 50 for ECG-JEPA. Both projection heads (g_{local} and g_{dyn}) are implemented as two-layer MLPs with hidden dimension 1024. For the locality-aware objective $\mathcal{L}_{\mathrm{local}}$, we adopt the negative L_2 distance as the segment-level similarity measure, following [28]. For the dynamics-aware objective, we first compute a segment-level affinity matrix from frozen teacher features using pairwise cosine similarity. Segment-wise affinity scores are then obtained by summing each row of the affinity matrix after excluding self-similarity. Based on these scores, we rank all 10-second segments within the input recording and retain the top 75% segments with the highest aggregate affinity as $\mathcal{I}_{\mathrm{bg}}$, *i.e.*, the most globally consistent windows. The remaining 25% segments are treated as relatively more dynamic and are excluded from consistency learning. The complete training procedure is summarized in Algorithm 1.

Algorithm 1 Long-Horizon Adaptation Framework

Input: Long signal x, frozen encoder f_θ, prompts $\mathbf{P}$, global positional embeddings $\mathbf{E}_{\text{global}}$, projectors g_{local} and g_{dyn}, classifier
Hyperparameters: Steps T, learning rate η, temperature τ, weights λ_1, λ_2
Output: Fine-tuned model f with long-horizon capability

for $t = 1$ to T **do**
  Slice x into 10s segments, extract teacher features $\mathbf{z}_T$ by f_θ ➢ *Locality-aware representation learning*
  Project segment feature $\mathbf{z}_S$ using g_{local}
  Calculate pairwise similarities $\mathbf{S}^T_{i,j}$, $\mathbf{S}^S_{i,j}$
  Calculate $\mathcal{L}_{\text{local}}$ via Eq. 2
  Calculate affinity matrix from $\mathbf{z}_T$, select $\mathcal{I}_{\text{bg}}$ ➢ *Dynamics-aware consistency learning*
  Project background segment features $\mathbf{z}_{\text{bg}}$ using g_{dyn}
  Calculate $\mathcal{L}_{\text{dyn}}$ via Eq. 3
  Predict label using classifier, calculate task loss $\mathcal{L}_{\text{task}}$
  Aggregate loss: $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{local}} + \lambda_2 \mathcal{L}_{\text{dyn}}$
  Update $\mathbf{P}$, $\mathbf{E}_{\text{global}}$, projectors, and classifier using gradient descent with rate η
end for

A.3 Experimental Configurations

The training settings are kept consistent across all ECG datasets. Unless otherwise specified, all models are trained for 60 epochs using AdamW with an initial learning rate of 5×10^{-4}, a cosine annealing schedule, and 5 warmup epochs. We optimize all experiments using asymmetric loss [48]. The default batch size is 8 per GPU, and is reduced to 4 only for experiments involving both ECG and PPG signals due to the higher memory usage of multimodal inputs. All experiments are conducted on 4 NVIDIA GeForce RTX 5090 GPUs with synchronized batch normalization. Code is available at `https://github.com/wetang7/Extending_ECG_FM`.

B Additional Main Results

B.1 Full Variable-Length Evaluation Results

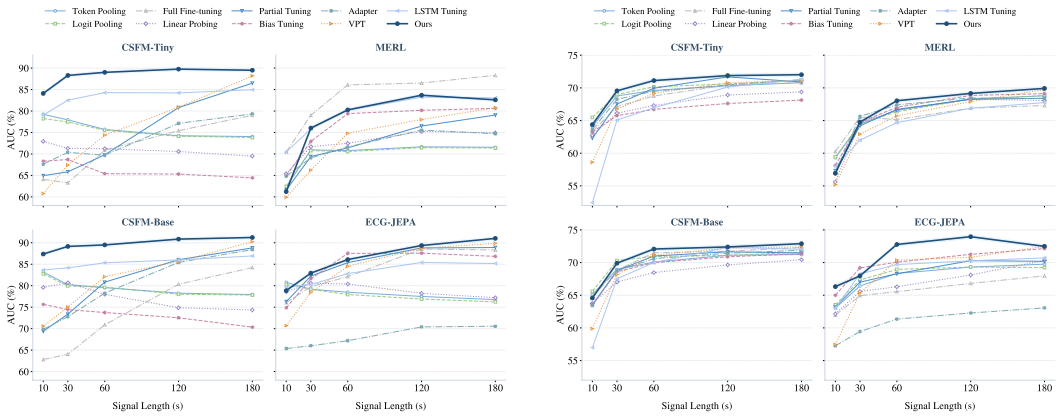
This section provides the complete per-length evaluation results across all inference durations. While the main paper reports the results at 180 s together with the average across durations, the full curves here allow a more direct view of how each method behaves under different test length. In all settings, models are trained on 3-minute recordings and evaluated on variable inference durations without retraining.

Figure A1 shows that the proposed method remains strong across a wide range of input lengths, while also revealing task- and backbone-specific differences in how performance changes with varying temporal context. In particular, these curves complement the main-table summary by showing not only the overall ranking of methods, but also their sensitivity to inference duration.

B.2 Full Results Across Input Modalities

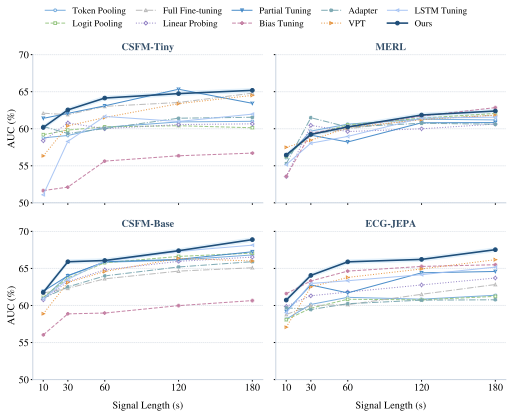
Table A1 reports the full per-length results of CSFM-Tiny on VTaC under three input modality settings: ECG-only, PPG-only, and ECG+PPG. Across all three settings, the proposed method remains among the strongest approaches, with particularly clear gains in the ECG-only and ECG+PPG settings. In the multimodal setting, it achieves the best performance at 30 s, 60 s, 120 s, and 180 s, indicating that the proposed extension strategy remains effective when longer-horizon modeling is applied to jointly available ECG and PPG signals.

The modality-specific results also reveal a consistent pattern in absolute performance. ECG+PPG generally yields the strongest results overall, followed by ECG-only, while PPG-only remains more

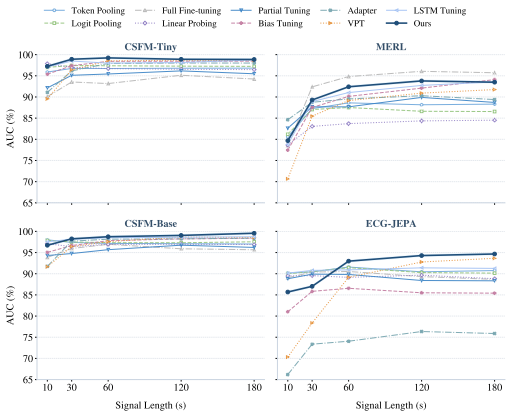


(a) VTaC

(b) MC-MED (ED)



(c) MC-MED (Acuity)



(d) CPSC2021

Figure A1: Per-length adaptation performance across four downstream tasks. Each subplot reports the performance trend from 10s to 180s for all adaptation strategies under different pretrained ECG foundation models. Our method is highlighted with a thicker dark-blue curve.

challenging for all methods. Empirically, these results suggest that learning information from PPG alone is more challenging than from ECG, although both modalities still benefit from the proposed adaptation strategy.

B.3 Additional Temporal Baselines on VTaC

A natural question is whether the gains of our method can already be recovered by applying standard temporal models to snapshot features. To test this, we compare our framework against three additional baselines on VTaC: a ResNet1D-50 model, a deeper 3-layer LSTM applied to concatenated 10-second CSFM-Tiny feature sequences, and an attention-based multi-instance learning (MIL) model operating on segment-level features. For the ResNet1D-50 and LSTM baselines, each 3-minute recording is first partitioned into 10-second segments, each segment is encoded into a 768-dimensional CSFM-Tiny feature, and the resulting sequence of 18 segment-level features is concatenated in temporal order into a one-dimensional feature sequence for downstream modeling. For MIL, the segment-level features are instead aggregated using a learned attention mechanism.

Table A2 shows that all three baselines remain below the proposed method across the full range of inference durations. Among them, the attention-based MIL baseline is the most competitive, but it still trails our method at every duration. These results indicate that simple temporal modeling over frozen snapshot features, even when paired with learned attention-based aggregation, is insufficient.

Table A1: **Detailed adaptation results of CSFM-Tiny under different input modality configurations on VTaC.** We report AUC (%) across five input signal lengths (10s–180s) for three settings: **ECG-only**, **PPG-only**, and **ECG+PPG**. **Best** is in bold and second best is underlined.

Method	ECG-only					PPG-only					ECG + PPG				
	10s	30s	60s	120s	180s	10s	30s	60s	120s	180s	10s	30s	60s	120s	180s
Snapshot Aggregation															
Logit Pooling	78.23	77.45	75.51	74.14	73.83	66.92	65.97	63.56	62.70	62.41	83.09	80.43	77.38	75.80	74.88
Token Pooling	79.23	77.94	75.71	74.25	74.05	65.82	65.47	64.17	63.74	63.13	82.93	78.54	75.40	73.82	72.75
Fine-tuning-based Adaptation															
Full Fine-tuning	64.10	63.26	70.89	75.42	79.01	61.52	61.20	72.14	76.69	78.73	62.36	62.36	70.54	79.38	82.98
Linear Probing	72.93	71.30	71.22	70.57	69.50	70.50	65.47	63.66	62.96	61.18	81.74	76.30	71.76	69.17	68.04
Partial Tuning	64.94	65.84	69.80	80.79	86.49	65.77	71.98	74.03	80.69	**85.41**	77.91	76.90	81.19	87.87	91.06
Bias Tuning	68.32	68.70	65.40	65.32	64.43	71.47	73.96	72.69	72.12	72.02	60.54	67.38	74.51	77.11	80.80
Adapter	67.63	70.37	69.70	77.12	79.33	67.31	67.96	64.50	62.58	61.23	81.82	76.55	74.03	72.20	71.61
VPT	60.78	67.44	74.37	80.86	88.18	72.28	70.19	70.32	76.17	81.44	57.78	65.65	73.86	87.48	88.96
LSTM Tuning	79.04	82.53	84.28	84.22	84.97	74.36	76.99	80.42	81.47	82.04	**84.18**	91.36	90.41	87.91	87.33
Ours	**84.10**	**88.29**	**89.01**	**89.76**	**89.50**	**76.36**	**81.45**	**83.50**	**83.97**	84.37	83.61	**92.26**	**92.53**	**92.94**	**91.81**

Table A2: **Additional temporal baselines on VTaC.** We compare the proposed method against ResNet1D-50, a 3-layer LSTM, and an attention-based multi-instance learning (MIL) baseline built on top of 10-second CSFM-Tiny features. Results are reported as AUC (%).

Method	10s	30s	60s	120s	180s
ResNet1D-50	54.01	68.20	64.16	60.87	73.67
3-layer LSTM	67.97	72.42	74.30	73.54	72.73
Attention-based MIL	80.01	80.31	79.56	77.37	77.40
Ours	**84.10**	**88.29**	**89.01**	**89.76**	**89.50**

B.4 Patient-Level Bootstrap Confidence Intervals

We quantify uncertainty estimates for comparisons on VTaC and MC-MED (ED) using patient-level bootstrap resampling. In each resample, test-set patients are sampled with replacement, with all recordings and segments from the same patient kept together. We recompute AUC in each of $5{,}000$ resamples while keeping the trained model fixed, and construct the confidence interval from the 2.5th and 97.5th percentiles of the resulting AUC estimates. Table A3 reports the mean bootstrap AUC together with the 95% percentile bootstrap interval defined above.

B.5 Evaluation at Additional Input Durations

Variable-length support refers to using a single model adapted at a maximum duration L to process different input durations up to L without further training. In the main experiments, $L = 180$ s and the evaluation durations are 10, 30, 60, 120, and 180 s. To examine performance at additional durations within this trained horizon, we evaluate the models adapted at 180 s on 15, 45, and 90 s inputs from VTaC (Table A4). We use the patient-level bootstrap protocol and reporting convention in Appendix B.4.

B.6 Extension Beyond 3 Minutes

We further evaluate whether the proposed framework remains effective when the temporal horizon extends beyond the 3-minute setting used in the main experiments. To this end, we conduct an additional stress test on VTaC by training models on longer recordings up to 5 minutes.

As shown in Table A5, the proposed method maintains stable performance across 1- to 5-minute test inputs, whereas full fine-tuning remains substantially weaker in this setting. These results suggest that the proposed adaptation strategy does not merely overfit to a fixed 3-minute horizon, but continues to provide effective long-horizon modeling when the recording duration is further increased.

B.7 Parameter Efficiency Across Backbones

For completeness, we also report the number of trainable parameters required by the proposed method and full fine-tuning across different backbone models. Compared with full fine-tuning, the

Table A3: **Patient-level bootstrap evaluation.** Entries report mean bootstrap AUC [95% CI] from 5,000 resamples. Δ is our mean bootstrap AUC minus the baseline mean in percentage points.

	VTaC		MC-MED (ED)	
Method	Mean AUC [95% CI]	Δ	Mean AUC [95% CI]	Δ
CSFM-Tiny				
Linear Probing	71.54 [57.14, 77.66]	+16.04	69.72 [67.23, 72.20]	+2.28
Full Fine-tuning	81.31 [75.36, 86.93]	+6.27	70.09 [68.59, 73.71]	+1.91
VPT	85.69 [78.78, 90.59]	+1.89	69.57 [67.11, 71.86]	+2.43
Logit Pooling	70.24 [55.48, 75.09]	+17.34	70.17 [66.19, 72.66]	+1.83
Ours	87.58 [82.82, 91.37]	–	72.00 [68.74, 73.15]	–
CSFM-Base				
Linear Probing	74.89 [65.34, 83.06]	+17.06	70.44 [67.80, 72.77]	+2.35
Full Fine-tuning	82.66 [75.11, 89.63]	+9.29	70.30 [69.02, 74.80]	+2.49
VPT	89.83 [85.13, 94.46]	+2.12	71.59 [69.95, 74.74]	+1.20
Logit Pooling	73.92 [61.48, 79.68]	+18.03	71.22 [70.06, 73.43]	+1.57
Ours	91.95 [88.08, 95.53]	–	72.79 [70.49, 75.42]	–
ECG-JEPA				
Linear Probing	77.21 [70.71, 83.32]	+13.62	70.85 [68.34, 73.32]	+2.20
Full Fine-tuning	88.68 [84.05, 92.79]	+2.15	68.31 [65.57, 71.38]	+4.74
VPT	89.39 [85.08, 95.18]	+1.44	72.11 [66.81, 73.50]	+0.94
Logit Pooling	74.12 [64.07, 78.10]	+16.71	68.91 [65.15, 72.79]	+4.14
Ours	90.83 [86.21, 95.03]	–	73.05 [67.32, 74.47]	–
MERL				
Linear Probing	72.97 [69.57, 81.26]	+8.26	68.05 [65.62, 70.58]	+2.16
Full Fine-tuning	87.46 [82.24, 92.23]	−6.23	68.33 [65.73, 70.86]	+1.88
VPT	82.83 [75.96, 86.39]	−1.60	69.38 [66.04, 71.78]	+0.83
Logit Pooling	71.39 [67.21, 72.32]	+9.84	68.12 [65.44, 70.12]	+2.09
Ours	81.23 [77.79, 87.74]	–	70.21 [66.61, 72.52]	–

Table A4: **Evaluation at additional input durations on VTaC.** Models are adapted at 180 s and evaluated at 15, 45, and 90 s without retraining.

Method	15 s AUC [95% CI]	45 s AUC [95% CI]	90 s AUC [95% CI]
CSFM-Tiny			
Logit Pooling	75.48 [69.59, 81.71]	76.45 [71.12, 82.31]	76.70 [71.97, 82.32]
Linear Probing	71.27 [63.83, 78.87]	70.54 [61.68, 79.44]	71.69 [58.27, 78.25]
Ours	79.81 [71.40, 88.27]	85.42 [75.68, 91.15]	86.19 [76.74, 91.50]
CSFM-Base			
Logit Pooling	78.40 [71.58, 84.99]	78.50 [69.57, 84.50]	77.74 [70.65, 80.11]
Linear Probing	79.59 [74.13, 84.82]	78.73 [71.38, 85.63]	76.33 [68.63, 83.38]
Ours	83.56 [74.05, 88.93]	86.25 [78.72, 90.33]	88.38 [83.38, 93.69]
ECG-JEPA			
Logit Pooling	79.35 [74.14, 83.96]	76.43 [70.21, 82.19]	77.47 [71.22, 83.21]
Linear Probing	77.67 [75.62, 86.92]	79.71 [73.60, 85.64]	80.82 [72.64, 84.48]
Ours	79.60 [74.38, 83.74]	83.30 [77.89, 88.37]	86.47 [81.93, 90.69]
MERL			
Logit Pooling	62.23 [56.95, 67.86]	68.77 [62.59, 74.95]	70.44 [65.17, 76.51]
Linear Probing	67.14 [60.34, 77.34]	73.76 [67.83, 79.95]	74.15 [68.10, 80.23]
Ours	64.41 [60.68, 78.47]	74.98 [68.54, 81.73]	80.51 [75.05, 86.05]

Table A5: **Extension beyond 3 minutes on VTaC.** We compare full fine-tuning and the proposed method on CSFM-Tiny when the temporal horizon is extended from 1 to 5 minutes. Results are reported as AUC (%).

Method	1 min	2 min	3 min	4 min	5 min
Full Fine-tuning	66.71	66.86	73.00	75.78	78.07
Ours	**85.55**	**86.24**	**85.97**	**86.09**	**85.45**

Table A6: **Cumulative parameter storage across four tasks.** Total stored parameters are normalized by the size of one CSFM-Base backbone.

CSFM-Base (157.39 M)		Total params	Scope: Input	Scope: Backbone	Scope: Task head	Extra modules
Total # of tasks	**4**					
(a)	Full Fine-tuning	4.00×		✓	✓	
(b)	Linear Probing	1.00×			✓	
	Partial Tuning	1.21×		✓	✓	
	Bias Tuning	1.01×		✓	✓	
(c)	Adapter	1.06×		✓	✓	✓
	VPT	1.09×	✓		✓	
	Ours	**1.11×**	✓		✓	✓

proposed method reduces the trainable parameter count from 91,270,969 to 3,891,201 for CSFM-Tiny, from 5,473,152 to 3,891,201 for MERL, from 157,389,625 to 4,352,001 for CSFM-Base, and from 85,374,720 to 3,891,201 for ECG-JEPA.

Table A6 reports the cumulative parameter storage across four tasks, normalized by the size of one CSFM-Base backbone (157.39 million parameters). Full fine-tuning stores a separate adapted backbone for each task, while parameter-efficient methods share the pretrained weights and store task-specific updates or additional components separately. The "Scope" columns indicate where adaptation is applied, and "Extra modules" denotes the presence of additional modules.

C Additional Ablations

C.1 Hyperparameter Sensitivity

We first study the sensitivity of the proposed framework to several key hyperparameters on VTaC using the CSFM-Tiny backbone. Table A7 summarizes the effect of varying the prompt length, the temperature τ used in $\mathcal{L}_{\text{local}}$, and the weighting coefficients of the two auxiliary losses across five input durations ranging from 10 seconds to 180 seconds. Based on the best average performance, we use prompt length $= 100$, temperature $\tau = 0.1$, and loss weights $(\lambda_1, \lambda_2) = (0.1, 1)$ as the default setting in all experiments unless otherwise specified.

We further examine the sensitivity of the proposed framework to the foreground ratio used in the foreground-background partition. Figure A2 shows that a moderate foreground ratio of 25% yields the best overall performance across inference durations on VTaC, while both smaller and larger ratios lead to weaker results. We use 25% as the default setting for our experiments.

C.2 Locality-Aware Contrastive Variants

The purpose of the locality-aware objective is to preserve the pretrained 10-second segment semantics of the foundation model during long-horizon extension, rather than to relearn local representations from scratch. Importantly, these pretrained representations are expected to be general rather than label-collapsed: even when two segments share the same downstream label, they may still differ due to subject-specific physiological variation. A desirable extension objective should therefore preserve this subject-induced variability when it is already encoded by the foundation model, instead of imposing overly restrictive contrastive structure.

Table A7: **Hyperparameter sensitivity on VTaC using CSFM-Tiny.** Results are reported as AUC (%) across different input durations. The default configuration adopted in the main experiments is shown in bold.

Group	Value	10s	30s	60s	120s	180s	Avg.
Prompt Length	1	78.19	80.06	79.86	78.57	78.91	79.12
	50	77.41	86.08	87.61	89.00	89.02	85.82
	100	**84.10**	**88.29**	**89.01**	**89.76**	**89.50**	**88.13**
	200	81.98	84.43	85.18	86.13	85.83	84.71
Temperature τ	0.05	82.18	87.95	89.12	89.80	89.19	87.65
	0.1	**84.10**	**88.29**	**89.01**	**89.76**	**89.50**	**88.13**
	0.5	81.76	86.52	87.29	88.38	87.94	86.38
	1.0	79.69	87.21	88.69	90.28	89.78	87.13
Loss Weights (λ_1, λ_2)	(1, 1)	79.12	84.98	86.01	85.58	85.08	84.15
	(1, 2)	82.14	86.10	87.43	87.11	86.62	85.88
	(0.1, 1)	**84.10**	**88.29**	**89.01**	**89.76**	**89.50**	**88.13**
	(0.1, 2)	82.46	86.97	87.78	87.76	86.75	86.34

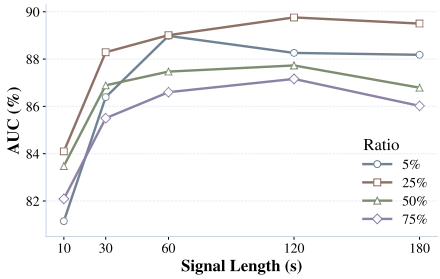


Figure A2: **Foreground ratio sensitivity on VTaC using CSFM-Tiny.** Results are reported as AUC (%) across five inference durations.

To examine this design choice, we compare three contrastive variants on MC-MED using CSFM-Tiny while keeping all other components fixed: standard batch-level contrastive learning, intra-subject contrastive learning, and inter-subject contrastive learning. For the inter-subject variant, segments are first shuffled within each subject, and contrastive matching is then performed across subjects without requiring positional alignment. This comparison allows us to test whether additional subject-aware pairing structure improves extension, or whether it instead interferes with the general segment-level representation bias learned by the frozen foundation model.

As shown in Table A8, standard batch-level contrastive learning achieves the strongest overall performance across nearly all inference durations. Intra-subject contrastive learning performs consistently worse, while inter-subject contrastive learning is slightly more competitive but still does not surpass the batch-level variant overall. Although the differences become smaller at longer durations, the overall pattern remains stable.

These results suggest that preserving the pretrained segment representations does not require additional subject-based contrastive constraints. A simple batch-level strategy is already sufficient to guide the extended model toward the teacher-defined local similarity structure. In contrast, imposing extra subject-aware pairing rules may introduce unnecessary bias and make it harder to preserve the general local representations learned by the foundation model.

Table A8: **Locality-aware contrastive variants on MC-MED (ED) using CSFM-Tiny.** We compare standard batch-level contrastive learning with intra-subject and inter-subject contrastive variants. Results are reported as AUC (%) across different inference durations. The best result in each column is shown in bold.

Contrastive Strategy	10s	30s	60s	120s	180s
Batch Contrast	**64.38**	**69.55**	**71.12**	**71.88**	72.02
Intra-subject Contrast	58.96	67.88	69.46	71.12	71.93
Inter-subject Contrast	59.93	66.48	70.49	71.57	**72.04**

D Assumption Analysis and Case Studies

D.1 Complementarity of the Two Segment Groups

We test whether the two segment groups identified by the affinity-based partition contain complementary predictive information, rather than one group being sufficient on its own. To this end, each long recording is partitioned into a high-affinity group and a low-affinity group according to the affinity ranking derived from frozen teacher features. We then construct three recording-level representations by averaging segment features from the high-affinity group only, the low-affinity group only, or both groups jointly, and train a separate linear classification head for each representation.

Table A9 summarizes the results across VTaC, MC-MED (ED), and CPSC under different partition ratios. Two consistent patterns emerge. First, combining both groups yields the strongest performance across all datasets and all ratios. Second, the high-affinity-only representation remains meaningfully predictive in every setting, indicating that this group retains useful information even when the dynamic segments are excluded.

Table A9: **Complementarity of the two affinity-based segment groups across datasets.** Each recording is partitioned into a high-affinity group and a low-affinity group according to affinity ranking under different partition ratios. Separate linear heads are trained on representations aggregated from the high-affinity group only (*high-only*), the low-affinity group only (*low-only*), and both groups jointly (*high+low*). All experiments are conducted on CSFM-Tiny using 3-minute ECG sequences for both training and testing. Results are reported as AUC (%).

	VTaC				MC-MED (ED)				CPSC			
Strategy	**10% low**	**25% low**	**50% low**	**75% low**	**10% low**	**25% low**	**50% low**	**75% low**	**10% low**	**25% low**	**50% low**	**75% low**
high-only	67.93	67.25	66.58	63.37	69.06	68.93	68.99	67.28	96.38	94.53	90.15	87.06
low-only	71.98	71.73	72.03	72.49	68.44	69.39	69.85	70.02	96.58	96.25	96.66	96.53
high+low	**74.05**	**74.05**	**74.05**	**74.05**	**70.80**	**70.80**	**70.80**	**70.80**	**96.81**	**96.81**	**96.81**	**96.81**

At the same time, the results also show that the two groups contribute differently. In general, the low-affinity-only representation is more predictive than the high-affinity-only representation, suggesting that the two groups capture different aspects of the recording. However, the consistent advantage of using both groups jointly over using either group alone shows that the most effective long-horizon representation is obtained by retaining information from both sides of the partition. We therefore interpret the affinity-based partition not as a separation between informative and uninformative segments, but as a practical decomposition of heterogeneous signal content.

D.2 Case Studies of the Affinity-Based Segment Partition

This section visualizes representative long ECG recordings to illustrate how the proposed affinity-based partition separates each recording into two segment groups over time. These examples complement the quantitative results in Appendix D.1 by providing a qualitative view of the resulting partition. In particular, they show that the selected segments are not distributed uniformly over time, and the partition is structured rather than arbitrary.

E Limitations and Future Work

This work studies how to extend snapshot ECG foundation models beyond their original 10-second operating regime. Our experiments focus on the 3-minute setting, rather than arbitrarily long inputs, as this horizon is already relevant for many practical clinical scenarios, including settings where rare events may occur; we further complement it with additional results up to 5 minutes. Together, these results support the central claim of the paper: pretrained 10-second ECG foundation models can be adapted to longer recordings without full retraining.

A remaining limitation is that we do not fully characterize the maximal temporal horizon supported by the proposed framework. In principle, the method can be extended further, but the practical upper bound will depend on computational budget, memory constraints, and the efficiency of the adaptation design. Exploring more memory-efficient strategies for even longer-horizon extension, such as token merging or other sequence compression methods, is an important direction for future work.

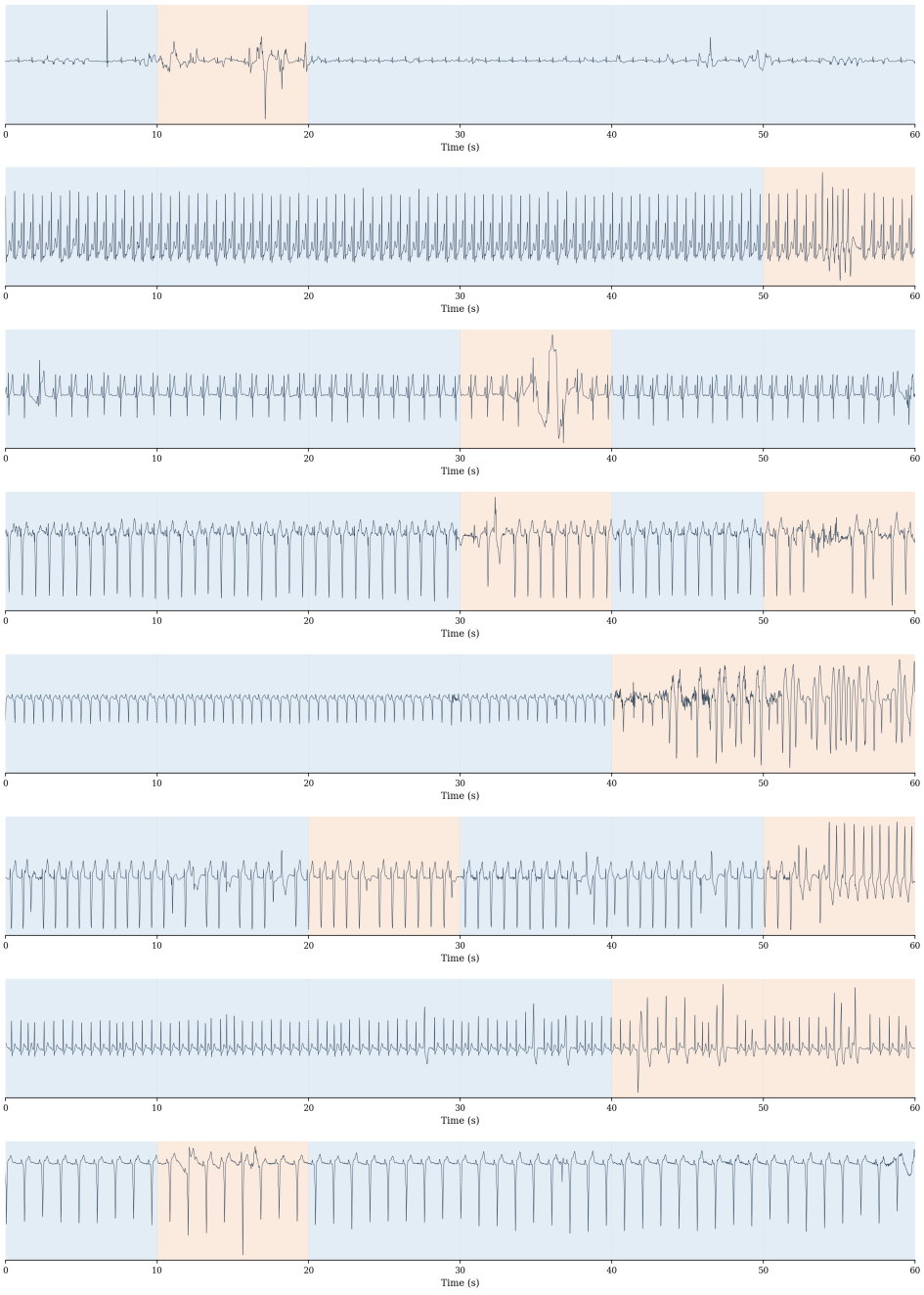


Figure A3: **Case visualization of ECG recordings with two segment groups identified by CSFM-Base latent affinity.** Blue and orange shaded regions indicate the higher-affinity and lower-affinity groups, respectively.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe the main contribution as adapting pretrained 10-second ECG foundation models to longer recordings without full retraining, and this scope matches the empirical study in the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations of our work and future directions in Appendix E in details.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: the paper does not include theoretical results that need theorems or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We discloses the information needed to reproduce the main experimental results, including the datasets and task definitions, evaluation setup, and pretrained backbone configurations in Section 4. Appendix A further provides implementation details and the complete training procedure.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will release the source code upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed experimental settings can be found in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Given the scale of the experimental study across several foundation-model backbones, we were unable to allocate sufficient computational resources to run all experiments three or five times for error-bar reporting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the computer resources used for the experiments in Appendix A.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper focuses on technical advances in representation adaptation for physiological time series and does not introduce a new deployment setting, decision policy, or real-world intervention. We therefore do not include a dedicated broader impact discussion beyond the general considerations that already apply to machine learning methods for medical signal analysis.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release a new high-risk generative model, scraped dataset, or similarly misuse-prone artifact of the kind referenced in this question. It studies adaptation methods for physiological signal models using existing datasets, so this item is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original sources of the datasets and pretrained backbone models used in the study. To the best of our knowledge, these assets were used in a manner consistent with their stated terms of use and intended research setting.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release the code for the proposed framework upon acceptance. The release will be accompanied by documentation describing the method, implementation details, and instructions for reproducing the main results.

    Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not include crowdsourcing experiments nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not include crowdsourcing experiments nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Any use of LLMs in this work was limited to writing and presentation assistance, such as grammar checking, editing, and formatting.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.