EVERY BEAT COUNTS: BEAT-ANCHORED PROBING OF FROZEN ECG ENCODERS FOR CROSS-SITE DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrocardiogram (ECG) diagnosis models that match cardiologists within one hospital degrade markedly at another, which hinders their clinical adoption. While existing efforts scale up encoder pretraining, we revisit this challenge from the readout side by asking where frozen encoders keep the information that cross-site diagnosis requires. Through phase-resolved probing of frozen tokens around each heartbeat, we find that such information is present and locked to the beat, yet mean pooling dilutes it for diagnoses confined to part of the beat. A multi-query attentive probe recovers part of it, but under diagnostic supervision alone its queries remain unassigned, covering only part of the beats and keeping to no phase of their own. Motivated by this, we propose the **Beat-Anchored Probe** (BAP), which expands one R-peak per beat into attention targets for the P, QRS, ST and T phases. Its training-only alignment loss allocates each query over every beat and registers it to one phase, leaving what the query reads to the diagnostic loss. With the encoder frozen, BAP achieves the highest mean transfer among the compared probes and improves on the linear probe on all four encoders, with the largest gains on the diagnoses that mean pooling dilutes, and its phase roles stay fixed across training seeds and decodable on an external benchmark. We further release **CPSC2018-R**, cardiologist-annotated R-peaks for CPSC2018, which provides the manual reference for phase-resolved probing and for verifying such probes. Code and data are at https://anonymous.4open.science/r/BAP-1490/.

1 INTRODUCTION

The electrocardiogram (ECG) records the electrical activity of the heart and is among the most widely used tests of cardiac health (Bhatia & Dorian, 2018; Yagi et al., 2024). Deep networks now match cardiologists on recordings from the site that trained them (Hannun et al., 2019; Ribeiro et al., 2020), yet degrade markedly at another site, where devices, patients and labeling practices differ (Reyna et al., 2021; Strodthoff et al., 2021), which hinders their clinical adoption. Recent studies have paid increasing attention to ECG foundation models, mainly improving the encoder by self-supervised pretraining on large corpora (Na et al., 2024; Liu et al., 2024; Jin et al., 2025; Xu et al., 2025; Li et al., 2025), and read the frozen encoder out with a linear classifier on its mean-pooled tokens (Liu et al., 2024). Despite the growing scale of pretraining, cross-site ECG diagnosis remains far from solved.

An ECG is read beat by beat. Each heartbeat traces the same sequence of deflections, namely the P wave as the atria activate, the QRS complex as the ventricles activate, and the ST segment and the T wave as the ventricles recover, and the sharp R-peak of the QRS complex marks the beat (Surawicz et al., 2009; Rautaharju et al., 2009). Many diagnoses are confined to part of this sequence. First-degree atrioventricular block (1AVB) delays the QRS complex after the P wave in every beat (Kusumoto et al., 2019), and premature atrial or ventricular contractions (PAC, PVC) appear only in occasional beats (Marcus, 2020). To read them, a model has to look at the right part of every beat of a record.

In this work, we revisit cross-site diagnosis from the readout side by asking where a frozen encoder keeps the information that it requires. Interestingly, we observe that the linear probe loses the most accuracy across sites on exactly these local diagnoses (cf. Figure 1a). Probing the frozen tokens by their time offset from each R-peak, which we term *phase-resolved probing*, we empirically find that the information is present and locked to the beat, e.g., 1AVB is read far better from the tokens over

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.

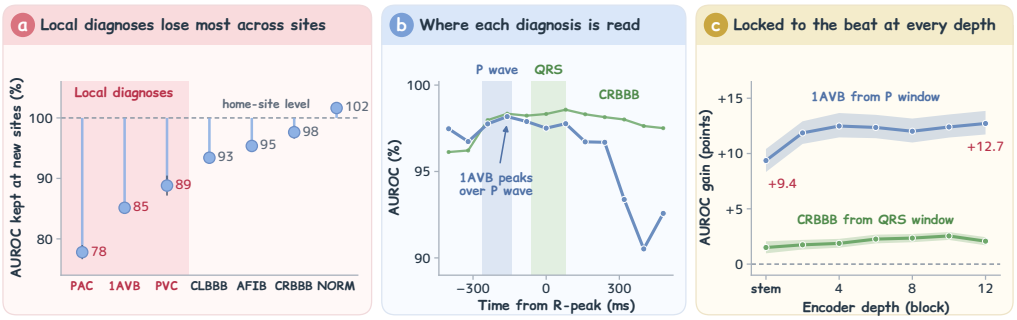


Figure 1: **Frozen tokens keep diagnostic information locked to the beat, and mean pooling dilutes it. (a)** Cross-site over in-domain AUROC of the linear probe per diagnosis. **(b)** In-domain AUROC of linear probes on block-12 tokens pooled around the R-peak. **(c)** Gain of the P window for 1AVB and the QRS window for CRBBB over shifted windows, by depth (95% bands).

the P wave of every beat than from the same window placed at another phase, at every depth of the encoder (cf. Figure 1b,c). Mean pooling averages these few informative tokens with all the others and dilutes them, which points to the readout as the place to intervene.

A natural remedy is a multi-query attentive probe (Bardes et al., 2024; Psomas et al., 2026), whose learnable queries each weight the tokens with their own attention, so that different queries can read different parts of the record. Such a probe indeed transfers better (cf. Table 1). Tracing its queries, we further observe that they remain unassigned, as no query settles on one phase and each attends to only some of the beats of a record (§2). The diagnostic loss favors no assignment of queries to phases, since it is unchanged when two queries swap roles together with their classifier weights, and it never asks a query to visit every beat. The information is thus present in the frozen tokens, yet unassigned to the queries.

Motivated by this, we propose the **Beat-Anchored Probe** (BAP), a simple yet effective probe that assigns each of four queries to one phase of the beat, i.e., P, QRS, ST or T, through *beat-anchored supervision* (Figure 2). As each phase lies at a typical offset from the R-peak (Surawicz et al., 2009), BAP expands one R-peak per beat into a target attention map for each query that covers its phase in every beat. A training-only alignment loss pulls each query toward its target. It spreads the query over every beat (*allocation*) and ties it to one phase (*registration*), and the diagnostic loss decides what the query reads. Unlike ECG models that build beats into the input or architecture (Hong et al., 2019; Plagwitz et al., 2026; Sameh et al., 2026), BAP trains only the probe and takes the ECG alone at inference. We also release **CPSC2018-R**, cardiologist-annotated R-peaks for CPSC2018 (Liu et al., 2018), as a manual reference for phase-resolved probing and for verifying probes.

Extensive experiments on five external cohorts and four encoders validate the effectiveness of BAP. With the encoder frozen, BAP achieves the highest mean transfer among the compared probes, exceeds low-rank adaptation of the encoder (Hu et al., 2022) and improves on the linear probe on all four encoders, with its largest gains on the local diagnoses that mean pooling dilutes (§4.2). Ablations trace this gain to allocation and the fixed query roles to registration (§4.3), and on the external LUDB benchmark (Kalyakulina et al., 2020) the registered queries label expert-delineated waves by their training names (§4.3). In summary, our contributions are threefold.

- We empirically reveal that frozen ECG tokens retain diagnostic information locked to the heartbeat and that mean pooling dilutes it for local diagnoses, and find that diagnosis-only attentive probes leave their queries unassigned.
- We propose BAP, a probe whose training-only beat-anchored targets allocate each query over every beat and register it to one phase, and release **CPSC2018-R**, cardiologist-annotated R-peaks for CPSC2018.
- With the encoder frozen, BAP transfers best among the compared probes, above all on the diluted local diagnoses, and ablations trace its gain to allocation and its fixed roles to registration.

2 PRESENT BUT UNASSIGNED: AN EMPIRICAL STUDY

In this section, we first reveal where frozen tokens keep diagnostic information and how mean pooling dilutes it, and then examine how a multi-query attentive probe reads these tokens.

Setup. We take MAE-Base as the example encoder, whose convolutional stem and 12 transformer blocks map a 10-s, 12-lead record to $T{=}125$ tokens of 80 ms (App. B). Probes on its frozen tokens $H \in \mathbb{R}^{T\times d}$ are trained on CPSC2018-R and tested on five external cohorts (§4.1) by the area under the ROC curve (AUROC) of seven diagnoses, i.e., normal (NORM), atrial fibrillation (AFIB), 1AVB, complete left and right bundle-branch block (CLBBB, CRBBB), PAC and PVC (App. A.1). To locate information within the beat, *phase-resolved probing* fits linear probes on tokens pooled by their offset from each R-peak, in short bins or in the four *phase windows* of §3.2 (App. B.2).

Frozen tokens retain diagnostic information locked to the beat. In-domain, linear probes on tokens pooled in 80-ms bins around the R-peak read 1AVB at up to 98.2 AUROC over the P wave and PR segment and lose up to 7.7 points in the bins more than 300 ms after the R-peak (Figure 1b). Across cohorts, reading 1AVB from the P window of every beat raises its AUROC at the last transformer block by 12.7 points over the same window moved to other phases, and the gain already appears at the convolutional stem (Figure 1c). Consistently, masking the P waves costs the linear probe (LP) on the mean-pooled token more on 1AVB than on any other diagnosis (Figure 3b).

Mean pooling dilutes the information of diagnoses confined to part of the beat. Across the five cohorts, LP keeps only 78 to 89% of its in-domain AUROC on the local diagnoses 1AVB, PAC and PVC, against 93% or more on the other diagnoses (Figure 1a, App. B.9).

Multi-query attentive probes read the tokens individually and leave their queries unassigned. A multi-query attentive probe (Bardes et al., 2024; Psomas et al., 2026) lets each learnable query q_k form its own attention-weighted average of the tokens,

$$z_k = \sum_{t=1}^{T} A_k(t)\, W_{\text{val}} H_t, \qquad A_k = \operatorname{softmax}\left(\frac{q_k^\top W_{\text{key}} H^\top}{\sqrt{d}}\right), \tag{1}$$

and a classifier reads $[z_1; \ldots; z_K]$. With $K{=}4$ and diagnosis as the only loss, it is the *unaligned probe* (UP), i.e., BAP (§3) without the alignment loss. Reading the tokens individually, UP recovers part of the diluted information and reaches a five-cohort mean AUROC of 86.9 on MAE-Base against 84.1 for LP, with its largest gains on 1AVB, PAC and PVC (Table 1, Figure 3a). Yet UP leaves its queries unassigned. The share of a query's attention inside its phase window, the *in-window share*, is 27.3% under the naming by index and only 32.4% under the naming most favorable to UP (App. D.1), close to the 25.0% of uniform attention. Each query gives at least half of its fair share of attention to 70.8% of the beats, the *beats covered*, against 99.9% for uniform (Table 4, App. B.11).

3 BEAT-ANCHORED PROBE

In this section, we first present BAP and its beat-anchored targets (§3.2), and then detail how the alignment loss allocates each query over every beat (§3.3) and registers it to one phase (§3.4). Finally, we introduce CPSC2018-R (§3.5).

3.1 OVERVIEW

Problem setup. Given a source corpus of ECG records x with multi-label diagnoses $y \in \{0,1\}^C$ over C classes, we train only a readout on a frozen pretrained encoder f_θ. The encoder maps x to T tokens $H = f_\theta(x) \in \mathbb{R}^{T\times d}$, each covering a short span of samples. For training only, each record carries the R-peaks $\{r_m\}_{m=1}^{M}$ of its M beats. Target cohorts supply the ECG alone.

Probe. As illustrated in Figure 2, BAP pools H with $K{=}4$ learnable queries $q_1, \ldots, q_4 \in \mathbb{R}^d$ by multi-head cross-attention (MHA). A classifier g then reads the pooled outputs,

$$z_k = W_{\text{out}}\,\text{MHA}(q_k, H), \qquad A_k = \frac{1}{4}\sum_{h=1}^{4} A_k^h, \qquad \hat{y} = g\big([z_1; z_2; z_3; z_4]\big), \tag{2}$$

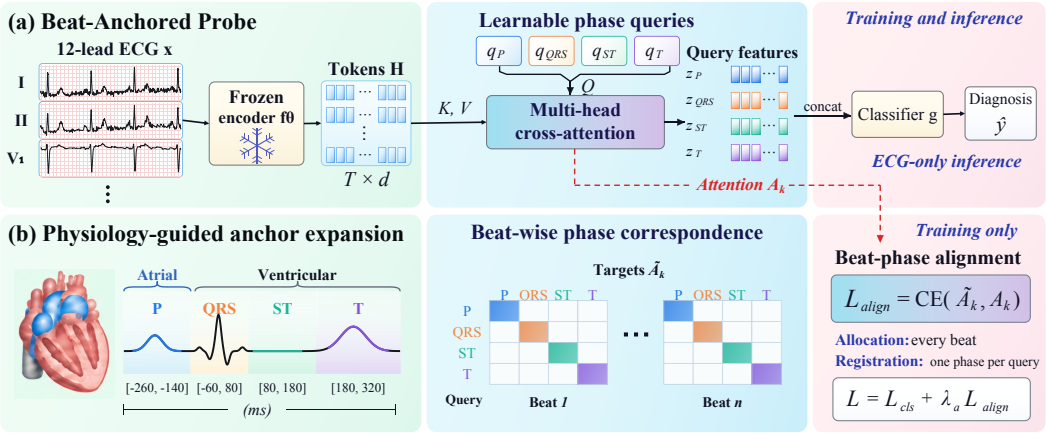


Figure 2: **Overview of BAP. (a)** The frozen encoder maps the ECG to tokens H, four learnable phase queries read H by multi-head cross-attention, and the classifier g reads their concatenated outputs; inference needs the ECG alone. **(b)** In training only, anchor expansion places the P, QRS, ST and T windows around every R-peak, which give each query its target $\widetilde{A}_k$ in every beat, and $\mathcal{L}_{\text{align}}$ pulls the attention A_k toward $\widetilde{A}_k$.

where MHA computes Eq. (1) at width $d/4$ in each of four heads, A_k^h is the attention of query k in head h, and W_{out} is a shared output projection. Each query thus yields one summary z_k and one attention map A_k. BAP rests on three design choices. *(i) Phase-indexed queries.* We name the four queries q_{P}, q_{QRS}, q_{ST} and q_{T}, one per cardiac phase. Every phase thus has a reader, and every query has one phase to read. *(ii) No bypass.* The classifier g reads only the four query outputs, with no mean-pooled path, so every diagnosis passes through the phase queries. *(iii) Training-only supervision.* The R-peaks enter only through a loss on the attention maps. The diagnoses thus decide what each query reads, and inference needs the ECG alone. BAP is trained with the joint objective

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_a \, \mathcal{L}_{\text{align}}, \tag{3}$$

where $\mathcal{L}_{\text{cls}}$ is the binary cross-entropy over the diagnoses, $\mathcal{L}_{\text{align}}$ is the beat-anchored alignment loss and λ_a is a trade-off weight. With λ_a=0, the probe is UP (§2).

3.2 Beat-anchored targets

As analyzed in §2, diagnostic information is locked to the beat. The target of each query should thus (R1) cover its phase in every beat, (R2) differ between queries and follow the record's own beats, and (R3) come from one landmark per beat, as diagnostic corpora carry no phase annotation. We take the R-peak as this landmark. It is the sharpest and most reliably located point of the beat (Pan & Tompkins, 1985; Martínez et al., 2004), and the four phases lie at typical offsets from it (Surawicz et al., 2009). Specifically, *anchor expansion* takes each R-peak r_m as the anchor of beat m and gives query k the *phase window* $W_{k,m} = [r_m + a_k,\ r_m + b_k]$. The offsets $[a_k, b_k]$ are $[-260, -140]$ (P), $[-60, 80]$ (QRS), $[80, 180]$ (ST) and $[180, 320]$ ms (T). Each window is clipped at the ends of the record and at the midpoints between neighboring R-peaks, so it never reaches into another beat. The target of query k weights each token by its overlap with these windows,

$$\widetilde{A}_k(t) \;\propto\; \sum_m \big|\, \tau(t) \cap W_{k,m} \,\big|, \qquad \textstyle\sum_t \widetilde{A}_k(t) = 1, \tag{4}$$

where $\tau(t)$ is the span of samples covered by token t. The target thus covers every beat (R1), follows the R-peaks of its record (R2) and needs one R-peak per beat (R3). The four targets stay distinct, as every encoder in this paper uses tokens no longer than the narrowest window (App. B).

3.3 Allocating each query over every beat

UP covers fewer beats than uniform attention (§2), so BAP aligns each query with its target,

$$\mathcal{L}_{\text{align}} = -\frac{1}{K} \sum_{k=1}^{K} \sum_{t=1}^{T} \widetilde{A}_k(t) \, \log A_k(t). \tag{5}$$

Algorithm 1 One BAP training step. At inference only lines 1, 3 and 4 run.

Require: frozen f_θ; record x, labels y, R-peaks $\{r_m\}$; offsets $\{(a_k, b_k)\}_{k=1}^{4}$; weight λ_a

1: $H \leftarrow f_\theta(x)$ ▷ no gradient
2: $\widetilde{A}_k \leftarrow \mathrm{normalize}\big(\sum_m |\tau(t) \cap W_{k,m}|\big)$ for $k = 1, \dots, 4$ ▷ Eq. (4)
3: $z_k,\ A_k \leftarrow$ cross-attention of q_k over H for $k = 1, \dots, 4$ ▷ Eq. (2)
4: $\hat{y} \leftarrow g([z_1; z_2; z_3; z_4])$
5: $\mathcal{L} \leftarrow \mathrm{BCE}(\hat{y}, y) - \lambda_a \frac{1}{4} \sum_k \sum_t \widetilde{A}_k(t) \log A_k(t)$ ▷ Eqs. (3), (5)
6: update $\{q_k\}$, W_{key}, W_{val}, W_{out} and g

As $\widetilde{A}_k$ is fixed, this cross-entropy equals the Kullback-Leibler (KL) divergence $\frac{1}{K} \sum_k \mathrm{KL}(\widetilde{A}_k \,\|\, A_k)$ up to a constant. It thus pulls each attention map toward its target. We split this pull into where the attention lands and how it is shared among beats. Let $S_k = \{t : \widetilde{A}_k(t) > 0\}$ be the tokens inside the windows of query k, and $A_k(S_k) = \sum_{t \in S_k} A_k(t)$ the attention placed there. As $\widetilde{A}_k$ vanishes outside S_k, the alignment term of query k splits exactly into

$$-\sum_t \widetilde{A}_k(t)\, \log A_k(t) \;=\; \underbrace{-\log A_k(S_k)}_{\text{placement}} \;+\; \underbrace{\mathrm{KL}\big(\widetilde{A}_k \,\|\, A_k(\cdot \mid S_k)\big)}_{\text{allocation}} \;+\; \mathbb{H}(\widetilde{A}_k), \tag{6}$$

where $A_k(t \mid S_k) = A_k(t)/A_k(S_k)$ is the attention renormalized inside the windows and $\mathbb{H}(\widetilde{A}_k)$ is the constant target entropy. The placement term asks the query to attend inside its windows. It is zero for any attention within them, even attention confined to one beat. The allocation term closes this gap. It asks the query to share this attention among all beats as the target does. Leaving out any beat, even an occasional abnormal one, thus raises the loss.

3.4 Registering each query to one phase

As analyzed in §2, UP leaves its queries unassigned. Permuting the queries with their classifier weights leaves $\mathcal{L}_{\text{cls}}$ unchanged, so this loss favors no assignment of queries to phases. In this subsection, we detail how the distinct targets break this symmetry. Specifically, $\mathcal{L}_{\text{align}}$ never falls below $\frac{1}{K} \sum_k \mathbb{H}(\widetilde{A}_k)$, and it nears this bound only when every A_k nears its target $\widetilde{A}_k$. At this bound, any permutation π that moves a query to another phase raises $\mathcal{L}_{\text{align}}$ by $\frac{1}{K} \sum_k \mathrm{KL}(\widetilde{A}_k \,\|\, \widetilde{A}_{\pi(k)})$, which is positive for distinct targets. $\mathcal{L}_{\text{align}}$ thus asks query k to attend to phase k in every record and every training run. This phase lies at the same offsets from every R-peak of the record, so the role of each query can be checked against the ECG.

Training and inference. The alignment loss depends only on the attention maps. Its gradient therefore reaches only their trainable inputs, the queries and W_{key} (Eq. (1)). The value and output projections and the classifier are trained by $\mathcal{L}_{\text{cls}}$ alone. A record without R-peaks receives no target and enters only $\mathcal{L}_{\text{cls}}$. The weight λ_a ramps up over the first three epochs (App. B.5). As the targets enter only the training loss, BAP needs the ECG alone at inference and has the same parameters and inference cost as UP. Algorithm 1 summarizes one training step.

3.5 CPSC2018-R

As CPSC2018 provides no R-peaks, we release **CPSC2018-R**, a manual R-peak layer over the public CPSC2018 12-lead database (Liu et al., 2018), with signals and labels unchanged. Two board-certified cardiologists independently annotated each of the 6,877 records at 500 Hz, and a senior cardiologist adjudicated their disagreements. In a 200-record audit, the two annotators agreed within 50 ms on more than 99% of beats. Against this layer, the NeuroKit2 detector (Makowski et al., 2021) used for the other corpora reaches a one-to-one F1 of 94.5 within 50 ms. It errs most on bundle-branch blocks and PVC (App. A.5). CPSC2018-R supplies the R-peaks of the training targets and a manual reference for locating attention and tokens relative to the beat.

4 Experiments

In this section, we first introduce the settings and cross-site results of BAP, then trace its transfer gain and query roles to its two components, confirming that BAP narrows both gaps of §2.

Table 1: Macro-AUROC (%) of LP, UP and BAP on four frozen encoders; bold marks the best mean.

Encoder	Probe	PTB-XL	Chapman	SPH	Ningbo	CODE-test	Mean
MAE-Base	LP	87.6 ± 0.5	82.4 ± 0.1	89.1 ± 0.4	83.8 ± 0.1	77.4 ± 1.0	84.1 ± 0.2
	UP	88.6 ± 0.4	83.7 ± 1.4	90.3 ± 0.3	86.4 ± 0.9	85.3 ± 1.9	86.9 ± 0.5
	BAP (Ours)	89.7 ± 0.9	84.5 ± 0.5	91.3 ± 0.9	87.9 ± 1.7	86.6 ± 0.4	**88.0** ± 0.7
LeJEPA	LP	77.5 ± 0.7	74.1 ± 0.2	78.6 ± 0.8	76.8 ± 0.4	75.8 ± 1.0	76.6 ± 0.5
	UP	81.5 ± 0.2	79.3 ± 1.0	83.5 ± 0.4	81.1 ± 0.9	75.7 ± 1.5	80.2 ± 0.6
	BAP (Ours)	82.9 ± 1.1	80.5 ± 1.6	85.0 ± 0.3	80.2 ± 0.5	78.3 ± 0.8	**81.4** ± 0.5
MERL	LP	89.9 ± 0.3	83.8 ± 0.2	91.8 ± 0.4	88.0 ± 0.5	91.9 ± 0.3	89.1 ± 0.3
	UP	91.9 ± 1.2	88.9 ± 2.1	94.3 ± 0.1	90.8 ± 0.9	93.7 ± 0.9	91.9 ± 1.0
	BAP (Ours)	92.2 ± 0.3	89.5 ± 1.0	94.2 ± 0.2	91.2 ± 0.8	94.0 ± 0.9	**92.2** ± 0.5
LEAST	LP	91.7 ± 0.5	86.6 ± 0.3	92.5 ± 0.3	87.5 ± 0.5	92.3 ± 0.2	90.1 ± 0.3
	UP	92.0 ± 0.5	85.8 ± 0.5	94.0 ± 0.2	88.6 ± 0.2	94.3 ± 1.6	90.9 ± 0.4
	BAP (Ours)	91.8 ± 0.8	87.0 ± 0.7	93.7 ± 0.3	88.5 ± 0.3	94.1 ± 0.4	**91.0** ± 0.4

4.1 Experimental settings

Every method is trained on CPSC2018-R (folds 1 to 8), selected on fold 9 and applied without adaptation to five cohorts from other health systems, i.e., PTB-XL (Wagner et al., 2020), Chapman (Zheng et al., 2020), SPH (Liu et al., 2022), Ningbo (Reyna et al., 2021) and CODE-test (Ribeiro et al., 2020). SPH, Ningbo and CODE-test were evaluated only after the design was fixed, and no choice of model or learning rate uses a target cohort (App. A). MAE-Base is the default encoder, and LeJEPA (Balestriero & LeCun, 2025), MERL (Liu et al., 2024) and LEAST (Xu et al., 2025) vary the pretraining objective, the architecture and the token rate.

We compare BAP with LP (Liu et al., 2024), the standard readout of ECG foundation models, and with UP (§2), its unaligned counterpart. To locate where the transfer gain arises, we further select three groups of alternatives, i.e., learned pooling into a single vector, namely generalized-mean (GeM) pooling (Radenović et al., 2019), attention-based multiple-instance learning (ABMIL) (Ilse et al., 2018), attentive statistics (Okabe et al., 2018), multihead attention pooling (MAP) (Zhai et al., 2022), SimPool (Psomas et al., 2023) and the V-JEPA attentive probe (Bardes et al., 2024), learned transforms applied after or before pooling, namely CLIP-Adapter (Gao et al., 2024) and a token-wise MLP, and efficient probing (Psomas et al., 2026), a multi-query probe whose queries receive no supervision. LoRA (Hu et al., 2022) instead adapts the encoder and keeps mean pooling. For fair comparisons, all heads share the encoder, source data, schedule and seeds, heads with a learning-rate grid select the rate on fold 9 (App. B.5), and we set λ_a=0.05 by default.

We report macro-AUROC over the diagnoses present in each cohort, with macro-AUPRC (area under the precision-recall curve) and F1 in Table 2, as mean ± SD over three seeds, with *Mean* averaging the five cohorts, and 95% intervals from record-level paired bootstraps (App. B.9). Besides the in-window share and beats covered of §2, we report the cross-record *consistency* of each query's attention profile around the R-peak on CPSC2018-R fold 9, and wave labeling on the 200 expert-delineated LUDB records (Kalyakulina et al., 2020) (App. B.10).

4.2 Main results

Across encoders. As shown in Table 1, BAP consistently improves frozen encoders over LP, e.g., lifting MAE-Base from 84.1 to 88.0 five-cohort mean AUROC. The gains hold across encoders that differ in pretraining objective, backbone and token rate, and are largest where linear transfer is poorest, reaching 4.8 points on LeJEPA. On every encoder, most of the gain comes from reading the tokens individually, and alignment adds 1.1 points ($[0.6, 1.7]$) on MAE-Base and 1.2 on LeJEPA, the two encoders with the weakest linear transfer. The largest gains fall on 1AVB, PAC and PVC, the local diagnoses that LP retains least across sites (Figure 3a). This corroborates §2, as the frozen tokens already hold the information and the readout decides how much of it survives.

Across readouts. As reported in Table 2, the compared readouts separate by where they process the tokens. Learned pooling into a single vector brings smaller gains over LP, as one weighting of the tokens still merges the phases of all beats into one summary. Interestingly, where a learned transform

Table 2: AUROC, AUPRC and F1 (%) of readouts grouped by how they treat the tokens, on MAE-Base (mean ± SD, three seeds). Bold and underline mark the best and second best frozen probe.

Method	Macro-AUROC per cohort					Five-cohort mean		
	PTB-XL	Chapman	SPH	Ningbo	CODE-test	AUROC	AUPRC	F1
Encoder adaptation, mean pooling								
LoRA + linear	88.5±0.6	85.9±1.3	91.2±0.5	86.1±0.3	80.7±1.9	86.5±0.1	58.0±0.3	44.5±0.6
Frozen encoder, pooling into a single vector								
Linear probe (LP)	87.6±0.5	82.4±0.1	89.1±0.4	83.8±0.1	77.4±1.0	84.1±0.2	53.3±0.6	39.7±0.5
GeM pooling	88.4±0.7	82.8±1.1	89.3±0.3	83.9±0.6	85.7±1.1	86.0±0.3	55.0±0.5	32.6±0.8
ABMIL	87.6±0.4	81.2±1.5	89.4±0.7	84.2±0.3	86.6±4.1	85.8±0.9	57.3±1.7	36.3±1.9
Attentive statistics	<u>88.8</u>±0.7	82.7±0.6	90.0±0.4	85.1±0.5	85.3±2.0	86.4±0.6	59.4±1.0	32.7±1.2
MAP pooling	88.0±0.2	82.2±0.9	89.7±0.7	85.1±1.5	85.1±1.5	86.0±0.8	57.6±2.7	38.6±3.1
SimPool	85.6±1.2	80.9±0.9	88.2±1.1	82.7±0.6	86.5±2.4	84.8±1.2	53.7±1.3	36.7±5.5
V-JEPA probe	88.0±2.2	**84.7**±1.7	89.6±1.4	84.8±2.6	85.2±1.1	86.5±1.3	59.6±1.6	<u>42.4</u>±2.4
Frozen encoder, learned transform after or before pooling								
CLIP-Adapter	88.6±0.5	81.7±0.8	89.4±0.3	84.3±0.3	79.2±0.6	84.6±0.2	54.2±0.5	39.5±1.9
Token-wise MLP	88.6±0.3	83.4±3.2	<u>90.4</u>±0.5	86.1±0.7	<u>87.1</u>±1.6	87.1±0.5	<u>60.6</u>±0.7	36.7±6.7
Frozen encoder, multi-query probes								
Efficient probing	88.7±0.3	84.0±2.0	<u>90.4</u>±0.7	86.3±1.0	**87.5**±1.0	<u>87.4</u>±0.7	<u>60.6</u>±0.6	38.8±5.3
Unaligned probe (UP)	88.6±0.4	83.7±1.4	90.3±0.3	<u>86.4</u>±0.9	85.3±1.9	86.9±0.5	60.0±1.3	42.3±2.2
BAP (Ours)	**89.7**±0.9	<u>84.5</u>±0.5	**91.3**±0.9	**87.9**±1.7	86.6±0.4	**88.0**±0.7	**61.4**±0.6	**46.1**±2.2

acts matters. A token-wise MLP before pooling reaches 87.1, and CLIP-Adapter on the pooled vector stays at the level of LP, as the information is already diluted once the tokens are averaged. Multi-query probes, i.e., efficient probing and UP, likewise read the tokens individually and transfer better. LoRA adapts the encoder yet keeps mean pooling and ends at the level of single-vector pooling. BAP keeps the encoder frozen and assigns each query to one phase of every beat, achieving the highest mean AUROC, AUPRC and F1 and exceeding LoRA by 1.5 points ($[0.6, 2.4]$).

Across source corpora. To examine whether the lead of BAP depends on the source corpus, we retrain the readouts on the larger PTB-XL and test them on the five other cohorts, CPSC2018 included (Table 3, App. C.1). Every readout reaches a higher five-cohort mean from this source, and BAP stays on top with the highest AUROC, AUPRC and F1, and alignment again adds 1.1 points ($[0.5, 1.6]$) over UP. LoRA gains only 1.0 point over LP and falls 3.5 points behind BAP, as adapting the encoder to one corpus still ends in mean pooling. Readouts that weight the tokens, i.e., MAP pooling, efficient probing and UP, close most of the gap to BAP, and BAP leads them by 0.6 to 1.1 points.

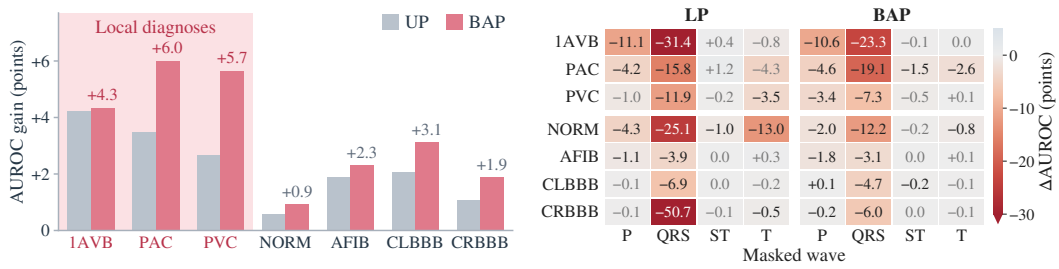
Table 3: Transfer with PTB-XL as the source; five-cohort means on MAE-Base.

Method	AUROC	AUPRC	F1
LP	86.9	51.7	46.4
LoRA + linear	87.9	61.1	53.3
MAP pooling	90.8	63.6	52.9
Efficient probing	90.6	64.7	48.6
UP	90.3	63.1	50.0
BAP (Ours)	**91.4**	**66.2**	**55.2**

4.3 Ablation studies and discussions

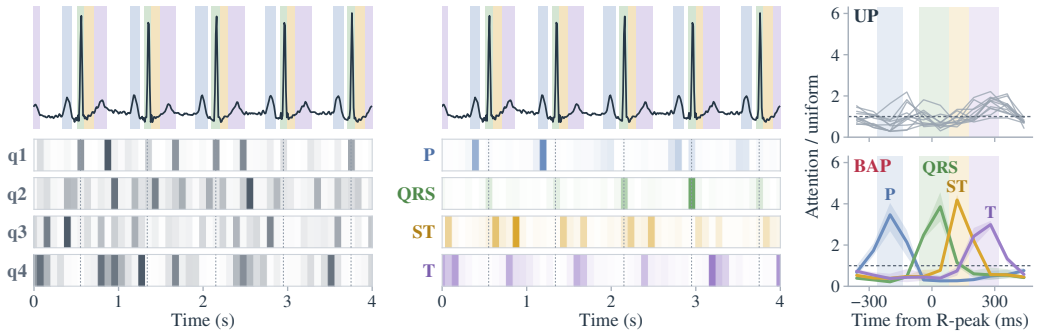
We hereby analyze and discuss BAP from the following aspects, based on ablation studies on MAE-Base with CPSC2018-R as the source unless stated otherwise.

BAP assigns each query to one phase of every beat. We first visualize the attention on unseen LUDB records. As shown in Figure 4, each BAP query stays at its own phase from beat to beat, and the queries of UP do not keep to the phase of their index. BAP narrows both gaps of §2, raising the beats covered from 70.8% to 88.7% and the in-window share from 27.3% to 65.1% (Table 4). Masking the P waves of the input also costs BAP more on 1AVB than on any other diagnosis, as it costs LP (Figure 3b, App. B.3), so BAP depends on the P wave that §2 locates for 1AVB.



(a) Gain over LP per diagnosis (b) AUROC change when one wave is masked

Figure 3: **Per-diagnosis analysis on MAE-Base. (a)** AUROC gain of UP and BAP over LP, averaged over the cohorts that score each diagnosis. **(b)** AUROC change on CPSC2018-R fold 10 when one wave is masked, against a control mask between beats; columns compare diagnoses under one mask, and gray numbers mark a 95% interval containing zero.



(a) UP on an unseen LUDB record (b) BAP on the same record (c) Around the R-peak

Figure 4: **Each BAP query attends to its own phase from beat to beat. (a, b)** Expert waves are shaded on lead II and each row is scaled to its maximum. **(c)** Attention around the R-peak over the 200 LUDB records relative to uniform attention (dashed line at 1), with phase windows shaded, per seed for UP and as mean and range over three seeds for BAP.

Effectiveness of the alignment objective. To identify which part of the objective yields the transfer gain, we remove the allocation term of Eq. (6), keep only the allocation of attention over beats, or remove the whole alignment loss, and compare with a label-free regularizer (Table 4). Without the allocation term, only the placement term remains, which is satisfied by attention inside the windows of a single beat (§3.3). As expected, the queries then hold as much attention in the windows as BAP yet cover only 59.0% of the beats, and the AUROC falls back to that of UP. Conversely, without registration, a beat-mass loss that asks each query for the same attention on every beat, with no phase and no constraint within the beat (App. B.7), covers every beat and reaches 87.6, close to BAP, yet its in-window share stays at 27.2%, near the 25.0% of uniform attention. A diversity term, which pushes the queries apart without R-peaks (App. B.7), likewise confines each query to fewer beats than UP and stays below BAP. Allocation thus carries the transfer gain, and registration adds the fixed phase roles at no cost in accuracy.

Table 4: Ablation of the alignment objective on MAE-Base; attention in % on fold 9.

Objective	AUROC	Beats covered	In-window
BAP	**88.0**	88.7	65.1
w/o allocation	86.9	59.0	68.7
w/o registration	87.6	99.4	27.2
$\mathcal{L}_{\text{align}} \rightarrow$ diversity	87.4	57.3	27.1
w/o $\mathcal{L}_{\text{align}}$ (UP)	86.9	70.8	27.3

Registration across encoders. To examine whether registration depends on the encoder, we compare UP and BAP on all four encoders (Table 5). Without alignment, the in-window share stays close to that of uniform attention on every encoder, as the diagnostic loss leaves the queries interchangeable (§3.4). With alignment, it rises to 59 to 74%, and each query keeps its placement across records, with a consistency of 0.80 to 0.85. Registration thus holds on all four encoders.

Table 5: Query roles on CPSC2018-R fold 9; Unif., uniform attention.

	In-window (%)			Consistency	
Encoder	Unif.	UP	BAP	UP	BAP
MAE-Base	25.0	27.3	**65.1**	0.48	**0.84**
LeJEPA	27.6	30.0	**63.3**	0.65	**0.80**
MERL	18.6	18.9	**59.2**	0.74	**0.84**
LEAST	27.6	31.9	**73.7**	0.62	**0.85**

Table 6: LUDB F1 (%) by training names and by a decoder over background, P, QRS and T (Dec.); LP has no attention map. †Detected R-peaks.

Attention	P	QRS	ST	T	Macro	Dec.
UP, best naming	38.5	31.3	28.9	39.0	34.4	26.2
BAP (Ours)	**81.9**	**78.8**	**60.7**	**74.6**	**74.0**	**50.8**
BAP targets†	85.7	86.8	71.0	82.7	81.5	62.0

Verifying the query roles on LUDB. As the in-window share is measured against our own windows, we further verify the roles against expert delineation on the external LUDB benchmark. Each token inside a P, QRS, ST or T segment takes the phase of the query that attends to it most, with the queries read by their training names (Table 6), and a decoder fitted on the attention maps also labels the background. BAP reaches 74.0 macro-F1, approaching the 81.5 that its target windows reach from detected R-peaks, against 34.4 for UP even under its most favorable naming. BAP reaches this without R-peaks at inference, and a LUDB decoder fitted on one seed reaches 49.0 macro-F1 on another, close to the 50.8 of the same seed and against 21.2 for UP (App. D.3), so its roles are fixed across seeds and verifiable against the ECG.

5 Related work

ECG foundation models and cross-site generalization. ECG foundation models pretrain encoders with self-supervision (Zhang et al., 2022; Wang et al., 2023; McKeen et al., 2025) on more than ten million recordings (Li et al., 2025) and read out mean-pooled tokens with a linear probe (Liu et al., 2024), leaving transfer across health systems (Reyna et al., 2021) rarely examined. Domain-shift methods retrain the encoder (Ganin et al., 2016; Cha et al., 2021; Ramé et al., 2022) or adapt it on unlabeled target data (Wang et al., 2021; Niu et al., 2022). We instead improve cross-site transfer through the readout of a frozen encoder, with no target data.

Attentive pooling and probing. Learned attention has replaced mean pooling in multiple-instance learning, speaker embedding and vision (Ilse et al., 2018; Okabe et al., 2018; Zhai et al., 2022; Psomas et al., 2023). Set Transformer pools with learnable seeds (Lee et al., 2019), slot attention lets slots compete for the input (Locatello et al., 2020), and multi-query attentive probes read frozen encoders with learnable queries (Bardes et al., 2024) that learn complementary maps without supervision (Psomas et al., 2026). Differently, BAP supervises each query to attend to one phase.

Supervised attention and ECG beats. Supervised attention fits attention or gradient maps to per-sample human annotations in language and vision (Liu et al., 2017; Li et al., 2018; Selvaraju et al., 2019; Ross et al., 2017), and concept bottleneck models supervise intermediate concepts (Koh et al., 2020). QRS detectors (Pan & Tompkins, 1985) and wave delineators (Martínez et al., 2004; Peimankar & Puthusserypady, 2021) locate beats and waves as an explicit output. ECG models such as MINA (Hong et al., 2019), Rlign (Plagwitz et al., 2026) and beat-synchronous tokenization (Sameh et al., 2026) build beats into the input or the architecture and compute beat structure at inference. BAP expands one R-peak per beat into dense phase targets used only in training, and, as the explanatory value of attention weights is debated (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), we verify its attention against expert annotations on LUDB (Kalyakulina et al., 2020).

6 Conclusion

Different from previous studies that scale up encoder pretraining, we asked where frozen encoders keep the information that cross-site diagnosis requires. We revealed that frozen tokens retain diagnostic information locked to the heartbeat, that mean pooling dilutes it for local diagnoses, and that diagnosis-only attentive probes leave their queries unassigned. Accordingly, we proposed BAP, whose training-only beat-anchored targets allocate each query over every beat and register it to one phase, achieving the highest mean transfer among the compared probes with fixed phase roles decodable on LUDB, and released CPSC2018-R. In the future, we plan to extend beat-anchored supervision to other quasi-periodic signals, e.g., photoplethysmography.

AI USE STATEMENT

We used LLM-based assistants to help with coding the training, analysis and plotting scripts, with experiment planning, and with editing the text. No generative tool was used to produce or annotate data. The authors checked all outputs and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work uses publicly released ECG corpora under their original ethics approvals and licenses (App. A), and no new patient data were collected. CPSC2018-R adds only cardiologist-annotated R-peak positions to records that are already public, introduces no patient-identifiable content and is released under the license of CPSC2018. A head-only readout that transfers across sites lowers the cost of deploying ECG foundation models at institutions without large labeled cohorts, and any clinical use requires validation on local data.

REPRODUCIBILITY STATEMENT

BAP, its targets and its loss are specified in §3 and Algorithm 1. The cohorts, the shared label space, the splits and the CPSC2018-R annotation protocol are given in App. A, and the encoders, the protocols of §2, the compared heads with their learning-rate selection and the attention measures in App. B. Every experiment except LoRA trains a head on a frozen encoder, and a 60-epoch run takes 2 to 3 minutes on one RTX PRO 6000 (App. B.12). Results of trained heads are means over three seeds, and intervals are paired bootstraps over the evaluated records (App. B.9). The training and evaluation code and CPSC2018-R, as a standalone annotation layer with its per-window quality score, are available at `https://anonymous.4open.science/r/BAP-1490/`.

REFERENCES

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

R Sacha Bhatia and Paul Dorian. Screening for cardiovascular disease risk with electrocardiography. *JAMA Internal Medicine*, 178(9):1163–1164, 2018.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22405–22418, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, et al. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. *PhysioNet*, 2023. doi: 10.13026/4nqg-sb35. Version 1.0.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 15979–15988. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01553.

Shenda Hong, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. MINA: Multilevel knowledge-guided attention for modeling electrocardiography signals. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5888–5894, 2019. doi: 10.24963/ijcai.2019/816.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ECG words and sentences via pre-training ECG language model. In *International Conference on Learning Representations (ICLR)*, 2025.

Alena Kalyakulina, Igor Yusipov, Viktor Moskalenko, Alexander Nikolskiy, Konstantin Kosonogov, Nikolai Zolotykh, and Mikhail Ivanchenko. Lobachevsky University Electrocardiography Database. PhysioNet, 2021. URL https://doi.org/10.13026/eegm-h675. Version 1.0.1.

Alena I Kalyakulina, Igor I Yusipov, Viktor A Moskalenko, Alexander V Nikolskiy, Konstantin A Kosonogov, Grigory V Osipov, Nikolai Yu Zolotykh, and Mikhail V Ivanchenko. LUDB: A new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, 8: 186181–186190, 2020. doi: 10.1109/ACCESS.2020.3029211.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348, 2020.

Fred M Kusumoto, Mark H Schoenfeld, Coletta Barrett, James R Edgerton, Kenneth A Ellenbogen, Michael R Gold, Nora F Goldschlager, Robert M Hamilton, José A Joglar, Robert J Kim, et al. 2018 ACC/AHA/HRS guideline on the evaluation and management of patients with bradycardia and cardiac conduction delay. *Journal of the American College of Cardiology*, 74(7):e51–e156, 2019.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753, 2019.

Jun Li, Aaron Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari D. Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7), 2025. doi: 10.1056/AIoa2401033.

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9215–9223, 2018. doi: 10.1109/CVPR.2018.00960.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31949–31963, 2024. URL https://proceedings.mlr.press/v235/liu24bg.html.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Ng Yin Kwee. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018. doi: 10.1166/jmihi.2018.2442.

Hui Liu, Dan Chen, Da Chen, Xiyu Zhang, Huijie Li, Lipan Bian, Minglei Shu, and Yinglong Wang. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Scientific Data*, 9(1):272, 2022.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1789–1798, 2017.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538, 2020.

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

Gregory M Marcus. Evaluation and management of premature ventricular complexes. *Circulation*, 141(17):1404–1418, 2020.

Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna. A wavelet-based ECG delineator: evaluation on standard databases. *IEEE Transactions on Biomedical Engineering*, 51(4):570–581, 2004.

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

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=WcOohbsF4H.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, 2022.

Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. Attentive statistics pooling for deep speaker embedding. In *Interspeech 2018*, pp. 2252–2256, 2018. doi: 10.21437/Interspeech.2018-993.

Jiapu Pan and Willis J Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3):230–236, 1985.

Abdolrahman Peimankar and Sadasivan Puthusserypady. DENS-ECG: A deep learning approach for ECG signal delineation. *Expert Systems with Applications*, 165:113911, 2021.

Lucas Plagwitz, Lucas Bickmann, Michael Fujarski, Alexander Brenner, Warnes Gobalakrishnan, Lars Eckardt, Antonius Büscher, and Julian Varghese. The Rlign algorithm for enhanced electrocardiogram analysis through heart rate-corrected ECG alignment for explainable classification and clustering. *European Heart Journal - Digital Health*, 7(5):ztag067, 2026. doi: 10.1093/ehjdh/ztag067.

Bill Psomas, Ioannis Kakogeorgiou, Konstantinos Karantzalos, and Yannis Avrithis. Keep it SimPool: Who said supervised transformers suffer from attention deficit? In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5327–5337, 2023. doi: 10.1109/ICCV51070.2023.00493.

Bill Psomas, Dionysis Christopoulos, Eirini Baltzi, Ioannis Kakogeorgiou, Tilemachos Aravanis, Nikos Komodakis, Konstantinos Karantzalos, Yannis Avrithis, and Giorgos Tolias. Attention, please! Revisiting attentive probing through the lens of efficiency. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=PXo0gtT7Al.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019. doi: 10.1109/TPAMI.2018.2846566.

Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, 2022.

Pentti M Rautaharju, Borys Surawicz, and Leonard S Gettes. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part IV: The ST segment, T and U waves, and the QT interval. A scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society. endorsed by the International Society for Computerized Electrocardiology. *Circulation*, 119(10):e241–e250, 2009.

Matthew A. Reyna, Nadi Sadr, Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D. Clifford. Will two do? varying dimensions in electrocardiography: The PhysioNet/Computing in cardiology challenge 2021. *Computing in Cardiology*, 48:1–4, 2021. doi: 10.23919/CinC53138.2021.9662687.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela M M Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton P S Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, Thomas B Schön, and Antonio Luiz P Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, 2020. doi: 10.1038/s41467-020-15432-4.

Antônio H. Ribeiro, Gabriela M. M. Paixão, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr., Thomas B. Schön, and Antônio Luiz P. Ribeiro. CODE-15%: A Large Scale Annotated Dataset of 12-Lead ECGs, 2021. URL https://doi.org/10.5281/zenodo.4916206. Dataset, Version 1.0.0.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371.

Ahmed Sameh, Nolan Wilson, Max Enderlein, and Yogatheesan Varatharajah. Beat-synchronous tokenization for ECG transformers. *arXiv preprint arXiv:2608.30367*, 2026.

Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2591–2600, 2019. doi: 10.1109/ICCV.2019.00268.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2021. doi: 10.1109/JBHI.2020.3022989.

Borys Surawicz, Rory Childers, Barbara J Deal, Leonard S Gettes, et al. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part III: Intraventricular conduction disturbances. A scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society. endorsed by the International Society for Computerized Electrocardiology. *Journal of the American College of Cardiology*, 53(11):976–981, 2009. doi: 10.1016/j.jacc.2008.12.013.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6.

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*, 2021.

Guoxin Wang, Qingyuan Wang, Ganesh Neelakanta Iyer, Avishek Nag, and Deepu John. Unsupervised pre-training using masked autoencoders for ECG analysis. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–5. IEEE, 2023.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002.

He-Yang Xu, Hongxiang Gao, Yuwen Li, Xiu-Shen Wei, and Chengyu Liu. Masked autoencoders that feel the heart: Unveiling simplicity bias for ECG analyses. *arXiv preprint arXiv:2506.22495*, 2025. doi: 10.48550/arXiv.2506.22495.

Ryuichiro Yagi, Yuichiro Mori, Shinichi Goto, Taku Iwami, and Kosuke Inoue. Routine electrocardiogram screening and cardiovascular disease events in adults. *JAMA Internal Medicine*, 184(9): 1035–1044, 2024.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Huaicheng Zhang, Wenhan Liu, Jiguang Shi, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. MaeFE: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2022.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48, 2020. doi: 10.1038/s41597-020-0386-x.

APPENDIX OVERVIEW

This appendix provides the data and annotations (App. A), the implementation details and protocols (App. B), additional results (App. C) and the validation of the phase roles (App. D) omitted from the main text due to space limitations.

A DATA AND ANNOTATIONS

A.1 COHORTS AND SHARED LABEL SPACE

Table 7: **Corpora and label counts.** Rows count records, except CPSC2018-R and SPH, which count 10-s windows. *Used* counts records with at least one shared label on the evaluated fold, all windows of folds 1 to 9 for CPSC2018-R, all usable exams for CODE-test and all records for LUDB. CODE-test has no normal class, and Ningbo has no AFIB.

			Records positive for each shared label						
Corpus	**Records**	**Used**	NORM	AFIB	1AVB	CLBBB	CRBBB	PAC	PVC
Encoder pretraining, labels unused									
CODE-15% (Ribeiro et al., 2021)	345,779	–	–	–	–	–	–	–	–
MIMIC-IV-ECG (Gow et al., 2023)	800,035	–	–	–	–	–	–	–	–
Head training and landmark supervision									
CPSC2018-R (ours)	9,372	8,413	1,088	1,438	807	265	2,111	914	1,121
Cross-site evaluation, no target labels used for training or selection, no adaptation									
PTB-XL (Wagner et al., 2020)	21,837	1,339	964	152	80	54	54	40	115
Chapman (Zheng et al., 2020)	10,247	448	201	165	27	27	47	17	28
SPH (Liu et al., 2022)	26,238	16,096	13,966	501	256	87	746	238	475
Ningbo (Reyna et al., 2021)	34,904	9,936	6,299	–	893	35	1,291	1,063	1,091
CODE-test (Ribeiro et al., 2020)	827	826	–	13	28	30	34	–	–
External delineation benchmark									
LUDB (Kalyakulina et al., 2020)	200	200	–	–	–	–	–	–	–

The shared label space is the seven diagnoses common to the source corpus and the two development cohorts after AHA and SNOMED codes are mapped to one ontology, i.e., NORM, AFIB, 1AVB (first-degree atrioventricular block), CLBBB and CRBBB (complete left and right bundle-branch block), PAC and PVC (premature atrial and ventricular contraction). Differently spelled codes are merged with the equivalence classes of the PhysioNet/CinC 2021 scoring function (Reyna et al., 2021), {RBBB, CRBBB}, {VPB, PVC} and {PAC, SVPB}. PTB-XL lists the SCP-ECG statements of a record together with a likelihood value, and a record is positive for a shared label whenever a statement of that label is listed, whatever its likelihood value; this rule gives the PTB-XL counts of Table 7 and the PTB-XL-source training labels (App. B.8). Atrial flutter keeps its own code, so Ningbo has no AFIB positive and is scored on six classes, and CODE-test is scored on the four classes it annotates (AFIB, 1AVB, CLBBB, CRBBB). Macro metrics average over the classes present in a cohort.

A.2 SPLITS, EVALUATED RECORDS AND PRETRAINING OVERLAP

Probes are trained and selected on CPSC2018-R (§4.1). PTB-XL and Chapman, the two cohorts seen during development, are evaluated on fold 10 of the recommended patient-stratified PTB-XL split (Wagner et al., 2020; Strodthoff et al., 2021) and on fold 10 of a ten-fold split of the Chapman shared-label records that we generate with a fixed seed, patient-disjoint since Chapman holds one record per patient. SPH, Ningbo and CODE-test were evaluated only after the design was fixed. The release that provides Ningbo holds 45,151 records, including the 10,247 Chapman records, and removing these leaves 34,904. A target record is evaluated if it carries at least one shared label, except on CODE-test, which has no normal class, so all 826 usable exams are kept.

MAE-Base and LeJEPA are pretrained without labels on MIMIC-IV-ECG and CODE-15%, which contain no CPSC2018 record and no target cohort. CODE-15% samples the CODE exams of 2010 to 2016 (Ribeiro et al., 2021) and CODE-test those of April to September 2018 (Ribeiro et al., 2020), so no evaluated exam enters their pretraining.

A.3 PREPROCESSING

Recordings are band-pass filtered (0.67 to 40 Hz, fourth-order Butterworth, zero phase), resampled to 100 Hz and standardized per record and lead, so the encoder input is $\mathbb{R}^{12\times1000}$. CPSC2018 recordings and SPH records longer than 10 s are cut into non-overlapping 10-s windows (26,238 SPH windows from 25,770 records). Shorter CPSC2018 recordings are zero-padded, CODE-test exams shorter than 10 s are extended to 10 s by repeating the recorded segment, and the other cohorts are used as the 10-s records they are.

A.4 CPSC2018-R ANNOTATION PROTOCOL

The annotation team, adjudication and audit are described in §3.5. Pacemaker spikes are marked only when their position is unambiguous. Of the 6,877 recordings, 6,867 pass quality control (no saturated lead, at least three detectable R-peaks in lead II, no flat-line span longer than 1 s) and give 9,372 windows, at most 14 per recording. Folds are disjoint at the level of recordings, and CPSC2018 releases no patient identifiers. Windows without a shared-label positive are kept as negatives. The layer is distributed as one file per window with the R-peak sample indices at the original 500 Hz and the quality score of App. A.5. Figure 5 shows three training windows.

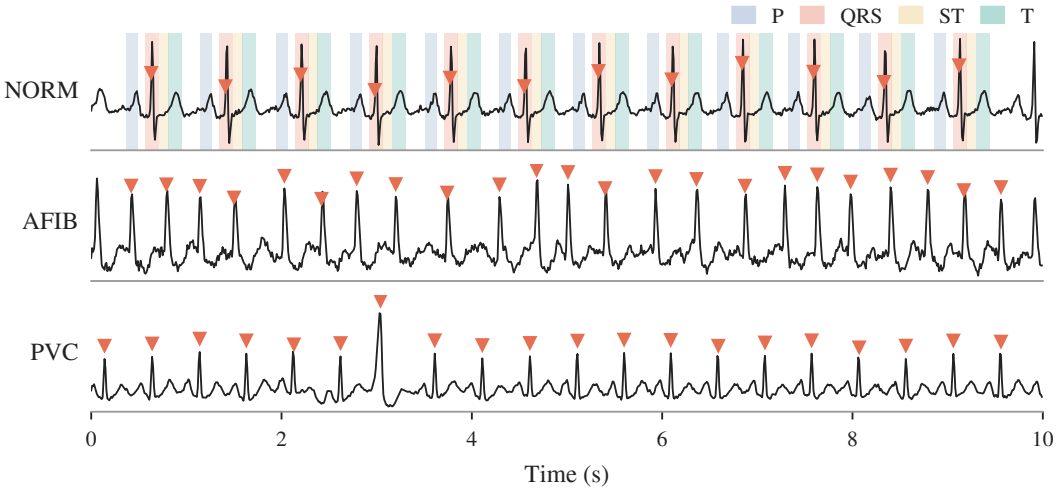


Figure 5: **CPSC2018-R.** Three training windows (lead II) with manual R-peaks; phase windows are shaded on the first.

A.5 LANDMARK VERIFICATION AND DETECTOR LANDMARKS

The detector audit compares 9,360 windows with NeuroKit2 (Makowski et al., 2021) (`ecg_peaks`, default method, cleaned lead II at 100 Hz). The per-window quality score, the fraction of manual R-peaks with a detector beat within 50 ms, has a mean of 0.94 and exceeds 0.5 on 96.5% of windows. Matched one to one within 50 ms, the detector reaches a precision of 95.4%, a recall of 93.7% and an F1 of 94.5, and matched beats differ by a median of 4 ms (90th percentile 18 ms). The F1 is 98.0 in NORM windows, against 89.6 in CLBBB, 90.9 in CRBBB and 92.6 in PVC windows. All alignment targets on CPSC2018-R derive from the manual R-peaks. The PTB-XL-source runs and the phase-resolved probing of the five target cohorts use detector R-peaks.

A.6 LICENSES

PTB-XL (v1.0.1), Chapman-Shaoxing, Ningbo, SPH, CODE-15% and CODE-test are released under Creative Commons Attribution 4.0, LUDB (v1.0.1) (Kalyakulina et al., 2021) under the Open Data Commons Attribution License v1.0, and MIMIC-IV-ECG (v1.0) under the Open Data Commons Open Database License v1.0. CPSC2018 is available under the public release terms of the challenge organizers, and CPSC2018-R is released under the license of CPSC2018.

B IMPLEMENTATION DETAILS AND PROTOCOLS

B.1 ENCODERS

MAE-Base is a ViT (width 768, depth 12, 12 heads) with a ResNet1d patch embedding that maps a 12×1000 input to $T{=}125$ tokens of 80 ms. It is pretrained with masked autoencoding (He et al., 2022) on 1,132,954 unlabeled records, 99% of the 799,289 MIMIC-IV-ECG and 345,109 CODE-15% records that pass preprocessing (mask ratio 0.6, 700 epochs, AdamW, base learning rate 1.5×10^{-4}, weight decay 0.05, 40 warm-up epochs). LeJEPA (Balestriero & LeCun, 2025) is a ViT pretrained on the same corpus with a joint-embedding predictive objective, and LEAST (Xu et al., 2025) uses its tokenization at width 256 and 8 layers. MERL (Liu et al., 2024) is the released ResNet-18, pretrained at 500 Hz, whose pre-pooling feature map gives $T{=}313$ frames of width 512; it receives our 100-Hz records upsampled by five, rescaled with its global min–max and reordered to its lead order. Against the narrowest phase window (ST, 100 ms), MAE-Base (80-ms tokens), LeJEPA and LEAST ($T{=}199$ tokens of 100 ms at a 50-ms stride) and MERL (32-ms frames) resolve every window. Token bins are expressed in the 100-Hz sample space of the targets.

B.2 PHASE-RESOLVED PROBING

Representations. The probes of §2 read the frozen tokens of MAE-Base at seven depths, i.e., the output of the convolutional stem, taken before the position embedding, any normalization and any self-attention, and the outputs of blocks $2, 4, \ldots, 12$, each normalized by the final LayerNorm of the encoder, with the class token dropped. Token t covers samples $[8t{-}8, 8t)$ at 100 Hz, and its center $c(t)$ lies at $80t{-}40$ ms, with R-peaks r_m also in ms.

Features. For a record with R-peaks $r_1, \ldots, r_M$, the offsets δ_j run from -400 to $+480$ ms in steps of 80 ms, and the bin feature b_j averages, over the beats, the mean of the tokens whose center lies within 40 ms of $r_m + \delta_j$,

$$b_j = \frac{1}{|\mathcal{M}_j|} \sum_{m \in \mathcal{M}_j} \frac{1}{|B_{m,j}|} \sum_{t \in B_{m,j}} H_t, \qquad B_{m,j} = \big\{ t : c(t) - r_m - \delta_j \in [-40, 40)\,\text{ms} \big\}, \tag{7}$$

where $\mathcal{M}_j$ holds the beats with a nonempty $B_{m,j}$. The phase feature of phase k weights the tokens by the target of Eq. (4),

$$h_k = \sum_{t=1}^{T} \widetilde{A}_k(t)\, H_t, \tag{8}$$

so it averages the tokens of the phase window of k over every beat, with the windows of §3.2 clipped at the midpoints between neighboring R-peaks. A feature that a record cannot define, e.g., a bin that no beat reaches, takes the mean of the training features.

Windows at the other beat phases. The anchors of a record are moved by s times its median RR interval, $s \in \{0.3, 0.5, 0.7\}$ (800 ms when the record has fewer than two anchors), anchors beyond the record are dropped, and the phase features $h_k^{(s)}$ are rebuilt with the same target builder. A separate probe is trained on each moved feature, and the phase specificity of diagnosis c read from phase k is

$$\Delta_{k,c} = \mathrm{AUROC}_c\big(h_k\big) - \frac{1}{3} \sum_{s \in \{0.3, 0.5, 0.7\}} \mathrm{AUROC}_c\big(h_k^{(s)}\big), \tag{9}$$

with P as the expected phase of 1AVB and PAC and QRS as that of CRBBB.

Probes. For each feature and diagnosis, an L2-regularized logistic regression with class-balanced weights is trained on CPSC2018-R folds 1 to 8 (7,442 windows), with the features standardized by the statistics of these folds and the inverse penalty chosen from $\{0.01, 0.1, 1\}$ by the AUROC on fold 9. These probes serve only §2, and LP in §4 is a trained head (App. B.6).

R-peaks and evaluation. Features on CPSC2018-R use the manual R-peaks, and the in-domain test set is fold 10 (959 windows). The five target cohorts use NeuroKit2 R-peaks (`ecg_peaks` on cleaned lead II at 100 Hz). Per-class five-cohort values average the cohorts that score the class, and the phase specificity is computed on the records of these cohorts pooled into one test set. Intervals are 95% percentile intervals of 1,000 paired bootstraps that resample records within each cohort.

B.3 INPUT MASKING

Wave segments. On the 955 windows of CPSC2018-R fold 10 with at least two manual R-peaks, the discrete-wavelet delineator of NeuroKit2 runs on cleaned lead II at 100 Hz, given the manual R-peaks, each moved to the largest absolute value of cleaned lead II within ± 40 ms. Each beat gives the P wave (P onset to P offset), the QRS complex (QRS onset to QRS offset), the ST segment (QRS offset to T onset) and the T wave (T onset to T offset). A segment is dropped when its boundaries are out of order or its duration falls outside 20 to 300 ms (P), 30 to 300 ms (QRS), 10 to 400 ms (ST) or 40 to 600 ms (T). A QRS onset that precedes the P offset of its beat is moved to that offset, and a T offset that follows the P onset of the next beat is moved to that onset.

Masking and control. Masking a wave replaces the samples strictly inside each of its segments, in all twelve leads of the standardized input, by the straight line between the two endpoint samples, and the tokens are recomputed. The control applies the straight-line replacement inside the TP intervals of the record, from the T offset of one beat to the P onset of the next, or to its QRS onset when that P wave is missing. Each segment is matched by a piece with the same number of replaced samples at a uniformly random free position, pieces never overlap, and when no free interval can hold a piece, the longest free interval takes the largest piece it can hold. Per window, masking replaces on average 104, 139, 111 and 118 samples for P, QRS, ST and T, and the control 92, 123, 106 and 111. Masking changes the input by a mean relative L2 norm of 0.13, 0.74, 0.11 and 0.23, and the control by 0.04 to 0.06. As masking the QRS complex changes the input most, the drops are compared across diagnoses within each masked wave.

Measure. For LP and BAP on MAE-Base, three seeds each, the control-corrected change of diagnosis c when wave w is masked is

$$\Delta_c(w) = \mathrm{AUROC}_c\big(x_w^{\mathrm{mask}}\big) - \mathrm{AUROC}_c\big(x_w^{\mathrm{ctrl}}\big), \tag{10}$$

averaged over seeds, with 95% percentile intervals from 1,000 record bootstraps shared by all heads, seeds and conditions.

B.4 BAP PROBE AND TARGETS

The queries of Eq. (2) enter the attention without a query projection and are initialized from a Gaussian with SD 0.02. The key, value and output projections are $d \times d$ with head width $d_h{=}d/4$ (192 on MAE-Base), dropout (0.1) acts on the attention weights and before the linear classifier, and $\mathcal{L}_{\mathrm{align}}$ uses the head-averaged attention before dropout. The phase windows are also clipped to the record. A record without detected anchors receives no target and enters only $\mathcal{L}_{\mathrm{cls}}$, and a query with an empty target is left out of the average in Eq. (5). Each head scales its logits by $\sqrt{d_h}$, the classifier g is a LayerNorm, dropout and a linear layer, and A_k is clamped at 10^{-6} inside the logarithm of Eq. (5).

B.5 TRAINING AND LEARNING-RATE SELECTION

All heads are trained for 60 epochs with AdamW, batch size 64 and no weight decay unless stated; the learning rate warms up over 5 epochs to a peak of 0.005, unless stated, and then follows a cosine schedule. The loss is binary cross-entropy with per-class positive weights (negative-to-positive ratio, capped at 20), and the epoch with the best macro-AUROC on the source validation fold is kept. Auxiliary loss weights are $1/3$, $2/3$ and 1 times their final value in the first three epochs, and the final λ_a of BAP is 0.05.

LP, UP, BAP, the loss conditions, CLIP-Adapter, GeM, ABMIL, attentive statistics and the token-wise MLP share the rate 0.005, at which all conditions and attention measures are computed. MAP pooling, SimPool, the V-JEPA attentive probe and efficient probing all select 0.002 from $\{0.002, 0.005, 0.010\}$ by the mean source-validation macro-AUROC of three seeds, and efficient probing selects $K_{\mathrm{EP}}{=}8$ from $\{4, 8\}$ queries in the same way; at the shared rate the V-JEPA attentive probe reaches 83.7, against 86.5 at its selected rate. No selection uses a target cohort. With this search applied to every probing head of Table 2 (0.010 for LP, BAP and the token-wise MLP, 0.002 for the others), BAP reaches 88.0, ahead of efficient probing (87.4) and ABMIL and UP (87.2), and LP reaches 83.3.

B.6 COMPARED METHODS

All probing rows of Table 2 share the encoder, source data, schedule and seeds. The linear probe (LP) classifies the mean-pooled token after a BatchNorm without affine parameters. GeM, ABMIL, attentive statistics, MAP pooling, SimPool and the V-JEPA attentive probe follow their papers with a linear classifier, and CLIP-Adapter is a bottleneck MLP on the pooled feature (residual ratio 0.2, reduction 4). The token-wise MLP applies a LayerNorm and a two-layer MLP ($d \to 1152 \to d$, GELU, dropout 0.1) to every token before mean pooling and the linear classifier. Efficient probing follows its paper, with K_{EP} queries that attend to the raw tokens without query, key or output projections and each aggregate a d/K_{EP}-dimensional subvector of one shared value projection, and uses BAP's classifier, with $K_{\mathrm{EP}}{=}8$. LoRA (Hu et al., 2022) adds rank-32 adapters ($\alpha/r{=}1$) to the fused query-key-value and attention output projections of all twelve blocks, trained with LP at an encoder learning rate of 10^{-4} and merged before evaluation.

B.7 LOSS CONDITIONS

The loss without the allocation term replaces Eq. (5) by the placement loss $-\frac{1}{K}\sum_k \log \sum_{t:\widetilde{A}_k(t)>0} A_k(t)$. The beat-mass loss replaces Eq. (5) by $-\frac{1}{K}\sum_k \frac{1}{M}\sum_m \log a_{k,m}$, with $a_{k,m}$ the attention of query k on the RR cell of beat m (App. B.11); it gives every query the same target, fixes only the attention mass of each beat and leaves the attention within a beat free. The diversity term is the mean similarity between the attention maps of different queries; it is minimized, uses no R-peak and is unchanged when the queries are permuted. Both are weighted like $\mathcal{L}_{\mathrm{align}}$, with the weight selected on the source validation fold from $\{0.05, 0.2, 1.0\}$, which selects 1.0 for the beat-mass loss and 0.05 for the diversity term.

B.8 PTB-XL AS THE SOURCE CORPUS

For Table 8, every method of Table 2 except the token-wise MLP is trained on folds 1 to 8 of PTB-XL (10,656 records with a shared label under the rule of App. A.1) and selected on fold 9 (1,338 records). As on the target cohorts, PTB-XL records without a shared label are left out. Training uses the code, schedule and λ_a of the CPSC2018-R runs. The four heads with a selected learning rate re-select it on PTB-XL fold 9, and all four select 0.002, and LoRA changes only the data. The training targets use R-peaks from NeuroKit2 `ecg_peaks` on cleaned lead II of every training record, a median of 11 beats per record. Evaluation uses the 8,119 CPSC2018-R windows of all ten folds that carry a shared label, the Chapman fold, SPH, Ningbo and CODE-test.

B.9 METRICS

AUROC and AUPRC are threshold free, and F1 uses a sigmoid threshold of 0.5 with no target tuning. SDs are sample SDs over training seeds. The 95% intervals come from at least $1{,}000$ paired bootstraps that resample the evaluated records of each cohort, i.e., windows on CPSC2018-R and SPH, and average the paired macro-AUROC difference over the seeds of each head. The retention of Figure 1a divides the cross-site AUROC of LP on a diagnosis, averaged over the cohorts that score it, by its in-domain AUROC on the 825 fold-10 windows of CPSC2018-R with a shared label.

B.10 ATTENTION MEASURES

In-window share is $A_k(S_k)$ of Eq. (6), averaged over queries and records of the CPSC2018-R validation fold with windows from the manual R-peaks, for all encoders and, on MAE-Base, for both sources. Uniform attention scores the share of tokens its window covers, 25.0% on MAE-Base, 27.6% on LeJEPA, 18.6% on MERL and 27.6% on LEAST. *Consistency* averages over the beats of a record, for each query, the attention on the tokens whose centers fall in each of twelve 80-ms bins centered from -400 to $+480$ ms around a manual R-peak, and normalizes it into a profile; it is the Pearson correlation between this profile and the corpus-mean profile of the same query, averaged over queries and records.

LUDB decoder F1. LUDB (Kalyakulina et al., 2020) has 200 records with expert P, QRS and T delineations on lead II, and each token takes the majority class of its samples (background, P, QRS, T). With encoder and head frozen, the four attention values of a token and their logarithms

feed a multinomial logistic regression with balanced class weights, fitted and scored by five-fold cross-validation over records; features shuffled in time give 17.8. The ST segment of a beat runs from its QRS offset to its T onset. For the labeling with training names, each token inside a P, QRS, ST or T segment takes the phase of the query that attends to it most, with the queries of UP named by their best one-to-one assignment to the phases, and the macro-F1 averages the four phases.

B.11 Beats covered

For a record with $M \geq 2$ R-peaks, the RR cell C_m of beat m runs from the midpoint between r_{m-1} and r_m to the midpoint between r_m and r_{m+1}, the first and last cells extend to the record edges, and a token belongs to the cell that contains its center. With $a_{k,m} = \sum_{t \in C_m} A_k(t)$ the attention of query k on cell m, the share of beats covered, i.e., of cells that receive at least half of their fair share $1/M$, is

$$\mathrm{Cov}_k = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\Big[a_{k,m} \geq \frac{1}{2M}\Big], \tag{11}$$

averaged over the records of the CPSC2018-R validation fold with its manual R-peaks and then over the four queries. Uniform attention covers 99.9% of the beats.

B.12 Compute and software

A head-only run takes 2 to 3 minutes on one RTX PRO 6000. We use PyTorch 2.9.1, scikit-learn 1.4, NeuroKit2 0.2.13 and `wfdb` 4.3.1.

C Additional results

C.1 PTB-XL as the source corpus

Table 8 repeats Table 2 with PTB-XL as the source corpus and detector R-peaks for the training targets (App. B.8). BAP reaches the highest five-cohort mean among the frozen-encoder probes (91.4), ahead of UP (90.3), and LoRA reaches 87.9.

Table 8: **Cross-site transfer with PTB-XL as the source.** MAE-Base with detector R-peaks; columns and marks as in Table 2.

	Macro-AUROC per cohort					Five-cohort mean		
Method	CPSC2018	Chapman	SPH	Ningbo	CODE-test	AUROC	AUPRC	F1
Encoder adaptation								
LoRA + linear	88.1±0.1	86.1±0.8	94.0±0.2	89.8±0.2	81.4±1.8	87.9±0.3	61.1±1.1	53.3±0.8
Probing (frozen encoder)								
Linear probe (LP)	87.2±0.5	86.4±0.4	93.5±0.2	87.7±0.8	79.6±0.9	86.9±0.3	51.7±0.4	46.4±1.3
GeM pooling	85.7±0.6	85.7±0.4	92.6±0.2	87.4±0.2	85.3±1.2	87.3±0.2	55.1±1.2	39.0±3.1
ABMIL	86.3±0.4	86.4±2.4	93.1±0.3	88.4±0.6	83.7±4.7	87.6±0.7	57.1±1.7	43.1±1.8
Attentive statistics	86.9±0.3	87.5±1.2	93.5±0.2	89.7±0.4	85.9±2.1	88.7±0.4	60.7±1.2	35.0±1.0
MAP pooling	87.9±0.6	90.3±1.7	**95.9**±0.4	**92.7**±0.5	87.0±2.2	90.8±0.7	63.6±1.2	52.9±2.2
SimPool	84.7±0.8	86.1±1.4	92.7±0.4	88.0±0.4	88.2±3.6	87.9±0.5	54.2±1.6	42.3±2.4
CLIP-Adapter	85.6±1.3	86.5±0.8	92.8±0.9	86.6±1.2	84.5±3.7	87.2±1.5	50.7±2.6	44.5±1.4
V-JEPA probe	83.6±1.0	89.5±0.7	95.0±0.2	91.3±0.3	87.5±2.0	89.4±0.2	60.3±1.5	44.5±2.4
Efficient probing	87.3±0.4	90.1±0.4	95.0±0.3	91.9±0.3	88.6±1.6	90.6±0.2	64.7±0.9	48.6±1.5
Unaligned probe (UP)	86.9±0.6	89.2±0.9	95.1±0.4	91.8±0.5	88.5±0.4	90.3±0.4	63.1±0.9	50.0±2.5
BAP (Ours)	**88.8**±0.3	**90.7**±0.8	95.7±0.3	92.5±0.3	**89.1**±0.3	**91.4**±0.2	**66.2**±0.4	**55.2**±0.6

D VALIDATION OF THE PHASE ROLES

D.1 QUERY IDENTITY UNDER THE BEST ASSIGNMENT

To test for specialization under another naming, we choose the best one-to-one assignment of queries to windows on one half of the validation windows and score it on the other. UP on MAE-Base then reaches an in-window share of 32.4%, against 66.2% for BAP under its training names in the same pass and 25.0% for uniform attention, and for BAP the nominal assignment is the best in every seed.

D.2 ATTENTION ON RECORDS WITH DIFFERENT DIAGNOSES

Figure 6 shows UP and BAP (MAE-Base, seed 0) on one CPSC2018-R test window per diagnosis. For each diagnosis, the window shown has the median BAP probability among the fold-10 windows whose only shared label is that diagnosis and that carry at least four manual R-peaks. In every row the BAP queries return to their phase in each beat, including the premature beats of PAC and PVC, and the UP queries do not.

D.3 REUSABLE READOUT ACROSS SEEDS

The LUDB decoder is fitted on the attention maps of seed A over four folds and applied unchanged to those of seed B on the held-out fold, for the six ordered seed pairs and the five folds, with the queries of UP renamed per pair by the one-to-one assignment that agrees best with seed A on the training folds. Averaged over these 30 combinations, the reused decoder reaches 49.0 macro-F1 with BAP, against 50.8 when fitted on the same seed, and 21.2 with UP, within 4 points of the time-shuffled floor of 17.8.

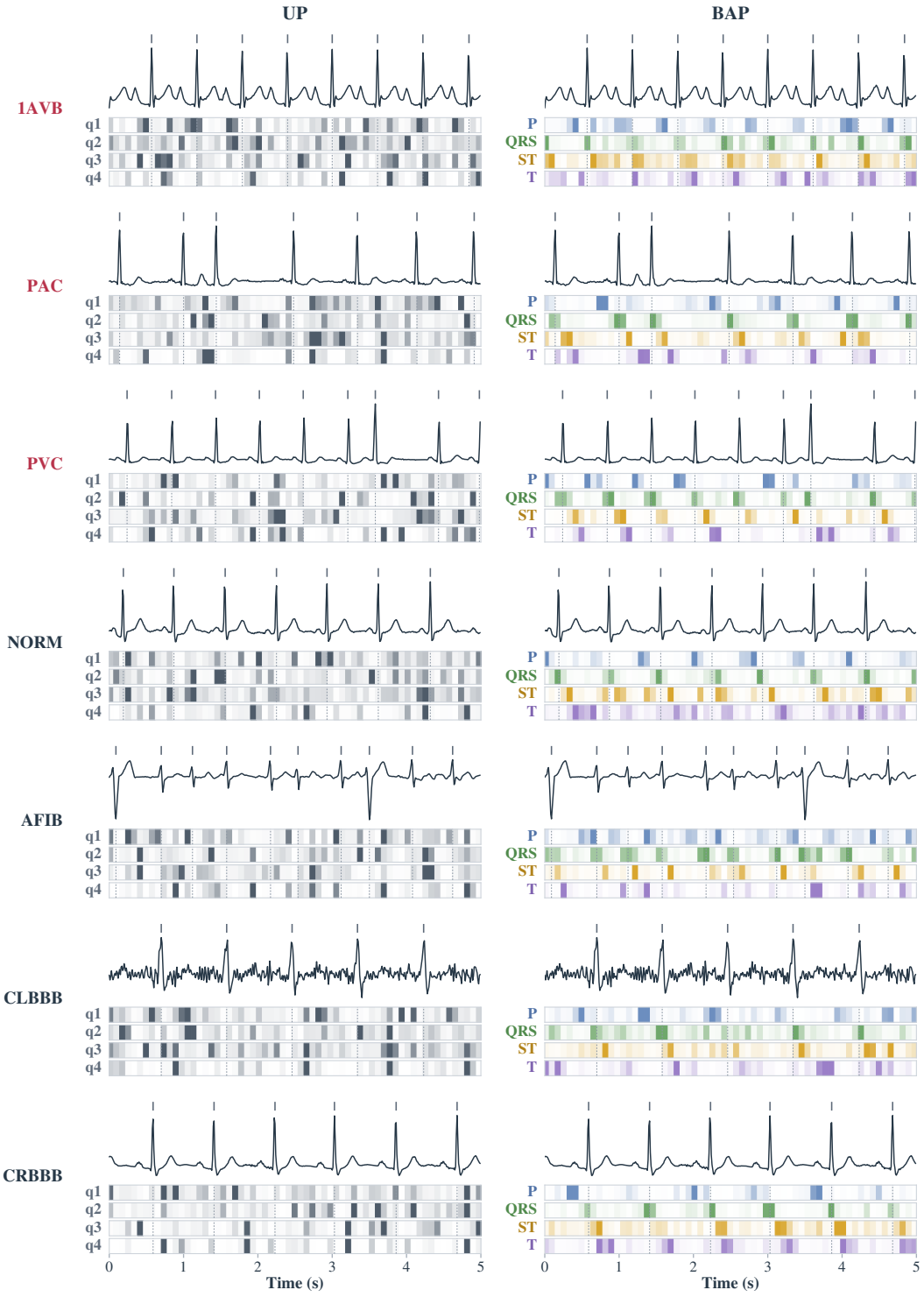


Figure 6: **UP and BAP attention on records with different diagnoses.** Each row shows lead II with manual R-peak ticks over the central 5 s of the window, and below it the attention of the four UP queries (q1 to q4) and the four BAP queries (P, QRS, ST, T) on the 80-ms tokens, with dotted lines at the R-peaks. Each strip is colored linearly from 0 to the 95th percentile of its own values.