# A PHYSIOLOGY-AWARE HIERARCHICAL FOUNDATION MODEL FOR LONG-CONTEXT WAVEFORM REPRESENTATION LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Physiological waveforms contain information at multiple timescales, from morphology within a cardiac cycle to changes across beats, minutes, and hours. Most existing waveform foundation models, however, divide signals into fixed-length patches and encode short windows independently. Consequently, their representation units may cut across meaningful cardiac structures, and their encoders cannot carry information from one window to the next. We introduce Physio-HNet, a causal hierarchical foundation model that addresses both limitations through physiology-aware dynamic chunking and continuous long-context modeling. Waveform landmarks and cardiac-cycle structure guide variable-length sub-beat and beat-level chunking, replacing fixed-length patches with physiologically informed units. Hierarchical compression and recurrent state preserve information across successive windows, allowing context to accumulate beyond isolated short segments. Following self-supervised waveform pretraining, routinely measured vital signs, including heart rate, respiratory rate, oxygen saturation, and blood pressure, supervise higher-level representations at successive measurement times along each recording. We evaluate Physio-HNet through frozen probes on two external clinical cohorts, context-length experiments, and controlled ablations. With 1.2 million parameters, the self-supervised encoder achieves performance comparable to existing waveform foundation models. Clinically anchored continued pretraining further improves blood-pressure estimation, yielding the highest $R^2$ and the lowest MAE in eleven of twelve blood-pressure comparisons across two cohorts and two probing protocols, including all six operating-room comparisons. Increasing available history to two hours improves estimation for most targets. Physiology-aware dynamic chunking improves several clinical estimates over fixed-length patching, while the segmentation achieving the best waveform reconstruction does not yield the best downstream representation. These findings support combining physiologically informed segmentation, continuous temporal context, and vital-sign supervision to learn transferable waveform representations.

## 1 INTRODUCTION

Physiological waveforms contain information at multiple temporal scales: morphology within a cardiac cycle, rhythm across successive beats, and physiological changes over minutes to hours (Charlton et al., 2023). A general-purpose representation should therefore preserve local morphology while accumulating information across an extended recording. Existing waveform foundation models have demonstrated the value of pretraining on large collections of short segments (Pillai et al., 2025; Saha et al., 2025; Nie et al., 2026; Guo et al., 2026; Chen et al., 2025; Li et al., 2025; Davies et al., 2024). Most of these models, however, organize signals on a fixed temporal grid and encode short windows independently, as do the patch-based general time-series models (Nie et al., 2023; Das et al., 2024). Fixed-length units can divide physiologically meaningful structures, such as cardiac cycles, across multiple representations, while independent window encoding prevents the encoder from carrying information from one window to the next (Figure 1, left). Consequently, any integration of longer histories must be performed by a downstream model. These limitations motivate three design choices that depend on one another: how to form representation units, how



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.

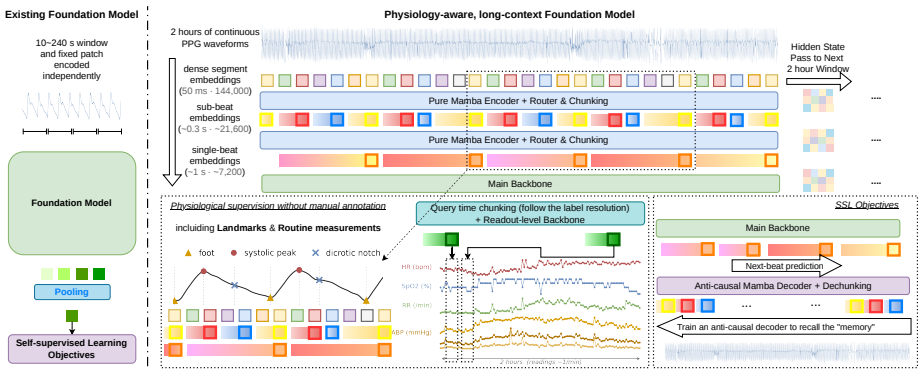



Figure 1: Comparison of independent-window encoding and the Physio-HNet streaming hierarchy. **Left:** fixed-length windows are encoded independently. **Right:** waveform tokens are progressively compressed into sub-beat and beat representations, with recurrent states carried across processing windows. **Bottom:** waveform landmarks supervise segmentation and routine clinical measurements supervise higher-level representations. Illustrative embedding counts assume a heart rate of 60 bpm. All signals shown are real recordings.

to accumulate temporal context across them, and how to supervise what that accumulated context retains.

We introduce Physio-HNet, a causal hierarchical waveform foundation model that adapts dynamic chunking (Hwang et al., 2025) to physiological structure and combines it with recurrent memory for long-context modeling. Automatically detected waveform landmarks (Goda et al., 2024) supervise sub-beat boundaries, while cardiac-cycle structure guides the chunking of beat-level segments. This produces variable-length chunks whose boundaries align with detected physiological structure rather than a fixed temporal grid. When a chunk closes, its final causal state becomes one input to the next level, progressively compressing the waveform into a sequence of beat-level representations. Causal state-space networks (Gu & Dao, 2024; Dao & Gu, 2024) maintain recurrent memory across successive processing windows, allowing these representations to incorporate preceding history (Figure 1, right). Dynamic chunking and recurrent memory therefore serve complementary roles: the hierarchy organizes and compresses local structure, while the state carries information over time. During self-supervised pretraining, a training-only decoder reconstructs the waveform through the hierarchy, encouraging the compressed representations to retain signal information.

Clinically anchored continued pretraining supervises the representation at every recorded measurement along a stream. In this work, heart rate, respiratory rate, oxygen saturation, and blood pressures are the targets. These targets are selected for this study; the framework can accommodate other time-aligned measurements, although their utility as supervision requires evaluation. Their values enter as targets and never as encoder inputs. The stage adapts the encoder under this supervision together with the retained self-supervised objectives, which regularize the representation and its segmentation rather than holding them fixed. The prediction heads are discarded before downstream probing, leaving a reusable encoder for evaluation across tasks and clinical settings.

We evaluate Physio-HNet on PPG using frozen-encoder probing on two external clinical cohorts, context-length experiments, and boundary and objective ablations. After self-supervised pretraining, the 1.2-million-parameter waveform encoder performs comparably to existing waveform foundation models, and clinically anchored continued pretraining further improves blood-pressure estimation across external emergency-department and operating-room cohorts. Increasing the available history to two hours improves estimation for most targets. We also compare physiology-aware dynamic chunking with fixed-length patching within the same hierarchy. Among the configurations selected for clinical continued pretraining, fixed-length patching achieves the best waveform reconstruction, yet landmark-supervised chunking achieves higher $R^2$ on all six clinical targets after that stage. Better reconstruction does not consistently correspond to better downstream performance across these configurations. These findings support evaluating segmentation and pretraining objectives through downstream utility alongside reconstruction quality. Our contributions are:

❶ **Physiology-aware dynamic chunking:** a hierarchical waveform encoder whose sub-beat and beat boundaries are guided by automatically derived physiological structure (§3.1, §3.2).
❷ **Continuous long-context modeling:** hierarchical compression combined with causal recurrent memory to accumulate waveform information across successive windows (§3.1, §5.2).
❸ **Clinically anchored continued pretraining:** repeated vital-sign supervision along individual recordings that shapes evolving representations and improves their transfer to external clinical cohorts (§3.4, §5.1).

## 2 Related work

**Physiological waveform foundation models.** Released PPG encoders (Pillai et al., 2025; Saha et al., 2025; Nie et al., 2026; Guo et al., 2026; Chen et al., 2025) and single-lead ECG encoders (Li et al., 2025; Davies et al., 2024) are pretrained on fixed-length windows ranging from seconds to minutes, using self-supervised or clinically supervised objectives. Although their architectures and objectives differ, these models generally process segments independently and form representations on grids not explicitly aligned with physiological events such as cardiac cycles. Physio-HNet instead uses a causal streaming hierarchy whose learned chunk boundaries are supervised by waveform landmarks and cardiac-cycle structure, and whose recurrent state is carried across successive processing windows.

**Physiology-guided representation learning.** Physiological knowledge has traditionally been incorporated upstream through handcrafted features or preprocessing (Goda et al., 2024; Akselrod et al., 1981; Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996; Charlton et al., 2023), while a parallel line of work uses cross-modal supervision from simultaneously recorded physiological signals (Shafiq et al., 2026; Chen et al., 2026). Physio-HNet instead uses waveform landmarks to supervise sub-beat segmentation and the times of clinical measurements to define the level above the beat during continued pretraining.

**Dynamic chunking.** H-Net (Hwang et al., 2025) introduces hierarchical, data-dependent chunking based on the similarity of adjacent hidden states. Physio-HNet adapts this mechanism to physiological waveforms by supervising the learned routers with automatically detected landmarks and cardiac-cycle structure during training; the detector is not required at inference. Additional discussion of these approaches appears in App. B.

## 3 Method

We introduce Physio-HNet, a causal hierarchical encoder for long-context physiological waveforms, trained through physiology-guided waveform pretraining and clinically anchored pretraining. The architecture is designed for quasi-periodic waveforms whose landmarks can be detected automatically. All experiments in this work use PPG; extension to other waveforms, such as ECG, remains to be evaluated. Figure 2 summarizes the framework; component details are in App. D.

### 3.1 Causal hierarchical encoding and compression

The upper path in Figure 2a shows the 1.2 M-parameter encoder used during waveform pretraining. PPG sampled at 40 Hz undergoes fixed dataset-level affine normalization, then every two non-overlapping samples are linearly projected into one token: 50 ms per token, or 20 tokens per second. The K0 encoder and router learn sub-beat units whose boundaries are encouraged to align with successive waveform landmarks; K1 groups these into beat-level units intended to follow cardiac cycles (§3.2), subject to the window boundary handling below. A beat backbone produces contextual embeddings $\{\mathbf{b}_i\}$. Time embeddings enter each encoder-path stack. All stacks use causal Mamba-2 blocks (Dao & Gu, 2024), modified to accept and return recurrent states in multi-token forward passes.

Both compression levels select chunk-end states (Figure 2b):

$$\mathbf{h}_t^{(\ell)} = \mathcal{E}_\ell(\mathbf{u}_{\le t}^{(\ell)}), \qquad \mathbf{u}_k^{(\ell+1)} = \mathbf{h}_{t_k^{(\ell)}}^{(\ell)}, \tag{1}$$

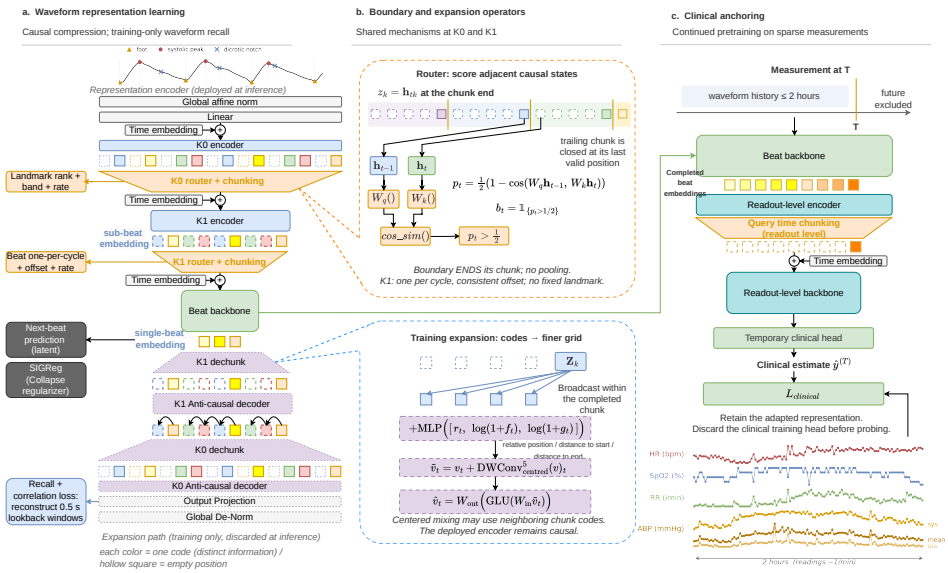



Figure 2: Physio-HNet architecture and training. **(a)** Causal hierarchical encoding with physiological boundary supervision and a training-only waveform reconstruction path. **(b)** Boundary scoring, chunk-end state selection and the expansion operators at the sub-beat and beat levels. **(c)** Query-aligned clinical supervision using waveform history up to each measurement time. The reconstruction and clinical heads are discarded before downstream probing. The selected A2 configuration omits next-beat prediction and retains the collapse regulariser.

where $t_k^{(\ell)}$ closes chunk $k$. Unlike the start-of-chunk convention of Hwang et al. (2025), the final state has processed the entire chunk, can retain earlier history, and aligns the coarse representation with the time it becomes available. No pooling is required.

At both levels, trailing chunks close at the last valid position of each processing window, so a cardiac cycle crossing a window boundary can be split into separate units. Recurrent states, however, persist across consecutive windows within an encounter during both training and inference, resetting between encounters. Training detaches these states at window boundaries to truncate backpropagation. Thus, waveform pretraining on five-minute windows can carry earlier context; the clinical stage trains and evaluates with up to two hours of history.

The waveform-pretraining hierarchy ends at the beat level. Clinical continued pretraining adds the query-aligned coarse level in Figure 2c (§3.4). The lower expansion path in Figure 2a is training-only (§3.3). Implementation and architecture details appear in App. D and App. H.

### 3.2 Physiology-aware dynamic chunking

K0 and K1 score adjacent causal states to identify chunk boundaries (Figure 2b):

$$p_t = \frac{1}{2}\left(1 - \cos(W_q \mathbf{h}_{t-1},\, W_k \mathbf{h}_t)\right), \qquad b_t = \mathbb{1}_{\{p_t > 1/2\}}. \tag{2}$$

Beyond the learned projections $W_q$ and $W_k$, there is no separate scorer or additional temporal receptive field: boundary supervision acts directly on the encoder representations. Automatically detected landmarks and cardiac cycles provide training targets (Goda et al., 2024); the detector is not required at inference.

At K0, pulse feet, systolic peaks, and dicrotic notches define boundary targets. A ranking loss assigns higher boundary scores to detected landmarks than to neighbouring non-landmark positions, a threshold loss calibrates scores to the operating point, and a rate loss controls boundary counts. Missing detections are masked. At K1, supervision encourages one boundary per cardiac cycle at a consistent within-cycle offset, without prescribing a particular landmark or enforcing equal beat durations.

During waveform pretraining, hard selection prevents representation losses from directly updating router parameters; boundary losses train the routers, while both objective types update the underlying encoder states. Window-end closure follows §3.1. Full objectives and detection details appear in App. D and App. I.

### 3.3 Self-supervised objectives

The training-only expansion path reconstructs the waveform from compressed beat representations (Figure 2a,b). At each level, dechunking broadcasts chunk codes onto the finer grid and adds within-chunk position features before anti-causal decoding. The decoder reads codes produced at chunk closure: it may combine later codes for reconstruction, but introduces no future inputs into the encoder's causal forward computation. No encoder skip connection bypasses the compression bottlenecks. A shared output head reconstructs the 0.5 s waveform window $w_t$ ending at each token:

$$\mathcal{L}_{\text{recall}} = \left\|\hat{w}_t - w_t\right\|^2_{\text{masked}}, \qquad \mathcal{L}_{\text{corr}} = 1 - \rho(\hat{w}_t, w_t), \tag{3}$$

where both terms are averaged over valid positions and $\rho$ is Pearson correlation within each reconstructed window. The correlation term measures morphological agreement independently of each window's positive amplitude scale and offset. The expansion path is discarded at inference.

We also evaluate latent next-beat prediction, $\mathcal{L}_{\text{nbp}} = 1 - \cos(\psi(\mathbf{b}_i),\, \text{sg}[\mathbf{b}_{i+1}])$, where $\psi$ is a predictor and sg denotes stop-gradient. A separate SIGReg collapse regularizer (Balestriero & LeCun, 2025) acts on beat embeddings and remains active when next-beat prediction is removed.

Pretraining combines these objectives with the boundary losses of §3.2. The main model retains recall, correlation, boundary supervision, and collapse regularization, omitting next-beat prediction based on the objective ablation (§5.4). Full definitions, loss weights, and the rationale for anti-causal decoding appear in App. D and App. H.

### 3.4 Clinically anchored continued pretraining

Clinical continued pretraining adds a query-aligned level above the beat backbone (Figure 2c). A causal readout-level encoder processes beat embeddings, with measurement times defining chunk boundaries in place of a learned router. At query time $T$, chunk-end selection uses the last available beat-level state at or before $T$, excluding future waveform samples. A readout-level backbone processes the resulting coarse sequence, and a temporary clinical head estimates the targets. Measurement values provide supervision only, never encoder inputs. The entire encoder, including its routers, is updated jointly with the added readout level and training-only expansion path.

We use heart rate, respiratory rate, oxygen saturation, and systolic, diastolic, and mean blood pressure as supervision:

$$\mathcal{L}_{\text{cont}} = \mathcal{L}_{\text{clinical}} + w_{\text{ssl}}\mathcal{L}_{\text{SSL}}, \qquad w_{\text{ssl}} = 0.1, \tag{4}$$

where $\mathcal{L}_{\text{clinical}}$ is an $L_1$ loss on observed, normalized targets at successive measurement times, and $\mathcal{L}_{\text{SSL}}$ retains the waveform pretraining objectives.

A straight-through gradient path is enabled for routing during this stage. The retained self-supervised objectives regularize the representations and segmentation during adaptation to clinical supervision. Each training window contains up to two hours of one patient's waveform, supervised at observed measurement times throughout. The temporary clinical head is subsequently discarded, while the adapted waveform encoder is frozen for downstream probing. Details appear in App. J and App. H.

## 4 Pretraining and datasets

**Pretraining data and configurations.** Physio-HNet is pretrained on approximately 24,800 hours of 40 Hz intensive-care PPG from 47,034 encounters and 39,481 patients at one academic center. The corpus contains 297,219 five-minute windows, split 90/5/5 by encounter; 555 patients appear across splits through repeated admissions. We apply no additional signal-quality filtering. Four arms share the corpus, representation backbone, and optimization budget of 400k steps with 96 concurrent streams: A1 uses a convolutional router with landmark supervision; A2 (Physio-HNet) uses the

physiologically supervised pairwise router; B uses the same router without boundary-location supervision; and C uses fixed-stride chunking. We evaluate waveform recall and next-beat prediction separately and jointly, retaining SIGReg throughout. Boundary settings and training hyperparameters appear in App. E and App. H.

**Clinically anchored continued pretraining.** We pair PPG with routine vital signs and arterial pressures in MIMIC-III (Johnson et al., 2016; Moody et al., 2020), using 3140/645/651 training/validation/test encounters and two-hour windows. Measurements are charted at a 60 s cadence; arterial pressures are available for approximately one third of test encounters. The training procedure follows §3.4, with each arm's objective configuration selected as described in §5.4. The held-out split supports the context and boundary evaluations below. Optimization settings, window counts, and target processing appear in App. J.

**External evaluation datasets.** We evaluate transfer on MOVER operating-room (Samad et al., 2023) and MC-MED emergency-department recordings (Kansal et al., 2025), neither used in either pretraining stage of Physio-HNet. Both provide PPG and measurements covering six target categories: heart rate, oxygen saturation, respiratory rate, and systolic, diastolic, and mean pressure. Clinical splits are formed at the encounter level. Preprocessing, target availability, and split details appear in App. L.

**Auxiliary benchmark datasets.** Objective ablations use four datasets from the benchmark suite of Kataria et al. (2025): WESAD, Stanford atrial fibrillation, Pulse Transit Time PPG, and Real-World PPG, covering affective state, arrhythmia, activity, and identity. All four datasets are held out from both stages of encoder training, but their probe scores also support objective configuration selection for clinical continued pretraining. Dataset references and splits appear in App. L; evaluation and selection criteria are specified in §5.4.

# 5 Results

## 5.1 Transfer of the frozen representation

We compare Physio-HNet, before and after clinical continued pretraining, with HeartGPT-PPG (Davies et al., 2024), PaPaGei-S (Pillai et al., 2025), AnyPPG (Nie et al., 2026), and SIGMA-PPG (Guo et al., 2026) on the two external cohorts, using static and recurrent probes trained on frozen embeddings and evaluated at the same measurement times (Table 1). The encoder checkpoints are the validation-selected ones and the probes themselves train to a fixed budget and are tested once; native encoder windows, embedding widths, and output rates remain model-specific (App. L).

With 1.2 million parameters and no pretraining data from either cohort, clinically anchored Physio-HNet achieves both the highest $R^2$ and the lowest MAE in eleven of the twelve blood-pressure comparisons. It ranks first in all six operating-room comparisons; none of the evaluated encoders was pretrained on that cohort. The remaining blood-pressure comparison is in the emergency-department cohort, which supplies approximately $72\%$ of AnyPPG's pretraining corpus. Physio-HNet also leads heart-rate estimation under the static probe in both cohorts and performs comparably to AnyPPG under the recurrent probe. Clinical continued pretraining improves blood-pressure estimation in all twelve comparisons relative to the self-supervised Physio-HNet encoder. Before that stage, the self-supervised encoder already attains the highest $R^2$ in six of the twelve, five of them in the operating room, and is comparable to the released encoders on heart rate. Clinical continued pretraining also lowers pressure MAE in every cohort–probe setting under all four boundary configurations (App. E.1, Table 6).

We exclude two columns from encoder comparisons: oxygen saturation is not directly determined by single-wavelength PPG, and operating-room respiratory rate is a sparsely charted ventilator parameter (App. L). Anchored Physio-HNet reaches $R^2 = .217$ on emergency-department respiratory rate under the recurrent probe.

## 5.2 Long context beyond a fixed window

To test whether the clinically adapted encoder benefits from longer context, we evaluate A2, pretrained with recall alone and given clinical continued pretraining on MIMIC-III, at seven history

Table 1: Frozen-encoder performance on MC-MED and MOVER under static and recurrent probing. Values are the mean and standard deviation of $R^2$ and MAE over five probe-training seeds. *SSL* and *anchored* are Physio-HNet before and after clinical continued pretraining. MAE is in bpm for heart rate, percentage points for oxygen saturation, breaths per minute for respiratory rate and mmHg for the pressures. Blue marks the best value within a cohort-probe block and shading identifies the anchored row; grey marks the two columns excluded from encoder comparisons (§5.1).

| | HR | | $SpO_2$ | | RR | | NBP-S | | NBP-D | | NBP-M | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE |
| *Static probe*, emergency department (MC-MED) | | | | | | | | | | | | |
| HeartGPT-PPG | .845±.012 | 3.62±0.33 | −.042±.076 | 1.57±0.14 | .053±.063 | 3.45±0.10 | .164±.028 | 16.31±0.27 | .087±.057 | 11.34±0.34 | .101±.057 | 11.54±0.31 |
| PaPaGei-S | .795±.023 | 4.53±0.47 | −.067±.083 | 1.57±0.10 | .075±.069 | 3.41±0.12 | .092±.025 | 17.06±0.18 | .055±.026 | 11.59±0.16 | .057±.026 | 11.85±0.13 |
| AnyPPG | .891±.010 | **2.72±0.26** | .059±.017 | 1.45±0.01 | .112±.032 | 3.32±0.05 | .254±.015 | 15.30±0.05 | .184±.026 | 10.66±0.12 | **.227±.018** | **10.63±0.07** |
| SIGMA-PPG | .879±.009 | 3.44±0.13 | .036±.030 | 1.45±0.02 | .194±.028 | 3.14±0.04 | .150±.030 | 16.34±0.18 | .073±.019 | 11.45±0.18 | .092±.024 | 11.56±0.09 |
| Physio-HNet (SSL) | **.900±.006** | 2.91±0.32 | .006±.052 | 1.49±0.07 | .066±.081 | 3.40±0.13 | .241±.027 | 15.16±0.33 | .173±.017 | 10.74±0.13 | .179±.014 | 11.04±0.16 |
| Physio-HNet (anchored) | .898±.007 | 3.07±0.29 | .022±.042 | 1.50±0.09 | **.204±.053** | **3.09±0.12** | **.269±.022** | **15.07±0.22** | **.195±.023** | **10.57±0.17** | .219±.024 | 10.67±0.08 |
| *Static probe*, operating room (MOVER) | | | | | | | | | | | | |
| HeartGPT-PPG | .638±.006 | 6.49±0.14 | −.068±.037 | 1.24±0.03 | .054±.024 | 2.92±0.06 | .040±.054 | 16.87±0.32 | .064±.048 | 8.79±0.14 | .059±.049 | 11.04±0.24 |
| PaPaGei-S | .622±.006 | 6.67±0.08 | −.065±.037 | 1.24±0.04 | .018±.030 | 2.99±0.06 | .026±.037 | 16.98±0.10 | .035±.019 | 8.88±0.02 | .041±.027 | 11.11±0.03 |
| AnyPPG | .644±.038 | 6.36±0.37 | −.073±.025 | 1.23±0.01 | .020±.009 | 2.98±0.02 | .097±.047 | 16.28±0.24 | .108±.070 | 8.48±0.17 | .117±.059 | 10.57±0.20 |
| SIGMA-PPG | .642±.004 | 6.51±0.09 | −.087±.025 | 1.21±0.01 | .034±.023 | 2.95±0.05 | .031±.012 | 16.77±0.09 | .019±.024 | 8.93±0.05 | .032±.018 | 11.07±0.07 |
| Physio-HNet (SSL) | .687±.009 | 6.04±0.24 | −.059±.033 | 1.29±0.23 | .004±.037 | 3.03±0.09 | .089±.082 | 16.51±0.60 | .140±.042 | 8.42±0.16 | .123±.060 | 10.64±0.45 |
| Physio-HNet (anchored) | **.696±.008** | **5.71±0.12** | −.059±.029 | 1.23±0.06 | .045±.021 | 2.92±0.05 | **.149±.012** | **15.80±0.33** | **.165±.010** | **8.22±0.11** | **.166±.008** | **10.28±0.24** |
| *Recurrent probe*, emergency department (MC-MED) | | | | | | | | | | | | |
| HeartGPT-PPG | .864±.008 | 3.17±0.12 | .046±.021 | 1.45±0.01 | .095±.023 | 3.35±0.03 | .218±.010 | 15.78±0.12 | .134±.008 | 11.09±0.13 | .165±.007 | 11.18±0.11 |
| PaPaGei-S | .845±.009 | 3.57±0.31 | .045±.020 | 1.44±0.01 | .126±.049 | 3.30±0.07 | .139±.012 | 16.61±0.09 | .091±.010 | 11.38±0.05 | .095±.008 | 11.63±0.04 |
| AnyPPG | **.920±.006** | **2.17±0.13** | .056±.013 | 1.46±0.02 | .134±.040 | 3.27±0.05 | .252±.024 | 15.35±0.21 | .198±.022 | 10.58±0.13 | .212±.020 | 10.77±0.13 |
| SIGMA-PPG | .829±.006 | 4.01±0.10 | .046±.022 | 1.45±0.01 | .190±.018 | 3.14±0.03 | .155±.014 | 16.30±0.09 | .091±.010 | 11.38±0.05 | .100±.014 | 11.57±0.06 |
| Physio-HNet (SSL) | .916±.003 | 2.38±0.15 | .061±.015 | 1.43±0.01 | .119±.030 | 3.29±0.04 | .269±.014 | 15.21±0.20 | .180±.025 | 10.75±0.26 | .203±.020 | 10.90±0.24 |
| Physio-HNet (anchored) | **.920±.004** | 2.27±0.07 | .068±.020 | 1.41±0.01 | **.217±.034** | **3.10±0.06** | **.303±.015** | **14.74±0.16** | **.220±.006** | **10.44±0.12** | **.242±.008** | **10.59±0.12** |
| *Recurrent probe*, operating room (MOVER) | | | | | | | | | | | | |
| HeartGPT-PPG | .660±.006 | 6.06±0.07 | −.068±.022 | 1.20±0.01 | .056±.019 | 2.92±0.03 | .088±.024 | 16.42±0.08 | .098±.028 | 8.60±0.03 | .106±.023 | 10.76±0.06 |
| PaPaGei-S | .664±.009 | 6.03±0.12 | −.066±.019 | 1.20±0.01 | .043±.009 | 2.94±0.02 | .079±.015 | 16.53±0.06 | .078±.017 | 8.71±0.04 | .091±.015 | 10.87±0.04 |
| AnyPPG | **.680±.006** | **5.66±0.05** | −.072±.036 | 1.22±0.01 | .031±.007 | 2.97±0.02 | .127±.036 | 15.97±0.16 | .146±.052 | 8.27±0.14 | .154±.041 | 10.37±0.15 |
| SIGMA-PPG | .586±.004 | 7.10±0.08 | −.074±.032 | 1.22±0.01 | .046±.012 | 2.94±0.02 | .031±.013 | 16.95±0.03 | .018±.023 | 9.03±0.04 | .031±.016 | 11.26±0.04 |
| Physio-HNet (SSL) | .671±.003 | 5.70±0.08 | −.055±.027 | 1.20±0.02 | .016±.023 | 3.01±0.04 | .138±.013 | 15.94±0.08 | .166±.018 | 8.24±0.03 | .162±.018 | 10.36±0.05 |
| Physio-HNet (anchored) | .671±.004 | 5.72±0.09 | −.059±.026 | 1.20±0.01 | .059±.006 | 2.91±0.02 | **.158±.012** | **15.74±0.10** | **.184±.016** | **8.16±0.06** | **.181±.015** | **10.23±0.08** |

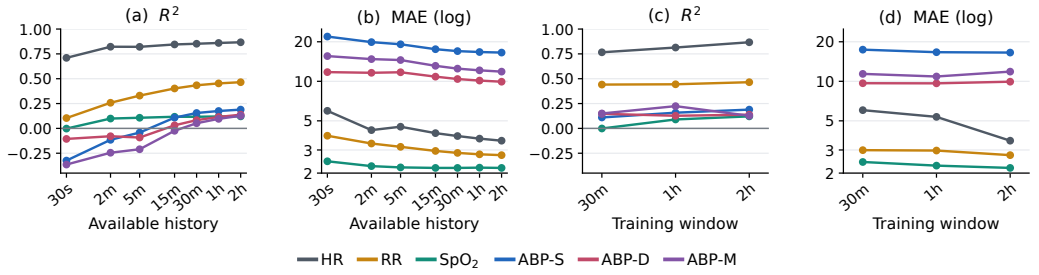



Figure 3: Clinical estimation on held-out MIMIC-III recordings, against the history available at evaluation **(a, b)** and against the processing window each model was trained with **(c, d)**. $R^2$ and MAE in each target's own unit, MAE on a logarithmic scale. The two-hour point of (c, d) is the checkpoint of (a, b), and the two horizontal axes are not comparable. Numerical values for (a, b) are in Table 10.

lengths on the held-out split. The validation-selected checkpoint and clinical head remain fixed. All lengths use the same observation anchors, with the recurrent state reset for each trial and inputs restricted to waveform samples at or before the query time (App. K). All targets improve from the shortest context to two hours, but their responses differ (Figure 3). Heart rate gains most between thirty seconds and two minutes, oxygen saturation largely plateaus after fifteen minutes, and respiratory rate continues improving with longer context. Pressure $R^2$ becomes positive by fifteen minutes for systolic and diastolic pressure and by thirty minutes for mean pressure, with further gains through two hours. A separate check confirms that feeding a trial as two windows with state carried across the join produces a different estimate from resetting state there.

We also examine the effect of the training-window length. Panels (c) and (d) compare models trained with thirty-minute, one-hour and two-hour processing windows, each carrying recurrent state over

Table 2: Clinical estimation by the four boundary configurations on held-out MIMIC-III recordings, after continued pretraining on that cohort. Each arm uses the configuration selected in §5.4; values are $R^2$ and MAE in each target's own unit, one run per arm. Blue marks the best value in a column and shading identifies A2.

| | HR | | $SpO_2$ | | RR | | ABP-S | | ABP-D | | ABP-M | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE |
| A1 | **.922** | 2.95 | **.123** | 2.37 | .277 | 3.10 | .081 | 18.01 | .106 | 10.00 | **.217** | 11.25 |
| A2 | .915 | 3.23 | .113 | **2.35** | **.468** | **2.88** | **.139** | 17.06 | **.212** | **9.61** | .210 | 11.24 |
| B | .917 | **2.90** | .089 | 2.44 | .375 | 3.20 | .024 | 18.47 | .118 | 9.88 | .210 | **11.07** |
| C | .912 | 2.95 | .017 | **2.35** | .157 | 3.93 | .100 | **16.73** | −.464 | 13.40 | −.088 | 12.71 |

Table 3: Ablation of waveform recall and next-beat prediction within each of the four boundary configurations. Reconstruction is MSE and Pearson $r$; the downstream scores are AUROC for emotion, atrial fibrillation and activity and macro-F1 for identity, from one probe-training seed; brackets are 95% bootstrap intervals over test examples. The shaded four-task mean selects each arm's objective configuration. Blue marks the best score within an arm and a red rule the best over the whole table, ties all marked. A dash marks a row trained without waveform recall, which has no reconstruction.

| | Recon. MSE | Recon. $r$ | WESAD emotion | DeepBeat AF | PTT-PPG activity | Real-World identity | mean |
|---|---|---|---|---|---|---|---|
| *A1*, supervised placement, convolutional scorer | | | | | | | |
| both | 0.0059 | .964 | .73 [.71, .75] | **.64 [.63, .65]** | .62 [.59, .66] | .99 [.99, .99] | .745 |
| recall only | **0.0039** | **.967** | **.77 [.75, .80]** | .63 [.62, .64] | .74 [.71, .77] | **1.00 [1.00, 1.00]** | **.784** |
| next-beat only | – | – | .71 [.69, .74] | .63 [.62, .64] | **.75 [.71, .78]** | .94 [.94, .95] | .760 |
| ***A2*, supervised placement, pairwise router** | | | | | | | |
| both | 0.0080 | .960 | **.77 [.75, .79]** | .68 [.67, .69] | .67 [.64, .70] | .99 [.98, .99] | .776 |
| recall only | **0.0053** | **.964** | **.77 [.75, .79]** | **.72 [.71, .73]** | **.77 [.74, .80]** | **1.00 [1.00, 1.00]** | **.812** |
| next-beat only | – | – | .66 [.64, .69] | .53 [.52, .54] | .54 [.51, .58] | .85 [.84, .86] | .647 |
| *B*, unsupervised placement, same router as A2 | | | | | | | |
| both | 0.0160 | .953 | .74 [.71, .76] | .67 [.66, .68] | .70 [.67, .73] | **.99 [.99, .99]** | .775 |
| recall only | **0.0039** | **.962** | **.76 [.74, .78]** | **.68 [.67, .69]** | **.76 [.73, .80]** | .98 [.98, .99] | **.797** |
| next-beat only | – | – | .74 [.71, .76] | .62 [.61, .63] | .69 [.66, .73] | .95 [.94, .95] | .751 |
| *C*, fixed stride, no router | | | | | | | |
| both | 0.0008 | .974 | **.72 [.69, .74]** | **.75 [.74, .76]** | **.70 [.66, .73]** | **.98 [.97, .98]** | **.785** |
| recall only | **0.0002** | **.975** | .62 [.59, .64] | .57 [.56, .58] | .69 [.66, .72] | .84 [.83, .86] | .680 |
| next-beat only | – | – | .54 [.51, .57] | .51 [.50, .52] | .54 [.50, .57] | .53 [.51, .55] | .529 |
| HeartGPT-PPG | – | – | .75 [.72, .77] | .59 [.58, .60] | .77 [.74, .80] | .96 [.95, .96] | .766 |
| PaPaGei-S | – | – | .67 [.65, .70] | .56 [.55, .57] | .72 [.68, .75] | .92 [.91, .93] | .715 |
| AnyPPG | – | – | .76 [.74, .78] | .69 [.68, .70] | .77 [.74, .80] | 1.00 [1.00, 1.00] | .804 |
| SIGMA-PPG | – | – | .66 [.63, .68] | .59 [.58, .59] | .55 [.52, .59] | .83 [.82, .85] | .657 |

six hours per stream, holding the state-carry duration constant while varying the backpropagation window. Each model is evaluated at its own window length. The sweep starts at thirty minutes because the clinical targets are charted once a minute, so shorter windows supply few anchors each. Heart rate, respiratory rate, oxygen saturation and systolic pressure improve with longer windows, heart-rate $R^2$ rising from .766 to .867, whereas diastolic and mean pressure read best at one hour, mean-pressure $R^2$ falling from .224 there to .127 at two hours. Each configuration is a single run.

Because the checkpoint of (a) and (b) was trained with two-hour windows, shorter evaluation histories also introduce a mismatch between training and evaluation context lengths; (c) and (d) vary the training window directly, but at three lengths and one run each. Each trial starts cold, so the results do not measure uninterrupted state carry across processing windows or separate retained history from recurrent-state settling.

### 5.3 Effect of boundary-location supervision

We compare the four arms of §4, pretrained on the same data at matched boundary rates and read after clinical continued pretraining (Table 2). The arms differ by more than boundary placement: A1 uses a non-causal convolutional scorer, and C carries a different selected objective (App. E). The A2–B contrast more directly isolates boundary-location supervision: both use the same router and objective configuration, differing in whether supervision specifies where boundaries belong.

By $R^2$, A2 outperforms B on respiratory rate, systolic and diastolic pressure, and oxygen saturation, and exceeds fixed-stride chunking on all six targets. The ordering is not uniform across metrics: on heart rate A2 has the highest error of the four arms, and on systolic pressure fixed-stride chunking

has the lowest. Fixed-stride chunking yields negative $R^2$ for diastolic and mean pressure and the lowest $R^2$ of the four arms on respiratory rate and oxygen saturation.

The A2–B comparison extends to the two external cohorts: after clinical continued pretraining A2 achieves lower blood-pressure MAE than B in all twelve cohort–probe–target settings and higher $R^2$ in eleven, with equal reported $R^2$ in the remaining setting (Table 6). A1 achieves both the higher $R^2$ and the lower MAE than A2 in five of the twelve.

After clinical continued pretraining, beat-level alignment against a rate-matched grid rises for A2 and falls for B, while A2's sub-beat margin falls toward that baseline (App. G.2, Table 8).

These results support boundary-location supervision within this hierarchy. App. M illustrates beat representations, boundary placement and waveform reconstruction before and after clinical continued pretraining.

### 5.4 Effect of pretraining objectives

Table 3 compares waveform recall and next-beat prediction on four auxiliary benchmarks, using AUROC for the first three tasks and macro-F1 for identification. Results use one probe-training seed with 95% intervals. Stanford atrial fibrillation is the primary task criterion; the unweighted four-task mean selects each arm's configuration for clinical continued pretraining.

Removing recall reduces the four-task mean by $0.17$ for A2 and $0.26$ for C, against $0.02$ for A1 and $0.05$ for B; the smaller differences remain unresolved by this single-seed evaluation. A2 declines across all four tasks, whereas C's largest decline is in identification, $0.98$ to $0.53$. The effect therefore varies across configurations that share either supervision or router architecture.

Removing next-beat prediction increases effective dimensionality two- to fourfold in every arm and reduces adjacent-embedding similarity where it was high (App. F, Table 7). It raises the four-task mean for A1, A2, and B but lowers it for C. We therefore retain recall without next-beat prediction for A1, A2, and B, and both objectives for C, with SIGReg active throughout. A2 reaches a mean of $0.812$ against AnyPPG's $0.804$, but the reported intervals do not resolve the difference. Transfer to independent clinical cohorts is evaluated in Table 1.

The ranking by reconstruction quality differs from the ranking by downstream performance. Fixed-stride chunking under recall alone achieves the lowest MSE, $0.0002$, but the third-lowest downstream mean, while the highest downstream mean accompanies an error more than twenty times larger (App. F, Table 7).

## 6 Limitations

Our evaluation is limited to PPG pretrained at a single center; transfer to ECG, to multimodal recordings and to broader acquisition conditions remains to be evaluated. Boundary supervision relies on automatically detected landmarks, and robustness to systematic detection errors is not established. Clinical anchoring uses basic vital signs, leaving other supervision targets and measurement frequencies unexplored. Context-length evaluation is limited to two hours; performance on longer uninterrupted streams remains untested.

## 7 Conclusion and future directions

Physio-HNet brings physiological structure, continuous temporal context and clinical supervision into one framework for waveform representation learning: chunking organizes local morphology, recurrent memory carries context across windows, and repeated clinical measurements supervise the representations along each recording. With 1.2 million parameters the self-supervised encoder is competitive, clinical anchoring further improves blood-pressure estimation on external cohorts, and reconstruction quality alone does not identify the most useful downstream representation.

Future work could extend the hierarchy to ECG using automatically detected ECG landmarks, and a multimodal extension could integrate the two at the beat level while preserving the timing differences between signals. Because segmentation and higher-level learning are separated, future work could also evaluate additional clinical targets as supervision for the coarse-level representations.

### AI USE STATEMENT

We used generative AI tools for language polishing and editorial revisions, including assistance with literature searches and reference checking. We have reviewed all AI-assisted work: every reference surfaced or checked this way was verified by the authors against the published record, and every revised passage was read against the results it reports. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work uses retrospective, de-identified physiological recordings and the routine clinical measurements charted alongside them. The intensive-care corpus used for self-supervised pretraining is private: it was analysed under an approved institutional review board protocol, it is not released with this paper and cannot be redistributed, and the contributing institution is left unnamed so that the submission remains anonymous. MIMIC-III (Johnson et al., 2016), MOVER (Samad et al., 2023), MC-MED (Kansal et al., 2025) and the auxiliary benchmark datasets are public and were used under their own credentialing requirements and data use agreements. No re-identification of any participant was attempted, and what we release, model weights and configuration files, carries no recordings and no patient identifiers.

Physio-HNet is a research model. Its estimates of heart rate, oxygen saturation, respiratory rate and arterial pressure are reported to characterise what the representation carries, not as a substitute for a monitored measurement, and nothing reported here is validated or intended for clinical decision-making or for deployment as a medical device. Two limits bear on equitable use. Photoplethysmography is sensitive to perfusion and to skin pigmentation, and the cohorts used here carry no annotation that would let us measure how the representation behaves across skin tone. And the pretraining corpus comes from a single academic center, so the two external cohorts test transfer across care settings rather than across health systems at large. App. L records which targets are scarce or device-derived in each cohort, including the two columns excluded from encoder comparisons.

### REPRODUCIBILITY STATEMENT

Model architecture and training objectives are defined in §3, the pretraining corpus and the continued-pretraining and evaluation cohorts in §4, and the evaluation protocol in §5. Component and loss definitions appear in App. D, the four boundary configurations and their objective weights in App. E, architecture sizes and the optimisation schedule in App. H, fiducial supervision in App. I, the clinical anchoring stage with its targets and anchors in App. J, the context sweep in App. K, and the external comparison protocol, under which every encoder is run without exception, in App. L. The pretraining and continued-pretraining configurations are frozen files, reported there in full down to every loss weight.

We will release the training side of the model: the encoder implementation, the two training stages, the configuration of every arm reported here, and the pretrained and clinically anchored checkpoints. Evaluation code is not part of that release; the protocol under which every encoder is run is specified in App. L. The pretraining recordings are private and cannot be redistributed, while clinical continued pretraining and every evaluation in this paper run on public data.

## REFERENCES

Solange Akselrod, David Gordon, F. Andrew Ubel, Daniel C. Shannon, A. Clifford Berger, and Richard J. Cohen. Power spectrum analysis of heart rate fluctuation: A quantitative probe of beat-to-beat cardiovascular control. *Science*, 213(4504):220–222, 1981. doi: 10.1126/science.6166045.

Abdul Fatir Ansari, Lorenzo Stella, Caner Türkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. arXiv:2403.07815.

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

Peter H. Charlton, John Allen, Raquel Bailón, Stephanie Baker, Joachim A. Behar, Fei Chen, Gari D. Clifford, David A. Clifton, Harry J. Davies, Cheng Ding, Xiaorong Ding, Jessilyn Dunn, Mohamed Elgendi, Munia Ferdoushi, Daniel Franklin, Eduardo Gil, Md Farhad Hassan, Jussi Hernesniemi, Xiao Hu, Nan Ji, Yasser Khan, Spyridon Kontaxis, Ilkka Korhonen, Panicos A. Kyriacou, Pablo Laguna, Jesús Lázaro, Chungkeun Lee, Jeremy Levy, Yumin Li, Chengyu Liu, Jing Liu, Lei Lu, Danilo P. Mandic, Vaidotas Marozas, Elisa Mejía-Mejía, Ramakrishna Mukkamala, Meir Nitzan, Tania Pereira, Carmen C. Y. Poon, Jessica C. Ramella-Roman, Harri Saarinen, Md Mobashir Hasan Shandhi, Hangsik Shin, Gerard Stansby, Toshiyo Tamura, Antti Vehkaoja, Will Ke Wang, Yuan-Ting Zhang, Ni Zhao, Dingchang Zheng, and Tingting Zhu. The 2023 wearable photoplethysmography roadmap. *Physiological Measurement*, 44(11):111001, 2023. doi: 10.1088/1361-6579/acead2.

Zhaoliang Chen, Cheng Ding, Saurabh Kataria, Runze Yan, Minxiao Wang, Randall Lee, and Xiao Hu. GPT-PPG: A GPT-based foundation model for photoplethysmography signals. *Physiological Measurement*, 46(5):055004, 2025. doi: 10.1088/1361-6579/add988.

Zhaoliang Chen, Saurabh Kataria, Minxiao Wang, and Xiao Hu. Bridging ECG and PPG: Latent-space prediction for robust physiological analysis. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 10693–10703, 2026. doi: 10.1145/3770855.3819000.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2405.21060.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2310.10688.

Harry J. Davies, James Monsen, and Danilo P. Mandic. Interpretable pre-trained transformers for heart time-series data. *arXiv preprint arXiv:2407.20775*, 2024.

Márton Á. Goda, Peter H. Charlton, and Joachim A. Behar. pyPPG: A python toolbox for comprehensive photoplethysmography signal analysis. *Physiological Measurement*, 45(4):045001, 2024. doi: 10.1088/1361-6579/ad33a2.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2312.00752.

Ramesh M. Gulrajani. The forward and inverse problems of electrocardiography. *IEEE Engineering in Medicine and Biology Magazine*, 17(5):84–101, 122, 1998. doi: 10.1109/51.715491.

Zongheng Guo, Tao Chen, Yang Jiao, Yi Pan, Xiao Hu, and Manuela Ferrario. SIGMA-PPG: Statistical-prior informed generative masking architecture for PPG foundation model. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2601.21031.

Arthur C. Guyton, Thomas G. Coleman, and Harris J. Granger. Circulation: Overall regulation. *Annual Review of Physiology*, 34(1):13–44, 1972. doi: 10.1146/annurev.ph.34.030172.000305.

Matti Hämäläinen, Riitta Hari, Risto J. Ilmoniemi, Jukka Knuutila, and Olli V. Lounasmaa. Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497, 1993. doi: 10.1103/RevModPhys.65.413.

Sukjun Hwang, Brandon Wang, and Albert Gu. Dynamic chunking for end-to-end hierarchical sequence modeling. *arXiv preprint arXiv:2507.07955*, 2025.

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(1):160035, 2016. doi: 10.1038/sdata.2016.35.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A. Kim. MC-MED, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12(1):1094, 2025. doi: 10.1038/s41597-025-05419-5.

Saurabh Kataria, Yi Wu, Zhaoliang Chen, Hyunjung Gloria Kwak, Yuhao Xu, Lovely Yeswanth Panchumarthi, Ran Xiao, Jiaying Lu, Ayca Ermis, Anni Zhao, Runze Yan, Alex Federov, Zewen Liu, Xu Wu, Wei Jin, Carl Yang, Jocelyn Grunwell, Stephanie R. Brown, Amit Shah, Craig Jabaley, Tim Buchman, Sivasubramanium V. Bhavani, Randall J. Lee, and Xiao Hu. Generalist vs specialist time series foundation models: Investigating potential emergent behaviors in assessing human health using PPG signals. *arXiv preprint arXiv:2510.14254*, 2025.

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

Philip Mehrgardt, Matloob Khushi, Simon Poon, and Anusha Withana. Pulse transit time PPG dataset (version 1.1.0). PhysioNet, 2022. doi: 10.13026/jpan-6n92.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III waveform database matched subset (version 1.0). PhysioNet, 2020. doi: 10.13026/c2294b.

Guangkun Nie, Xiaocheng Fang, Gongzheng Tang, Yujie Xiao, Jun Li, Bo Liu, Hongyan Li, and Shenda Hong. AnyPPG: An ECG-guided PPG foundation model trained on over 100,000 hours of recordings for holistic health profiling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 11717–11727, 2026. doi: 10.1145/3770855.3818911.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2211.14730.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.20542.

Yulia Rubanova, Ricky T.Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1907.03907.

Mithun Saha, Maxwell A. Xu, Wanting Mao, Sameer Neupane, James M. Rehg, and Santosh Kumar. Pulse-PPG: An open-source field-trained PPG foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–35, 2025. doi: 10.1145/3749494. Article 126.

Muntaha Samad, Mirana Angel, Joseph Rinehart, Yuzo Kanomata, Pierre Baldi, and Maxime Cannesson. Medical informatics operating room vitals and events repository (MOVER): A public-access operating room database. *JAMIA Open*, 6(4):ooad084, 2023. doi: 10.1093/jamiaopen/ooad084.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *ACM International Conference on Multimodal Interaction (ICMI)*, pp. 400–408, 2018. doi: 10.1145/3242969.3242985.

Hamza Shafiq, Hung Manh Pham, Bin Zhu, Pan Zhou, Jun Hu, and Aaqib Saeed. CardioState-JEPA: Delay-aware cross-modal learning of a shared cardiac representation. *arXiv preprint arXiv:2608.12944*, 2026.

Ali I. Siam, Atef Abou Elazm, Nirmeen A. El-Bahnasawy, Ghada M. El Banby, and Fathi E. Abd El-Samie. PPG-based human identification using mel-frequency cepstral coefficients and neural networks. *Multimedia Tools and Applications*, 80(17):26001–26019, 2021. doi: 10.1007/s11042-021-10781-8.

Ali I. Siam, Fathi Abd El-Samie, Atef Abu Elazm, Nirmeen El-Bahnasawy, and Ghada Elbanby. Real-world PPG dataset (version 3). Mendeley Data, 2022. doi: 10.17632/yynb8t9x3d.3.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996. doi: 10.1161/01.CIR.93.5.1043.

Jessica Torres-Soto and Euan A. Ashley. Multi-task deep learning for cardiac rhythm detection in wearable devices. *npj Digital Medicine*, 3(1):116, 2020. doi: 10.1038/s41746-020-00320-4.

## A Extended discussion

**Additional controls.** Two experiments would sharpen the claims of §5 and are not reported here. A fixed window read against the carried recurrent state at matched context would indicate how much of §5.2 reflects the state rather than the amount of signal admitted. And a reconstruction probe, a decoder trained on each frozen embedding, would give a complementary assessment of how much waveform information the representation retains, which a task probe alone cannot separate from what its head can extract.

The two levels of the hierarchy may also differ in what they need from the data. Landmark supervision for the fine level is produced offline by a detector, while the coarse level sees only the patients, devices and clinical states the corpus contains. The context-length measurement (§5.2) and the external comparison (§5.1) are consistent with this reading and neither isolates it: the context sweep varies history rather than data, and the external comparison varies pretraining corpora that differ in more than one respect. It is a reading of the results, not a claim they test.

Hierarchical organization may also allow different levels to receive different forms of supervision: waveform landmarks constrain local segmentation, while clinical measurements supervise representations over longer timescales. Whether the levels can be adapted independently is not tested here; the clinical stage updates the whole encoder.

The objective a waveform model optimizes is defined on the trace, while the quantities the representation is asked about are properties of a process the trace only measures. If that gap is why loss and quality can move apart, the effect would not be specific to chunking, though our evidence covers one comparison on one modality.

Four limitations bound these claims. First, the fine scales depend on reliable fiducial detection; where landmarks are less established, or where detectors fail systematically, the same construction would need another source of boundaries. Second, pretraining data come from a single-vendor acquisition setting, so device diversity is the axis this work argues for and does not yet have. Third, the coarse supervision assumes the availability of time-aligned clinical measurements, which exist in monitored care but not in every setting we would like to serve. Fourth, this work covers PPG; extending the same principles to ECG and to multimodal physiological streams is future work.

## B Positioning and extended related work

Approaches to physiological representation learning differ in the supervision they use to bridge the gap between a waveform measurement and the physiological state it reflects (Hämäläinen et al., 1993; Gulrajani, 1998; Akselrod et al., 1981; Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996; Rubanova et al., 2019). General time-series foundation models (Ansari et al., 2024; Das et al., 2024) do not face that gap, since the observed series is itself the forecasting target, which is why a waveform model cannot simply inherit their recipe. The families below are grouped by the supervision they draw on.

**Waveform foundation models: tokenization.** Released PPG encoders (PaPaGei (Pillai et al., 2025), Pulse-PPG (Saha et al., 2025), AnyPPG (Nie et al., 2026), SIGMA-PPG (Guo et al., 2026), and generative models such as GPT-PPG (Chen et al., 2025)) and single-lead ECG encoders (ECG-Founder (Li et al., 2025), HeartGPT (Davies et al., 2024)) are pretrained on fixed-length windows on the order of seconds to minutes, either fed to a convolutional encoder or split into fixed patches. Either way the grid is fixed with respect to the signal and slices through the cardiac cycle at arbitrary

phase, so the semantic unit, the beat, is split across the units the model encodes, and downstream consumers inherit window-level embeddings rather than physiological objects. Windows are also encoded independently: none of these models carries state from one window to the next, so context does not accumulate across a recording even where the encoder itself accepts inputs of other lengths, and inference remains a batch operation over windows. This is a structural mismatch with monitoring, where the value of the signal accumulates over hours. Our encoder replaces the grid with dynamic chunking at physiological units and carries recurrent state across an unbounded stream (§3.1).

**Waveform foundation models: objectives.** The dominant pretraining objectives are contrastive: views of the same participant or window are pulled together under augmentation- and subject-level invariances. Such invariances can be undesirable for monitoring when they suppress changes within a recording that downstream tasks need to preserve. Generative and masked alternatives model the surface signal, whose energy is dominated by baseline wander and amplitude nuisance rather than by the latent state (§2). Neither family measures retention directly: these objectives encourage the representation to keep what their own training target needs, which does not guarantee that the information a downstream clinical task needs is preserved, yet pretraining loss is commonly used for optimization, checkpoint selection, and scaling comparisons. A recent third family applies joint-embedding prediction: CardioState-JEPA (Shafiq et al., 2026) predicts masked cardiac states across ECG, PPG, and PCG in a shared token space, using a learned delay aligner to place all modalities on a common cardiac clock. We share the latent-prediction principle (our next-beat objective is likewise a joint-embedding prediction) but apply it on a semantically chunked, streaming hierarchy rather than a masked fixed-token grid; and we note that the electro-mechanical delay such alignment removes is itself a hemodynamic biomarker (pulse-transit timing), so a shared cardiac clock is not an information-free coordinate change. We evaluate waveform reconstruction through the chunk bottleneck and latent next-beat prediction within the same hierarchy; the selected configuration retains reconstruction and collapse regularization and omits next-beat prediction (§5.4).

**Dynamic chunking.** H-Net (Hwang et al., 2025) showed that tokenization can be learned inside the model, with cosine-similarity routing discovering byte-level chunks end-to-end for language modeling. The difference that matters here is what each modality already provides. Text has no comparable event detector, so H-Net's routers are trained without boundary supervision; a physiological waveform comes with reliable detectors at its fine scales, so our routers receive boundary supervision from automatically detected landmarks and cardiac-cycle structure, and the unsupervised router is kept as a rate-matched comparison arm (§5.3). Sections 3.1 to 3.4 describe how the operations are adapted for a causal physiological stream.

**Latent-state views of biosignals.** That biosignals are observations of a hidden dynamical system is an old and well-supported position: the inverse-problem formulation of MEG/EEG (Hämäläinen et al., 1993) and of body-surface ECG (Gulrajani, 1998), heart-rate variability as a quantitative probe of autonomic control (Akselrod et al., 1981; Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996), and latent-ODE models for irregular clinical series (Rubanova et al., 2019). General time-series FMs (Ansari et al., 2024; Das et al., 2024) occupy a different informational role: their input series is the trusted coordinate system whose own future is the target, whereas a waveform is a proxy whose targets are measured by *other* instruments. Our contribution is to build the training recipe (units, objectives, labels, and deployment shape) around these properties rather than to restate them.

## C Waveforms and the pretraining recipe: the full argument

The main text states the premise in a paragraph; this appendix gives the argument behind it. It has four parts: what the other foundation-model domains hand the model without being asked, what a physiological waveform is as an object of inference, the design principles that follow, and why existing waveform models succeed anyway.

**What the other domains provide.** Language, vision and protein modelling share three properties that a physiological waveform does not. The unit is given: a token, a pixel patch, an amino acid. The freely available self-supervised target is informative: the next word carries information, the next residue constrains structure. And the signal is itself the object of interest: text is the medium of thought, an image is what the world looks like, a sequence is what folds. Table 4 sets the properties

side by side. A PPG or ECG trace has none of the three. Its nearest candidate unit, the beat, is ill-defined under arrhythmia and artifact and has to be detected; its next sample is closely determined by the last few, so next-sample prediction is near interpolation, while next-*beat* prediction presupposes the detector; and its information is concentrated at a handful of landmarks per cycle, spread over timescales from milliseconds to days, and modulated by a device whose response is neither known nor constant.

Table 4: Qualitative comparison of language, vision, protein sequences and physiological waveforms as inputs for representation learning: the given representation unit, the freely available self-supervised target, the timescales a task spans and the dependence on acquisition. The entries locate the design problem and are not measurements.

| | Language | Vision | Protein | PPG / ECG |
|---|---|---|---|---|
| Given unit | word | none; objects emerge | residue | none; beat is inferred |
| Free label | next word | invented (mask) | next / masked residue | next sample highly predictable; next beat needs a detector |
| Information density | high | redundant | high | low; concentrated at landmarks |
| Timescales spanned | token–document | pixel–image | residue–fold | beat–days |
| Acquisition dependence | none | camera, lighting | none | sensor, contact, filter chain |
| Identity in signal | n/a | a feature | n/a | a feature, but of the projection |
| Object of interest | the text | the image | the fold | a physiological state measured indirectly |

**A waveform is an inverse problem.** The last row is the one that matters. A PPG or ECG trace is rarely what one wants to know. It is a noisy, low-bandwidth, device-distorted projection of the cardiovascular, autonomic and respiratory control loops (Guyton et al., 1972; Akselrod et al., 1981; Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996), modulated by posture, movement, sleep, drugs and disease. Heart rate and its variability are statistics of a latent rhythm generator; an arrhythmia is a state of the conduction system; blood pressure is haemodynamic state projected through PPG morphology. In each case the reference measurement comes from a separate instrument, an arterial line, a polysomnograph, an echocardiogram, and the model is asked to invert a poorly characterised, time-varying observation operator, the structure the biomagnetic and electrocardiographic inverse problems also have (Hämäläinen et al., 1993; Gulrajani, 1998). The other domains are not posed this way: language predicts text from text, vision labels what is in the image, and protein models map a sequence to the structure it causes.

Three consequences follow, and they are the reasons for the design rather than results of it. Modelling the next sample of the surface signal models the projection rather than the source, so a loss on the surface need not track the quality of the state estimate; the differing reconstruction and downstream rankings in §5.4 are consistent with this, but do not establish the mechanism. Evaluation should therefore reference a separate measurement of the latent state rather than a reconstruction or prediction score on the waveform itself. And subject identity is carried by the projection as much as by the source, so the identity cohort in Table 3 reads how much individuating detail survives the chunk bottleneck rather than how well the latent state is estimated.

**Design principles.** These considerations motivate a hierarchical architecture whose unit boundaries are placed where the physiology places them, which learns surface structure and longer-timescale state at different levels, and which is trained over spans long enough to contain the state it is meant to recover. Where external supervision is used, routine clinical measurements reference the latent state more directly than surface labels do. Physio-HNet instantiates these choices: landmark-placed units (§3.2), a beat-level state carried across the recording (§3.1), recall across a span as the pretraining objective (§3.3), and clinically anchored continued pretraining with the self-supervised terms retained (§3.4). Whether each choice is necessary is not established here; the experiments compare alternatives for boundary placement and for the objective, and do not vary the remaining choices.

**Relation to existing waveform models.** The recipe imported from language and vision (Chen et al., 2025; Saha et al., 2025; Nie et al., 2026; Guo et al., 2026) produces useful models, for reasons that should be stated. The projection is systematic rather than random: electrical conduction and optical absorption are physics, so surface morphology correlates with latent state and a model of the surface acquires some state as a side effect. Many clinically useful tasks are surface-level tasks, beat detection and rhythm classification among them, and for these a surface objective is matched to the target. Scale compensates in part for architectural mismatch, and covers rare morphologies. And transfer holds while the observation operator is approximately constant, which standardised acquisition often provides inside one setting. The limits we expect are where these stop: tasks that require inverting the projection, settings where the operator changes, and timescales longer than a window. Those are what §§5.1–5.2 are chosen to probe.

## D Components of Physio-HNet

This section specifies the components of Physio-HNet and their behavior during waveform pretraining and clinically anchored continued pretraining. Controlled boundary variants are described in App. E; architecture settings and optimization details are provided in App. H.

**Input and temporal encoding.** PPG sampled at 40 Hz is clipped and normalized using fixed dataset-level statistics:

$$x'_n = \frac{\mathrm{clip}(x_n, 0, x_{\max}) - \mu}{\sigma}, \tag{5}$$

where $x_{\max}$ is the clipping bound specified by the preprocessing cache. The statistics are shared across processing windows rather than re-estimated within each window, preserving a consistent input scale when recurrent states are carried forward. Every two non-overlapping samples are linearly projected into one token, giving a token rate of 20 Hz.

Time features are added on the encoder path before the stack at each resolution. They encode temporal position and the interval since the preceding embedding, accounting for the irregular grids produced by chunking. Their projection is zero-initialized, so these features initially contribute zero to the encoder inputs. The expansion path uses within-chunk position features, defined below, without adding the encoder-path time embeddings.

**Causal encoding and state propagation.** Each encoder stack uses causal Mamba-2 blocks modified to accept an initial recurrent state and return the updated state during multi-token forward passes. Writing $\mathbf{u}_t^{(\ell)}$ for the input at level $\ell$, the encoder produces

$$\mathbf{h}_t^{(\ell)} = \mathcal{E}_\ell\big(\mathbf{u}_{\le t}^{(\ell)}\big), \tag{6}$$

with history from preceding windows supplied through the carried state. K0 processes tokens, K1 processes sub-beat units, and the beat backbone processes the resulting beat-level sequence to produce $\{\mathbf{b}_i\}$.

States are carried between consecutive windows in the same encounter stream and detached at training-window boundaries. A new stream starts with reset states. Consequently, the backpropagation span is bounded by the processing window, whereas the forward state can incorporate preceding windows. The stream construction is specified in App. H.

**Pairwise boundary router.** For either compression level, the router scores adjacent encoder states:

$$p_t = \mathrm{clamp}\big(\tfrac{1}{2}\left[1 - \cos(W_q\mathbf{h}_{t-1},\, W_k\mathbf{h}_t)\right], 0, 1\big), \qquad b_t = \mathbb{1}_{\{p_t > 1/2\}}, \tag{7}$$

where $W_q$ and $W_k$ are learned linear projections initialized to the identity, and $p_1 = 1$ by convention. Beyond these projections, the router has no separate scorer or additional temporal receptive field. Boundary losses therefore supervise the encoder representations as well as the router projections. The reported configuration caps K0 boundary counts at 3000 per sequence, compared with approximately 1100 boundaries in a typical five-minute window.

**Chunk-end selection and window boundaries.** Let $t_1 < \cdots < t_K$ denote chunk endpoints and set $t_0 = 0$. Chunk $k$ contains positions $t_{k-1} + 1, \ldots, t_k$ and passes one state to the next level:

$$z_k = \mathbf{h}_{t_k}. \tag{8}$$

This differs from the start-of-chunk convention of Hwang et al. (2025). The selected state has processed all inputs in the chunk and may also contain earlier history. Its timestamp is the time at which the chunk closes; no pooling is applied.

At both K0 and K1, the final valid position of a processing window is included as an endpoint, closing any trailing chunk. Thus, a cardiac cycle crossing a window boundary can produce separate units on either side. Padding beyond the valid prefix is masked. Chunk boundaries are local to each processing window, while recurrent states can cross the window boundary.

**Sub-beat boundary supervision.** An offline detector supplies pulse feet, systolic peaks, and dicrotic notches as K0 targets. Let $\mathcal{P}$ denote detected landmark positions and $\mathcal{N}(i)$ the valid non-landmark positions within $\pm 0.5$ s of landmark $i$. A ranking loss with margin $m = 0.15$ encourages each landmark score to exceed nearby non-landmark scores:

$$\mathcal{L}_{\text{rank}} = \left\langle \left\langle \max\left(0, m - (p_i - p_j)\right) \right\rangle_{j \in \mathcal{N}(i)} \right\rangle_{i \in \mathcal{P}}. \tag{9}$$

The loss first averages over valid negatives for each landmark, then over landmarks, giving landmarks equal weight regardless of their valid neighborhood sizes.

A separate threshold loss calibrates scores around the operating point:

$$\mathcal{L}_{\text{band}} = \left\langle \max(0, \tfrac{1}{2} + m - p_i) \right\rangle_{i \in \mathcal{P}} + \left\langle \max(0, p_j - \tfrac{1}{2} + m) \right\rangle_{j \in \mathcal{Q}}, \tag{10}$$

where $\mathcal{Q}$ contains valid negative positions. Positive and negative terms are averaged separately. Missing detections are masked rather than imputed; detector and mask construction are described in App. I.

A rate term controls boundary counts. Let $F$ be the fraction of valid positions selected as boundaries, $G$ the mean boundary probability, and $\tau$ the target boundary fraction derived from the window's detected landmarks. With $N = 1/\tau$,

$$\mathcal{L}_{\text{rate}} = \frac{N}{N-1}\left[(1-F)(1-G) + FG(N-1)\right], \tag{11}$$

averaged across the batch.

**Beat-level boundary supervision.** At K1, detected cardiac cycles provide grouping constraints without prescribing which landmark must terminate a beat. Let $\gamma(k)$ identify the cardiac cycle containing sub-beat unit $k$. Its within-cycle offset is assigned to one of four bins, $o(k) \in \{0, 1, 2, 3\}$, with offsets beyond the fourth bin merged into the last bin. The objectives encourage one boundary per cycle and a consistent offset:

$$\mathcal{L}_{\text{one}} = \left\langle \left( \sum_{k:\gamma(k)=c} p_k - 1 \right)^2 \right\rangle_c, \qquad \mathcal{L}_{\text{cons}} = 1 - \sum_{o=0}^{3} q_o^2, \tag{12}$$

where

$$q_o = \frac{\sum_{k:o(k)=o} p_k}{\sum_k p_k}. \tag{13}$$

A threshold term uses each cycle's highest-scoring position as its positive, and a rate term controls the boundary count. No objective enforces equal durations between successive boundaries, allowing beat-to-beat timing variability.

**Training-only expansion.** The expansion path maps beat codes back to sub-beat positions and then to tokens. At either level, let $k(t)$ identify the chunk containing position $t$, $f_t$ and $g_t$ be its

distances from the chunk start and end, and $r_t \in [0, 1]$ its relative position. Expansion broadcasts the chunk code and adds within-chunk position features:

$$v_t = z_{k(t)} + \mathrm{MLP}\left([r_t, \log(1 + f_t), \log(1 + g_t)]\right), \tag{14}$$

$$\tilde{v}_t = v_t + \mathrm{DWConv}^{5}_{\mathrm{centered}}(v)_t, \tag{15}$$

$$(a_t, c_t) = W_{\mathrm{in}}\tilde{v}_t, \qquad \hat{v}_t = W_{\mathrm{out}}\big(a_t \odot \sigma(c_t)\big). \tag{16}$$

An anti-causal Mamba-2 decoder then processes the expanded sequence:

$$\mathrm{Dec}_{\mathrm{rev}}(\hat{v}) = \mathrm{flip}\left(\mathrm{Dec}(\mathrm{flip}(\hat{v}))\right). \tag{17}$$

Reversal is restricted to each row's valid prefix. The two decoder stacks have independent weights; between expansion levels, width reduction takes the leading coordinates rather than applying a learned projection.

The centered convolution can mix neighboring chunk codes, and reverse-time decoding can use later codes to reconstruct an earlier position. Neither operation supplies future inputs to the causal encoder. No encoder skip connection bypasses the compression bottlenecks. The entire expansion path is discarded at inference; the rationale for anti-causal decoding is discussed in App. H.

**Waveform reconstruction objectives.** A shared linear head predicts the raw 0.5 s waveform window $w_t \in \mathbb{R}^{20}$ ending at token $t$. Unlike the non-overlapping input tokens, these reconstruction windows overlap. With $m_t$ indicating valid reconstruction positions,

$$\mathcal{L}_{\text{recall}} = \frac{\sum_t m_t \|\hat{w}_t - w_t\|_2^2}{20 \sum_t m_t}, \qquad \mathcal{L}_{\text{corr}} = \frac{\sum_t m_t [1 - \rho(\hat{w}_t, w_t)]}{\sum_t m_t}. \tag{18}$$

Pearson correlation is computed separately within each 20-sample window before masked averaging. It measures morphological agreement independently of the window's positive amplitude scale and offset.

**Latent prediction and collapse regularization.** The optional next-beat objective predicts the next backbone embedding from the current one:

$$\mathcal{L}_{\text{nbp}} = \left\langle 1 - \cos\left(\psi(\mathbf{b}_i),\, \mathrm{sg}[\mathbf{b}_{i+1}]\right)\right\rangle_i, \tag{19}$$

over valid adjacent beat pairs. The predictor $\psi$ is a two-layer network with a GELU nonlinearity, and sg denotes stop-gradient.

SIGReg (Balestriero & LeCun, 2025) separately regularizes the beat embeddings against collapse. Embeddings are centered and projected onto random directions, and the projected distributions are compared with a standard normal using an Epps–Pulley statistic. This regularizer remains active when next-beat prediction is omitted.

**Combined pretraining objective.** For the main Physio-HNet configuration,

$$\mathcal{L}_{\mathrm{SSL}} = \mathcal{L}_{\text{recall}} + 0.3\,\mathcal{L}_{\text{corr}} + \mathcal{L}_{\text{SIGReg}} + \mathcal{L}_{\mathrm{boundary}}, \tag{20}$$

where

$$\mathcal{L}_{\mathrm{boundary}} = \mathcal{L}^{\mathrm{K0}}_{\text{rank}} + 0.5\,\mathcal{L}^{\mathrm{K0}}_{\text{band}} + \mathcal{L}^{\mathrm{K0}}_{\text{rate}} \tag{21}$$

$$+ \mathcal{L}^{\mathrm{K1}}_{\text{one}} + \mathcal{L}^{\mathrm{K1}}_{\text{cons}} + 0.5\,\mathcal{L}^{\mathrm{K1}}_{\text{band}} + \mathcal{L}^{\mathrm{K1}}_{\text{rate}}. \tag{22}$$

The next-beat weight is zero in the main configuration. Objective ablations introduce $\mathcal{L}_{\text{nbp}}$ with weight one and remove the corresponding reconstruction terms when evaluating prediction alone (§5.4).

**Gradient paths across training stages.** During waveform pretraining, chunk selection is discrete and the expansion operators consume boundary indices and chunk lengths without a differentiable routing path. Representation losses therefore do not directly update $W_q$ and $W_k$. Boundary losses update both these projections and their input encoder states; representation losses also update the encoder through selected chunk states.

During clinically anchored continued pretraining, the entire encoder, routers, expansion path, and added readout level remain trainable. At both K0 and K1, each selected chunk state is multiplied by its boundary probability through a straight-through estimator:

$$z_k \leftarrow z_k \, \mathrm{ste}(p_{t_k}), \qquad \mathrm{ste}(p) = 1 + p - \mathrm{sg}[p]. \tag{23}$$

The estimator has forward value one and backward derivative $\partial \, \mathrm{ste}(p)/\partial p = 1$. Thus, this operation leaves the forward chunk codes unchanged while allowing the clinical loss to update $W_q$ and $W_k$ through the selected boundary probabilities. Each probability is obtained from the same chunk selection that produces its corresponding state, ensuring their alignment.

The retained self-supervised objectives regularize adaptation:

$$\mathcal{L}_{\mathrm{cont}} = \mathcal{L}_{\text{clinical}} + 0.1 \, \mathcal{L}_{\mathrm{SSL}}. \tag{24}$$

Clinical readout construction and target normalization are described in App. J.

# E Controlled boundary variants

The four arms share the waveform corpus, representation backbone, chunk-end convention, and optimization budget. They differ in boundary mechanisms and supervision, which Table 5 summarizes. Objective ablations are conducted within each arm. The configurations carried into clinical continued pretraining use waveform recall without next-beat prediction for A1, A2, and B, and both objectives for C. Consequently, comparisons among all four clinically adapted arms do not isolate boundary placement alone.

**Convolutional router (A1).** A1 uses a convolutional scorer with kernel width 15, corresponding to a receptive field of $\pm 7$ tokens,

$$p_t = \sigma\Big(\mathrm{Conv}^1\big(\mathrm{GELU}(\mathrm{Conv}^{15}([\mathbf{h};\, \Delta\mathbf{h};\, \Delta^2\mathbf{h}]))\big)\Big)_t, \qquad b_t = \mathbb{1}_{\{p_t > 1/2\}}, \tag{25}$$

where $\sigma$ is the logistic function. Both K0 and K1 receive per-position binary cross-entropy supervision from detected physiological boundaries. Neither level uses rate regularization. The centered scorer makes A1's boundary placement non-causal; A1 is an experimental control rather than the proposed deployment configuration.

**Physiologically supervised pairwise router (A2).** A2 is the configuration proposed in §3.2. At K0, ranking, threshold, and rate losses have weights 1.0, 0.5, and 1.0. Ranking uses margin 0.15 and negatives within $\pm 10$ tokens. At K1, the one-per-cycle, consistency, threshold, and rate losses have weights 1.0, 1.0, 0.5, and 1.0, respectively. Full definitions appear in App. D.

**Undirected pairwise router (B).** B uses the same pairwise cosine router as A2 but receives no boundary-location supervision. A symmetric hinge margin depends on $|p_t - 1/2|$, with margin 0.15 and weight 16.0. It encourages confident decisions without specifying where boundaries should occur. Rate regularization is retained at both levels. A2 and B both cap K0 boundary counts at 3000 per sequence.

**Fixed-stride chunking (C).** C replaces learned boundaries with deterministic strides derived from each window's reference landmark and beat rates,

$$b_t = \mathbb{1}_{\{t \equiv 0 \pmod{S}\}}, \qquad S = \mathrm{round}(1/\tau), \tag{26}$$

where $\tau$ is the window's landmark rate and the same construction is applied at the beat level with $S'$ from its beat rate. Positions are selected at regular intervals, without using landmark locations. Neither level uses boundary losses. A router is instantiated for architectural consistency but is not consulted by chunk selection and receives no gradient.

## E.1 Transfer of the four arms on the external cohorts

Table 1 reads the deployed arm against the released encoders. Table 6 places all four arms on that same benchmark, each in its frozen self-supervised form and in its clinically anchored form, so that

Table 5: Router mechanisms and boundary-training objectives of the four configurations at the sub-beat (K0) and beat (K1) levels. BCE is binary cross-entropy; the individual loss terms are defined in App. D and App. E.

| Arm | Router | K0 objectives | K1 objectives |
|---|---|---|---|
| A1 | Convolutional | BCE | BCE |
| A2 | Pairwise cosine | Rank, band, rate | One, consistency, band, rate |
| B | Pairwise cosine | Symmetric margin, rate | Symmetric margin, rate |
| C | Fixed stride | None | None |

the boundary comparison of §5.3, which is read on MIMIC-III, can also be read on the two cohorts that supplied no pretraining data. The A2 rows repeat the Physio-HNet rows of Table 1; the other arms are encoded and probed under the same anchors, seeds, and probe-training budget (App. L). The caveats of this section carry over: the arms differ by more than boundary placement, so the A2–B contrast is the one that isolates boundary-location supervision.

The clinical stage improves pressure prediction under every boundary mechanism. Over the four arms, the four cohort-probe blocks and the three pressures, MAE falls in all 48 settings and $R^2$ rises in 47, the exception being B on the emergency-department static probe, where it is level at $.242$. Improvements after clinical continued pretraining are therefore observed under all four boundary configurations.

Anchored A2 holds a consistent pressure advantage over B and C. Against B it takes the higher $R^2$ in eleven of the twelve pressure settings and is level in the twelfth, and the lower MAE in all twelve; against C it takes both in all twelve.

A1 achieves higher $R^2$ and lower MAE than A2 in five of the twelve settings, and the five are not scattered: all three pressures under the operating-room static probe, and the diastolic pressure under both emergency-department probes. The non-causal scorer is competitive on this benchmark, and what A2 offers against it is a causal boundary rule, not a uniformly better arm.

Table 6: Frozen-encoder transfer of the four boundary configurations on MC-MED and MOVER, before (*SSL*) and after (*anchored*) clinical continued pretraining. Each arm uses the objective configuration selected in §5.4. Probes, anchors, protocol and MAE units are those of Table 1 (App. L); values are the mean and standard deviation over five probe-training seeds. Blue marks the best value within a cohort-probe block, shading identifies A2, and grey marks the columns excluded from encoder comparisons (§5.1). The A2 rows also appear in Table 1.

| | HR | | $SpO_2$ | | RR | | NBP-S | | NBP-D | | NBP-M | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE |
| *Static probe*, emergency department (MC-MED) | | | | | | | | | | | | |
| A1 (SSL) | .900±.007 | **2.91±0.19** | .001±.034 | 1.53±0.07 | .043±.092 | 3.45±0.14 | .230±.031 | 15.54±0.22 | .173±.019 | 10.74±0.09 | .178±.032 | 10.97±0.12 |
| A1 (anchored) | .897±.008 | 2.94±0.23 | .010±.087 | 1.51±0.06 | .172±.086 | 3.17±0.16 | .249±.030 | 15.35±0.25 | **.205±.013** | **10.47±0.10** | .201±.047 | 10.82±0.22 |
| A2 (SSL) | .900±.006 | **2.91±0.32** | .006±.052 | 1.49±0.07 | .066±.081 | 3.40±0.13 | .241±.027 | 15.16±0.33 | .173±.017 | 10.74±0.13 | .179±.014 | 11.04±0.16 |
| A2 (anchored) | .898±.007 | 3.07±0.29 | .022±.042 | 1.50±0.09 | **.204±.053** | **3.09±0.12** | **.269±.022** | **15.07±0.22** | .195±.023 | 10.57±0.17 | **.219±.024** | **10.67±0.08** |
| B (SSL) | .895±.008 | 3.07±0.32 | −.026±.044 | 1.54±0.07 | .065±.082 | 3.38±0.14 | .242±.029 | 15.46±0.33 | .157±.019 | 10.91±0.13 | .172±.029 | 11.10±0.16 |
| B (anchored) | **.901±.010** | 2.98±0.31 | .001±.050 | 1.52±0.06 | .137±.070 | 3.27±0.14 | .242±.026 | 15.37±0.25 | .175±.012 | 10.70±0.15 | .196±.027 | 10.86±0.17 |
| C (SSL) | .882±.006 | 3.05±0.19 | .022±.044 | 1.51±0.06 | .084±.038 | 3.38±0.04 | .064±.012 | 17.15±0.22 | .043±.018 | 11.70±0.11 | .019±.015 | 12.08±0.07 |
| C (anchored) | .897±.007 | 2.99±0.31 | −.009±.041 | 1.51±0.06 | .036±.092 | 3.46±0.14 | .201±.021 | 15.89±0.19 | .156±.022 | 10.93±0.23 | .176±.030 | 11.07±0.24 |
| *Static probe*, operating room (MOVER) | | | | | | | | | | | | |
| A1 (SSL) | .673±.015 | 6.16±0.22 | −.062±.033 | 1.29±0.08 | .001±.041 | 3.03±0.10 | .074±.123 | 16.64±1.07 | .123±.070 | 8.45±0.18 | .106±.101 | 10.74±0.50 |
| A1 (anchored) | .692±.004 | 5.75±0.08 | −.069±.040 | 1.23±0.05 | .027±.013 | 2.96±0.03 | **.159±.010** | **15.73±0.22** | **.172±.017** | **8.19±0.03** | **.172±.011** | **10.25±0.15** |
| A2 (SSL) | .687±.009 | 6.04±0.24 | −.059±.033 | 1.29±0.23 | .004±.037 | 3.03±0.09 | .089±.082 | 16.51±0.60 | .140±.042 | 8.42±0.16 | .123±.060 | 10.64±0.45 |
| A2 (anchored) | **.696±.008** | **5.71±0.12** | −.059±.029 | 1.23±0.06 | .045±.021 | 2.92±0.05 | .149±.012 | 15.80±0.33 | .165±.010 | 8.22±0.11 | .166±.008 | 10.28±0.24 |
| B (SSL) | .683±.009 | 6.05±0.20 | −.095±.074 | 1.35±0.23 | .004±.031 | 3.03±0.09 | .076±.080 | 16.63±0.60 | .131±.044 | 8.47±0.16 | .101±.086 | 10.79±0.45 |
| B (anchored) | .692±.007 | 5.79±0.10 | −.066±.038 | 1.25±0.05 | .038±.018 | 2.93±0.03 | .132±.014 | 15.98±0.26 | .165±.018 | 8.25±0.09 | .159±.012 | 10.35±0.16 |
| C (SSL) | .653±.007 | 6.07±0.08 | −.080±.025 | 1.26±0.04 | .015±.016 | 3.01±0.05 | .032±.018 | 16.82±0.03 | .024±.029 | 8.97±0.02 | .037±.021 | 11.12±0.03 |
| C (anchored) | .647±.021 | 6.35±0.43 | −.047±.020 | 1.29±0.11 | −.003±.045 | 3.05±0.10 | .071±.066 | 16.64±0.48 | .093±.047 | 8.63±0.17 | .080±.066 | 10.88±0.31 |
| *Recurrent probe*, emergency department (MC-MED) | | | | | | | | | | | | |
| A1 (SSL) | .919±.004 | 2.38±0.08 | .062±.015 | 1.42±0.01 | .093±.035 | 3.34±0.04 | .243±.028 | 15.23±0.25 | .192±.006 | 10.63±0.11 | .196±.012 | 10.84±0.10 |
| A1 (anchored) | **.924±.003** | **2.13±0.08** | .064±.017 | 1.42±0.02 | **.247±.035** | **3.01±0.06** | .281±.023 | 15.01±0.16 | **.224±.014** | **10.42±0.14** | .232±.020 | 10.67±0.12 |
| A2 (SSL) | .916±.003 | 2.38±0.15 | .061±.015 | 1.43±0.01 | .119±.030 | 3.29±0.04 | .269±.014 | 15.21±0.20 | .180±.025 | 10.75±0.26 | .203±.020 | 10.90±0.24 |
| A2 (anchored) | .920±.004 | 2.27±0.07 | .068±.020 | 1.41±0.01 | .217±.034 | 3.10±0.06 | **.303±.015** | **14.74±0.16** | .220±.006 | 10.44±0.12 | **.242±.008** | **10.59±0.12** |
| B (SSL) | .911±.006 | 2.37±0.12 | .049±.017 | 1.43±0.02 | .090±.024 | 3.35±0.03 | .257±.013 | 15.22±0.17 | .184±.019 | 10.69±0.17 | .208±.015 | 10.84±0.14 |
| B (anchored) | .923±.002 | 2.24±0.09 | .051±.024 | 1.43±0.02 | .209±.029 | 3.11±0.05 | .299±.016 | 14.87±0.18 | .203±.010 | 10.60±0.14 | .236±.009 | 10.67±0.13 |
| C (SSL) | .919±.002 | 2.31±0.26 | .044±.019 | 1.47±0.00 | .105±.019 | 3.33±0.02 | .078±.021 | 17.11±0.22 | .051±.013 | 11.69±0.11 | .034±.011 | 12.06±0.11 |
| C (anchored) | .912±.005 | 2.25±0.10 | .018±.020 | 1.46±0.02 | .091±.022 | 3.35±0.02 | .212±.008 | 15.68±0.08 | .162±.017 | 10.89±0.22 | .179±.019 | 11.04±0.18 |
| *Recurrent probe*, operating room (MOVER) | | | | | | | | | | | | |
| A1 (SSL) | .658±.006 | 5.84±0.08 | −.057±.037 | 1.20±0.03 | .012±.021 | 3.01±0.03 | .141±.016 | 15.89±0.09 | .160±.019 | 8.27±0.06 | .158±.016 | 10.36±0.07 |
| A1 (anchored) | .673±.006 | 5.66±0.09 | −.056±.031 | 1.19±0.02 | .036±.022 | 2.94±0.04 | .157±.012 | 15.75±0.06 | .178±.016 | 8.18±0.05 | .174±.014 | 10.26±0.06 |
| A2 (SSL) | .671±.003 | 5.70±0.08 | −.055±.027 | 1.20±0.02 | .016±.023 | 3.01±0.04 | .138±.013 | 15.94±0.08 | .166±.018 | 8.24±0.03 | .162±.018 | 10.36±0.05 |
| A2 (anchored) | .671±.004 | 5.72±0.09 | −.059±.026 | 1.20±0.01 | .059±.006 | 2.91±0.02 | **.158±.012** | **15.74±0.10** | **.184±.016** | **8.16±0.06** | **.181±.015** | **10.23±0.08** |
| B (SSL) | .666±.005 | 5.84±0.12 | −.059±.038 | 1.20±0.03 | .005±.024 | 3.02±0.04 | .143±.015 | 15.88±0.11 | .163±.017 | 8.25±0.05 | .165±.014 | 10.32±0.06 |
| B (anchored) | .666±.006 | 5.76±0.09 | −.057±.021 | 1.20±0.01 | .036±.010 | 2.94±0.02 | .153±.018 | 15.78±0.11 | .176±.020 | 8.18±0.05 | .173±.021 | 10.27±0.08 |
| C (SSL) | .669±.009 | 5.68±0.08 | −.055±.032 | 1.22±0.02 | .020±.023 | 3.00±0.04 | .064±.029 | 16.72±0.10 | .058±.016 | 8.91±0.05 | .070±.024 | 11.08±0.06 |
| C (anchored) | **.680±.007** | **5.62±0.06** | −.052±.033 | 1.20±0.01 | .013±.016 | 3.01±0.03 | .126±.021 | 16.05±0.09 | .155±.019 | 8.31±0.04 | .155±.021 | 10.42±0.07 |

# F RECONSTRUCTION AND GEOMETRY UNDER EACH OBJECTIVE

Table 7 reports reconstruction and embedding-geometry metrics for the checkpoints evaluated downstream in Table 3, so the two readings are taken on one set of checkpoints.

Table 7: Waveform reconstruction and beat-embedding geometry for the objective ablations of Table 3. MSE and Pearson $r$ measure reconstruction; *adj* is the adjacent-embedding cosine similarity and *dim* the effective dimensionality. A dash marks a row trained without waveform recall, which has no reconstruction. Blue marks, within each arm, the best reconstruction, the lowest adjacent similarity and the highest effective dimensionality.

| | Reconstruction | | Geometry | |
|---|---|---|---|---|
| | MSE | $r$ | adj | dim |
| *A1*, supervised placement, convolutional scorer | | | | |
| both | 0.0059 | .964 | .93 | 17.5 |
| recall only | **0.0039** | **.967** | .30 | 51.7 |
| next-beat only | – | – | .95 | 12.2 |
| *A2*, supervised placement, pairwise router | | | | |
| both | 0.0080 | .960 | .90 | 20.7 |
| recall only | **0.0053** | **.964** | .31 | 66.8 |
| next-beat only | – | – | .97 | 13.1 |
| *B*, unsupervised placement, same router as A2 | | | | |
| both | 0.0160 | .953 | .16 | 30.3 |
| recall only | **0.0039** | **.962** | .28 | 63.1 |
| next-beat only | – | – | .98 | 3.5 |
| *C*, fixed stride, no router | | | | |
| both | 0.0008 | .974 | .80 | 28.5 |
| recall only | **0.0002** | **.975** | .06 | 106.5 |
| next-beat only | – | – | .98 | 3.9 |

The reconstruction columns do not order the arms the way the downstream columns of Table 3 do. The best reconstruction, MSE $0.0002$ at $r$ $0.975$, belongs to the fixed-stride arm under recall alone, whose downstream mean is third-lowest of the twelve pretraining rows, while the best downstream mean comes with a reconstruction error more than twenty times larger. Selecting an objective by reconstruction and selecting it by downstream performance therefore give different answers on these runs.

The geometry columns separate the two objectives. Removing the next-beat term raises effective dimensionality by a factor of two to four in every arm and lowers adjacent-embedding similarity wherever it was high. Including the next-beat term is therefore associated with higher adjacent-embedding similarity and lower effective dimensionality, a change of geometry that is not a uniform gain or loss downstream.

# G EFFECTS OF CLINICALLY ANCHORED CONTINUED PRETRAINING

## G.1 ANALYSIS SETS AND MEASURES

The four arms of §4 are read twice on the same windows, once at the end of waveform pretraining and once after clinical continued pretraining, over 600 held-out windows of the emergency-department cohort that neither stage trained on. Each arm carries the objective §5.4 selected for it. Two null models appear below and they are not the same one.

A *margin* is a boundary F1 against the detector, at a $\pm 100$ ms tolerance on a 50 ms token grid, minus the F1 of a uniform grid emitting the same number of boundaries in the same window, averaged over eight grid phases. The subtraction is needed because sub-beat landmarks lie about 250 ms apart, so a grid of the right rate scores about $0.78$ with no localisation at all and a raw F1 cannot be read on its own.

A trajectory excess, written $\Delta r_t$, is a correlation minus the same correlation after the beat order is shuffled within that window.

Table 8: Reconstruction, representation geometry and boundary alignment before and after clinical continued pretraining, over 600 held-out MC-MED windows. Entries are medians, before → after. The K0 and K1 margins are boundary F1 against a count-matched uniform grid, defined in App. G.1.

| | recall MSE before → after | eff. dim before → after | adj. cos before → after | K0 margin before → after | K1 margin before → after |
|---|---|---|---|---|---|
| *A1*, convolutional scorer | .0027 → .0054 | 49.6 → 18.2 | .175 → .413 | +.201 → +.241 | +.646 → +.675 |
| ***A2***, pairwise router | **.0144 → .0049** | 52.0 → **24.1** | .195 → **.356** | +.085 → +.019 | +.583 → +.626 |
| *B*, same router, unplaced | .0023 → .0067 | 56.1 → 19.8 | .170 → .421 | −.125 → −.162 | +.480 → +.219 |
| *C*, fixed stride | .0007 → .0106 | 9.9 → 5.8 | .699 → .231 | +.003 → +.003 | −.001 → −.001 |

Both floors move from window to window, so both are formed inside the window and only then reduced across windows. One limit applies throughout: each arm is read at its own checkpoint of the clinical stage, selected on validation, and those fell between 139.4k and 191.7k steps of that stage against the 400k of the pretraining budget, a 1.4-fold spread among the arms that the readings do not control for. Table 8 gives the medians.

### G.2 Boundary alignment and waveform reconstruction

**The two levels respond differently, and the direction depends on whether the boundaries were supervised.** At the beat level the alignment margin rises for A2, $+.583$ to $+.626$, and falls for B, $+.480$ to $+.219$. The two share a router and an objective and differ only in whether anything told the router where a boundary belongs. A2's sub-beat margin also falls, $+.085$ to $+.019$, toward the rate-matched baseline, while its sub-beat boundary rate rises from 231 to 277 per minute. The fixed-stride arm sits at zero at both levels and does not move, which is what the baseline is for: a grid of the right rate scores about $0.78$ against these labels with no localisation at all, so an agreement rate cannot be read on its own.

Reconstruction separates A2 from the rest. It is the only arm whose waveform error falls across the stage, from $.0144$ to $.0049$, while the other three rise; the fixed-stride arm, which reconstructs best of all before the stage at $.0007$, ends worst at $.0106$.

**Agreement with the detector does not order the arms the way transfer does.** A1 holds the highest margin at both levels, before and after, and the lowest four-task mean of the three arms that have a router, $.784$ against $.812$ and $.797$ in Table 3. The dissociation App F reports between reconstruction and downstream utility holds one level up, between boundary agreement and downstream utility, and neither quantity is the one the model is selected on. A1 also differs from A2 in two ways at once, a centered kernel of 15 tokens that lets the decision see $\pm 350$ ms and a per-token cross-entropy in place of a ranking loss, so its margin is an upper reference rather than a measurement of what a causal router costs.

The segmentations and trajectories these readings summarise are drawn window by window in App. M.

### G.3 Changes in representation geometry

**The stage compresses the representation and strengthens its slow component.** From one beat to the next the codes show no measured order: the cosine between successive displacement vectors is $-.49$ in each of the three arms with a router, the value a sequence of independent points gives, which is consistent with independence at that scale without establishing it. Averaged over fifteen beats a slow component does appear, but a moving average manufactures one from independent points on its own, and the reading is therefore the excess over a within-window shuffle of the beat order. That excess rises in every arm with a router, $\Delta r_t$ from $+.145$ to $+.252$ in A1, $+.195$ to $+.250$ in A2 and $+.116$ to $+.189$ in B, the path straightens against its own shuffle in step, and effective dimension halves over the same windows, $48.1$ to $23.3$ in A2 and falls further in the other two (Figure 4; Table 8 reports the same quantity on the windows of the boundary analysis). The baseline all three excesses are taken over is a shuffle of the beat order within the window, which removes order of any kind. It does not separate a slow component that carries state from one that drifts with elapsed time, and no reading given here makes that separation.

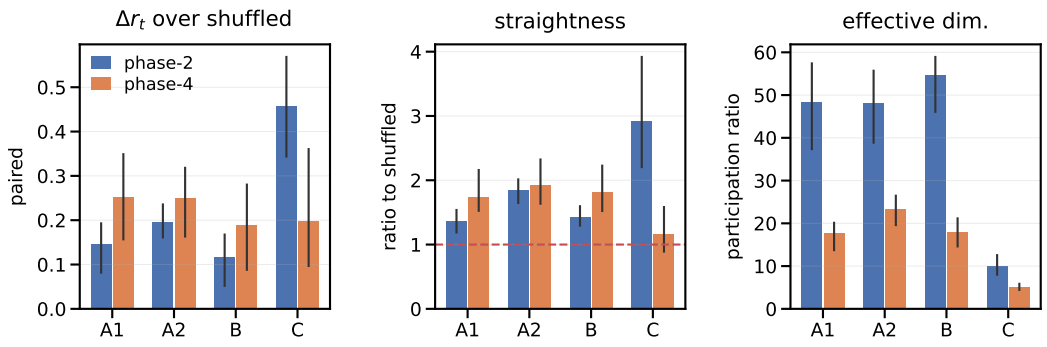



Figure 4: Beat-embedding trajectory statistics before and after clinical continued pretraining over 64 MC-MED encounters. **Left:** the temporal correlation of the 15-beat smoothed trajectory above a within-window shuffled-order baseline, $\Delta r_t$. **Centre:** trajectory straightness relative to the same baseline; the dashed line is a ratio of one. **Right:** effective dimensionality as the participation ratio. Bars are medians and whiskers inter-quartile ranges.

What the compression leaves can still be read out linearly. On a held-out wrist cohort a linear probe holds a three-class signal-quality macro-F1 of .78 to .80 across the stage, unchanged within its spread. The same probe reads atrial fibrillation at .53 to .58 AUROC before the stage as well as after, so on that cohort this probe is not a sensitive test of whether the stage changes rhythm-related information. It is also a different recording from the ones above, wrist photoplethysmography resampled from 32 Hz and read through segment-mean codes rather than per-beat codes, and the two should not be read as one experiment.

# H ARCHITECTURE AND TRAINING DETAILS

**Naming and widths.** Each encoder stack and router is named after the level it produces: the K0 encoder runs on tokens and feeds the K0 router; the K1 encoder runs on K0 embeddings (padded to $d_1$) and feeds the K1 router; the main backbone is a final causal stack on the K1 grid. Of the 1.22 M parameters that pretraining updates on the compression path, 1.14 M are the token-to-beat representation pathway and the rest the boundary generator; the anti-causal decoder is a further 0.44 M that training uses and deployment discards, and the convolutional boundary generator of the A1 arm (App. D) raises the compression-path total to 1.71 M. Every stack is two Mamba-2 layers: one on the token grid, one on the sub-beat grid, one in the main network on the beat grid, and one on each grid the expansion path returns to.

**Pretraining streams and optimization.** All four arms are trained for 400k optimization steps with batch size 96, where the batch counts concurrent streams rather than individual windows. Each stream contains up to five consecutive five-minute windows from one encounter. A window contains 6000 tokens, approximately 1100 sub-beat units and 390 beat-level units. Recurrent states are carried between windows within a stream, detached at window boundaries, and reset when a new stream begins. One pass over the corpus corresponds to approximately 2.8k optimization steps. The corpus was recorded between 2010 and 2020, and its 90/5/5 encounter-level split is drawn from a fixed seed.

Training uses AdamW with weight decay 0.01. After 1000 warmup steps, the learning rate follows a cosine schedule from $5.196 \times 10^{-4}$ to one tenth of that value. Gradient norms are clipped at 5.0, and computation uses bfloat16. These optimization settings are shared across arms.

**Shared representation settings.** All arms use $d_0 = 64$ and $d_1 = 128$, two-sample input tokens, chunk-end selection, and time encoding on the token, sub-beat, and beat grids. The common configuration specifies `seg_len=2`, `inner_target_rate=1/3`, and the stack layout `m2[m2m2m2]m2`. Boundary-specific settings are given in App. E.

**Objective settings.** When waveform recall is enabled, its MSE and correlation weights are 1.0 and 0.3, respectively. Next-beat prediction has weight 1.0 when enabled. SIGReg remains active in all objective configurations, with weight 1.0, 256 projection directions, 256 samples, and $\gamma = 0.5$. The main A2 configuration omits next-beat prediction. Objective definitions are provided in App. D.

**Why the decoder runs anti-causally.** Causality at inference is a property of the compression path, which reads no future whichever way the decoder runs, and the decoder is discarded before inference; the direction is therefore a choice about what the objective measures rather than a condition the causal claim imposes. Under the end-of-chunk convention the expansion already places $z_{k(t)}$ at $t$, and that code was computed over an interval containing $t$, so neither direction has to reconstruct $t$ from codes computed without it. What the direction changes is which other codes contribute. Codes of chunks that ended before $t$ were computed without it, and a causal decoder reads those; codes at positions $\geq t$ were each computed over an interval containing $t$, so an anti-causal decoder reconstructs $t$ from representations that encoded it, which is what the objective is meant to test. The centred depthwise convolution mixes a few neighbouring positions either way. A bidirectional decoder measures a mixture of the two; it is the harder result to read, not an inadmissible one.

**Normalisation and state.** Input normalization applies fixed dataset-level statistics (a global affine map). Per-window statistics would inject discontinuities at window seams and break the recurrence that carries state between windows; transferring to a new dataset re-pins these statistics from unlabelled data only. States are detached between windows (truncated backpropagation through time) and reset only at stream starts.

**Time encoding.** Sinusoidal features of absolute time within the stream and of the interval to the previous embedding are added to the embeddings of each newly formed grid, before the stack that consumes that grid; the projection is zero-initialised. This is an architectural component for irregular grids and gaps, and we make no claim that it improves any downstream measure.

# I FIDUCIAL SUPERVISION DETAILS

The landmarks come from a derivative-based detector run offline over the pretraining corpus, which returns the pulse foot, the systolic peak and the dicrotic notch of each detected cycle; it is not the pyPPG toolbox, which we cite for the definitions of the landmarks themselves. The cardiac period is taken from the spectral peak between 0.5 and 4 Hz. Feet and systolic peaks are found by peak detection against a fraction of the local dynamic range, and the notch as the first local maximum of the smoothed first derivative on the downslope, between 10 and 75% of the cycle.

Over the pretraining corpus the detector returns about 3.7 sub-beat and 1.3 beat boundaries per second, and these are the rates the boundary terms of every arm are matched to.

No manual review is applied. Positions the detector does not return are masked rather than imputed, so a cycle whose notch is not found contributes its other two landmarks and nothing at the missing position, and a window the detector fails on contributes no boundary supervision at all while still contributing to the reconstruction terms. That happens on $1.48\%$ of windows, which stay in training with their routing targets empty. Of the detected cycles $85.2\%$ carry all three landmarks; the per-window notch detection rate has median $0.94$ and falls below one half on $8.0\%$ of windows, and masking the missing notches removes $5.2\%$ of token positions from the placement terms.

Negatives for the ranking term are the non-landmark tokens within ten tokens of a landmark, half a second at the token rate, and the ranking margin is $0.15$. The rate term takes its target from the window's own landmark count rather than from a global constant, so a window recorded at a faster heart rate is asked for proportionally more boundaries and the arms remain rate-matched window by window rather than on average.

Beat-level supervision uses no landmark identity. The terms ask for one boundary per detected cardiac cycle and for a consistent offset within the cycle, which leaves the choice of which landmark ends a beat to the model; the detector supplies the cycles, not the answer. Successive beat intervals are not regularized toward a constant duration: beat-to-beat timing variability is itself physiologically informative, and constraining the spacing would remove the variation the beat representation should retain.

# J CLINICAL ANCHORING DETAILS

**Continued-pretraining configuration.** The reported MIMIC-III configuration uses two-hour waveform windows, 62,324 for training, 12,062 for validation and 11,778 for test, and trains for 30 epochs with batch size 6. Optimization uses AdamW with weight decay 0.01, constant learning rates, and no warmup. Parameters belonging to the loaded encoder use $3 \times 10^{-5}$; all remaining trainable parameters use $3 \times 10^{-4}$. No additional parameter-name-based learning-rate groups are applied.

The entire encoder, including the input projection, K0 and K1 encoders, time embeddings, and routers, remains trainable alongside the expansion path and added readout level. Routing uses the straight-through path described in App. D. The clinical objective is a per-target masked $L_1$ loss on normalized measurements. Retained self-supervised objectives are weighted by $w_{\text{ssl}} = 0.1$, which scales the balance the encoder was pretrained with rather than retuning it; the boundary supervision of §3.2 is carried unchanged into this stage. The readout hidden dimension is 128, and query times determine coarse chunk boundaries. For the context-length experiment, the checkpoint is selected by validation performance before held-out evaluation.

**Causal dechunking.** When a target reads below the query grid, the dechunk broadcasts only *completed* chunk embeddings: positions before the first completed chunk are marked invalid, and position features are limited to the age of the embedding. The standard dechunk of Hwang et al. (2025) leaks future information through the embedding of the chunk still in progress, through a distance-to-end feature, and through centred convolution, so it cannot be used for a causal readout. An encoder residual is added at each resolution, without which a fine resolution carries no fine information.

**Scoring.** $R^2$ is computed over the scored readings of the cohort and MAE is in the target's own unit. Targets are min–max normalised by their declared physiological range before a masked $L_1$ loss over the observed readings; no target-specific loss weights are applied, and errors are not rescaled by how

Table 9: Training hyperparameters and readout configuration for clinically anchored continued pretraining on MIMIC-III.

| Setting | Value |
|---|---|
| Optimizer | AdamW |
| Loaded-encoder learning rate | $3 \times 10^{-5}$ |
| Remaining-parameter learning rate | $3 \times 10^{-4}$ |
| Learning-rate schedule | Constant |
| Warmup | None |
| Weight decay | 0.01 |
| Batch size | 6 |
| Training epochs | 30 |
| Waveform window | 2 hours |
| Self-supervised weight | 0.1 |
| Clinical objective | Per-target masked $L_1$ |
| Boundary supervision | As in pretraining |
| Readout hidden dimension | 128 |
| Readout chunking | Query-time boundaries |
| Encoder adaptation | All parameters trainable |

widely a target varies. **Cohort and preprocessing.** Splits are formed at the encounter level and hold 3203, 656 and 663 encounters; 3140, 645 and 651 of them are present in the waveform store and are the ones used, giving 62,324, 12,062 and 11,778 windows. A window is 240 consecutive 30 s segments of 40 Hz photoplethysmogram, two hours in all, and windows do not overlap. Segments are admitted whenever their timestamps strictly increase, so a window may span a recording gap rather than being discarded for one. A per-record gain correction is applied once in the store rather than at load time, over a manifest of 2231 records with a divisor of $4.0029$. Waveform normalisation uses constants held with the dataset, mean $0.4605$ and standard deviation $0.1467$ for this channel, which replace the constants the pretrained encoder carries from its own corpus ($4.4746$ and $1.3451$).

**Targets and anchors.** Six targets are read: heart rate, oxygen saturation, respiratory rate and systolic, diastolic and mean arterial pressure. Each is min-max scaled by the physiological range the variable registry declares for it, which is 40 to 300, 20 to 200 and 30 to 250 mmHg for the three pressures. The charting cadence is 60.0 s from the fifth to the ninety-ninth percentile, so the coarse grid is regular in practice even though nothing enforces it. Of the candidate anchors, 3475 charted readings fall inside a recording gap and have no waveform at the instant they are charted, so they cannot be anchored and are not scored.

## K VARYING CONTEXT AT EVALUATION

The encoder is the pairwise arm at 400k pretraining steps, continued on the intensive-care cohort with the encoder trainable throughout and the self-supervised terms retained as an anchor, read at its best validation epoch. Continued pretraining uses two-hour windows on the training split of this cohort, and the sweep is inference on its held-out split, so the comparison is in-domain and no weights change between rows. Because the clinical head is trained at two hours, the short rows sit further from its training regime; they are not outside it, as every training stream begins with a cold state and short effective history.

**Anchors and trials.** An anchor is a charted observation instant $T$ in the held-out split with at least two hours of continuous recording before it. For a context length $C$, one trial is the $C$ seconds preceding the anchor's own segment together with that segment, fed as a single window whose recurrent state is zeroed at its first sample. The state therefore integrates exactly the recording inside that window, so a row labelled $C$ carries $C$ seconds of history; the store is segmented at 30 s, so the realised history falls in $[C, C + 30\,\text{s}]$ and the label is its lower bound. The same anchors are scored at every context length. Only real charted readings are scored: the cadence grid is filled to preserve the coarse segmentation, but filled instants supervise nothing.

**Cutting the window at the label.** No signal after $T$ is admitted. The sample array is kept whole-segment, because the beat axis is derived by dividing the sample count by the segment count and a

Table 10: Numerical values for the history sweep of Figure 3(a, b) on held-out MIMIC-III recordings. Results are reported as $R^2$ and MAE, with MAE units as in Table 1. Every history length uses the same checkpoint, clinical head and observation anchors, with the recurrent state reset for each trial.

| History | HR | | RR | | $SpO_2$ | | ABP-S | | ABP-D | | ABP-M | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE | $R^2$ | MAE |
| 30 s | .710 | 5.96 | .104 | 3.85 | −.002 | 2.46 | −.324 | 21.92 | −.106 | 11.75 | −.364 | 15.56 |
| 2 min | .822 | 4.25 | .258 | 3.36 | .099 | 2.26 | −.115 | 19.88 | −.079 | 11.61 | −.245 | 14.76 |
| 5 min | .821 | 4.51 | .330 | 3.17 | .107 | 2.21 | −.042 | 19.14 | −.092 | 11.72 | −.210 | 14.52 |
| 15 min | .845 | 4.03 | .402 | 2.95 | .117 | 2.19 | .108 | 17.56 | .031 | 10.85 | −.024 | 13.12 |
| 30 min | .852 | 3.83 | .434 | 2.85 | .117 | 2.19 | .155 | 16.98 | .083 | 10.41 | .052 | 12.49 |
| 1 h | .860 | 3.66 | .453 | 2.78 | .120 | 2.20 | .175 | 16.71 | .115 | 10.13 | .096 | 12.12 |
| 2 h | .867 | 3.53 | .465 | 2.74 | .121 | 2.19 | .189 | 16.55 | .138 | 9.93 | .127 | 11.85 |

ragged array would shift that axis for every segment in the window; the *valid* length is set instead to the sample index of $T$. The encoder turns it into a token mask and each router intersects its boundaries with that mask. The last valid position is itself an endpoint (§3.1), so the chunk in progress at $T$ closes there rather than being dropped: under the end convention it covers the span since the previous boundary, and what the model holds at $T$ therefore includes that final unit, shorter than a cardiac cycle. This is the readout of §3.4. The query grid is truncated at $T$ as well.

**Verification.** Four assertions run before the sweep and are recorded with it. The configuration is compared field by field against the one stored in the checkpoint. The window builder must reproduce the dataset's own windows where the two overlap. Overwriting the samples after $T$ with a different recording's, at unchanged length, must leave the prediction at $T$ bit-identical, and it does on every probed trial; this verifies that the estimates use only signal at or before $T$. And feeding one trial as two windows with state carried across the join must differ from feeding it with the state reset there, which confirms that the recurrent state rather than the windowing is what carries the history.

**What the anchor requirement selects.** Requiring two hours of preceding recording places anchors later in each stay and favours longer encounters. A further 3475 readings, about a fifth of the candidates, fall inside a recording gap and have no waveform at their own instant; they are dropped. The final evaluation set contains 12 880 anchors over 651 encounters for heart rate, 12 655 and 12 312 over the same encounters for respiratory rate and oxygen saturation, and near 3530 over 227 encounters for each arterial pressure, the pressures being scarcer because they require an arterial line.

**Effect of context length.** From the shortest context to the longest, $R^2$ rises by 0.157 for heart rate, 0.361 for respiratory rate and 0.123 for oxygen saturation, and by 0.513, 0.244 and 0.491 for systolic, diastolic and mean pressure. Mean absolute error falls by 41% for heart rate, 29% for respiratory rate, 11% for oxygen saturation and 15 to 24% across the pressures. Oxygen saturation is the one target that stops early: it gains almost all of that by fifteen minutes and is flat from there, its error moving by $0.01$ over the longer context lengths.

**What is not measured.** Every trial starts cold, so the sweep does not exercise state carried *across* window boundaries, and the right edge of the ladder is two hours from a cold start rather than the deployed condition. §5.2 states the other limit: a rising curve does not separate retained history from a state that has merely settled.

## L External baseline protocol

**How the two cohorts' photoplethysmograms differ.** They are not the same signal recorded in two places. Both reach us at 40 Hz, from 125 Hz in the emergency department and 100 Hz in the operating room. The emergency-department monitor autoscales its pleth: measured over 60 encounters, the pulse amplitude spans $1.25\times$ from the tenth to the ninetieth percentile of encounters and varies within an encounter with a coefficient of variation of $0.06$, and the baseline sits within five counts of mid-scale for every encounter. The operating-room store is not autoscaled and keeps that variation, $8.4\times$ across encounters in its earlier subset and a coefficient of variation of $0.19$. The

Table 11: Encoder parameter counts, native input context lengths and input waveform sampling rates of the evaluated models. Each encoder reads the waveform resampled to the rate it was pretrained for.

| Model | params | context | input rate |
|---|---|---|---|
| HeartGPT-PPG | 0.44 M | 10 s | 50 Hz |
| Physio-HNet | 1.2 M | 2 h | 40 Hz |
| PaPaGei-S | 5.79 M | 10 s | 125 Hz |
| AnyPPG | 4.04 M | 10 s | 125 Hz |
| SIGMA-PPG | 28.2 M | 240 s | 50 Hz |

two are filtered differently as well: relative spectral power between 10 and 15 Hz is $8 \times 10^{-5}$ in the emergency department and $1.2 \times 10^{-2}$ in the operating room, so the first is band-limited well inside our sampling rate and the second is not. Whatever a representation carries about pulse amplitude is therefore available in one cohort and largely removed in the other, before any encoder is applied.

**Inherited filtering.** We add no signal-quality filter on any cohort, but MC-MED is not released raw: its authors remove any waveform segment whose first difference is zero for ten seconds or longer, which drops detached probes and flatlines before we see the data (Kansal et al., 2025). MOVER and the pretraining corpus carry no such step.

**Probe cohorts.** The four public cohorts of §4 are WESAD (Schmidt et al., 2018), 15 subjects under the suite's own segmentation; Stanford wearable atrial fibrillation (Torres-Soto & Ashley, 2020), 163 subjects in 25 s segments; Pulse Transit Time PPG (Mehrgardt et al., 2022), 22 subjects in 5 s segments of sitting, walking and running; and Real-World PPG (Siam et al., 2022; 2021), 35 subjects in 6 s segments under the authors' own split. The first is scored on affective state, the second against non-atrial-fibrillation, and the third on one activity against the other two. The identification cohort's validation split is re-drawn from its training segments, because the released one duplicates them.

The two criteria named in §4 act at different places. Atrial fibrillation is the primary task criterion, the single cohort a claim about representation quality on these four tasks is read on. The unweighted mean over the four is the selection criterion of §5.4, the number that chooses which objective configuration each arm carries into the clinical stage. Neither selects a checkpoint.

**Models included.** The four released encoders, HeartGPT-PPG (Davies et al., 2024), PaPaGei-S (Pillai et al., 2025), AnyPPG (Nie et al., 2026) and SIGMA-PPG (Guo et al., 2026), span two orders of magnitude in parameter count, from 0.44 M to 28.2 M (Table 11). The smallest, HeartGPT-PPG, is included as a floor on what a released PPG encoder recovers rather than as a competitor. A fifth was considered and excluded because the per-subject normalisation its inputs assume was not released, so it cannot be run under a protocol shared with the others.

**The two probes.** Every encoder emits unit embeddings at its own rate together with their times; the units are mean-pooled into 60 s bins, the charting cadence, and both probes read that one cached bin sequence. The recurrent probe is a single-layer LSTM of width 128 with dropout 0.2 and a linear map to the targets, each label read from the hidden state at its own bin, so nothing after a label enters its prediction. The static probe is the same head without the recurrence, a two-layer network over the 240 s of bins ending at the query. The encoder is frozen and only the probe trains, under the same $L_1$ readout and a fixed epoch budget, with no probe validation: the reported number is one test of the final weights, at five probe seeds. Our own rows are the K1 beat tap of the frozen checkpoint, so the readout level that §3.4 adds during the clinical stage is not part of what is probed here, and the 1.22 M of App. H is the whole of what encodes for these tables.

**Resolution of the comparison.** Both probes are run at five seeds and the table reports the mean and standard deviation over them. Released weights carry no pretraining seeds, so the spread reflects probe-training variability and not pretraining variability. Small differences relative to that spread are interpreted cautiously rather than as rankings: that covers the emergency-department heart-rate and oxygen-saturation ties with AnyPPG under the recurrent probe, and, under the static probe in that cohort, both the heart-rate cell and the diastolic cell that §5.1 counts among the eleven. AnyPPG's emergency-department rows are also in-domain, that cohort supplying about 72% of its pretraining corpus, so an encounter there may also not be novel to that model.

**Whether a patient crosses a split.** The splits are formed at the encounter level, which leaves open whether one person appears on both sides of one. Patient identifiers were used to check whether any patient appears in more than one split. In the emergency department 10,001 encounters resolve to 9,039 patients, 7,239 of them in training, 908 in validation and 892 in test; in the operating room 8,553 cases resolve to 8,463 patients, 5,925, 1,269 and 1,269 across the three. No patient lies in two of them in either cohort.

**Probe against window.** Both probes read one cached bin sequence per model, produced once by that encoder at its native window, so between the two tables the head and its pooling change and nothing else does, for every model and not only for ours. The direction of the difference is not the same in the two settings: in the emergency department the recurrent head raises heart rate for every encoder and raises our systolic pressure, while in the operating room it lowers all three pressures for AnyPPG, SIGMA-PPG and ours. We have no controlled explanation for that reversal.

**The two greyed columns.** Oxygen saturation is not determined by this input: oximetry reads it from absorbance at two wavelengths and these recordings carry one, so an encoder can recover only what correlates with the quantity. The best cell reaches $0.07$. Operating-room respiratory rate is a property of that cohort rather than of the encoders. Under general anaesthesia it is a ventilator setting rather than an observed signal, and the repository's own data dictionary lists it among ventilator parameters rather than among the monitored vital signs (Samad et al., 2023); it reaches the canonical store only through the flowsheets of that repository's later subset. It is therefore scored on 249 cases where heart rate and oxygen saturation are scored on 1232 and 1233, against 619 of 738 in the emergency department. What is left to explain is correspondingly narrow: predicting the training mean costs $3.02$ breaths per minute in the operating room against $3.58$ in the emergency department, and the best model improves on it by $0.10$ rather than $0.45$, so the denominator is small while the absolute error is the lower of the two cohorts.

**What is matched and what is not.** Architectures that emit embeddings at different rates cannot be compared at a common window without changing one of them, so each released encoder is run at the window it was trained for and its output is brought to the shared 60 s probe grid by pooling, never by re-cutting its input; a bin with no embedding carries the last filled one forward. Table 11 states what that leaves unmatched: parameter count, native context and input sampling rate.

A recoverability audit, training a decoder to reconstruct the waveform from each frozen embedding, would separate what a representation retains from what a probe head can extract. It is not reported in this paper.

# M Qualitative examples before and after clinical continued pretraining

The pages below draw, window by window, what App. G reads as medians over hundreds of windows. They are illustrations of quantities reported there and carry no reading of their own.

## M.1 Qualitative visualization of beat representations

Four pages, Figures 5 to 8, draw the same recording excerpts as each arm represents them before and after the clinical stage, two arms to a page. Dots are beat embeddings and the grey line joining them their order in time, colour runs over the excerpt, and the heavy curve is a 15-beat moving average over the embeddings. The heart rate and pulse amplitude beside each excerpt are the detector's, drawn on the same time axis for reference. The principal components are fitted separately in each panel, so the directions of the axes and the size and shape of a path belong to that panel alone and are not comparable between arms or between stages.

Rows are drawn under two rules, one ordering encounters by how far the pulse amplitude moves within a window and one spreading them evenly over charted heart rate, in both cases among the encounters whose detector quality index reaches $0.10$. The two schemes give complementary views of variation in pulse amplitude and in heart rate.

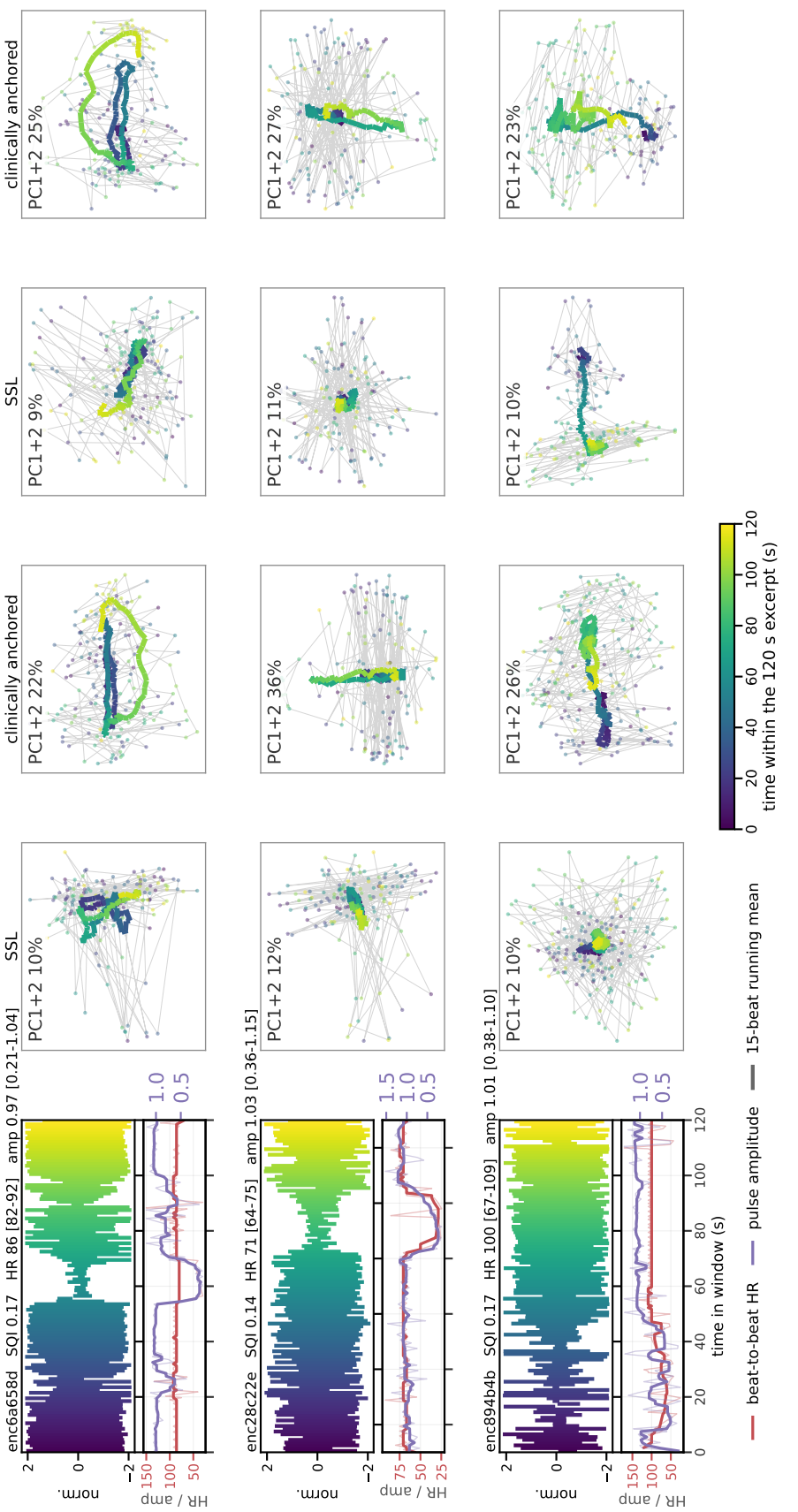



Figure 5: Beat-embedding trajectories of the supervised arms A1 and A2, before and after clinical continued pretraining, in three MC-MED encounters. The two columns within an arm are that arm after self-supervised pretraining and after the clinical stage. **Left:** the input over the same span, a per-second envelope above and the detector's beat-to-beat heart rate and pulse amplitude below, drawn on the same time axis for reference. **Right:** dots and grey lines are the beats and their order and the heavy coloured curve is the 15-beat smoothed trajectory, with colour running over the 120 s excerpt. The principal components are fitted separately in each panel, so PC1+2 is that panel's own explained variance and no panel is comparable with another by the direction, size or shape of its path. Rows are drawn in order of how far the pulse amplitude moves.

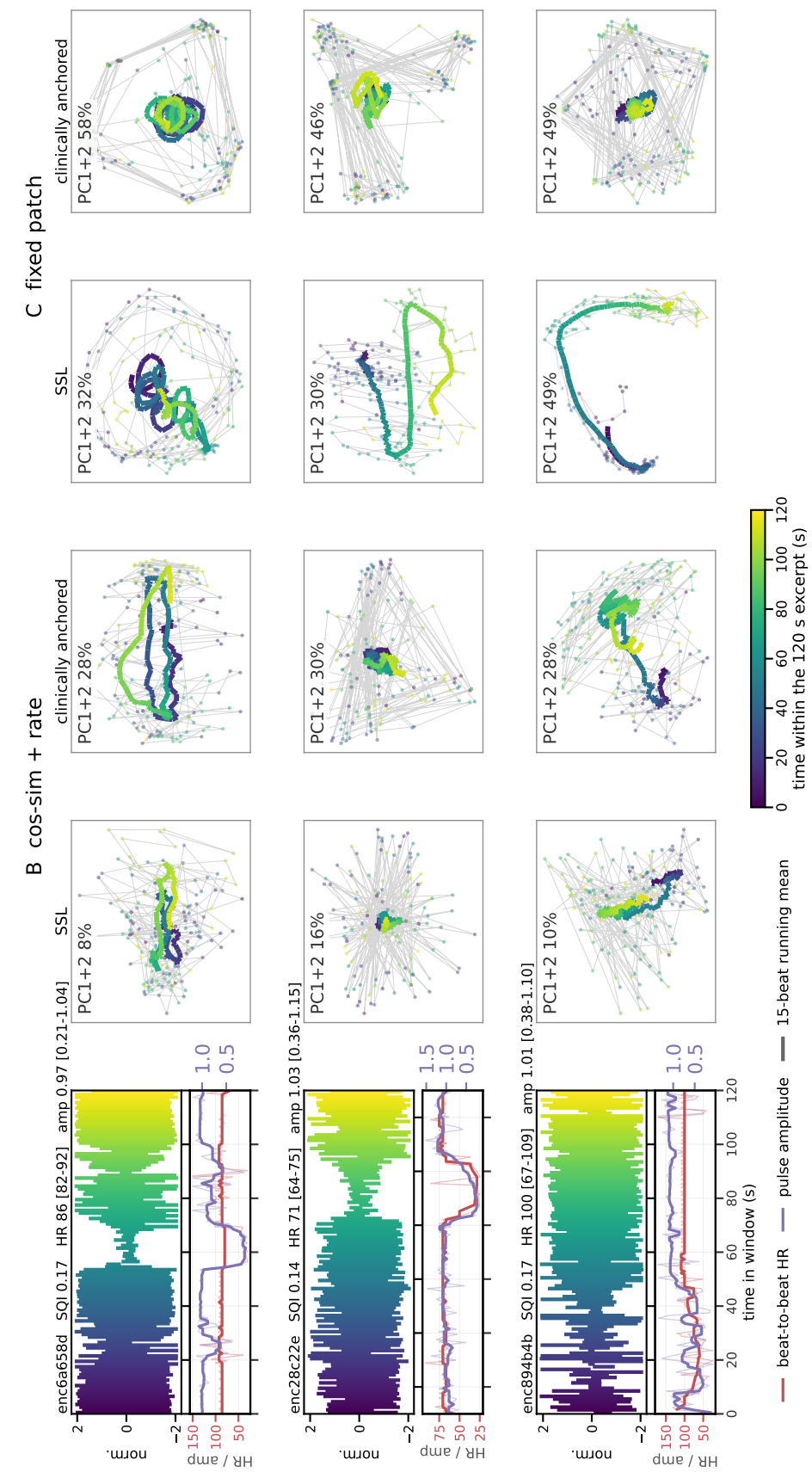



Figure 6: Beat-embedding trajectories of the unsupervised pairwise-router arm B and the fixed-stride arm C, before and after clinical continued pretraining. Encounters, input traces, projection and conventions are those of Figure 5.

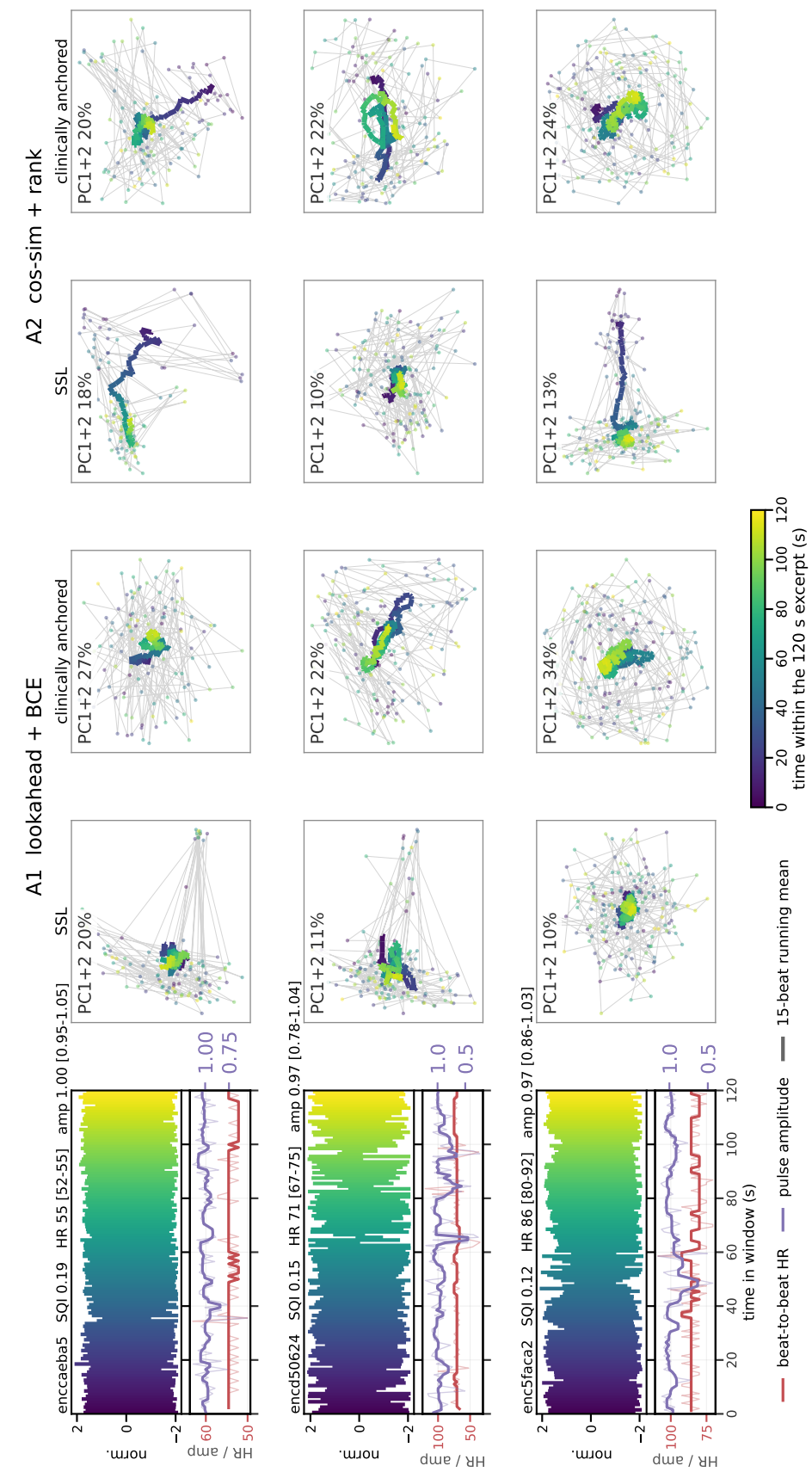



Figure 7: Beat-embedding trajectories of the supervised arms A1 and A2 in three further MC-MED encounters, drawn as in Figure 5 but with rows spread evenly over charted heart rate rather than ordered by amplitude movement.

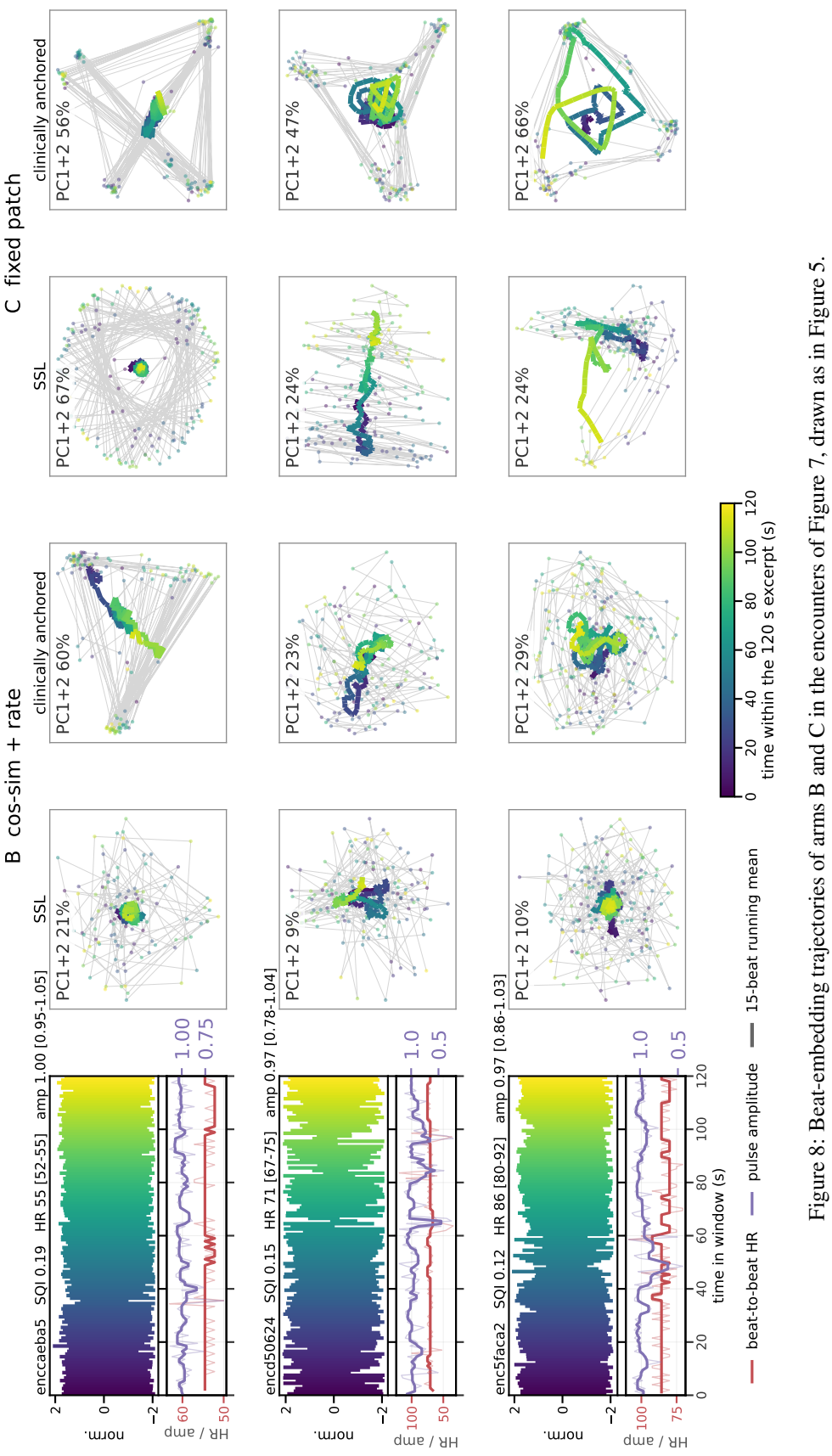



Figure 8: Beat-embedding trajectories of arms B and C in the encounters of Figure 7, drawn as in Figure 5.

### M.2 BOUNDARY PLACEMENT AND WAVEFORM RECONSTRUCTION

**What the segmentations look like.** Figures 9 to 12 draw three windows per arm on the same cohort, sampled across heart-rate and signal-quality ranges from the 21 of 66 encounters whose detector quality index reaches $0.10$; the sampling uses that index alone and no model output. Each panel carries the input, the two reconstructions and the two sets of K0 boundaries, with the detector's landmarks marked on the trace; the reconstruction is the overlap-add average of the decoder's per-token windows, since every sample is predicted ten times and the average is the model's claim. A1 places one boundary of each landmark class per cycle. A2 covers the classes with the dicrotic notch intermittent. B carries no boundary at the dicrotic notch at all and spends two or three on the systolic peak instead, which is the shape a rate that is met while a class is missed takes. C is an even grid, unrelated to the trace, and its reconstruction overshoots the systolic peak and flattens the notch.

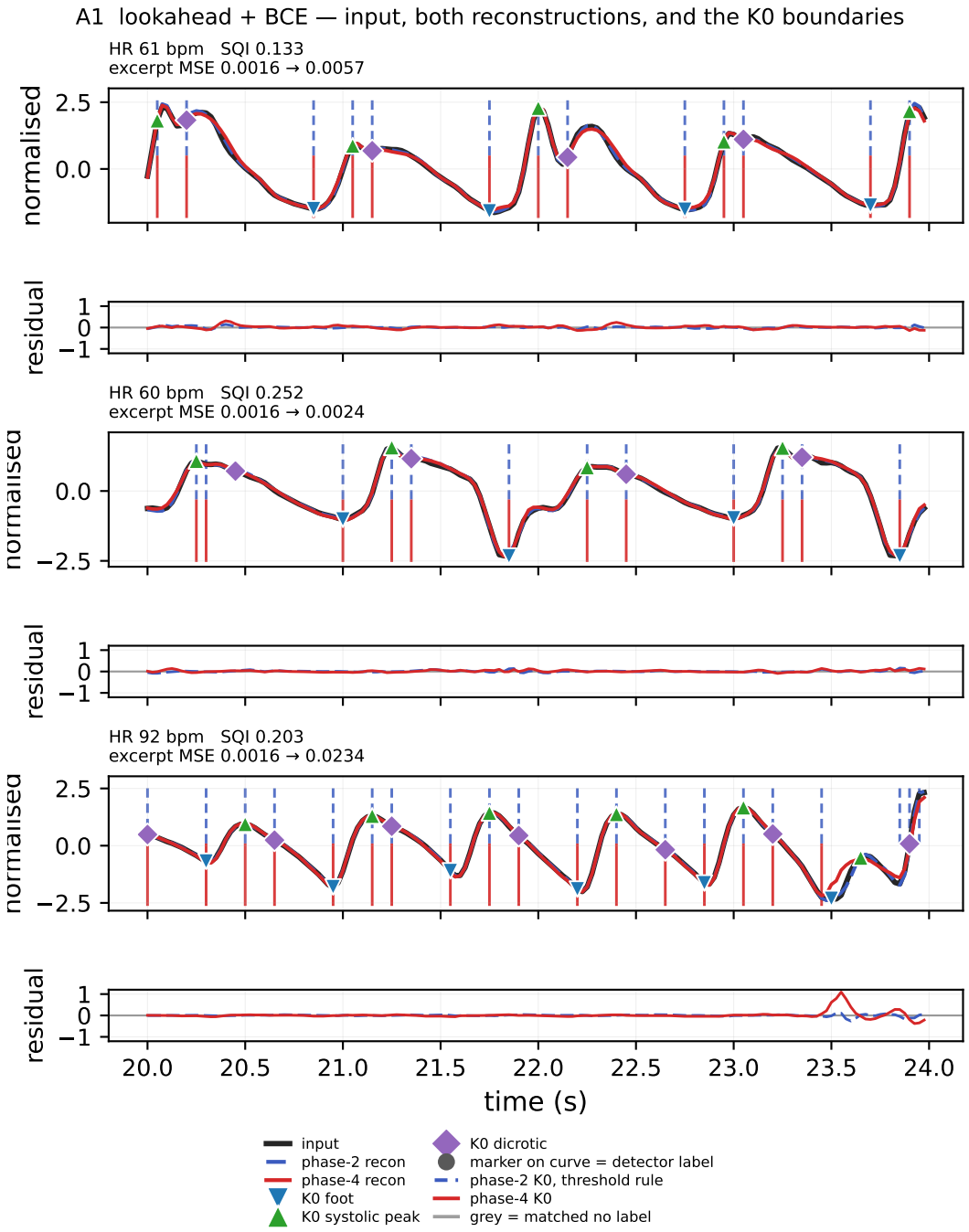



Figure 9: Waveform reconstruction and sub-beat boundaries of arm A1 in three MC-MED excerpts. Black is the input and dashed blue and solid red the reconstruction before and after clinical continued pretraining, overlap-added from the decoder's per-token windows. Markers on the trace are the detector's pulse feet, systolic peaks and dicrotic notches. Vertical lines are K0 boundaries, dashed above the axis before adaptation and solid below it after, offset so that coincident boundaries stay visible and grey where no landmark falls within $\pm 100$ ms. The narrow panel under each excerpt is the reconstruction residual on a scale shared by every row.

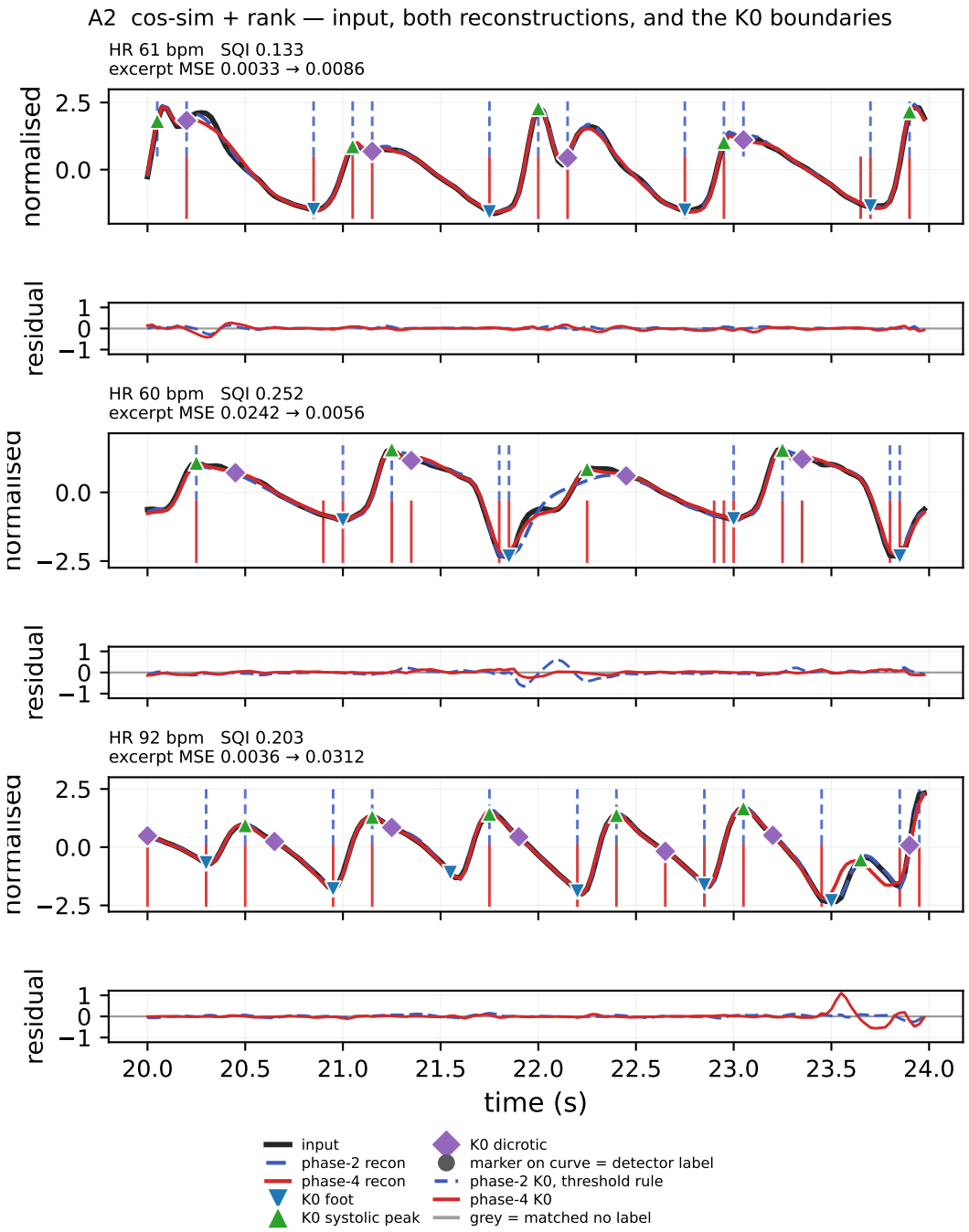



Figure 10: Waveform reconstruction and sub-beat boundaries of arm A2, the deployed configuration, in the conventions of Figure 9.

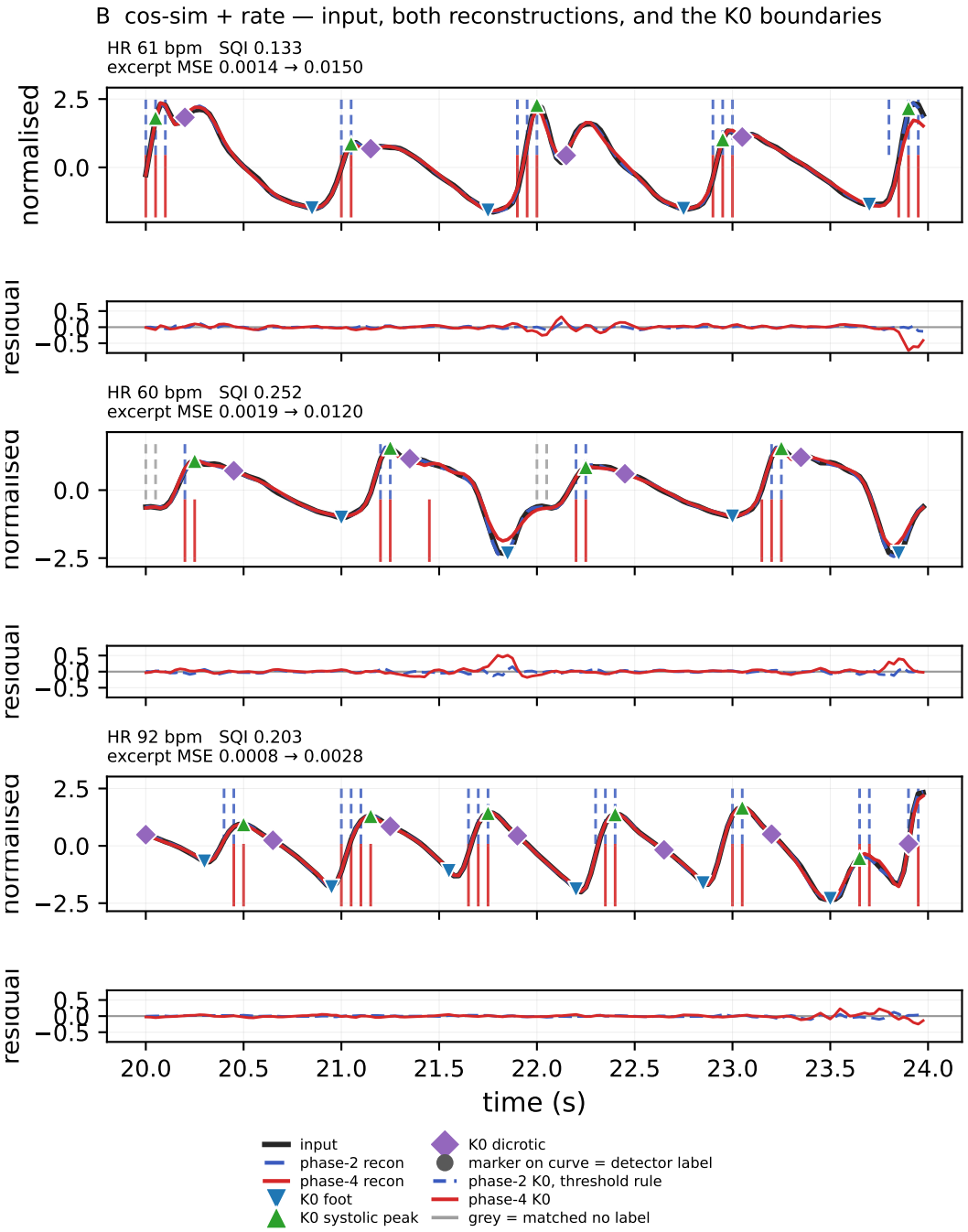



Figure 11: Waveform reconstruction and sub-beat boundaries of arm B, which carries the pairwise router without boundary-location supervision, in the conventions of Figure 9.

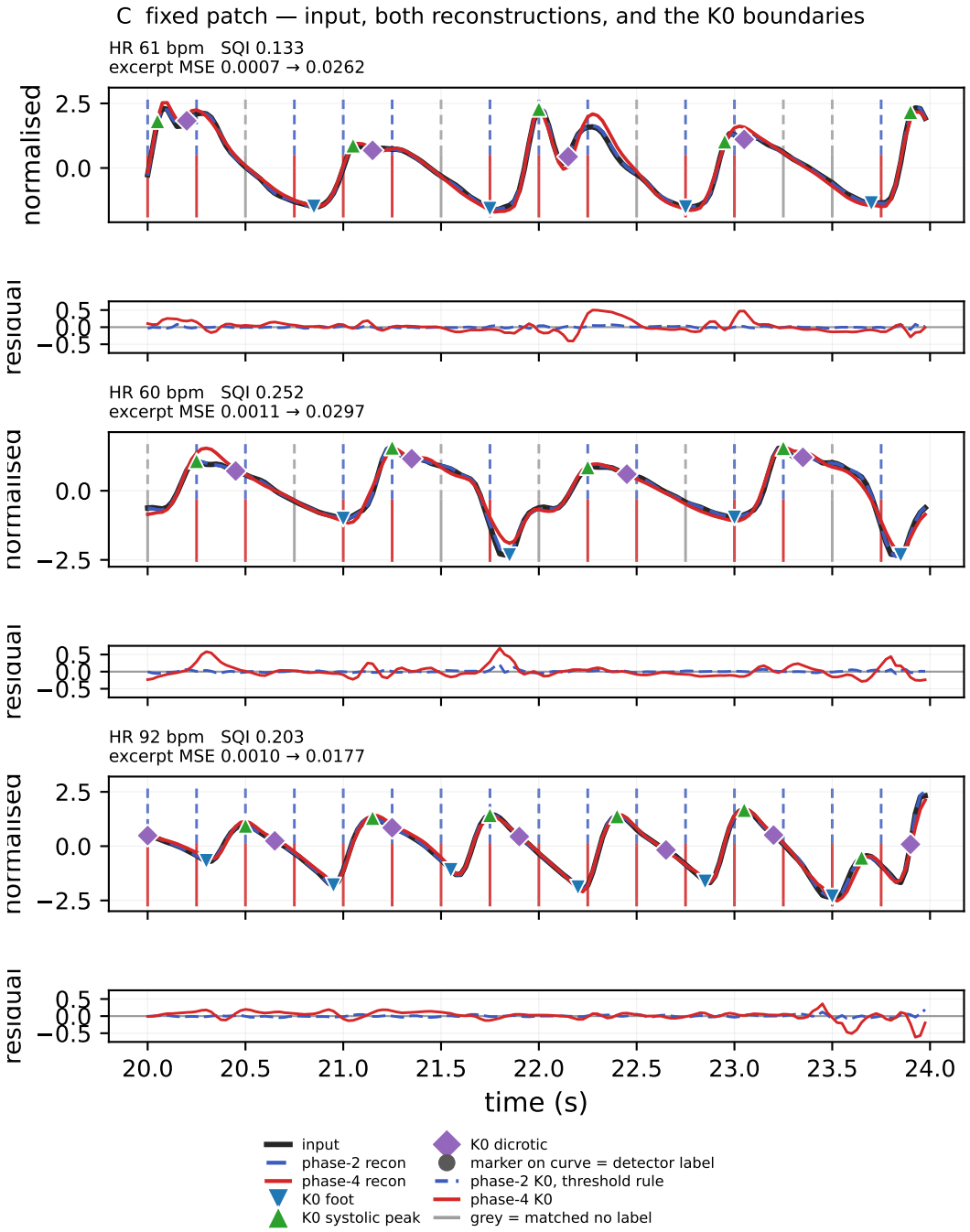



Figure 12: Waveform reconstruction and fixed-stride sub-beat boundaries of arm C, in the conventions of Figure 9.