EPISODIC CONTEXT MEMORY: PATIENT-STATE ADAPTATION FOR LONGITUDINAL CARDIOVASCULAR MONITORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Physiological classifiers trained on short, independently sampled recordings are often deployed on continuous streams from unseen individuals, where observations become temporally correlated and patient-specific. We study this mismatch as longitudinal personalization and introduce Episodic Context Memory (ECM), a stateful adaptation framework that keeps a pretrained population predictor fixed while learning how to construct and retrieve patient-specific context. Using labeled streams from separate calibration patients, ECM learns shared rules for writing preceding observations into a compact multi-slot memory and retrieving context relevant to the current query. At deployment on an unseen patient, only the memory state evolves, without target-patient labels or gradient updates. We evaluated atrial fibrillation (AF) detection from the electrocardiogram (ECG) on patient-disjoint internal and external cohorts against supervised fine-tuning, test-time adaptation, prototype methods, recurrent conditioning, and neural memory. Under external dataset shift, the primary ECM variant increased AUROC by +1.1% and decreased AF burden error by -1.2% relative to the best comparator, and an AF-specific extension increased macro-F1 by +2.7% and AUROC by +2.1%. In longitudinal analyses, ECM approaches achieved the highest macro-F1 improvement of up to +15.1%. Ablations identified patient-specific context and query-conditioned retrieval as the main contributors, while evaluations on additional cardiovascular tasks and sensing modalities provided promising evidence beyond AF detection. These results show that patient-state adaptation can provide robust, label-free personalization for longitudinal deployment under distribution shift.

1 INTRODUCTION

Neural classifiers for cardiovascular disease detection from physiological signals are commonly trained on large collections of independent, identically distributed (i.i.d.), short, and often class-balanced recordings. Clinical deployment instead produces hours- to days-long streams from individual patients, in which adjacent windows are correlated and event prevalence may be low and time-varying (Alqudah & Moussavi, 2025; Lee et al., 2024). This creates a hierarchical distribution shift: acquisition hardware and recording conditions vary across datasets and over time (Bläsing et al., 2022; Ansari et al., 2024), patient demographics and rhythm composition differ across populations (Biton et al., 2023), and inter-patient variability limits the generalization of population-trained models (Ding et al., 2024). Consequently, the high classification accuracy achieved on i.i.d. short windows may not translate into reliable longitudinal monitoring.

Existing strategies address this challenge only partially: domain generalization does not adapt to the deployed patient, fine-tuning and test-time adaptation (TTA) modify population-model parameters, and static references or single-state context models offer limited representations of heterogeneous patient history. This raises the question: how can a frozen population model exploit recurring patient-specific states to improve predictions during longitudinal monitoring?

We introduce Episodic Context Memory (ECM), an instantiation of patient-state adaptation that separates population knowledge, shared adaptation rules, and patient-specific state. We use a frozen

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

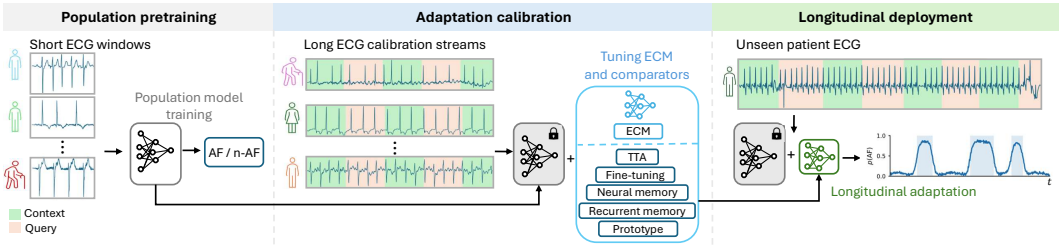


Figure 1: Overview of our proposed approach for longitudinal adaptation of physiological classifiers. A population classifier is pretrained on short signal windows and then frozen. ECM is calibrated on longitudinal recordings from separate patients. At deployment, ECM uses preceding patient context to adapt predictions for unseen patients and is evaluated against competing adaptation methods.

atrial fibrillation (AF) detector pretrained on short electrocardiogram (ECG) windows as a population model. Using labeled streams from separate calibration patients, ECM learns shared rules for writing preceding observations into a compact multi-slot memory, retrieving query-relevant patient context, and applying a residual prediction correction. At deployment on a new patient, only the patient state evolves without gradient updates. Fig. 1 schematizes our approach. Our contributions are the following:

1. We introduce ECM, a neural architecture for longitudinal biosignal personalization through an evolving patient-specific latent memory. ECM learns to accumulate and retrieve patient context to condition a frozen population predictor, enabling causal, label-free deployment.
2. Across multiple ECG datasets, we benchmark ECM on binary AF detection against supervised fine-tuning, source-free TTA, prototype-based methods, recurrent conditioning, and neural memory. We show that patient-state adaptation is more robust under external dataset shift, where ECM achieves the best overall performance. We further evaluate ECM on PPG–ECG hypotension detection and four-class ECG ectopy classification, demonstrating its benefits beyond AF detection from ECG.
3. Through controlled ablations and memory analyses, we identify patient-specific context and query-conditioned retrieval as key contributors to ECM's gains, and characterize how the learned memory stores and uses context over extended monitoring horizons.

2 Related Work

Deep learning for AF detection from the ECG. AF is the most common sustained cardiac arrhythmia and can occur in paroxysmal and asymptomatic episodes (Brundel et al., 2022; Linz et al., 2024). Neural AF detectors are commonly developed from short ECG windows (Hammer et al., 2025; Si et al., 2025; Kraft & Rumm, 2025), whereas ambulatory and wearable systems are intended to monitor patients continuously (Francisco et al., 2025). Long-term use therefore requires not only segment discrimination, but also accurate AF burden estimation across patient-specific morphology and recording conditions (Doehner et al., 2025). Other studies approached long-term AF detection by training directly on large datasets of continuous Holter recordings (Ben-Moshe et al., 2022; 2024). Nevertheless, these studies rarely consider how a population detector trained on short, i.i.d. recordings should adapt to previously unseen patients during longitudinal deployment.

Test-time and domain adaptation. Physiological TTA has been studied for cross-subject electroencephalogram (EEG) analysis (Wimpff et al., 2024; Peng et al., 2025; Li et al., 2026b; Lee et al., 2026), as well as for sleep staging and electromyogram gesture recognition (Jia et al., 2024; Touko et al., 2026). In continuous patient streams, online adaptation has been applied to seizure prediction and personalized blood pressure estimation (Mao et al., 2023; Jo et al., 2025), and in combination with retrieved cross-patient examples for hypotension prediction (Li et al., 2026a). However, parameter adaptation and reference retrieval alone do not explicitly model evolving patient-specific context while preserving a frozen population predictor.

Supervised and few-shot patient personalization. Deep patient-specific ECG classifiers combine population data with labeled target-patient beats or initial recording segments (Kiranyaz et al., 2016; Luo et al., 2017). Such methods can directly capture subject-specific characteristics, but require labeled target-patient data.

Patient-reference and memory-based conditioning. Patient-reference approaches condition predictions on representative ECG waveforms or expert-selected normal beats (Kim et al., 2024; An et al., 2025), while distributionally adaptive ECG models address intra- and inter-patient variability (Wu et al., 2026). Neural memory has also been incorporated into EEG seizure forecasting (Pham & Tran, 2026). Fixed references, however, do not evolve with the deployed patient stream, while end-to-end memory architectures typically learn memory jointly with the predictive model.

Long-horizon physiological modeling. A closely related work by Tang et al. (2026) extended pretrained 10-s ECG foundation models to variable-length inputs. Their focus is classification of extended ECG inputs, whereas ours is sequential prediction over hours-to-days patient streams.

The limitations of previous approaches motivate patient-state adaptation: conditioning a frozen population predictor on evolving patient context without target-patient labels or deployment-time parameter updates.

3 Background

Physiological classifier training. Let $\mathcal{X} \subseteq \mathbb{R}^{C\times L}$ denote physiological signal windows with C channels and L samples per channel, and let $\mathcal{Y} = \{1, \dots, M\}$ denote the label space for M classes. A population model comprises an encoder $f_\phi : \mathcal{X} \to \mathbb{R}^d$ and a classifier $g_\omega : \mathbb{R}^d \to \mathbb{R}^M$, producing class probabilities $q_\theta(\boldsymbol{x}) = \mathrm{softmax}(g_\omega(f_\phi(\boldsymbol{x})))$, where $\theta = (\phi, \omega)$. Given $\mathcal{D}_\mathrm{s} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, with source distribution $P_\mathrm{s}(X, Y)$, the population model is trained by minimizing a supervised classification loss over short labeled windows, yielding shared parameters $\hat{\theta} = (\hat{\phi}, \hat{\omega})$.

Clinical longitudinal deployment. Clinical deployment involves ordered patient trajectories $\mathcal{T}^{(p)} = \{(\boldsymbol{x}_n^{(p)}, y_n^{(p)})\}_{n=1}^{T_p}$, where T_p is the number of windows, and reference labels are unavailable to the model during deployment. Adjacent windows may be dependent, and physiological states can persist, change, or recur within a recording. Let a index an acquisition domain and $P_{a,n}(X, Y)$ denote the joint distribution among patients observed at trajectory position n in that domain. Differences in acquisition conditions, patient composition, and window sampling can produce a distribution shift between source and acquisition domains: $P_{a,n}(X, Y) \neq P_\mathrm{s}(X, Y)$.

Test-time adaptation and personalization. To address the distribution shift at deployment, TTA modifies a pretrained model using unlabeled deployment observations, without accessing source training data (Wang et al., 2021a). At deployment update b, let $\mathcal{B}_b$ denote the currently observed batch of unlabeled windows. A gradient-based TTA update can be written as

$$\theta_b = \theta_{b-1} - \eta_b \nabla_\theta \mathcal{L}_b(\theta; \mathcal{B}_b)\big|_{\theta=\theta_{b-1}}, \qquad \theta_0 = \hat{\theta}, \tag{1}$$

where η_b is the step size and $\mathcal{L}_b$ is an unsupervised adaptation objective. TTA approaches differ in their supervision and update rules, but generally adapt the parameters of the predictor rather than the patient-specific information used to condition its predictions. We investigate the latter regime, learning shared adaptation rules from separate calibration patients and updating only an explicit patient state during deployment.

4 Episodic Context Memory

Patient-state adaptation. We separate the parameters learned across patients from the state maintained for an individual patient. The population encoder $f_{\hat{\phi}}$ and classifier $g_{\hat{\omega}}$ remain frozen, while shared adaptation parameters ψ are learned separately, yielding $\hat{\psi}$. Starting from an initial state $\boldsymbol{S}_0$, K preceding context windows $\boldsymbol{x}_t^c$ update the state sequentially:

$$\boldsymbol{S}_t = U_{\hat{\psi}}\left(\boldsymbol{S}_{t-1}, f_{\hat{\phi}}(\boldsymbol{x}_t^c)\right), \qquad t = 1, \dots, K. \tag{2}$$

A subsequent query window $\boldsymbol{x}_j^q$ is classified using the resulting state $\boldsymbol{S}_K$:

$$\boldsymbol{p}_j = H_{\hat{\psi}}\left(\boldsymbol{x}_j^q, \boldsymbol{S}_K; f_{\hat{\phi}}, g_{\hat{\omega}}\right). \tag{3}$$

Here, U_ψ denotes the state-update function and H_ψ the state-conditioned prediction function, which returns probabilities over the M classes. We refer to this regime as patient-state adaptation.

Episodic calibration. ECM implements U_ψ through a learned memory write and H_ψ through a query-conditioned memory read followed by a residual correction to the frozen classifier's input. During calibration, each contiguous episode e is divided into K context windows followed by Q query windows. The frozen source encoder maps them to context embeddings $\boldsymbol{h}_t^c = f_{\hat{\phi}}(\boldsymbol{x}_t^c) \in \mathbb{R}^d$ and query embeddings $\boldsymbol{h}_j^q = f_{\hat{\phi}}(\boldsymbol{x}_j^q) \in \mathbb{R}^d$. Here, t indexes context windows and j query windows. Context embeddings construct the patient state, which then conditions prediction of the query. A query is predicted before it can enter the context of a later episode and does not update the memory used for its own prediction. Only the shared ECM parameters ψ are calibrated, while the source encoder and classifier remain frozen. Fig. 2 provides a schematic overview of ECM.

Context-conditioned memory write. ECM represents a patient's state after the first t context windows by a multi-slot memory $\boldsymbol{S}_t = [\boldsymbol{s}_{t,1}, \ldots, \boldsymbol{s}_{t,R}]^\top \in \mathbb{R}^{R\times d}$ of R slots. Within an episode, we abbreviate $\boldsymbol{S}_t^{(e)}$ as $\boldsymbol{S}_t$; independent calibration episodes start from learnable $\boldsymbol{S}_0$. At step t, the current context embedding $\boldsymbol{h}_t^c$ is compared with every previous slot $\boldsymbol{s}_{t-1,i}$. The learned write-query and write-key projections $\boldsymbol{W}_q^{\mathrm{w}}$ and $\boldsymbol{W}_k^{\mathrm{w}}$ map them into a common d-dimensional space. Scaled dot-product attention then produces the write-attention weights $a_{t,i}$:

$$a_{t,i} = \mathrm{softmax}_i\left(\frac{\langle \boldsymbol{W}_q^{\mathrm{w}}\boldsymbol{h}_t^c, \boldsymbol{W}_k^{\mathrm{w}}\boldsymbol{s}_{t-1,i}\rangle}{\sqrt{d}}\right). \tag{4}$$

Each scalar $a_{t,i}$ therefore measures how strongly the current context is routed to slot i. To determine what that slot should contain after observing the context, a write $\mathrm{MLP}_{\mathrm{w}} : \mathbb{R}^{2d} \to \mathbb{R}^d$ receives the concatenation of the previous slot and the context embedding and proposes the candidate content

$$\widetilde{\boldsymbol{s}}_{t,i} = \mathrm{MLP}_{\mathrm{w}}([\boldsymbol{s}_{t-1,i}; \boldsymbol{h}_t^c]). \tag{5}$$

A separate learned map $\boldsymbol{W}_g \in \mathbb{R}^{d\times 2d}$ followed by a sigmoid σ produces the feature-wise write gate $\boldsymbol{g}_{t,i} \in (0,1)^d$,

$$\boldsymbol{g}_{t,i} = \sigma(\boldsymbol{W}_g[\boldsymbol{s}_{t-1,i}; \boldsymbol{h}_t^c]), \tag{6}$$

which selects how strongly each candidate feature may modify the slot. $a_{t,i}$ allocates the observation across slots, while $\boldsymbol{g}_{t,i}$ controls the update within each selected slot. Combining them gives

$$\boldsymbol{s}_{t,i} = \boldsymbol{s}_{t-1,i} + a_{t,i}\,\boldsymbol{g}_{t,i} \odot (\widetilde{\boldsymbol{s}}_{t,i} - \boldsymbol{s}_{t-1,i}). \tag{7}$$

Thus, a slot changes only when the current context attends to it and its feature-wise gate is open. This update is designed to allow different slots to retain distinct recurring patient states rather than forcing the full history into a single running embedding. Appendix A.1.4 provides additional details.

Query-conditioned memory read. After the K context writes, the query reads the final memory $\boldsymbol{S}_K = [\boldsymbol{s}_{K,1}, \ldots, \boldsymbol{s}_{K,R}]^\top$. The learned read-query and read-key projections $\boldsymbol{W}_q^{\mathrm{r}}$ and $\boldsymbol{W}_k^{\mathrm{r}}$ compare the query embedding $\boldsymbol{h}_j^q$ with each slot. Scaled dot-product attention with temperature $T_{\mathrm{r}} > 0$ produces the read-attention weights

$$b_{j,i} = \mathrm{softmax}_i\left(\frac{\langle \boldsymbol{W}_q^{\mathrm{r}}\boldsymbol{h}_j^q, \boldsymbol{W}_k^{\mathrm{r}}\boldsymbol{s}_{K,i}\rangle}{T_{\mathrm{r}}\sqrt{d}}\right). \tag{8}$$

The weights $b_{j,i}$ select the slots relevant to the current query. Their value-projected contents form the retrieved patient context

$$\boldsymbol{c}_j = \sum_{i=1}^{R} b_{j,i}\boldsymbol{W}_v^{\mathrm{r}}\boldsymbol{s}_{K,i}, \tag{9}$$

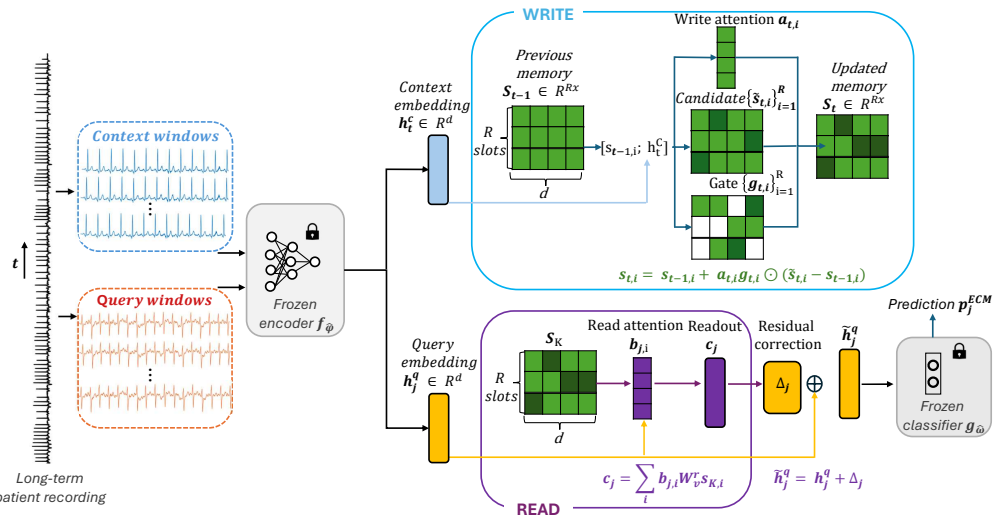


Figure 2: Schematic overview of ECM. Patient ECGs are divided into context and query windows, which are embedded by a frozen population encoder. Context embeddings update the memory (write), while query embeddings retrieve relevant context (read). The retrieved context produces a residual correction to the query embedding before prediction by the frozen classifier.

where $\boldsymbol{W}_v^{\mathrm{r}} \in \mathbb{R}^{d\times d}$ is the learned read-value projection. A learned map $\boldsymbol{W}_c \in \mathbb{R}^{d\times 2d}$ produces the feature-wise correction gate

$$\boldsymbol{g}_j^{\mathrm{r}} = \sigma\big(\boldsymbol{W}_c[\boldsymbol{h}_j^q; \boldsymbol{c}_j]\big)\,. \tag{10}$$

Using learned correction projection $\boldsymbol{W}_p \in \mathbb{R}^{d\times d}$ and learned scalar α_m, ECM adds the gated memory contribution to the source embedding,

$$\widetilde{\boldsymbol{h}}_j^q = \boldsymbol{h}_j^q + \alpha_m \boldsymbol{g}_j^{\mathrm{r}} \odot \mathrm{LayerNorm}(\boldsymbol{W}_p \boldsymbol{c}_j). \tag{11}$$

The frozen classifier finally produces

$$\boldsymbol{p}_j^{\mathrm{ECM}} = \mathrm{softmax}(g_{\hat{\omega}}(\widetilde{\boldsymbol{h}}_j^q)). \tag{12}$$

The residual path preserves the source representation, while query-dependent attention and gating select which patient context influences each prediction.

Learning and deployment. The parameter set ψ comprises the initial slots, write and read modules, write MLP, gates, and correction scale. Over the calibration episodes $\mathcal{E}_{\mathrm{cal}}$, these parameters are learned by unrolling the context writes and minimizing query cross-entropy ℓ together with a weighted memory regularizer $\mathcal{R}$ (see Appendix A.1.4). Denoting the label and ECM prediction for query window j of episode e by $y_{e,j}$ and $\boldsymbol{p}_{e,j}^{\mathrm{ECM}}$, respectively, the objective is

$$\mathcal{L}_{\mathrm{ECM}} = \frac{1}{|\mathcal{E}_{\mathrm{cal}}|} \sum_{e\in\mathcal{E}_{\mathrm{cal}}} \left[\frac{1}{Q} \sum_{j=1}^{Q} \ell(y_{e,j}, \boldsymbol{p}_{e,j}^{\mathrm{ECM}}) + \mathcal{R}(\boldsymbol{S}_K^{(e)}, \boldsymbol{S}_0) \right]. \tag{13}$$

Here, $\boldsymbol{S}_K^{(e)}$ is the memory after the K context writes of episode e. At deployment, context writes start from the preceding carried state $\boldsymbol{S}_0^{(e)} = \overline{\boldsymbol{S}}^{(e-1)}$ and produce $\boldsymbol{S}_K^{(e)}$. With $\overline{\boldsymbol{S}}^{(0)} = \boldsymbol{S}_0$ and $\rho \in [0, 1)$, we update

$$\overline{\boldsymbol{S}}^{(e)} = \rho\, \overline{\boldsymbol{S}}^{(e-1)} + (1-\rho)\boldsymbol{S}_K^{(e)}. \tag{14}$$

Queries read $\overline{\boldsymbol{S}}^{(e)}$, which is retained for the next episode and reset to $\boldsymbol{S}_0$ between patients.

Architectural instantiations. ECM-single (ECM-S) applies one memory bank to the final source embedding. We additionally study transition-aware ECM (t-ECM-S), an AF-specific extension that augments ECM-S with a fast state, transition-oriented supervision, and a causal prediction router to target abrupt rhythm changes. Further implementation details are provided in Appendix A.1.4.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

In our experiments, we evaluated whether patient-state adaptation improves longitudinal AF detection. Using a pretrained population detector, we calibrated ECM and comparator methods on long-term ECGs, and tested them on held-out patients and independent external datasets.

Population pretraining. We used a CNN–transformer population detector developed by Anonymous (2026) and trained on 10-s Einthoven lead II ECGs from a class-balanced subset of MIMIC-IV-ECG (Gow et al., 2023; Pollard et al., 2026), comprising 150,785 recordings from 77,261 patients. The dataset was partitioned by patient, with 15% reserved for source-domain testing.

Adaptation calibration. We calibrated ECM and other comparators using continuous long-term ECGs from IRIDIA-AF (Gilon et al., 2023), LTAFDB (Petrutiu et al., 2007), and CPSC2021 (Wang et al., 2021b). For each of five seeds, patients within each dataset were stratified by AF burden and assigned 30% to calibration and 70% to held-out internal evaluation. Calibration patients were further partitioned into 80% adaptation-training and 20% validation subsets. Thus, each split defined a distinct train/validation/internal test triplet. For every calibration-trained method, a separate model instance was selected exclusively using the validation subset and evaluated on its corresponding held-out cohort. This produced five independently calibrated models per method. We compared ECM with four source-free TTA methods: TENT (Wang et al., 2021a), CoTTA (Wang et al., 2022), EATA (Niu et al., 2022), and SAR (Niu et al., 2023). As a supervised comparator, full fine-tuning updated the complete population model on the labeled long-term calibration recordings. We additionally considered two learned memory baselines: long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) and neural long-term memory (NLTM) (Behrouz et al., 2025). Motivated by previous reference-conditioning approaches (Kim et al., 2024), we constructed a context-mean prototype (CMP) by averaging the embeddings of each current context, as well as an exponential moving average (EMA).

Longitudinal deployment. We evaluated the calibrated methods on two deployment regimes. Internal deployment used the held-out 70% of patients from IRIDIA-AF, LTAFDB, and CPSC2021. External deployment used MIT-BIH AF (Moody & Mark, 2000), MIT-BIH NSR (Moody, 1999), and SHDB-AF (Tsutsui et al., 2025a;b). Each of the five split-specific calibrated models was evaluated on its matched internal-test cohort and on the same fully independent external test cohort. All recordings were processed chronologically. Table 1 reports cohort composition. Complete dataset, optimization, and evaluation details are provided in Appendix A.1.

Table 1: ECG dataset composition for AF detection. Calibration and internal-test values are mean ± standard deviation across five patient-disjoint splits.

Set	n patients	n 10-s ECG windows	AF prevalence [%]
Pretraining	77,261	150,785	51.2
Calibration	101	$1{,}065{,}188_{\pm 58{,}189}$	$30.9_{\pm 0.9}$
Internal deployment	239	$2{,}549{,}456_{\pm 58{,}189}$	$31.2_{\pm 0.4}$
External deployment	133	1,501,463	20.3

5.2 MAIN RESULTS

Population-model performance. We re-evaluated the pretrained CNN-transformer on its patient-disjoint MIMIC-IV-ECG source test set, obtaining an F1 score of 96.8% and an AUROC of 99.6%, establishing the source-domain performance of the population detector.

Calibration-validation performance. Calibration-validation results used for model and hyperparameter selection are reported in Appendix Table 9. Fine-tuning performed best, reaching 89.2% macro-F1 and 3.0% burden error (BE), while NLTM and fine-tuning showed the highest AUROC

and AUPRC. ECM-S achieved intermediate validation performance (85.4% F1, 3.3% BE), and t-ECM-S shifted the operating regime toward higher sensitivity.

Deployment performance. We next evaluated the central question of the study: how the adaptation strategies generalize to unseen patients during longitudinal deployment. On held-out internal patients, fine-tuning achieved the highest macro-F1 (88.0%) and AUROC (97.6%) and the lowest BE (3.9%). ECM-S and t-ECM-S achieved macro-F1 scores of 84.7% and 83.1%, respectively, exceeding the population model (74.8%) but trailing fine-tuning by -3.3% and -4.9%. Source-referenced distance analyses (see Appendix A.2) further revealed heterogeneous adaptation gains across patients, with substantial improvements of ECM in the high-distance internal stratum. The ranking reversed under external dataset shift (Table 2): fine-tuning lost substantial sensitivity, whereas the ECM variants occupied complementary operating regimes. ECM-S yielded the lowest false AF (FA) rate (1.6%), improving it by -0.4% over fine-tuning. Relative to EMA, ECM-S improved macro-F1 by +0.8% and AUROC by +1.1% while reducing BE by -1.2%. t-ECM-S achieved the highest macro-F1 (80.4%), AUROC (93.7%), and AUPRC (80.8%), and the lowest BE (7.6%). These correspond to gains of +2.7% in macro-F1 and +2.1% in AUROC and a -1.5% reduction in BE relative to EMA, as well as a +1.4% improvement in AUPRC over fine-tuning. These results suggest that modifying the population model was advantageous in-distribution, whereas preserving population parameters and adapting patient state was more robust under the evaluated external dataset shift. Extended results are reported in Appendix Table 4.

Table 2: Patient-level ECG AF-detection performance on external deployment datasets. Best scores are shown in bold.

Method	**AF-containing** ($n = 95$)				**non-AF-only** ($n = 38$)
	F1↑	BE↓	AUROC↑	AUPRC↑	FA↓
Population	63.6	24.0	91.0	77.4	22.8
TENT	65.2	22.7	91.6	77.5	22.8
CoTTA	64.0	24.0	91.9	77.1	22.8
EATA	64.0	24.0	91.9	77.1	22.9
SAR	64.2	23.8	91.9	77.3	22.9
Fine-tuning	$75.2_{\pm 2.0}$	$11.1_{\pm 1.8}$	$90.0_{\pm 2.7}$	$79.4_{\pm 3.0}$	$2.0_{\pm 1.0}$
CMP	$76.8_{\pm 2.8}$	$9.7_{\pm 2.3}$	$90.8_{\pm 3.2}$	$76.9_{\pm 3.2}$	$3.7_{\pm 3.3}$
EMA	$77.7_{\pm 2.6}$	$9.0_{\pm 2.4}$	$91.6_{\pm 1.6}$	$77.8_{\pm 2.0}$	$3.4_{\pm 3.4}$
LSTM	$76.5_{\pm 3.9}$	$10.3_{\pm 3.1}$	$90.0_{\pm 3.1}$	$77.8_{\pm 4.1}$	$3.1_{\pm 2.0}$
NLTM	$67.2_{\pm 4.2}$	$16.2_{\pm 2.6}$	$90.3_{\pm 1.3}$	$74.1_{\pm 2.8}$	$3.4_{\pm 2.7}$
ECM-S	$78.5_{\pm 2.4}$	$7.8_{\pm 1.5}$	$92.7_{\pm 0.7}$	$78.4_{\pm 1.3}$	$\mathbf{1.6}_{\pm 1.7}$
t-ECM-S	$\mathbf{80.4}_{\pm 3.1}$	$\mathbf{7.6}_{\pm 3.1}$	$\mathbf{93.7}_{\pm 1.0}$	$\mathbf{80.8}_{\pm 1.6}$	$5.4_{\pm 4.0}$

5.3 Longitudinal and transition analysis

Performance over the monitoring horizon. We next examined whether improvements persist over long deployment horizons. Fig. 3 shows a paired F1 gain over the population model of +15.1% for t-ECM-S beyond 10 h, compared with +13.8% for ECM-S, +12.6% for EMA, and +11.9% for LSTM. In this interval, ECM-S reduced BE by -16.5% and FA by -25.3%, the largest reductions among the evaluated methods. The gains support a persistent benefit over the population model during prolonged monitoring. Complete metrics are reported in Appendix Table 5.

Performance around rhythm transitions. Fig. 4 reports external patient-macro performance following AF onset and offset, stratified by the duration of the ensuing rhythm episode. The results illustrate the complementary operating behavior of the two ECM variants. For AF episodes lasting at least 10 s, t-ECM-S achieved 74.1% sensitivity, exceeding the best comparator by +7.8%, whereas for episodes lasting at least 60 min, it reached 86.0%, versus 79.1% achieved by EMA. Following AF offset, ECM-S achieved 95.4% specificity for non-AF episodes lasting at least 60 min, matching fine-tuning and exceeding t-ECM-S (92.8%).

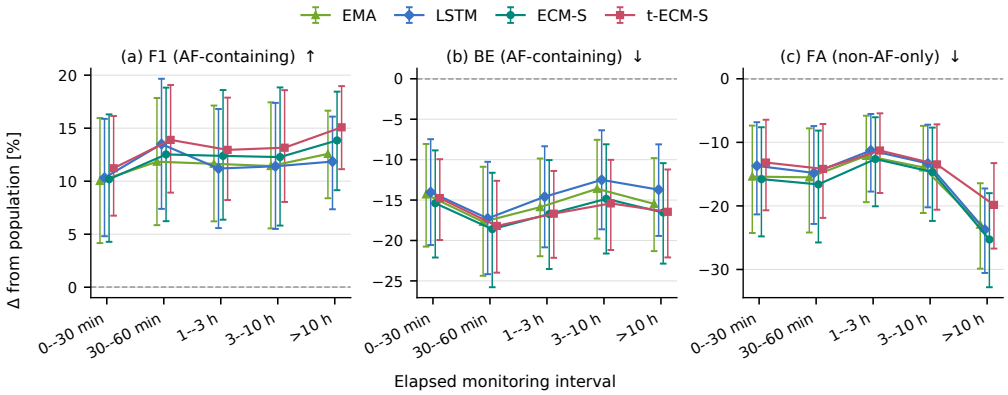


Figure 3: External ECG performance over elapsed monitoring intervals across 5 seeds.

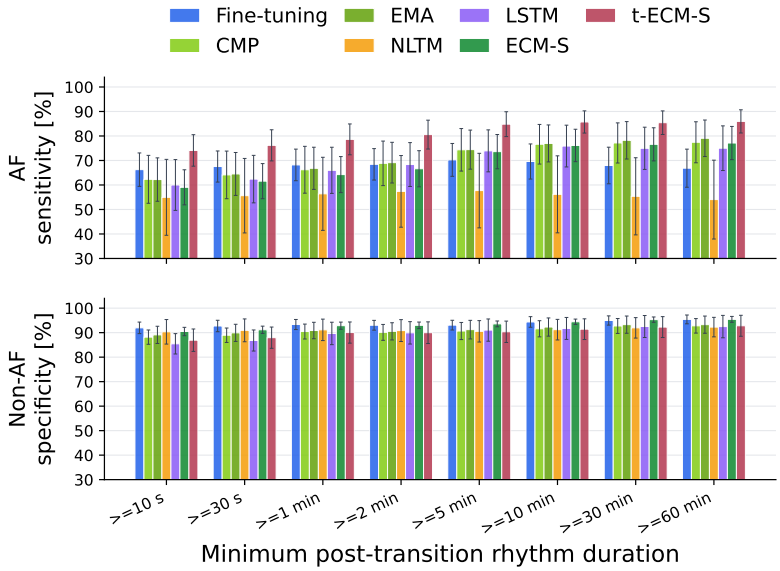


Figure 4: External transition performance across 5 seeds.

5.4 Mechanistic analysis of ECM

Component-level mechanism ablations. Figure 5 shows whether ECM-S benefits from its learned correction, persistent state, selective retrieval, and patient-specific context. Removing the correction increased sensitivity but reduced specificity by -23.2% and F1 by -14.6%. Thus, the principal role of the memory correction is to suppress unsupported AF evidence. Replacing the memory with that of another patient reduced every metric, including F1 by -7.6%, indicating that the correction depends on patient-specific context rather than acting only as a generic population-level adjustment. Replacing query-conditioned retrieval with a mean over slots also impaired performance, whereas resetting the carried state had a comparatively small effect. Coupled with further analyses (see Appendix A.4), these results support a mechanism based on patient-specific, sparse writes and selective reads. We additionally evaluated temporal smoothing, a patient-independent learned correction, and a parameter-matched causal-attention adapter. Across all external patients in the three-seed comparison, ECM-S achieved higher mean macro-F1 and AUROC and lower BE than these controls (see Appendix Table 11).

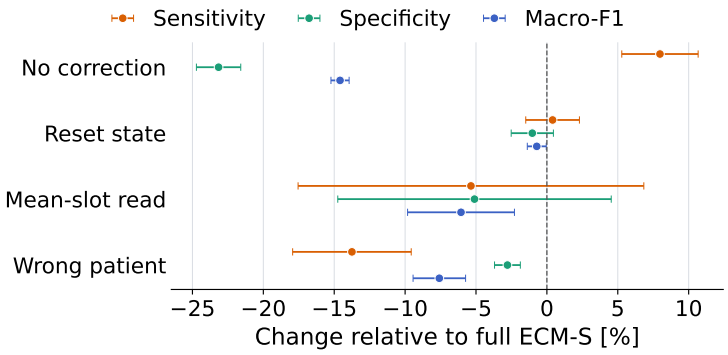


Figure 5: Change in performance after disabling individual ECM-S mechanisms.

5.5 Generalization beyond AF detection

To investigate the generalizability of our approach, we evaluated ECM and the comparators on two additional tasks. Complete details and results are reported in Appendix A.6.

PPG–ECG hypotension detection. For binary hypotension detection, we pretrained a separate population model using 30-s simultaneous ECGs and PPGs from MIMIC-III (Moody et al., 2020; Johnson et al., 2016), and calibrated ECM-S and the comparators on held-out long-term recordings across 3 seeds. We then evaluated external performance on 200 patient recordings from VitalDB (Lee & Jung, 2022; Lee et al., 2022). ECM-S improved macro-F1 (+0.8%) and BE (-1.4%) relative to the population model, while preserving its AUROC, and outperformed all comparators in F1 and BE.

Multiclass ECG ectopy classification. For classification of ectopic beats, we pretrained a population model on 10-s ECGs from Icentia11k (Tan et al., 2022) and calibrated each adaptation method on a held-out long-term subset. The external evaluation was performed on 84 recordings from LTAFDB. While fine-tuning performed best for ventricular beats (AUROC=92.8%), ECM-S obtained the highest improvements with respect to the population model for normal beats (+0.7% AUROC), supraventricular beats (+7.8% AUROC), and mixed beats (+2.6% AUROC). These evaluations support the applicability of the framework to multimodal inputs and multiclass classification.

6 Conclusion

We introduced ECM for longitudinal adaptation of short-window physiological classifiers to unseen long-term patient streams. ECM keeps the population model fixed and personalizes predictions through a compact patient-specific state learned from preceding observations, enabling label-free deployment without parameter updates. Across long-term ECG cohorts, the main benefit emerged under external distribution shift, where ECM was more robust than the comparators. ECM-S provided the more direct evaluation of patient-state adaptation, whereas t-ECM-S achieved stronger AF-specific performance with additional transition-oriented supervision. In longitudinal AF detection, ECM outperformed the evaluated adaptation methods, highlighting its potential relevance for prolonged monitoring, while the hypotension and multiclass ectopy results support the broader applicability of patient-state adaptation. The mechanistic analysis of ECM indicates that its gains depend on storing context from the correct patient and selectively retrieving state relevant to the current query. However, the smaller reset effect suggests that maintaining state across episodes is not equally important in every evaluated setting. Label-free deployment also depends on adaptation rules learned from labeled calibration data. Moreover, our retrospective evaluation did not simulate recording interruptions and unreliable context, which may affect realistic prospective scenarios. Quality-aware memory updates and prospective evaluation could clarify whether ECM's benefits persist under practical monitoring constraints. Future work could investigate event forecasting, adaptation of physiological foundation models, and transfer across further modalities.

AI USE STATEMENT

Generative AI tools were used to improve spelling, grammar, and clarity of the manuscript, and to assist with editing and polishing software code. They were not used for research ideation, experimental design, analysis, or generating scientific content. All AI-assisted text and code were carefully reviewed and verified by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This study involved only secondary analysis of previously collected, de-identified retrospective datasets made publicly available or provided under authorized research access. No new participants were recruited, and no interventions or data collection were conducted. All datasets were used in accordance with their respective licenses, data-use conditions, and original ethical approvals.

REPRODUCIBILITY STATEMENT

We provide anonymized code as supplementary material, including data preparation, model training and calibration, deployment evaluation, ablation studies, and statistical and visualization scripts. Dataset descriptions, patient-level partitioning, hyperparameter selection, and evaluation protocols are detailed in the appendix. Dataset sources are cited in the paper. Access remains subject to their respective licenses and data use requirements.

REFERENCES

Ali Mohammad Alqudah and Zahra Moussavi. A review of deep learning for biomedical signals: Current applications, advancements, future prospects, interpretation, and challenges. *Computers, Materials & Continua*, 83(3):3753–3841, 2025. doi: 10.32604/cmc.2025.063643. URL https://doi.org/10.32604/cmc.2025.063643.

Junho An, Kwanglo Lee, Sunghoon Jung, and Jinkook Kim. Personalized ECG arrhythmia classification using single normal beat guidance: A minimal supervision approach for atypical morphology cases. Research Square preprint, 2025. URL https://doi.org/10.21203/rs.3.rs-7966938/v1. Version 1, posted November 12, 2025.

Anonymous. Title redacted for review, 2026. Personal correspondence.

Mohammed Yusuf Ansari, Marwa Qaraqe, Raffaella Righetti, Erchin Serpedin, and Khalid Qaraqe. Enhancing ECG-based heart age: impact of acquisition parameters and generalization strategies for varying signal morphologies and corruptions. *Frontiers in Cardiovascular Medicine*, 11: 1424585, 2024. doi: 10.3389/fcvm.2024.1424585. URL https://doi.org/10.3389/fcvm.2024.1424585.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113506–113543. Curran Associates, Inc., 2025. doi: 10.52202/085713-3786. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4ca07aa108036f80cbb5b82285fd4b1-Paper-Conference.pdf.

Noam Ben-Moshe, Shany Biton, and Joachim A. Behar. ArNet-ECG: Deep learning for the detection of atrial fibrillation from the raw electrocardiogram. In *2022 Computing in Cardiology (CinC)*, volume 49, pp. 1–4, 2022. doi: 10.22489/CinC.2022.061. URL https://www.cinc.org/archives/2022/pdf/CinC2022-061.pdf.

Noam Ben-Moshe, Kenta Tsutsui, Shany Biton Brimer, Eran Zvuloni, Leif Sörnmo, and Joachim A. Behar. RawECGNet: Deep learning generalization for atrial fibrillation detection from the raw ECG. *IEEE Journal of Biomedical and Health Informatics*, 28(9):5180–5188, 2024. doi: 10.1109/JBHI.2024.3404877. URL https://doi.org/10.1109/JBHI.2024.3404877.

Shany Biton, Mohsin Aldhafeeri, Erez Marcusohn, Kenta Tsutsui, Tom Szwagier, Adi Elias, Julien Oster, Jean Marc Sellal, Mahmoud Suleiman, and Joachim A. Behar. Generalizable and robust deep learning algorithm for atrial fibrillation diagnosis across geography, ages and sexes. *npj Digital Medicine*, 6(1):44, 2023. doi: 10.1038/s41746-023-00791-1. URL https://doi.org/10.1038/s41746-023-00791-1.

Dominic Bläsing, Anja Buder, Julian Elias Reiser, Maria Nisser, Steffen Derlien, and Marcus Vollmer. ECG performance in simultaneous recordings of five wearable devices using a new morphological noise-to-signal index and Smith-Waterman-based RR interval comparisons. *PLOS ONE*, 17(10):e0274994, 2022. doi: 10.1371/journal.pone.0274994. URL https://doi.org/10.1371/journal.pone.0274994.

Bianca J. J. M. Brundel, Xun Ai, Mellanie True Hills, Myrthe F. Kuipers, Gregory Y. H. Lip, and Natasja M. S. de Groot. Atrial fibrillation. *Nature Reviews Disease Primers*, 8 (1):21, 2022. doi: 10.1038/s41572-022-00347-9. URL https://doi.org/10.1038/s41572-022-00347-9.

Cheng Ding, Tianliang Yao, Chenwei Wu, and Jianyuan Ni. Deep learning for personalized electrocardiogram diagnosis: A review. arXiv preprint arXiv:2409.07975, 2024. URL https://arxiv.org/abs/2409.07975.

Wolfram Doehner, Giuseppe Boriani, Tatjana Potpara, Carina Blomstrom-Lundqvist, Rod Passman, Luciano A. Sposato, Dobromir Dobrev, Ben Freedman, Isabelle C. Van Gelder, Taya V. Glotzer, Jeff S. Healey, Theodore Karapanayiotides, Gregory Y. H. Lip, Jose Luis Merino, George Ntaios, Renate B. Schnabel, Jesper H. Svendsen, Emma Svennberg, Rolf Wachter, Karl Georg Haeusler, and A. John Camm. Atrial fibrillation burden in clinical practice, research, and technology development: a clinical consensus statement of the European Society of Cardiology Council on Stroke and the European Heart Rhythm Association. *EP Europace*, 27(3):euaf019, 2025. doi: 10.1093/europace/euaf019. URL https://doi.org/10.1093/europace/euaf019.

Antonino Francisco, Capela Pascoal, Pedro Lamborne, Humberto Morais, and Mauer Gonçalves. Wearables and atrial fibrillation: Advances in detection, clinical impact, ethical concerns, and future perspectives. *Cureus*, 17(1):e77404, January 2025. doi: 10.7759/cureus.77404. URL https://doi.org/10.7759/cureus.77404.

Cédric Gilon, Jean-Marie Grégoire, Marianne Mathieu, Stéphane Carlier, and Hugues Bersini. IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database. *Scientific Data*, 10(1):714, 2023. doi: 10.1038/s41597-023-02621-1. URL https://doi.org/10.1038/s41597-023-02621-1.

Marc Goettling, Alexander Hammer, Hagen Malberg, and Martin Schmidt. xECGArch: a trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features. *Scientific Reports*, 14(1):13122, June 2024. doi: 10.1038/s41598-024-63656-x. URL https://doi.org/10.1038/s41598-024-63656-x.

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.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Alexander Hammer, Marc Goettling, Hagen Malberg, Axel Linke, Sergio Richter, Norman Mangner, and Martin Schmidt. Fusion of automatically learned rhythm and morphology features matches diagnostic criteria and enhances AI explainability. *npj Artificial Intelligence*, 1 (1):19, 2025. doi: 10.1038/s44387-025-00022-w. URL https://doi.org/10.1038/s44387-025-00022-w.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Ziyu Jia, Xihao Yang, Chenyang Zhou, Haoyang Deng, and Tianzi Jiang. ATTA: Adaptive test-time adaptation for multi-modal sleep stage classification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5882–5890, 2024. doi: 10.24963/ijcai.2024/650. URL https://www.ijcai.org/proceedings/2024/650.

Yong-Yeon Jo, Byeong Tak Lee, Jeong-Ho Hong, Hak Seung Lee, Joon-myoung Kwon, and Beom Joon Kim. Test-time calibration: A framework for personalized test-time adaptation in real-world biosignals. In Xuhai Orson Xu, Edward Choi, Pankhuri Singhal, Walter Gerych, Shengpu Tang, Monica Agrawal, Adarsh Subbaswamy, Elena Sizikova, Jessilyn Dunn, Roxana Daneshjou, Tasmie Sarker, Matthew McDermott, and Irene Chen (eds.), *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 381–394. PMLR, 25–27 Jun 2025. URL https://proceedings.mlr.press/v287/jo25a.html.

Alistair E. W. 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*, 3:160035, 2016. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Sangkyu Kim, Jiwoo Lim, and Jaeseong Jang. Patient-adaptive beat-wise temporal transformer for atrial fibrillation classification in continuous long-term cardiac monitoring. *IEEE Access*, 12: 172358–172367, 2024. doi: 10.1109/ACCESS.2024.3498043.

Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 63(3): 664–675, 2016. doi: 10.1109/TBME.2015.2468589. URL https://doi.org/10.1109/TBME.2015.2468589.

Dimitri Kraft and Peter Rumm. Atrial fibrillation and atrial flutter detection using deep learning. *Sensors*, 25(13):4109, 2025. doi: 10.3390/s25134109. URL https://doi.org/10.3390/s25134109.

Gabriel Jason Lee, Jathurshan Pradeepkumar, and Jimeng Sun. Test-time adaptation for EEG foundation models: A systematic study under real-world distribution shifts. In Rahul G. Krishnan, Wouter A. C. van Amsterdam, Sumit Chopra, Shauna Overgaard, Michael Hughes, Erkin Ötleş, Yiqiu Shen, Divya Shanmugam, Madhur Nayan, Matthew Engelhard, Jim Fackler, and Michael Oberst (eds.), *Proceedings of the 11th Machine Learning for Healthcare Conference*, volume 340 of *Proceedings of Machine Learning Research*, pp. 948–984. PMLR, 12–14 Aug 2026. URL https://proceedings.mlr.press/v340/lee26a.html.

Hyung-Chul Lee and Chul-Woo Jung. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *PhysioNet*, September 2022. doi: 10.13026/czw8-9p62. URL https://doi.org/10.13026/czw8-9p62. Version 1.0.0.

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9:279, 2022. doi: 10.1038/s41597-022-01411-5. URL https://doi.org/10.1038/s41597-022-01411-5.

Yoon Jae Lee, Cheoljeong Park, Hodam Kim, Seong J. Cho, and Woon-Hong Yeo. Artificial intelligence on biomedical signals: technologies, applications, and future directions. *Med-X*, 2(1):25, 2024. doi: 10.1007/s44258-024-00043-1. URL https://doi.org/10.1007/s44258-024-00043-1.

Kanxue Li, Yibing Zhan, Hua Jin, Chongchong Qi, Xu Lin, and Baosheng Yu. Cross-sample augmented test-time adaptation for personalized intraoperative hypotension prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(27):22958–22966, 2026a. doi: 10.1609/aaai.v40i27.39460. URL https://doi.org/10.1609/aaai.v40i27.39460.

Siyang Li, Jiayi Ouyang, Zhenyao Cui, Ziwei Wang, Tianwang Jia, Feng Wan, and Dongrui Wu. Backpropagation-free test-time adaptation for lightweight EEG-based brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–13, 2026b. doi: 10.1109/JBHI.2026.3725326. URL https://doi.org/10.1109/JBHI.2026.3725326.

Dominik Linz, Monika Gawalko, Konstanze Betz, Jeroen M. Hendriks, Gregory Y.H. Lip, Nicklas Vinter, Yutao Guo, and Søren Johnsen. Atrial fibrillation: epidemiology, screening and digital health. *The Lancet Regional Health - Europe*, 37:100786, 2024. ISSN 2666-7762. doi: 10.1016/j.lanepe.2023.100786. URL https://www.sciencedirect.com/science/article/pii/S2666776223002053.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Kan Luo, Jianqing Li, Zhigang Wang, and Alfred Cuschieri. Patient-specific deep architectural model for ECG classification. *Journal of Healthcare Engineering*, 2017:4108720, 2017. doi: 10.1155/2017/4108720. URL https://doi.org/10.1155/2017/4108720.

Tingting Mao, Chang Li, Yuchang Zhao, Rencheng Song, and Xun Chen. Online test-time adaptation for patient-independent seizure prediction. *IEEE Sensors Journal*, 23(19):23133–23144, 2023. doi: 10.1109/JSEN.2023.3307223. URL https://doi.org/10.1109/JSEN.2023.3307223.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III Waveform Database. *PhysioNet*, April 2020. doi: 10.13026/c2607m. URL https://doi.org/10.13026/c2607m. Version 1.0.

George B. Moody. MIT-BIH normal sinus rhythm database. PhysioNet, 1999. URL https://doi.org/10.13026/C2NK5R. Version 1.0.0.

George B. Moody and Roger G. Mark. MIT-BIH atrial fibrillation database. PhysioNet, 2000. URL https://doi.org/10.13026/C2MW2D. Version 1.0.0.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Yiheng Peng, Jingjing Luo, Hongbo Wang, Shijie Guo, Yuzhu Guo, Dongsheng Xu, and Yang Li. Test-time adaptation for cross-subject motor imagery EEG classification using information-aggregation and source-guided weighting. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2025. doi: 10.1109/IJCNN64981.2025.11228030. URL https://doi.org/10.1109/IJCNN64981.2025.11228030.

Simona Petrutiu, Alan V. Sahakian, and Steven Swiryn. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. *EP Europace*, 9(7):466–470, 2007. doi: 10.1093/europace/eum096. URL https://doi.org/10.1093/europace/eum096.

Tien-Dat Pham and Xuan-The Tran. EEG-Titans: Long-horizon seizure forecasting via dual-branch attention and neural memory. arXiv preprint arXiv:2601.13748, 2026. URL https://arxiv.org/abs/2601.13748.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, August 2026. ISSN 3005-0693. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Jiajia Si, Yiliang Bao, Fengling Chen, Yue Wang, Meimei Zeng, Nongyue He, Zhu Chen, and Yuan Guo. Research on atrial fibrillation diagnosis in electrocardiograms based on CLA-AF model. *European Heart Journal - Digital Health*, 6(1):82–95, 2025. doi: 10.1093/ehjdh/ztae092. URL https://doi.org/10.1093/ehjdh/ztae092.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8: 985–1005, 2007. URL https://www.jmlr.org/papers/volume8/sugiyama07a/sugiyama07a.pdf.

Shawn Tan, Guillaume Androz, Satya Ortiz-Gagné, Ahmad Chamseddine, Pierre Fecteau, Aaron Courville, Yoshua Bengio, and Joseph Paul Cohen. Icentia11K: An unsupervised representation learning dataset for arrhythmia subtype discovery. In *2021 Computing in Cardiology (CinC)*, 2021. URL https://www.cinc.org/2021/Program/accepted/229_Preprint.pdf.

Shawn Tan, Satya Ortiz-Gagné, Nicolas Beaudoin-Gagnon, Pierre Fecteau, Aaron Courville, Yoshua Bengio, and Joseph Paul Cohen. Icentia11k Single Lead Continuous Raw Electrocardiogram Dataset. *PhysioNet*, April 2022. doi: 10.13026/kk0v-r952. URL https://doi.org/10.13026/kk0v-r952. Version 1.0.

Wei Tang, Jinpei Han, Kangning Cui, Mattia Carletti, Fredrik K. Gustafsson, Shreyank N Gowda, Patitapaban Palo, Anshul Thakur, Lei Clifton, Jean-michel Morel, Raymond H. Chan, David A. Clifton, and Xiao Gu. Extending pretrained 10-second ECG foundation models to longer horizons. arXiv preprint arXiv:2605.16975, 2026. URL https://arxiv.org/abs/2605.16975.

Nia Touko, Matthew O A Ellis, Cristiano Capone, Alessio Burrello, Elisa Donati, and Luca Manneschi. Lightweight test-time adaptation for EMG-based gesture recognition. arXiv preprint arXiv:2601.04181, 2026. URL https://arxiv.org/abs/2601.04181.

Kenta Tsutsui, Shany Biton Brimer, and Joachim Behar. SHDB-AF: a Japanese Holter ECG database of atrial fibrillation. PhysioNet, April 2025a. URL https://doi.org/10.13026/n6yq-fq90. Version 1.0.1.

Kenta Tsutsui, Shany Biton Brimer, Noam Ben-Moshe, Jean Marc Sellal, Julien Oster, Hitoshi Mori, Yoshifumi Ikeda, Takahide Arai, Shintaro Nakano, Ritsushi Kato, and Joachim A. Behar. SHDB-AF: a Japanese holter ECG database of atrial fibrillation. *Scientific Data*, 12 (1):454, 2025b. doi: 10.1038/s41597-025-04777-4. URL https://doi.org/10.1038/s41597-025-04777-4.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7201–7211, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

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 2021b. URL https://doi.org/10.13026/ksya-qw89. Version 1.0.0.

Martin Wimpff, Mario Döbler, and Bin Yang. Calibration-free online test-time adaptation for electroencephalography motor imagery decoding. In *2024 12th International Winter Conference on Brain-Computer Interface (BCI)*, pp. 1–6. IEEE, 2024. doi: 10.1109/BCI60775.2024.10480468. URL https://doi.org/10.1109/BCI60775.2024.10480468.

Xinyi Wu, Zouheir Rezki, Bhavan Balusu, Jiaming Zhou, and Jun Bai. One model, dual tasks: a novel distributionally adaptive learning framework for ECG classification and generation addressing intra- and inter-patient variability. *Expert Systems with Applications*, 299:130096, 2026. doi: 10.1016/j.eswa.2025.130096. URL https://doi.org/10.1016/j.eswa.2025.130096.

A Appendix

A.1 Additional implementation details

A.1.1 Population model architecture

The convolutional front end of the CNN-transformer begins with a one-dimensional projection to 16 channels, followed by three residual convolutional blocks with 32, 64, and 128 output channels. Each block applies a dilated convolution, 8-group normalization, ReLU activation, and factor-two average-pooling. A pooled 1×1 projection forms the residual path when the channel dimension changes. The input projection uses kernel size 16, and the 3 blocks use kernel sizes $(51, 25, 15)$ and dilations $(1, 4, 20)$. The resulting 128-dimensional sequence is processed by 6 pre-normalized transformer blocks with 4 attention heads, a fourfold MLP expansion, sinusoidal positional encodings, and dropout of 0.15. Global temporal average-pooling and a linear two-class head produce the final prediction. The model was optimized with AdamW using learning rate 10^{-3} and batch size 512; checkpoint selection maximized validation F1. The population model was developed and pretrained by Anonymous (2026) and is not a contribution of the present study.

A.1.2 Extended dataset details

Dataset description and preprocessing. IRIDIA-AF (Gilon et al., 2023) contains 167 Holter recordings from 152 patients with paroxysmal AF. The recordings span 19–95 hours, were sampled at 200 Hz, and include AF onset and offset annotations made by an expert cardiologist and a specialist cardiac nurse. LTAFDB (Petrutiu et al., 2007) contains 84 recordings of 24–25 hours from subjects with paroxysmal or sustained AF. Its automatically generated rhythm annotations were manually verified by ECG technicians. CPSC2021 (Wang et al., 2021b) contains variable-duration Holter ECGs from 105 patients with persistent or paroxysmal AF labels. Recordings include persistent and paroxysmal AF rhythm labels, as well as non-AF. In each patient-disjoint split, IRIDIA-AF contributed 45 calibration and 106 internal-test patients, LTAFDB contributed 25 and 59 patients, and CPSC2021 contributed 31 and 74 patients, respectively. SHDB-AF (Tsutsui et al., 2025b;a) contains approximately 24-hour Holter recordings. We extracted 92 recordings with manual beat-level annotations by certified cardiologists, including AF, atrial flutter, atrial tachycardia, and other supraventricular tachycardias. MIT-BIH AF (Moody & Mark, 2000) comprises 25 recordings with 10-hour duration from subjects with predominantly paroxysmal AF; 23 records contain two ECG channels and manually prepared rhythm annotations. MIT-BIH NSR (Moody, 1999) contains 18 long-term recordings from subjects without significant arrhythmias and was treated as a non-AF-only cohort. External evaluation used fixed cohorts of 92 SHDB-AF, 23 MIT-BIH AF, and 18 MIT-BIH NSR patients, for 133 external patients in total. We extracted Einthoven lead II from each dataset, using the corresponding NASA lead in SHDB-AF. Following Goettling et al. (2024), signals were resampled to 500 Hz, divided into non-overlapping 10-s windows, high-pass filtered at 0.3 Hz with a fourth-order Butterworth filter, tapered with a Tukey window ($\alpha = 0.06$), and normalized to [0,1]. Windows were labeled positive if containing AF intervals, and negative if not.

A.1.3 Detailed evaluation protocol

Chronological context. For an episode beginning at window s, ECM writes the context $[s, s+K)$ and predicts the immediately following query $[s+K, s+K+Q)$. Context and query therefore never overlap within an episode. At deployment, episode starts advance by Q windows. Consequently,

query blocks do not overlap, contexts may overlap when $K > Q$, and a query window may become part of a later context only after its own prediction has been made.

Memory lifetime. ECM memory is initialized from the learned state $\boldsymbol{S}_0$ at the beginning of each patient. Under the selected carry policy, it is retained between chronologically adjacent episodes of that patient and blended according to Eq. 14. It is reset before the next patient and before every independent evaluation run.

Patient-level metrics. For each patient and seed, all scored windows are concatenated in chronological order and converted to binary predictions at the fixed 0.5 threshold. The operating threshold was not optimized on either validation or test data. Validation performance was used only to select hyperparameters and checkpoints. The resulting patient-specific confusion counts determine sensitivity, specificity, precision, accuracy, and F1 scores. Absolute BE is the absolute difference between the predicted and reference fractions of AF windows. For a non-AF-only patient, FA burden is the predicted AF window fraction. Patient AUROC and AUPRC are computed from continuous probabilities. Within each split seed, patient-level metrics are averaged with equal weight across patients in the corresponding evaluation cohort. We report the mean and sample standard deviation of these patient-macro averages across the five seeds.

A.1.4 ECM VARIANTS AND REGULARIZATION

Diversity-aware allocation. The write MLP contains two layers with GELU between them. The learned correction scale α_m is initialized to 0.01. The diversity-aware ECM-S implementation refines the content-based write in Eqs. 4–7. Let $z_{t,i} = \langle \boldsymbol{W}_q^{\mathrm{w}} \boldsymbol{h}_t^c, \boldsymbol{W}_k^{\mathrm{w}} \boldsymbol{s}_{t-1,i} \rangle / \sqrt{d}$ denote its content score. In the same learned addressing space, observation novelty and its bounded write multiplier are

$$\nu_t = 1 - \max_i \cos\big(\boldsymbol{W}_q^{\mathrm{w}} \boldsymbol{h}_t^c, \boldsymbol{W}_k^{\mathrm{w}} \boldsymbol{s}_{t-1,i}\big)\,, \tag{15}$$

$$\gamma_t = \gamma_{\min} + (1 - \gamma_{\min})\sigma((\nu_t - \tau_\nu)/T_\nu)\,. \tag{16}$$

Here, $\gamma_{\min} \in [0, 1]$ is the minimum write scale, τ_ν the novelty threshold, and $T_\nu > 0$ its temperature. Let $\boldsymbol{u}_t = (u_{t,1}, \ldots, u_{t,R})$ track slot usage, initially $\boldsymbol{u}_0 = \mathbf{0}$. With usage-penalty weight $\lambda_u \geq 0$, the write logits penalize overused slots,

$$\zeta_{t,i} = z_{t,i} - \lambda_u \gamma_t \left(\frac{R u_{t-1,i}}{\max(\sum_{r=1}^{R} u_{t-1,r}, \epsilon)} - 1 \right), \tag{17}$$

where $\epsilon = 10^{-6}$ stabilizes the denominator. The write-attention vector $\boldsymbol{a}_t = (a_{t,1}, \ldots, a_{t,R})$ is the softmax of $\zeta_{t,i}$ restricted to the $k_{\mathrm{w}} \in \{1, \ldots, R\}$ slots with largest unpenalized scores $z_{t,i}$. The increment in Eq. 7 is multiplied by γ_t, and usage follows $\boldsymbol{u}_t = \beta_u \boldsymbol{u}_{t-1} + (1 - \beta_u)\boldsymbol{a}_t$, where $\beta_u \in [0, 1)$ is its decay coefficient. Usage accompanies the slots through carry and patient resets. Thus, novel observations write more strongly, while sparse allocation and usage balancing distribute patient states across the memory.

Let $\bar{\boldsymbol{s}} = R^{-1} \sum_{i=1}^{R} \boldsymbol{s}_i$ be the mean slot vector of $\boldsymbol{S}$ and $\boldsymbol{S}^{\mathrm{c}} = \boldsymbol{S} - \mathbf{1}\bar{\boldsymbol{s}}^\top$ its centered slots, with $\mathbf{1} \in \mathbb{R}^R$ the all-ones vector. $\mathrm{n}(\cdot)$ divides each row by its Euclidean norm clamped below at 10^{-6}, and offdiag sets the diagonal to zero. Slot collapse is discouraged using weak raw and centered cosine-similarity penalties,

$$\mathcal{L}_{\mathrm{div}}^{\mathrm{c}} = \frac{\big\|\mathrm{offdiag}\big(\mathrm{n}(\boldsymbol{S}^{\mathrm{c}})\,\mathrm{n}(\boldsymbol{S}^{\mathrm{c}})^\top\big)\big\|_F^2}{R(R-1)}, \tag{18}$$

$$\mathcal{L}_{\mathrm{div}}^{\mathrm{r}} = \frac{\big\|\mathrm{offdiag}\big(\mathrm{n}(\boldsymbol{S})\,\mathrm{n}(\boldsymbol{S})^\top\big)\big\|_F^2}{R(R-1)}. \tag{19}$$

Both penalties are zero for $R = 1$. The weighted regularizer in Eq. 13 is

$$\mathcal{R}(\boldsymbol{S}, \boldsymbol{S}_0) = \lambda_a \frac{\|\boldsymbol{S} - \boldsymbol{S}_0\|_F^2}{Rd} + \lambda_c \mathcal{L}_{\mathrm{div}}^{\mathrm{c}} + \lambda_r \mathcal{L}_{\mathrm{div}}^{\mathrm{r}}, \tag{20}$$

where $\lambda_a, \lambda_c, \lambda_r \geq 0$ weight the initial-state anchor and the centered and raw diversity penalties evaluated on $\boldsymbol{S}$.

Transition-aware variant. t-ECM-S augments the slow slot memory with a fast state $\boldsymbol{f}_j \in \mathbb{R}^d$ available before query j. Each episode initializes $\boldsymbol{f}_1$ to the mean embedding of its context tail; with decay $\beta_f \in [0, 1)$, it updates only after the current query is predicted,

$$\boldsymbol{f}_{j+1} = \beta_f \boldsymbol{f}_j + (1 - \beta_f)\boldsymbol{h}_j^q. \tag{21}$$

A learned scalar authority gate scales the slow correction, while a separate feature-wise gate and learned scale control an additive fast correction computed from $[\boldsymbol{h}_j^q; \boldsymbol{f}_j]$. This fast–slow model is warm-started from ECM-S and trained on transition-enriched episodes with query classification, onset/offset auxiliary losses, and KL retention away from transition boundaries. A subsequent causal router is supervised from disagreements between the frozen source and fast–slow models:

$$m_j^{\mathrm{tECM}} = m_j^0 + r_j \min(m_j^E - m_j^0, 0), \qquad r_j \in [0, 1], \tag{22}$$

where m_j^0 and m_j^E are the AF-minus-non-AF logit margins of the frozen source and fast–slow expert, respectively, and r_j is the learned router's sigmoid gate. The final AF probability is $\sigma(m_j^{\mathrm{tECM}})$.

Parallel integration. ECM-parallel (ECM-P) associates a separate memory bank with each selected encoder block. Let $\boldsymbol{H}_l$ denote the input token sequence to frozen block B_l, and $M_l(\boldsymbol{H}_l; \boldsymbol{S}^{[l]})$ its memory-adapted counterpart using that block's state $\boldsymbol{S}^{[l]}$. The parallel output is

$$\boldsymbol{H}_{l+1} = B_l(\boldsymbol{H}_l) + \sigma(\xi_l)\left(M_l(\boldsymbol{H}_l; \boldsymbol{S}^{[l]}) - \boldsymbol{H}_l\right), \tag{23}$$

where ξ_l is a learned scalar gate logit. Both branches receive the same pre-block tokens. The source-block computation is therefore preserved and the memory contributes a gated residual branch. The deployment results of ECM-P are reported in Appendix A.3.1.

A.1.5 ADAPTATION HYPERPARAMETER SEARCH SPACE

The complete hyperparameters tuned during calibration and the corresponding selected values are shown in Table 3. For calibration-trained methods, each hyperparameter configuration was trained for up to 150 epochs, with early stopping after 15 epochs without validation F1 improvement. TTA configurations were instead evaluated by running online adaptation on the calibration-validation streams, without supervised calibration training. For both groups, hyperparameters were selected by maximizing mean validation F1 minus its standard deviation across three seeds. The selected TTA configuration was then applied to the common pretrained population model during deployment.

Table 3: Evaluated adaptation hyperparameters and selected ECG configurations. K: number of context windows; LN: layer normalization; lr: learning rate; Q: number of query windows.

Method	Evaluated parameters	Selected values
TENT	Adaptable parameters / lr.	LN / 0.0001.
CoTTA	Student lr / restoration probability / teacher-confidence threshold / Teacher EMA decay.	0.00005 / 0.01 / 0 / 0.999.
EATA	lr / entropy-reliability margin / Fisher regularization weight / redundancy threshold.	0.0001 / 0.4 / 0.01 / 0.05.
SAR	lr / entropy-reliability margin / sharpness-aware radius / entropy-EMA decay / reset threshold / optimizer momentum.	0.0001 / 0.6 / 0.1 / 0.9 / 0.2 / 0.9.
Fine-tuning	lr / batch size / weight decay.	0.0005 / 512 / 0.05
NLTM	Memory depth / memory dimension / surprise lr / memory weight decay / calibration lr / memory layers.	1 / 32 / 0.00005 / 0.05 / 0.0005 / all.
LSTM	K / Q / hidden dimension / dropout / lr / optimizer weight decay.	50 / 50 / 128 / 0.10 / 0.0005 / 0.0001.
CMP	Prototype dimension / MLP depth / K / Q / lr / optimizer weight decay.	128 / 2 / 50 / 50 / 0.0005 / 0.0001.
EMA	Prototype dimension / MLP depth / EMA coefficient / K / Q / anchor weight / lr / optimizer weight decay.	128 / 2 / 0.5 / 50 / 50 / 0.03 / 0.0005 / 0.0001.
ECM-S	K / Q / write top-k / novelty threshold / usage penalty / read temperature / carry coefficient / raw-slot diversity weight / anchor weight / memory weight decay / lr.	50 / 50 / 2 / 0.50 / 0.05 / 0.75 / 0.25 / 0.0005 / 0.03 / 0.05 / 0.0005.
ECM-P	Parallel-branch gate initialization / K / Q / write top-k / novelty threshold / usage penalty / read temperature / carry coefficient / raw-slot diversity weight / anchor weight / memory weight decay / lr (transferred).	0.10 / 50 / 50 / 2 / 0.50 / 0.05 / 0.75 / 0.25 / 0.0005 / 0.03 / 0.05 / 0.0005
t-ECM-S	gate EMA / retained-KL weight / supervision weight / preservation weight / loss margin / router lr.	0.95 / 1 / 0.25 / 0.25 / 0.05 / 0.0005.

A.2 Distribution shift and adaptation gains

In this section, we provide quantitative analyses of distribution shift between source and deployment cohorts in waveform, label prevalence, and frozen embedding space, and examine their association with patient-level adaptation gains. We retain the notation $P_{\mathrm{s}}(X, Y)$ and $P_{a,n}(X, Y)$ from Section 3.

A.2.1 Covariate shift.

In its strict form, covariate shift assumes (Sugiyama et al., 2007)

$$P_{a,n}(X) \neq P_{\mathrm{s}}(X), \qquad P_{a,n}(Y \mid X) = P_{\mathrm{s}}(Y \mid X). \tag{24}$$

We examined the input-marginal component without establishing the conditional invariance required for pure covariate shift. Here, X denotes the preprocessed input ECG. We used squared maximum mean discrepancy (MMD^2; Gretton et al., 2012) with Gaussian kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_F^2}{2CL\sigma_{\mathrm{ker}}^2}\right), \tag{25}$$

where $C = 1$, $L = 5000$, and σ_{ker}^2 is the median pairwise mean-squared waveform difference among 256 separate source-training patients. For split seed 42, we sampled one window per patient and matched the numbers of source-training, source-held-out, and comparison patients. At the source-fitted bandwidth ($\sigma_{\mathrm{ker}}^2 = 0.235$), source-to-cohort MMD^2 was 0.009 and 0.024 for calibration training and validation, compared with matched source vs source-held-out references of 0.007 and 0.024. Internal and external deployment yielded 0.038 and 0.059, compared with 0.003 and 0.008 for their respective source references (Fig. 6). The point estimates therefore suggest larger waveform differences at deployment.

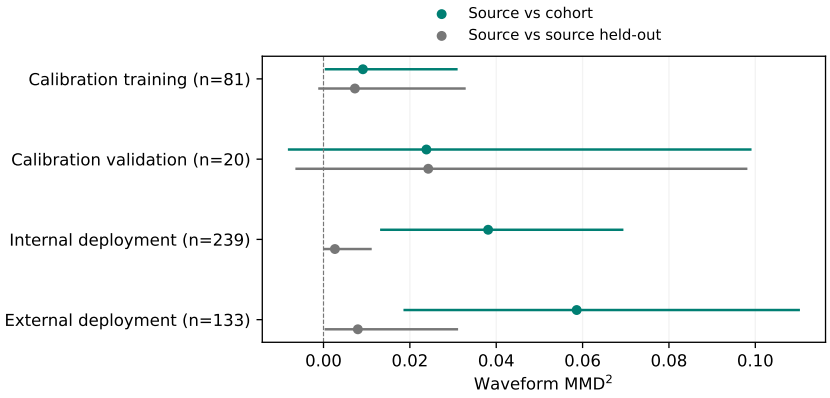
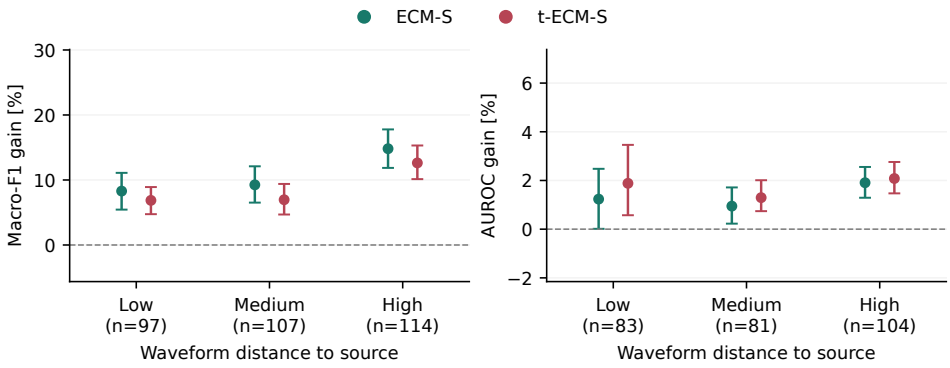
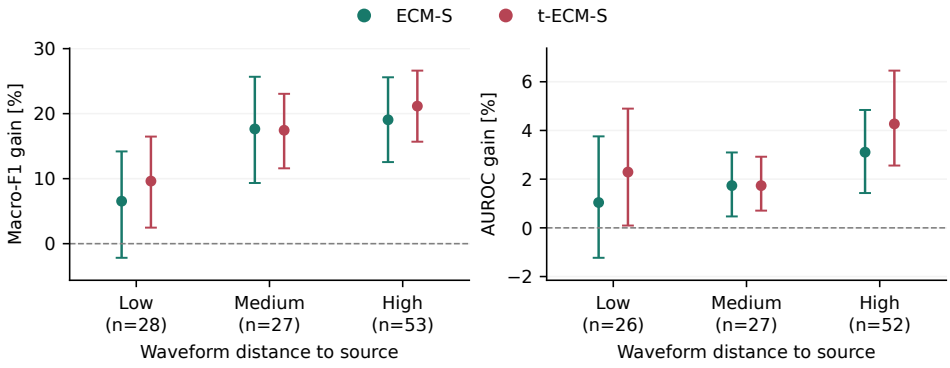


Figure 6: Input marginal differences in preprocessed waveform space. Points show MMD^2 between source training and each cohort, alongside a matched source training versus source held-out reference. Bars show pointwise 95% bands from 500 patient bootstrap resamples. Within-source references are independently subsampled for each cohort to match its patient count.

For all stratified gain analyses below, we used only held-out deployment observations from each split. The internal cohort contains 239 patients per split and 340 distinct patients across the five splits. F1 analyses include AF-containing recordings (288 distinct internal and 95 external patients), whereas AUROC requires both classes in the scored windows (243 internal and 92 external patients). Repeated seed observations were averaged within each patient and stratum before computing patient-mean gains. For patient-level associations across split seeds, we compared each patient's empirical waveform distribution with a fixed reference of 1,000 source-training patients, using one window per source patient and equal total weight for each class. Patient distance estimates were averaged across 3 sampling seeds of up to 64 scored windows. Low, medium, and high distances used calibration patient tertiles fitted separately for each seed. In external deployment, macro-F1 gains in the high-distance group were +19.1% and +21.2% for ECM-S and t-ECM-S, respectively, versus +6.5% and +9.6% in the low-distance group. High-distance AUROC gains were +3.1% and +4.3% (Fig. 7). The positive association is thus clearest for external macro-F1.



(a) Internal deployment.



(b) External deployment.

Figure 7: Paired adaptation gains by waveform distance to source. Let δ_p^x denote patient p's waveform distance to source. Low, medium, and high denote $\delta_p^x < q_{1/3}$, $q_{1/3} \leq \delta_p^x < q_{2/3}$, and $\delta_p^x \geq q_{2/3}$, respectively, with quantiles fitted on calibration patients separately for each seed. Points show patient-mean gains over the frozen population model, whereas vertical bars show 95% dataset-stratified patient-bootstrap intervals after averaging seed observations within patient and stratum.

A.2.2 LABEL SHIFT.

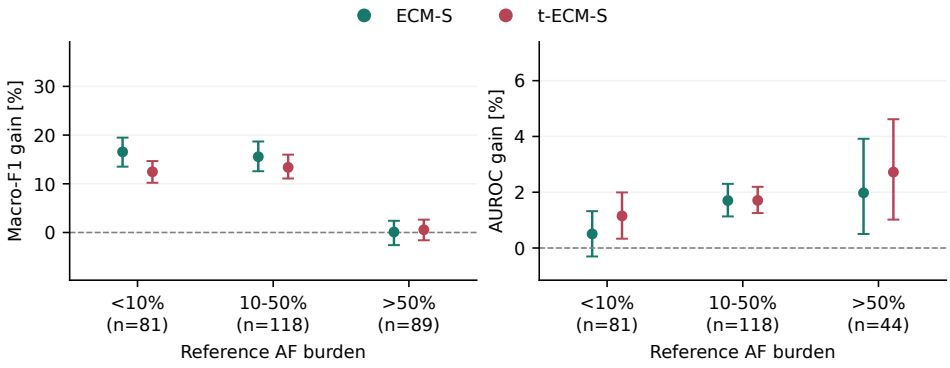
Pure label shift assumes a changed label marginal but unchanged class-conditional input distributions (Lipton et al., 2018):

$$P_{a,n}(Y) \neq P_{\mathrm{s}}(Y), \qquad P_{a,n}(X \mid Y) = P_{\mathrm{s}}(X \mid Y). \tag{26}$$

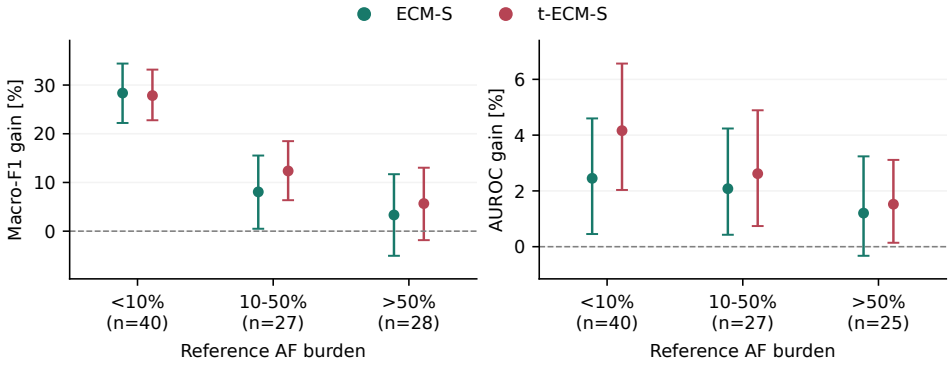
We quantified AF prevalence differences without establishing the class-conditional invariance required for pure label shift. Since population pretraining used class-balanced sampling, its effective AF prior is $\pi_{\mathrm{s}} = P_{\mathrm{s}}(Y = \mathrm{AF}) = 0.5$. For patient p, let $\mathcal{I}_p$ be the shared set of scored windows. We define

$$\hat{\pi}_p = \frac{1}{|\mathcal{I}_p|} \sum_{n \in \mathcal{I}_p} \mathbf{1}\{y_n^{(p)} = \mathrm{AF}\}, \qquad \delta_p^{\mathrm{prior}} = |\hat{\pi}_p - \pi_{\mathrm{s}}|. \tag{27}$$

In split seed 42, pooled-window AF prevalence was 30.5% in calibration training, 30.2% in calibration validation, 31.4% in internal deployment, and 20.3% in external deployment. The corresponding patient-mean prevalences were 31.3%, 38.7%, 34.0%, and 21.8%, illustrating the effect of patient versus window weighting. Figure 8 shows stratified gains by observed AF burden. Among the 40 external AF-containing patients with burden below 10%, ECM-S and t-ECM-S improved macro-F1 by +28.4% $[22.2, 34.4]$ and +27.8% $[22.8, 33.2]$, respectively. For burden from 10% to 50% (inclusive), gains were +8.1% and +12.4%. Above 50%, including all-AF recordings, they were +3.3% $[-5.1, 11.7]$ and +5.7% $[-1.8, 13.0]$, indicating heterogeneous benefits across AF prevalence.



(a) Internal deployment.



(b) External deployment.

Figure 8: Paired adaptation gains by reference AF burden in the shared scored windows: below 10%, 10–50%, and above 50%. Points show patient-mean gains over the frozen population model, whereas vertical bars show 95% dataset-stratified patient-bootstrap intervals after averaging seed observations within patient and stratum.

A.2.3 DISTRIBUTION SHIFT IN EMBEDDING SPACE.

Let $Z = f_{\hat{\phi}}(X)$ denote the representation produced by the frozen population encoder. For any measurable set $B \subseteq \mathbb{R}^d$,

$$P_{a,n}(Z \in B) = P_{a,n}\left(f_{\hat{\phi}}(X) \in B\right), \qquad P_{a,n}(Z) \neq P_{\mathrm{s}}(Z) \tag{28}$$

describes a shift in representation space. Figure 9 compares the distribution of distances to the MIMIC-IV centroid in the frozen population encoder embedding space after z-normalization. The long-term ECG cohorts exhibit dataset-specific differences from the source distribution: SHDB-AF and IRIDIA-AF are shifted toward larger distances, whereas LTAFDB shows a broader distribution and CPSC2021 and the MIT-BIH cohorts overlap more strongly with MIMIC-IV. These distributions indicate heterogeneous differences in the frozen population representation. To quantify differences between source, calibration, and deployment cohorts, we additionally computed each window's minimum Mahalanobis distance to the two source class centroids, using source-training z-normalization and a shared, class-balanced covariance with shrinkage 0.1.

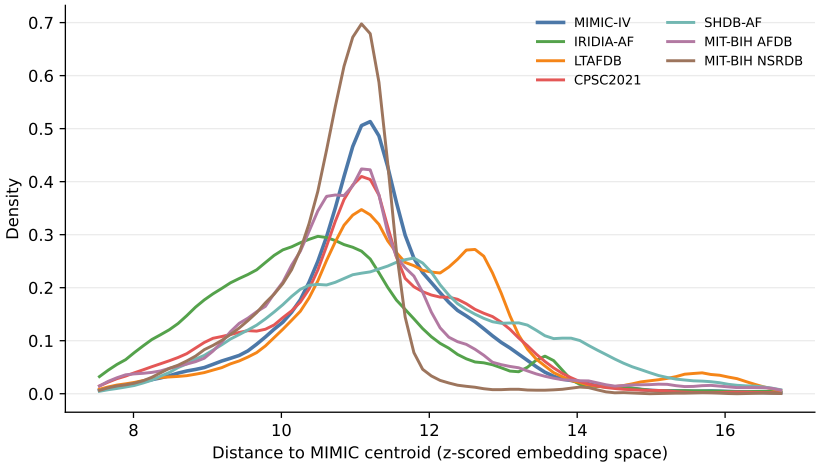
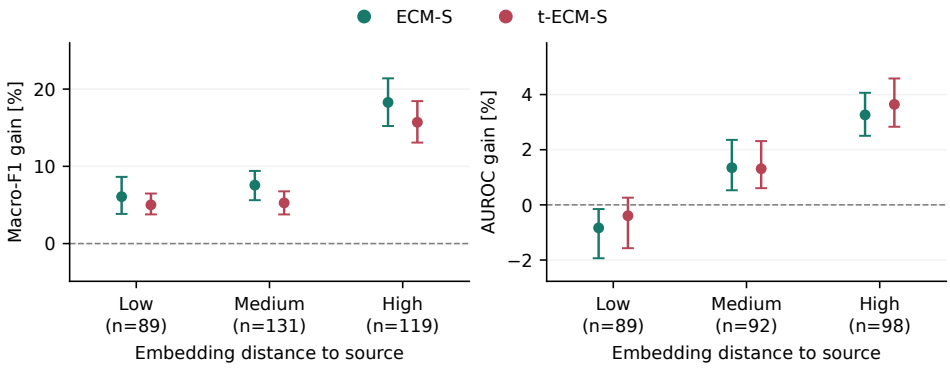
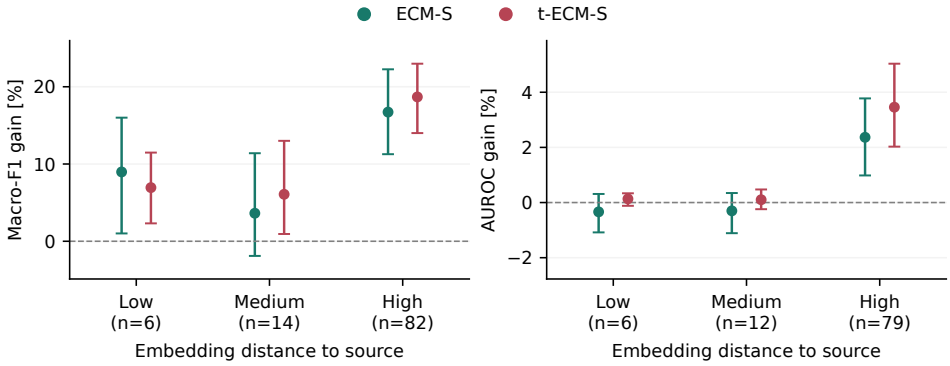


Figure 9: Embedding-domain shift between MIMIC-IV and the long-term ECG datasets.

Figure 10 reports paired gains over the frozen population predictor in distance strata defined by calibration-patient tertiles separately for each seed. In the low-distance internal stratum, ECM-S changed AUROC by -0.8%, whereas in the high-distance internal stratum, ECM-S and t-ECM-S improved macro-F1 by +18.3% and +15.7%, respectively. The corresponding external gains were +16.7% and +18.7%, with AUROC gains of +2.4% and +3.5%. These results complement the covariate shift analysis, suggesting that ECM's gains over the population model tended to be larger for deployment recordings farther from the source distribution.



(a) Internal deployment.



(b) External deployment.

Figure 10: Paired adaptation gains by embedding distance to source. Each patient's distance is the median nearest-class Mahalanobis distance across evaluated windows. Let δ_p^z denote patient p's embedding distance to source. Low, medium, and high denote distances $\delta_p^z < q_{1/3}$, $q_{1/3} \leq \delta_p^z < q_{2/3}$, and $\delta_p^z \geq q_{2/3}$, respectively, where $q_{1/3}$ and $q_{2/3}$ are calibration-patient distance quantiles computed separately for each seed and applied to both deployment cohorts. Points show patient-mean paired gains over the population predictor, whereas vertical bars indicate 95% dataset-stratified patient-bootstrap intervals after averaging seed observations within patient and stratum.

A.3 Additional experimental results

A.3.1 Extended deployment results

Complete deployment results. Table 4 shows the complete patient-level deployment results on internal held-out patients and external patients, supplementing the descriptions in Section 5.2.

Table 4: Patient-level ECG AF-detection performance on held-out internal and external datasets compared to the pretrained population model. Best scores are shown in bold.

(a) Held-out internal datasets.

Method	AF-containing ($n = 203$)						non-AF-only ($n = 36$)	
	Se.↑	Sp.↑	F1↑	BE↓	AUROC↑	AUPRC↑	Sp.↑	FA↓
Population	95.7	65.0	74.8	14.6	93.7	84.9	82.6	17.4
TENT	**95.8**	69.1	76.7	14.5	93.4	83.8	83.0	17.0
CoTTA	**95.8**	69.2	76.4	14.5	93.5	84.0	82.7	17.3
EATA	95.7	69.2	76.4	14.5	93.5	84.0	82.7	17.3
SAR	**95.8**	69.2	76.4	14.4	93.5	84.0	82.7	17.3
Fine-tuning	$92.2_{\pm 1.0}$	$\mathbf{93.6}_{\pm 1.1}$	$\mathbf{88.0}_{\pm 1.3}$	$\mathbf{3.9}_{\pm 0.6}$	$\mathbf{97.6}_{\pm 0.2}$	$\mathbf{92.4}_{\pm 1.7}$	$93.0_{\pm 3.4}$	$7.0_{\pm 3.4}$
CMP	$89.1_{\pm 3.4}$	$82.1_{\pm 4.1}$	$84.0_{\pm 2.4}$	$5.0_{\pm 1.8}$	$93.9_{\pm 0.7}$	$87.4_{\pm 1.6}$	$92.2_{\pm 2.8}$	$7.8_{\pm 2.8}$
EMA	$88.2_{\pm 3.6}$	$81.9_{\pm 3.7}$	$83.4_{\pm 1.4}$	$5.1_{\pm 1.3}$	$93.4_{\pm 1.5}$	$86.6_{\pm 1.6}$	$92.5_{\pm 2.3}$	$7.5_{\pm 2.3}$
LSTM	$89.8_{\pm 2.2}$	$83.2_{\pm 2.3}$	$86.3_{\pm 1.8}$	$4.6_{\pm 0.8}$	$94.5_{\pm 0.6}$	$89.2_{\pm 1.3}$	$93.7_{\pm 2.7}$	$6.3_{\pm 2.7}$
NLTM	$83.2_{\pm 10.7}$	$92.3_{\pm 2.4}$	$82.9_{\pm 3.2}$	$7.2_{\pm 1.6}$	$96.8_{\pm 0.7}$	$90.1_{\pm 2.0}$	$\mathbf{94.0}_{\pm 1.2}$	$\mathbf{6.0}_{\pm 1.2}$
ECM-S	$87.4_{\pm 3.8}$	$82.9_{\pm 2.7}$	$84.7_{\pm 1.2}$	$4.6_{\pm 0.9}$	$94.2_{\pm 0.5}$	$87.0_{\pm 1.7}$	$93.4_{\pm 2.4}$	$6.6_{\pm 2.4}$
t-ECM-S	$92.6_{\pm 1.6}$	$79.3_{\pm 2.8}$	$83.1_{\pm 1.3}$	$6.3_{\pm 0.8}$	$94.3_{\pm 0.1}$	$87.4_{\pm 1.5}$	$91.4_{\pm 3.1}$	$8.6_{\pm 3.1}$

(b) External datasets.

Method	AF-containing ($n = 95$)						non-AF-only ($n = 38$)	
	Se.↑	Sp.↑	F1↑	BE↓	AUROC↑	AUPRC↑	Sp.↑	FA↓
Population	94.5	65.0	63.6	24.0	91.0	77.4	77.2	22.8
TENT	94.2	66.7	65.2	22.7	91.6	77.5	77.2	22.8
CoTTA	**94.6**	65.8	64.0	24.0	91.9	77.1	77.2	22.8
EATA	**94.6**	65.8	64.0	24.0	91.9	77.1	77.1	22.9
SAR	**94.6**	66.1	64.2	23.8	91.9	77.3	77.1	22.9
Fine-tuning	$69.5_{\pm 6.5}$	$\mathbf{94.6}_{\pm 1.9}$	$75.2_{\pm 2.0}$	$11.1_{\pm 1.8}$	$90.0_{\pm 2.7}$	$79.4_{\pm 3.0}$	$98.0_{\pm 1.0}$	$2.0_{\pm 1.0}$
CMP	$72.2_{\pm 8.7}$	$89.9_{\pm 3.7}$	$76.8_{\pm 2.8}$	$9.7_{\pm 2.3}$	$90.8_{\pm 3.2}$	$76.9_{\pm 3.2}$	$96.3_{\pm 3.3}$	$3.7_{\pm 3.3}$
EMA	$72.7_{\pm 7.9}$	$90.4_{\pm 4.1}$	$77.7_{\pm 2.6}$	$9.0_{\pm 2.4}$	$91.6_{\pm 1.6}$	$77.8_{\pm 2.0}$	$96.6_{\pm 3.4}$	$3.4_{\pm 3.4}$
LSTM	$71.4_{\pm 9.3}$	$89.9_{\pm 5.3}$	$76.5_{\pm 3.9}$	$10.3_{\pm 3.1}$	$90.0_{\pm 3.1}$	$77.8_{\pm 4.1}$	$96.9_{\pm 2.0}$	$3.1_{\pm 2.0}$
NLTM	$56.5_{\pm 15.3}$	$92.1_{\pm 3.8}$	$67.2_{\pm 4.2}$	$16.2_{\pm 2.6}$	$90.3_{\pm 1.3}$	$74.1_{\pm 2.8}$	$96.6_{\pm 2.7}$	$3.4_{\pm 2.7}$
ECM-S	$69.9_{\pm 6.9}$	$92.8_{\pm 1.7}$	$78.5_{\pm 2.4}$	$7.8_{\pm 1.5}$	$92.7_{\pm 0.7}$	$78.4_{\pm 1.3}$	$\mathbf{98.4}_{\pm 1.7}$	$\mathbf{1.6}_{\pm 1.7}$
t-ECM-S	$82.5_{\pm 5.2}$	$89.9_{\pm 4.6}$	$\mathbf{80.4}_{\pm 3.1}$	$\mathbf{7.6}_{\pm 3.1}$	$\mathbf{93.7}_{\pm 1.0}$	$\mathbf{80.8}_{\pm 1.6}$	$94.6_{\pm 4.0}$	$5.4_{\pm 4.0}$

Performance across elapsed monitoring intervals. To characterize performance over the monitoring horizon, we divided each external ECG stream into non-overlapping intervals of 0–30 minutes, 30–60 minutes, 1–3 hours, 3–10 hours, and more than 10 hours. F1 and BE use AF-containing recordings, FA uses non-AF-only recordings. Metrics are calculated per patient and interval, averaged equally across patients within each seed, and summarized as mean and sample SD across five seeds in Table 5.

Table 5: External ECG performance across non-overlapping elapsed intervals.

		AF-containing		non-AF-only
Interval	Method	F1 ↑	BE ↓	FA ↓
0–30 min	Population	79.1	28.5	16.4
	EMA	$88.4_{\pm 1.3}$	$14.2_{\pm 1.7}$	$1.0_{\pm 1.1}$
	LSTM	$\mathbf{89.9}_{\pm 3.3}$	$14.5_{\pm 3.3}$	$2.7_{\pm 3.4}$
	ECM-S	$88.8_{\pm 0.6}$	$\mathbf{13.1}_{\pm 0.9}$	$\mathbf{0.6}_{\pm 0.6}$
	t-ECM-S	$88.8_{\pm 2.8}$	$13.7_{\pm 3.6}$	$3.3_{\pm 4.0}$
30–60 min	Population	77.6	30.3	17.0
	EMA	$88.7_{\pm 1.7}$	$12.7_{\pm 1.8}$	$1.5_{\pm 2.1}$
	LSTM	$\mathbf{90.8}_{\pm 3.5}$	$13.1_{\pm 4.0}$	$2.2_{\pm 2.7}$
	ECM-S	$89.5_{\pm 0.4}$	$\mathbf{11.7}_{\pm 0.5}$	$\mathbf{0.4}_{\pm 0.4}$
	t-ECM-S	$90.4_{\pm 1.5}$	$12.1_{\pm 3.4}$	$2.9_{\pm 3.6}$
1–3 h	Population	76.1	27.7	12.8
	EMA	$87.1_{\pm 1.9}$	$11.8_{\pm 2.3}$	$0.7_{\pm 0.7}$
	LSTM	$87.0_{\pm 3.4}$	$13.1_{\pm 4.0}$	$1.5_{\pm 1.8}$
	ECM-S	$\mathbf{88.1}_{\pm 0.8}$	$\mathbf{11.0}_{\pm 1.4}$	$\mathbf{0.2}_{\pm 0.2}$
	t-ECM-S	$87.9_{\pm 1.8}$	$11.0_{\pm 3.4}$	$1.5_{\pm 1.9}$
3–10 h	Population	73.9	24.3	15.2
	EMA	$84.2_{\pm 2.4}$	$10.7_{\pm 2.5}$	$1.2_{\pm 1.3}$
	LSTM	$84.7_{\pm 3.2}$	$11.8_{\pm 3.7}$	$1.9_{\pm 1.8}$
	ECM-S	$85.5_{\pm 1.5}$	$9.5_{\pm 1.8}$	$\mathbf{0.5}_{\pm 0.3}$
	t-ECM-S	$\mathbf{86.5}_{\pm 2.4}$	$\mathbf{8.9}_{\pm 3.1}$	$1.7_{\pm 2.1}$
$>$ 10 h	Population	68.9	24.6	27.6
	EMA	$81.6_{\pm 3.1}$	$9.2_{\pm 2.4}$	$4.6_{\pm 4.9}$
	LSTM	$81.1_{\pm 3.8}$	$10.9_{\pm 3.0}$	$3.8_{\pm 2.4}$
	ECM-S	$82.9_{\pm 1.7}$	$\mathbf{8.1}_{\pm 1.1}$	$\mathbf{2.3}_{\pm 2.7}$
	t-ECM-S	$\mathbf{84.0}_{\pm 3.0}$	$8.2_{\pm 2.8}$	$7.7_{\pm 5.4}$

Performance of ECM-P. Table 6 reports the patient-level results for ECM-P on the internal and external deployment datasets. On internal datasets, ECM-P obtained comparable F1, BE, AUROC and AUPRC to those of ECM-S, increasing mean sensitivity by 3.4%. The external evaluation highlights a higher AF sensitivity (+6.4%) with respect to ECM-S, while the remaining metrics favor the latter ECM variant. Thus, the addition of memory states at each transformer block did not result in clear improvements, while adding substantial computational burden (see Appendix Table 10).

Table 6: Patient-level ECM-P performance across deployment cohorts.

Signal	**Evaluation**	**AF-containing**						**non-AF-only**	
		Se.↑	Sp.↑	F1	BE↓	AUROC↑	AUPRC↑	Sp.↑	FA↓
ECG	Internal	$90.8_{\pm 1.5}$	$81.4_{\pm 1.9}$	$83.6_{\pm 1.3}$	$5.0_{\pm 0.9}$	$93.7_{\pm 0.9}$	$87.8_{\pm 1.2}$	$92.7_{\pm 2.1}$	$7.3_{\pm 2.1}$
ECG	External	$76.3_{\pm 5.3}$	$90.3_{\pm 2.3}$	$77.9_{\pm 1.5}$	$9.0_{\pm 1.0}$	$91.3_{\pm 0.7}$	$77.7_{\pm 1.3}$	$96.2_{\pm 2.0}$	$3.8_{\pm 2.0}$

Paired external patient uncertainty The standard deviations across calibration seeds in the main tables quantify training variability, but not uncertainty due to sampling the external patient population. We therefore treated the five fitted models as fixed, first averaged each patient's metric across seeds, and then performed 20,000 paired patient-bootstrap resamples at 95% confidence within external dataset strata. The same sampled patients were used for both methods in every replicate. We compared ECM-S and t-ECM-S with EMA and LSTM, the best-performing non-ECM longitudinal comparators in the pooled external ECG analysis, and directly compared the two ECM variants. Table 7 shows that ECM-S reduced absolute BE and FA rate and improved AUROC relative to both

comparators. F1 improvement was supported against LSTM, whereas the interval relative to EMA narrowly included zero. t-ECM-S improved patient-macro F1, BE, and AUROC relative to both comparators, but increased FA rate among non-AF-only patients. Relative to ECM-S, t-ECM-S improved F1 and AUROC, showed no clear difference in BE, and increased FA rate.

Table 7: Paired external patient bootstrap differences in absolute percentage points.

Target − reference	ΔF1 ↑	ΔBE ↓	ΔAUROC ↑	ΔFA ↓
ECM-S − EMA	$0.8\,[-0.1, 1.7]$	$-1.2\,[-2.1, -0.4]$	$1.1\,[0.3, 2.0]$	$-1.8\,[-2.8, -0.9]$
ECM-S − LSTM	$2.0\,[0.9, 3.2]$	$-2.4\,[-3.5, -1.5]$	$2.6\,[1.3, 4.1]$	$-1.5\,[-2.6, -0.6]$
t-ECM-S − EMA	$2.7\,[1.6, 3.8]$	$-1.5\,[-2.5, -0.6]$	$2.1\,[1.1, 3.2]$	$2.0\,[1.3, 2.7]$
t-ECM-S − LSTM	$3.9\,[2.4, 5.4]$	$-2.7\,[-3.8, -1.7]$	$3.6\,[2.2, 5.3]$	$2.2\,[1.4, 2.9]$
t-ECM-S − ECM-S	$1.9\,[0.6, 3.1]$	$-0.3\,[-1.5, 0.9]$	$1.0\,[0.6, 1.5]$	$3.8\,[2.9, 4.7]$

Table 8: Segment-level performance across deployment cohorts.

(a) Held-out internal datasets.

Method	**AF-containing** ($n = 203$)				**non-AF-only** ($n = 36$)	
	Se.↑	Sp.↑	F1	BE↓	Sp.↑	FA↓
Population	98.7	74.1	82.0	16.7	84.2	15.8
TENT	**98.8**	74.1	82.0	16.7	84.6	15.4
CoTTA	98.7	74.2	82.1	16.6	84.3	15.7
EATA	98.7	74.2	82.1	16.6	84.3	15.7
SAR	98.7	74.3	82.1	16.5	84.3	15.7
Fine-tuning	$95.0_{\pm 1.5}$	**96.9**$_{\pm 0.7}$	**95.8**$_{\pm 0.6}$	**0.7**$_{\pm 0.5}$	$93.1_{\pm 3.4}$	$6.9_{\pm 3.4}$
CMP	$94.6_{\pm 2.0}$	$93.6_{\pm 4.4}$	$93.4_{\pm 2.4}$	$2.6_{\pm 3.2}$	$92.3_{\pm 2.8}$	$7.7_{\pm 2.8}$
EMA	$94.4_{\pm 2.7}$	$93.3_{\pm 3.5}$	$93.1_{\pm 1.6}$	$3.1_{\pm 2.3}$	$92.4_{\pm 2.5}$	$7.6_{\pm 2.5}$
LSTM	$95.0_{\pm 1.3}$	$95.0_{\pm 2.1}$	$94.5_{\pm 1.1}$	$2.1_{\pm 0.9}$	$93.4_{\pm 2.7}$	$6.6_{\pm 2.7}$
NLTM	$86.0_{\pm 9.8}$	$95.8_{\pm 1.6}$	$91.4_{\pm 3.1}$	$3.4_{\pm 3.1}$	**93.9**$_{\pm 1.1}$	**6.1**$_{\pm 1.1}$
ECM-S	$94.3_{\pm 1.7}$	$94.6_{\pm 2.6}$	$93.9_{\pm 1.3}$	$1.8_{\pm 2.0}$	$93.0_{\pm 2.4}$	$7.0_{\pm 2.4}$
ECM-P	$95.0_{\pm 1.4}$	$92.4_{\pm 2.2}$	$92.7_{\pm 1.2}$	$3.3_{\pm 1.7}$	$92.0_{\pm 2.4}$	$8.0_{\pm 2.4}$
t-ECM-S	$96.4_{\pm 1.5}$	$90.7_{\pm 2.4}$	$92.1_{\pm 1.1}$	$4.9_{\pm 2.0}$	$91.3_{\pm 2.6}$	$8.7_{\pm 2.6}$

(b) External datasets.

Method	**AF-containing** ($n = 95$)				**non-AF-only** ($n = 38$)	
	Se.↑	Sp.↑	F1	BE↓	Sp.↑	FA↓
Population	**93.5**	62.5	70.3	25.0	74.8	25.2
TENT	**93.5**	64.9	71.8	23.4	74.8	25.2
CoTTA	**93.5**	62.5	70.3	25.0	74.7	25.3
EATA	**93.5**	62.5	70.2	25.1	74.7	25.3
SAR	93.4	63.0	70.5	24.7	74.7	25.3
Fine-tuning	$62.1_{\pm 9.6}$	**95.1**$_{\pm 2.2}$	$80.7_{\pm 3.2}$	$7.2_{\pm 4.1}$	$98.2_{\pm 0.9}$	$1.8_{\pm 0.9}$
CMP	$76.0_{\pm 7.5}$	$90.9_{\pm 4.5}$	$83.5_{\pm 3.6}$	$3.8_{\pm 1.5}$	$96.1_{\pm 3.2}$	$3.9_{\pm 3.2}$
EMA	$77.7_{\pm 7.2}$	$91.6_{\pm 5.0}$	$84.8_{\pm 3.3}$	$3.9_{\pm 2.4}$	$96.5_{\pm 3.5}$	$3.5_{\pm 3.5}$
LSTM	$74.1_{\pm 9.6}$	$90.6_{\pm 6.2}$	$82.5_{\pm 4.9}$	$4.3_{\pm 3.5}$	$96.5_{\pm 2.5}$	$3.5_{\pm 2.5}$
NLTM	$48.6_{\pm 17.5}$	$93.0_{\pm 3.8}$	$72.1_{\pm 6.1}$	$10.5_{\pm 5.8}$	$96.8_{\pm 2.9}$	$3.2_{\pm 2.9}$
ECM-S	$75.9_{\pm 6.1}$	$94.5_{\pm 1.9}$	$86.3_{\pm 2.2}$	$3.3_{\pm 2.0}$	**98.5**$_{\pm 1.5}$	**1.5**$_{\pm 1.5}$
ECM-P	$78.0_{\pm 4.4}$	$91.3_{\pm 2.9}$	$84.6_{\pm 1.1}$	**2.1**$_{\pm 2.1}$	$96.2_{\pm 1.9}$	$3.8_{\pm 1.9}$
t-ECM-S	$84.1_{\pm 4.5}$	$91.6_{\pm 5.9}$	**87.3**$_{\pm 3.7}$	$3.7_{\pm 3.4}$	$94.9_{\pm 4.2}$	$5.1_{\pm 4.2}$

Segment-level performance. Table 8 reports the segment-level results. Segment-level metrics pool predictions across patients, giving longer recordings greater influence. The segment-level results mirror the patient-level scores shown in Section 5.2, with fine-tuning performing best internally, while t-ECM-S and ECM-S lead the external evaluation.

A.3.2 CALIBRATION-VALIDATION PERFORMANCE.

Table 9: Patient-level ECG performance across the five calibration-validation cohorts.

Method	**AF-containing**						**non-AF-only**	
	Se.↑	Sp.↑	F1↑	BE↓	AUROC↑	AUPRC↑	Sp.↑	FA↓
Population	98.3	59.8	75.3	13.4	91.8	88.0	75.7	24.3
TENT	**98.4**	59.6	75.3	13.5	92.3	87.9	76.4	23.6
CoTTA	**98.4**	59.8	75.3	13.4	91.8	88.0	75.7	24.3
EATA	**98.4**	59.8	75.3	13.4	91.8	88.0	75.6	24.4
SAR	**98.4**	59.8	75.3	13.4	91.8	88.0	75.7	24.3
Fine-tuning	$96.0_{\pm 2.2}$	**93.3**$_{\pm 3.3}$	**89.2**$_{\pm 2.4}$	**3.0**$_{\pm 1.6}$	**98.4**$_{\pm 0.6}$	**97.5**$_{\pm 2.1}$	**91.8**$_{\pm 13.4}$	**8.2**$_{\pm 13.4}$
CMP	$91.5_{\pm 3.6}$	$74.8_{\pm 8.5}$	$83.8_{\pm 4.8}$	$4.5_{\pm 3.8}$	$91.5_{\pm 2.6}$	$91.1_{\pm 4.5}$	$85.8_{\pm 21.9}$	$14.2_{\pm 21.9}$
EMA	$90.4_{\pm 3.8}$	$74.7_{\pm 8.2}$	$83.4_{\pm 4.0}$	$4.5_{\pm 3.5}$	$91.0_{\pm 2.4}$	$90.7_{\pm 4.7}$	$84.3_{\pm 24.6}$	$15.7_{\pm 24.6}$
LSTM	$95.7_{\pm 2.2}$	$78.2_{\pm 9.8}$	$86.7_{\pm 4.2}$	$3.2_{\pm 1.3}$	$94.1_{\pm 2.2}$	$93.8_{\pm 2.8}$	$87.9_{\pm 24.0}$	$12.1_{\pm 24.0}$
NLTM	$94.1_{\pm 2.5}$	$92.0_{\pm 4.9}$	$86.8_{\pm 3.7}$	$4.2_{\pm 2.6}$	$98.1_{\pm 1.0}$	$97.0_{\pm 2.0}$	$91.2_{\pm 16.8}$	$8.8_{\pm 16.8}$
ECM-S	$93.5_{\pm 2.6}$	$77.3_{\pm 9.2}$	$85.4_{\pm 4.1}$	$3.3_{\pm 1.4}$	$93.3_{\pm 2.3}$	$91.7_{\pm 2.4}$	$87.3_{\pm 25.7}$	$12.7_{\pm 25.7}$
t-ECM-S	$96.8_{\pm 2.6}$	$73.8_{\pm 11.6}$	$83.7_{\pm 4.5}$	$5.4_{\pm 1.9}$	$92.9_{\pm 2.2}$	$91.0_{\pm 3.3}$	$85.9_{\pm 25.5}$	$14.1_{\pm 25.5}$

A.3.3 COMPUTATIONAL EFFICIENCY

Table 10 reports a computational efficiency analysis of the adaptation methods. Under a controlled CUDA-synchronized SHDB-AF streaming benchmark, ECM-S used 216.6k learned adaptation parameters and 8.06 KiB of persistent state per patient, while adding only $0.53_{\pm 0.35}$ ms of online processing per 10-s window relative to the frozen model. t-ECM-S retained the same patient-state footprint with 279.0k parameters and $0.48_{\pm 0.42}$ ms incremental latency. ECM-P increased the adapter to 1.30M parameters and the state to 48.38 KiB. TENT incurred the largest peak evaluation-time GPU allocation, whereas NLTM required the largest persistent patient state (1.57 MiB). ECM-S therefore provided a favorable efficiency–performance compromise among the comparators.

Table 10: Computational efficiency under a controlled SHDB-AF streaming benchmark.

Method	Trainable parameters↓	Online adapt. [ms/window]↓	Inference [ms/window]↓	Peak GPU [MiB]↓	State/patient [KiB]↓
Population	0	–	$3.65_{\pm 0.38}$	$867.9_{\pm 0.0}$	0.00
Fine-tuning	1.82M	offline	$3.69_{\pm 0.48}$	$867.9_{\pm 0.0}$	0.00
TENT	3.8k	$1.09_{\pm 0.46}$	$4.75_{\pm 0.27}$	$6661.0_{\pm 0.0}$	44.25
NLTM	423.2k	$0.14_{\pm 0.16}$	$3.65_{\pm 0.04}$	$3843.6_{\pm 0.0}$	1603.62
CMP	99.2k	$0.55_{\pm 0.34}$	$4.21_{\pm 0.04}$	$4589.5_{\pm 0.0}$	0.00
EMA	99.2k	$0.51_{\pm 0.43}$	$4.17_{\pm 0.05}$	$4589.5_{\pm 0.0}$	0.50
ECM-S	216.6k	$0.53_{\pm 0.35}$	$4.19_{\pm 0.10}$	$4589.5_{\pm 0.0}$	8.06
ECM-P	1.30M	$0.60_{\pm 0.42}$	$4.25_{\pm 0.16}$	$4589.5_{\pm 0.0}$	48.38
t-ECM-S	279.0k	$0.48_{\pm 0.42}$	$4.14_{\pm 0.08}$	$4589.5_{\pm 0.0}$	8.06

A.4 ADDITIONAL MECHANISTIC ANALYSES

Memory utilization and dynamics. Figure 11 provides a complementary view of the learned memory. Although ECM-S contains 16 slots, their centered effective rank was 9.2 ± 0.3, while only

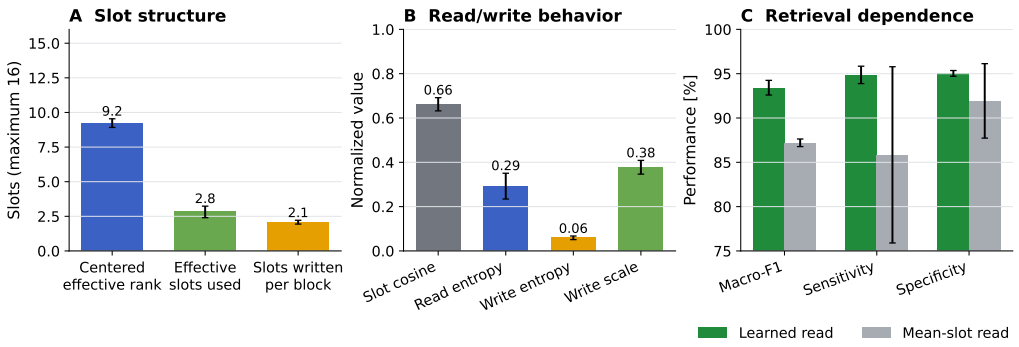


Figure 11: Detailed ECM-S memory behavior.

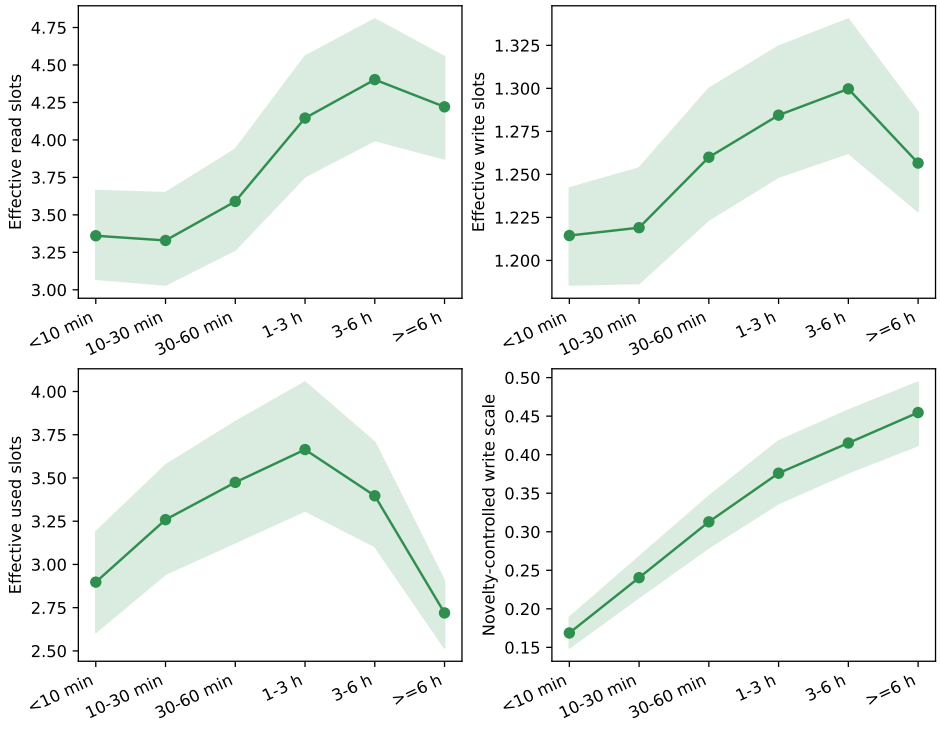


Figure 12: Slot utilization and memory dynamics across 3 seeds.

2.1 ± 0.1 slots were meaningfully written per context block. These measurements indicate that the learned memory maintains non-redundant slot structure while concentrating each context update on a small subset of slots. Low write entropy and the small number of active slots likely indicate sparse assignment, whereas the performance gap between learned and mean-slot retrieval suggests that the query-dependent attention selects useful content at inference.

Memory behavior across the monitoring horizon. Fig. 12 summarizes ECM-S memory behavior across successive monitoring-duration intervals. The number of effectively retrieved slots increased with monitoring duration, whereas each write remained concentrated on approximately one slot. As patient history accumulated, predictions drew from a broader set of stored states without updating the entire memory indiscriminately. The number of effectively used slots increased during the first hours and subsequently declined, suggesting slot reuse. Meanwhile, the novelty-controlled write scale increased steadily, indicating that novel observations received progressively stronger updates during longer recordings.

A.5 Ablation study

Contribution of patient context. To distinguish patient-state adaptation from a learned correction of the frozen population representation, we evaluated a calibration-trained stateless residual adapter on all external patients (Table 11). The stateless adapter conditioned its correction on the current query without patient history. Across seeds 42, 67, and 89, it achieved 81.0% macro-F1, 90.9% AUROC, and 10.1% BE, compared with 84.6%, 93.7%, and 6.2% for ECM-S. We also evaluated a parameter-matched causal-attention adapter over frozen embeddings, with a causal mask and no state carried between episodes. The adapter achieved 83.0% macro-F1, 92.5% AUROC, and 6.5% BE. ECM-S had better mean scores in this three-seed comparison.

Causal probability smoothing. To test whether temporal smoothing alone explains ECM's gains, we evaluated a causal EMA of the frozen population model's AF probabilities.
Let $q_n^{\mathrm{AF}} = [q_{\hat{\theta}}(\boldsymbol{x}_n)]_{\mathrm{AF}}$ be the frozen model's AF probability.
Its causal EMA is $\bar{q}_n^{\mathrm{AF}} = \alpha_{\mathrm{prob}} q_n^{\mathrm{AF}} + (1 - \alpha_{\mathrm{prob}})\bar{q}_{n-1}^{\mathrm{AF}}$, initialized with $\bar{q}_1^{\mathrm{AF}} = q_1^{\mathrm{AF}}$ and reset for each patient. We selected $\alpha_{\mathrm{prob}} = 0.02$ using calibration-validation patient-macro F1 and fixed it for deployment. Table 11 reports the results on external patients. Probability EMA improved macro-F1 from 69.7% to 73.2% over the population model but reduced AUROC from 91.9% to 88.1% and only slightly reduced BE, from 23.7% to 23.2%. ECM-S achieved higher macro-F1 and AUROC and substantially lower BE than the smoother, indicating that its benefits are not explained by this simple causal probability-smoothing baseline.

Table 11: Context and temporal-smoothing controls on all external patients across 3 seeds.

Configuration	Macro-F1↑	AUROC↑	BE↓
Population	69.7	91.9	23.7
Probability EMA	73.2	88.1	23.2
Stateless residual adapter	$81.0_{\pm 1.7}$	$90.9_{\pm 2.2}$	$10.1_{\pm 1.6}$
Causal-attention adapter	$83.0_{\pm 1.6}$	$92.5_{\pm 1.7}$	$6.5_{\pm 1.8}$
ECM-S	$\mathbf{84.6}_{\pm 1.9}$	$\mathbf{93.7}_{\pm 0.7}$	$\mathbf{6.2}_{\pm 1.3}$

Contributions of slow memory and routing. Table 12 separates the contribution of slow memory from that of the causal router. We evaluated a recent-context-only variant, which initializes its state from the last 6 context window embeddings and combines that state with the current query embedding, which is used to produce a gated residual correction. This variant is matched in terms of capacity and parameters to ECM-S. After predicting each query, the state is updated by exponential moving average with decay coefficient $\beta_f = 0.5$. The fast plus slow memory variant uses the same mechanism combined with ECM-S's persistent memory. Moreover, we evaluated t-ECM-S without slow memory and without the router. Across all external patients, adding slow memory to the recent-context-only model increased mean macro-F1 from 83.4% to 84.7%, and AUROC from 92.5% to 93.8%, while reducing BE from 7.6% to 7.1%. Full t-ECM-S likewise achieved higher mean macro-F1 (84.9% versus 81.3%) and AUROC (94.4% versus 93.0%) and lower BE (8.1% versus 10.8%) than the trained router model without slow memory. In contrast, bypassing the router increased mean macro-F1 to 85.3% and reduced BE to 6.8%, with nearly unchanged AUROC (94.3%). Thus, routing did not improve macro-F1 or BE in this comparison.

Table 12: Slow-memory and routing ablations on all external patients across 3 seeds.

Configuration	Macro-F1↑	AUROC↑	BE↓
Recent-context-only	$83.4_{\pm 2.0}$	$92.5_{\pm 1.6}$	$7.6_{\pm 1.6}$
Fast + slow memory	$84.7_{\pm 1.9}$	$93.8_{\pm 1.0}$	$7.1_{\pm 2.6}$
t-ECM-S without slow memory	$81.3_{\pm 2.9}$	$93.0_{\pm 0.8}$	$10.8_{\pm 3.3}$
t-ECM-S without router	$\mathbf{85.3}_{\pm 2.5}$	$94.3_{\pm 1.4}$	$\mathbf{6.8}_{\pm 3.0}$
Full t-ECM-S	$84.9_{\pm 3.2}$	$\mathbf{94.4}_{\pm 1.0}$	$8.1_{\pm 3.8}$

Context length, query length, and memory capacity. We conducted further ablations to determine the effects of context length, query length (Table 13), and memory capacity (Table 14). The results of the first analysis favored intermediate context lengths, while the preferred temporal scale varied somewhat across metrics. At $K = 10$, increasing Q from 5 to 50 raised external F1 by 8.8% and reduced BE by 6.6%, suggesting lower stability as Q decreases. With $Q = 50$, no single context length dominated all external metrics: $K = 25$ gave the highest external F1 and lowest FA rate, whereas $K = 50$ gave the best external AUROC and BE. Longer contexts did not consistently improve the reported metrics. Memory capacity showed a similarly non-monotonic pattern: 16 slots gave the best internal F1 and BE and the best external AUROC and FA rate, whereas increasing capacity to 64 slots degraded all reported external metrics.

Table 13: Patient-level ECM-S context/query-length ablation across 3 seeds.

		Held-out internal				**External**			
K	Q	F1↑	AUROC↑	BE↓	FA↓	F1↑	AUROC↑	BE↓	FA↓
3	50	$81.0_{\pm 4.7}$	$93.7_{\pm 0.6}$	$7.3_{\pm 3.8}$	$9.6_{\pm 1.4}$	$75.9_{\pm 3.9}$	$91.1_{\pm 0.2}$	$10.6_{\pm 4.2}$	$6.3_{\pm 7.2}$
6	50	$79.7_{\pm 3.0}$	$92.6_{\pm 0.5}$	$7.7_{\pm 3.0}$	$10.6_{\pm 1.9}$	$75.0_{\pm 2.3}$	$90.9_{\pm 0.2}$	$11.7_{\pm 2.5}$	$8.1_{\pm 5.0}$
10	5	$75.2_{\pm 9.1}$	$92.6_{\pm 1.3}$	$12.0_{\pm 8.2}$	$7.9_{\pm 0.7}$	$69.4_{\pm 7.3}$	$90.6_{\pm 2.7}$	$15.5_{\pm 6.1}$	$5.1_{\pm 4.4}$
10	10	$81.9_{\pm 4.6}$	$93.4_{\pm 0.7}$	$6.6_{\pm 3.6}$	$9.1_{\pm 2.0}$	$76.7_{\pm 3.5}$	$91.4_{\pm 0.6}$	$10.6_{\pm 3.0}$	$5.6_{\pm 6.4}$
10	50	$82.0_{\pm 2.2}$	$92.4_{\pm 1.3}$	$5.9_{\pm 0.8}$	$8.6_{\pm 2.7}$	$78.2_{\pm 1.9}$	$91.8_{\pm 0.9}$	$8.9_{\pm 1.4}$	$4.7_{\pm 1.9}$
25	50	$\mathbf{85.0}_{\pm 1.2}$	$93.6_{\pm 1.4}$	$4.4_{\pm 0.0}$	$6.5_{\pm 2.1}$	$\mathbf{78.3}_{\pm 1.7}$	$91.9_{\pm 1.0}$	$8.4_{\pm 1.5}$	$\mathbf{0.6}_{\pm 0.2}$
50	50	$84.9_{\pm 0.8}$	$\mathbf{94.1}_{\pm 0.7}$	$\mathbf{4.2}_{\pm 0.8}$	$\mathbf{5.9}_{\pm 2.3}$	$77.8_{\pm 2.7}$	$\mathbf{92.5}_{\pm 0.7}$	$\mathbf{8.2}_{\pm 2.0}$	$1.1_{\pm 0.8}$
100	50	$81.3_{\pm 5.0}$	$93.9_{\pm 0.8}$	$7.1_{\pm 4.4}$	$8.9_{\pm 2.8}$	$75.1_{\pm 4.0}$	$91.9_{\pm 0.9}$	$10.9_{\pm 4.6}$	$5.8_{\pm 8.4}$
360	50	$82.7_{\pm 2.5}$	$93.3_{\pm 0.6}$	$5.6_{\pm 2.1}$	$9.4_{\pm 1.3}$	$76.7_{\pm 0.9}$	$91.7_{\pm 0.4}$	$9.3_{\pm 1.1}$	$4.3_{\pm 3.2}$

Table 14: Patient-level ECM-S memory-capacity ablation at $K = Q = 50$ across 3 seeds.

	Held-out internal				**External**			
Slots	F1↑	AUROC↑	BE↓	FA↓	F1↑	AUROC↑	BE↓	FA↓
2	$84.4_{\pm 0.4}$	$93.6_{\pm 0.1}$	$4.6_{\pm 0.3}$	$6.6_{\pm 3.7}$	$78.2_{\pm 2.0}$	$91.3_{\pm 1.7}$	$8.3_{\pm 1.8}$	$1.8_{\pm 0.9}$
4	$84.3_{\pm 0.5}$	$93.4_{\pm 0.8}$	$4.7_{\pm 0.2}$	$7.0_{\pm 3.3}$	$\mathbf{78.4}_{\pm 3.3}$	$91.8_{\pm 1.5}$	$8.2_{\pm 1.7}$	$1.4_{\pm 0.8}$
8	$83.0_{\pm 2.1}$	$\mathbf{94.1}_{\pm 1.1}$	$5.7_{\pm 1.4}$	$8.4_{\pm 1.5}$	$77.9_{\pm 1.2}$	$91.8_{\pm 0.2}$	$9.0_{\pm 1.0}$	$3.3_{\pm 2.2}$
16	$\mathbf{84.9}_{\pm 0.8}$	$\mathbf{94.1}_{\pm 0.7}$	$\mathbf{4.2}_{\pm 0.8}$	$\mathbf{5.9}_{\pm 2.3}$	$77.8_{\pm 2.7}$	$\mathbf{92.5}_{\pm 0.7}$	$8.2_{\pm 2.0}$	$\mathbf{1.1}_{\pm 0.8}$
32	$83.4_{\pm 1.5}$	$92.5_{\pm 0.1}$	$5.1_{\pm 0.9}$	$7.2_{\pm 2.8}$	$78.2_{\pm 1.9}$	$92.3_{\pm 0.5}$	$\mathbf{8.1}_{\pm 1.0}$	$3.0_{\pm 1.9}$
64	$81.1_{\pm 2.9}$	$92.7_{\pm 1.1}$	$7.5_{\pm 4.0}$	$9.9_{\pm 2.6}$	$73.0_{\pm 3.3}$	$90.4_{\pm 0.8}$	$13.8_{\pm 4.4}$	$9.2_{\pm 7.2}$

Calibration-cohort composition. Table 15 reports per-dataset external performance when varying the calibration cohort's proximity to the source distribution. We compared the original calibration cohort with source-near and source-far cohorts (distances defined as in Section A.2.3) at a matched calibration budget, keeping the validation cohort fixed. Across these resampled cohorts, ECM-S obtained the highest AUROC and lowest BE on SHDB-AF, while LSTM had the highest F1 and AUROC on MIT-BIH AF.

Table 15: External performance by calibration-cohort composition for selected comparators across 3 seeds.

Calibration composition	Method	SHDB-AF			MIT-BIH AF			MIT-BIH NSR
		F1↑	AUROC↑	BE↓	F1↑	AUROC↑	BE↓	FA↓
Original	Fine-tuning	$70.0_{\pm 2.0}$	$86.4_{\pm 4.5}$	$11.9_{\pm 0.8}$	$85.0_{\pm 1.3}$	$98.7_{\pm 0.2}$	$3.8_{\pm 0.3}$	$2.7_{\pm 1.5}$
	LSTM	$70.1_{\pm 6.5}$	$88.3_{\pm 4.1}$	$12.2_{\pm 4.3}$	$\mathbf{89.4}_{\pm 4.4}$	$\mathbf{98.8}_{\pm 0.3}$	$2.5_{\pm 1.0}$	$2.4_{\pm 3.9}$
	ECM-S	$\mathbf{73.9}_{\pm 1.8}$	$\mathbf{92.2}_{\pm 0.8}$	$\mathbf{8.1}_{\pm 1.9}$	$86.6_{\pm 1.4}$	$98.4_{\pm 0.3}$	$\mathbf{2.3}_{\pm 1.1}$	$\mathbf{1.3}_{\pm 1.5}$
Source-near	Fine-tuning	$63.6_{\pm 9.5}$	$80.5_{\pm 8.3}$	$15.8_{\pm 7.4}$	$83.2_{\pm 2.2}$	$97.8_{\pm 0.8}$	$6.2_{\pm 1.7}$	$2.5_{\pm 1.4}$
	LSTM	$\mathbf{74.3}_{\pm 3.8}$	$89.0_{\pm 4.0}$	$9.2_{\pm 1.5}$	$\mathbf{89.5}_{\pm 3.1}$	$\mathbf{98.6}_{\pm 0.1}$	$\mathbf{1.8}_{\pm 0.4}$	$6.2_{\pm 1.5}$
	ECM-S	$72.2_{\pm 1.5}$	$\mathbf{91.7}_{\pm 0.9}$	$\mathbf{8.3}_{\pm 0.1}$	$86.2_{\pm 1.9}$	$98.5_{\pm 0.2}$	$2.2_{\pm 0.6}$	$\mathbf{2.2}_{\pm 1.2}$
Source-far	Fine-tuning	$65.5_{\pm 2.7}$	$87.2_{\pm 3.3}$	$13.1_{\pm 2.9}$	$86.0_{\pm 2.6}$	$98.2_{\pm 0.2}$	$3.9_{\pm 2.5}$	$\mathbf{2.1}_{\pm 1.6}$
	LSTM	$73.1_{\pm 4.1}$	$91.3_{\pm 2.6}$	$9.2_{\pm 2.4}$	$\mathbf{91.0}_{\pm 1.9}$	$\mathbf{98.9}_{\pm 0.1}$	$\mathbf{2.2}_{\pm 0.1}$	$6.7_{\pm 5.9}$
	ECM-S	$\mathbf{74.9}_{\pm 3.0}$	$\mathbf{91.8}_{\pm 0.8}$	$\mathbf{8.0}_{\pm 1.0}$	$86.1_{\pm 3.2}$	$98.0_{\pm 0.9}$	$2.8_{\pm 1.0}$	$4.4_{\pm 4.9}$

A.6 Additional cardiovascular tasks

We provide implementation details and full results of the evaluation on the additional tasks described in Section 5.5. We followed the protocol used in the main experiments: pretraining a population model on short windows, calibrating ECM and its comparators, and evaluating longitudinal adaptation on long-term recordings from an external dataset.

A.6.1 Hypotension detection from PPG-ECG

Population pretraining and adaptation calibration. We evaluated binary hypotension detection from simultaneous 30-s PPG and Einthoven lead II ECG windows using a two-channel instantiation of the CNN-transformer. The population development cohort comprised 112,072 class-balanced windows from 2,139 MIMIC-III patients (Moody et al., 2020; Johnson et al., 2016). Adaptation calibration used 91,361 windows from 111 held-out patients. The dataset composition is summarized in Table 16. Windows with systolic blood pressure (SBP) < 85 mmHg or diastolic blood pressure (DBP) < 55 mmHg were assigned the hypotensive label, whereas negatives required SBP ≥ 95 and DBP ≥ 65. Intermediate windows remained unlabeled for classification. For the population development subset, window-level SBP and DBP were obtained by taking the median of the three released 10-s pressure summaries within each 30-s window. Both input channels were sampled at 250 Hz. PPG preprocessing comprised fourth-order Butterworth bandpass filtering at 0.5–10 Hz and z-normalization, whereas ECGs were standardized. Calibration patients were divided into approximately 80% training and 20% validation partitions using 3 seeds.

Table 16: ECG–PPG dataset composition for hypotension detection.

Set	n patients	n 30-s ECG–PPG windows	hypotension prevalence [%]
Pretraining	2,139	112,072	50.0
Calibration	111	91,361	48.3
External deployment	200	93,274	38.3

Longitudinal deployment. We evaluated the adaptation methods on 200 external patient recordings from VitalDB (Lee & Jung, 2022; Lee et al., 2022). VitalDB reference pressures were derived from arterial waveform systolic peaks and diastolic troughs and summarized by their medians within each window. Table 17 reports the deployment results of adaptation. ECM-S obtained the highest macro-F1 and lowest BE across all adaptation methods, increasing macro-F1 by +0.8% and decreasing BE by 1.37% relative to the population model, while approximately preserving population-model AUROC.

Table 17: Patient-level performance for PPG–ECG hypotension detection on the external deployment cohort across 3 seeds.

Method	Macro-F1↑	AUROC↑	BE↓
Population	57.5	**73.1**	23.1
TENT	42.1	61.3	39.6
CoTTA	57.5	**73.1**	23.1
EATA	57.4	**73.1**	23.1
SAR	52.3	69.1	28.8
Fine-tuning	$45.8_{\pm 1.7}$	$61.4_{\pm 2.2}$	$33.6_{\pm 1.9}$
EMA	$55.7_{\pm 2.3}$	$69.4_{\pm 2.0}$	$23.2_{\pm 2.4}$
CMP	$56.2_{\pm 1.5}$	$69.0_{\pm 2.3}$	$\mathbf{21.7}_{\pm 1.0}$
LSTM	$53.9_{\pm 4.1}$	$69.4_{\pm 4.3}$	$24.6_{\pm 2.5}$
NLTM	$53.7_{\pm 1.0}$	$67.1_{\pm 3.7}$	$23.6_{\pm 2.0}$
ECM-S	$\mathbf{58.3}_{\pm 1.1}$	$\mathbf{73.1}_{\pm 0.2}$	$\mathbf{21.7}_{\pm 1.0}$

A.6.2 Ectopy classification from ECG

Population pretraining and adaptation calibration. We formulated ectopy detection as four-class classification of 10-s ECG windows. We extracted single-lead ECGs from a class-balanced subset of Icentia11k (Tan et al., 2022; 2021), comprising 115,380 modified Einthoven lead I windows from 4,952 patients, with 28,845 windows in each class. We derived the labels from reference beat annotations according to the presence within the 10 seconds of supraventricular (S) and ventricular (V) ectopic beats: normal-only (N), S-containing, V-containing, and S+V. The ECGs were preprocessed following Goettling et al. (2024). The dataset composition is summarized in Table 18. We pretrained a new instantiation of the CNN-transformer on Icentia11k, allocating 10% of the patient-disjoint samples for testing and 10% of the remaining development data to validation. For calibration, we used a held-out subset comprising 133,792 windows from 140 patients. Each adaptation method was trained and validated using 3 seeds.

Table 18: ECG dataset composition for ectopy classification.

Set	n patients	n 10-s ECG windows	S / V / S+V prevalence [%]
Pretraining	4,952	115,380	25.0 / 25.0 / 25.0
Calibration	140	133,792	14.9 / 9.6 / 0.6
External deployment	84	679,029	7.1 / 9.4 / 1.6

Longitudinal deployment. For external longitudinal deployment, we extracted Einthoven lead II ECGs from LTAFDB (Petrutiu et al., 2007), which contains annotations of ectopic beats, and applied the same preprocessing protocol as for the pretraining data. Table 19 shows the results of deployment evaluation. ECM-S achieved the highest F1 and AUROC for N-only windows (94.0% and 92.0%) and S-containing windows (60.3% and 93.1%) among the reported methods. Fine-tuning performed best for V-containing windows, achieving an F1 of 71.0% compared with 67.8% for ECM-S. For mixed S+V windows, ECM-S achieved the highest mean AUROC (95.0%), whereas EMA attained the highest F1 (44.6%). ECM-S's mixed-class F1 remained variable across seeds ($33.3 \pm 18.3\%$). The results of these analyses show that ECM's benefits are not limited to binary AF detection from single-channel ECGs, suggesting its applicability across biosignal modalities and classification tasks.

Table 19: External deployment results of classification of ECGs containing normal beats only (N), supraventricular beats (S), ventricular beats (V), and mixed beats (S+V).

	N		S		V		S+V	
Method	F1 ↑	AUROC ↑	F1 ↑	AUROC ↑	F1 ↑	AUROC ↑	F1 ↑	AUROC ↑
Population	81.3	91.3	36.5	85.3	45.8	81.8	16.5	92.4
TENT	88.3	81.1	57.6	88.0	56.1	78.2	11.9	79.7
CoTTA	81.9	91.4	37.2	85.6	46.3	82.2	16.9	92.6
EATA	81.3	91.3	36.5	85.4	45.8	81.8	16.5	92.4
SAR	84.2	91.2	38.6	86.0	50.4	83.9	19.0	93.5
Fine-tuning	$88.9_{\pm 2.1}$	$89.5_{\pm 0.6}$	$41.2_{\pm 4.3}$	$88.6_{\pm 0.5}$	$\mathbf{71.0}_{\pm 0.5}$	$\mathbf{92.8}_{\pm 0.5}$	$38.2_{\pm 3.1}$	$94.9_{\pm 0.6}$
EMA	$93.0_{\pm 0.6}$	$90.6_{\pm 0.8}$	$55.6_{\pm 2.9}$	$92.6_{\pm 1.1}$	$65.3_{\pm 3.5}$	$89.4_{\pm 1.5}$	$\mathbf{44.6}_{\pm 3.5}$	$94.7_{\pm 0.6}$
LSTM	$90.8_{\pm 1.0}$	$90.5_{\pm 0.8}$	$48.0_{\pm 3.0}$	$92.7_{\pm 0.6}$	$67.0_{\pm 0.5}$	$91.3_{\pm 0.4}$	$34.5_{\pm 7.3}$	$92.5_{\pm 0.4}$
NLTM	$89.3_{\pm 0.2}$	$88.8_{\pm 0.3}$	$41.5_{\pm 0.2}$	$87.4_{\pm 0.1}$	$67.0_{\pm 0.4}$	$92.0_{\pm 0.1}$	$35.9_{\pm 1.5}$	$93.4_{\pm 0.1}$
ECM-S	$\mathbf{94.0}_{\pm 1.1}$	$\mathbf{92.0}_{\pm 0.6}$	$\mathbf{60.3}_{\pm 6.7}$	$\mathbf{93.1}_{\pm 1.1}$	$67.8_{\pm 2.4}$	$92.3_{\pm 0.3}$	$33.3_{\pm 18.3}$	$\mathbf{95.0}_{\pm 0.5}$