BEATFORMER: A DELAY-AWARE CARDIOVASCULAR FOUNDATION MODEL FOR BEAT-RESOLVED ECG–PPG REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrocardiography and photoplethysmography record the same heartbeat at two ends of the cardiovascular system, separated by a beat-to-beat delay that reflects ejection dynamics, blood pressure and arterial stiffness. Existing foundation models tokenize the signals into fixed-length patches, obscuring this delay variation and conflating within-beat morphology with across-beat rhythm. We introduce BEATFORMER, a multimodal foundation model that organizes synchronized ECG and PPG as sequences of paired heartbeats anchored on shared R peaks. A delay-aware fusion module estimates a posterior over the beat-specific ECG–PPG delay and uses its confidence to gate bidirectional cross-modal attention. An orthogonal beat–phase Transformer then models within-beat morphology and across-beat dynamics using physical-time and rhythm-aware positional biases. Pretrained on 84.76 million valid cardiac cycles, BEATFORMER outperforms existing foundation models on 9 of 15 physiological-signal classification and regression tasks, and predicts pulse arrival time (PAT) with a mean absolute error of 11.48 ms, 24.6% below the strongest baseline. Across 11 masked-signal reconstruction conditions, it reduces mean RMSE by 11.3% relative to the respective strongest baselines and achieves the highest median signal-to-noise ratio improvement in 13 of 15 denoising conditions. These results support beat-resolved ECG-PPG modeling for diverse physiological prediction and waveform recovery tasks. Code is available at: `https://anonymous.4open.science/r/anonymous-model-release-6541/`.

1 INTRODUCTION

Cardiovascular disease is the leading cause of death worldwide (Vaduganathan et al., 2022), and its monitoring increasingly relies on peripheral biosignals acquired outside clinical settings (Charlton et al., 2022). Electrocardiography (ECG) and photoplethysmography (PPG) are two of the most widely available of these signals (Bayoumy et al., 2021). ECG records myocardial electrical activation, whereas PPG captures the peripheral blood-volume response generated by the ensuing pressure pulse, which typically reaches the fingertip roughly 150–$300\,\mathrm{ms}$ later in healthy adults (Tripathi et al., 2017). Importantly, recording the two signals together exposes the delay between each electrical event and its peripheral response (Franklin et al., 2023). Pulse arrival time (PAT; Figure 1) quantifies this delay from the ECG R peak to a fiducial point on the corresponding PPG pulse (Finnegan et al., 2021). It comprises the pre-ejection period and the vascular pulse-transit time, and is influenced by ejection dynamics, blood pressure, arterial stiffness and vascular tone (Mukkamala et al., 2015). These factors act on timescales from years to single beats, so the delay varies from beat to beat (Finnegan et al., 2021; Ajtay et al., 2023). Assuming a fixed ECG–PPG alignment can blur physiological correspondence and discard informative timing variation. Joint ECG–PPG representations should instead preserve the correspondence between each electrical event and its delayed peripheral response, while separating within-beat morphology from the rhythm and haemodynamic dynamics that emerge across beats.

Recent foundation models (FMs) have begun to learn transferable representations from large-scale ECG and PPG data. ECG-only models, such as ECG-FM (McKeen et al., 2025) and ECGFounder (Li et al., 2025), lack the peripheral response, whereas PPG-only models, including PaPaGei (Pillai et al., 2025), Pulse-PPG (Saha et al., 2025), GPT-PPG (Chen et al., 2025) and SIGMA-PPG (Guo

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.

et al., 2026), lack an electrical reference. Neither can therefore directly represent cross-modal timing measures such as PAT. ECG-guided and multimodal approaches, including AnyPPG (Nie et al., 2025), xMAE (Zhou et al., 2026), VCT-M2AE (Hu et al., 2025), CSFM (Gu et al., 2026) and CardioFM (Hassanuzzaman et al., 2026), introduce electrical supervision or jointly encode the two signals, while UniCardio (Chen et al., 2026) additionally supports restoration and translation among PPG, ECG and blood-pressure waveforms (Appendix A). Although VCT-M2AE and xMAE explicitly acknowledge that ECG and PPG events are asynchronous, these approaches still establish cross-modal correspondence over complete **windows** or uniformly sampled **patches** (Figure 1(a)) and leave the alignment to be learned implicitly. Their contrastive, reconstruction and translation objectives do not explicitly represent the beat-specific delay between an ECG event and its peripheral PPG response. Delay information therefore becomes entangled with waveform morphology and can only be recovered post hoc through task-specific regression (Hassanuzzaman et al., 2026). Moreover, modelling both signals along a single temporal axis does not explicitly distinguish relationships among **phases** within a beat from the evolution of homologous phases across irregularly spaced beats. These limitations motivate a beat-resolved representation that explicitly models the beat-specific ECG–PPG delay while separating within-beat morphology from across-beat dynamics.

We introduce BEATFORMER, a multimodal ECG–PPG foundation model that organizes synchronized cardiovascular signals as sequences of paired heartbeats. We choose ECG R peaks as shared cardiac anchors (Figure 1(b)) because they provide salient and relatively reliable electrical landmarks, and encode each R-anchored cycle as modality-specific phase tokens. Because peripheral PPG fiducials can be uncertain under motion, low perfusion, and sensor-dependent distortion, a delay-aware fusion module models beat-specific ECG–PPG timing uncertainty and uses it to regulate cross-modal information exchange. After fusion, the orthogonal beat–phase Transformer separates within-beat waveform morphology from across-beat rhythm and haemodynamic variation. Cumulative beat-time encoding and physiological relative-position biases preserve physical timing across irregular interbeat intervals. Together, the delay-aware fusion, beat–phase Transformer, and temporal encodings yield representations that remain relatively stable under these perturbations.

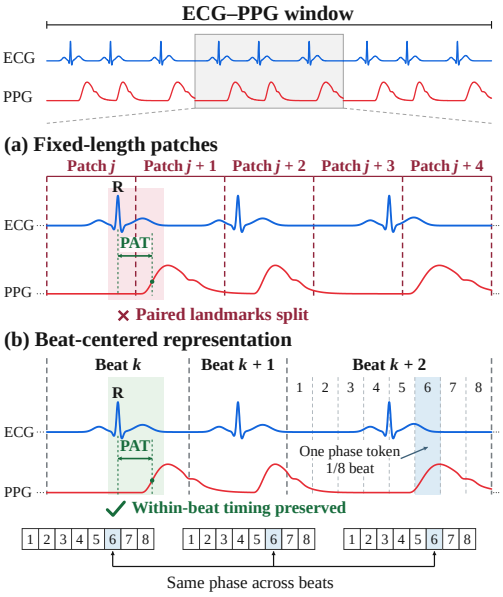


Figure 1: (a) Fixed-length patches can split paired ECG–PPG events and within-beat timing. (b) Beat-centered units preserve paired beats, within-beat timing and corresponding phases across beats.

We pretrain BEATFORMER using approximately 84.76 million valid cardiac cycles from 5 data sources. We compare it with six foundation models on 15 prediction tasks from 9 datasets, and evaluate waveform recovery on 11 masked-reconstruction conditions and 15 denoising conditions spanning 5 noise types and 3 severity levels. The most direct evidence for delay modeling comes from PAT estimation, where BEATFORMER achieves a mean absolute error of $11.48\,\mathrm{ms}$, 24.6% lower than the strongest baseline. Overall, it ranks first on 9 of 15 prediction tasks, attains the lowest RMSE in 10 of 11 reconstruction conditions, and yields the largest median signal-to-noise ratio (SNR) improvement in 13 of 15 denoising conditions. We summarize our contributions as follows:

- **Modeling the electrical-to-peripheral delay as a physiological quantity.** To our knowledge, BEATFORMER is the first multimodal cardiovascular foundation model to explicitly formulate paired ECG–PPG representations and incorporate beat-specific PAT supervision directly into its pretraining objective. This shared beat structure preserves the relative temporal displacement between each ECG event and its delayed PPG response.
- **Cross-modal fusion under uncertain vascular timing.** Rather than imposing a fixed delay or relying on a noisy PPG fiducial, BEATFORMER estimates a posterior over the delay at every beat and weights bidirectional information exchange by its confidence, which propagates timing information only where the observation supports it.

- **Disentangling waveform morphology from cardiovascular dynamics.** Morphology is defined within a cardiac cycle and rhythm emerges across cycles. BEATFORMER separates them through orthogonal phase and beat attention, and encodes cumulative beat time so that irregular intervals and transient rhythm changes are represented as elapsed time rather than ordinal position.

2 METHOD

BEATFORMER is a beat-resolved multimodal foundation model for learning transferable representations from synchronized ECG and PPG signals (Figure 2). The model consists of three stages. A beat-centered input module segments ECG and PPG into paired beats anchored on shared R peaks and encodes each beat as a sequence of modality-specific phase tokens. A delay-aware fusion module estimates a beat-wise ECG–PPG delay posterior and uses its confidence to gate bidirectional cross-modal attention. An orthogonal beat–phase Transformer applies phase-wise attention to capture within-beat morphology and beat-wise attention to model rhythm-dependent haemodynamic dynamics across homologous phases, followed by a two-level readout that produces beat- and window-level representations.

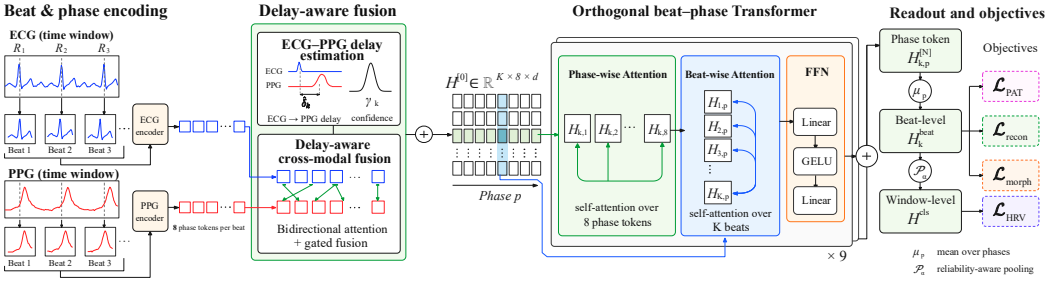


Figure 2: Overview of BEATFORMER.The model combines R-peak-anchored phase tokenization, delay-aware cross-modal fusion, orthogonal beat–phase contextualization, and a reliability-aware two-level readout. The resulting representations are optimized using physiology-informed pertaining objectives.

2.1 BEATFORMER ARCHITECTURE

2.1.1 BEAT-CENTERED INPUT AND PHASE TOKENIZATION.

Beat-centered input representation. To define an input unit in which ECG and PPG describe the same cardiac cycle, we use ECG R peaks as shared anchors. We first interpolate missing samples and filter ECG and PPG separately, and detect R peaks on the processed ECG. For each interior R peak R_i, the beat boundaries are the midpoints between adjacent R peaks:

$$s_i = \left\lfloor \frac{R_{i-1} + R_i}{2} \right\rfloor, \qquad e_i = \left\lfloor \frac{R_i + R_{i+1}}{2} \right\rfloor.$$

Applying the same interval to ECG and PPG yields paired beat segments that share the ECG R peak R_i as their cardiac anchor. Each segment is resampled to $L = 128$ samples and standardized independently within each beat and channel. Then we group K consecutive beats into one input window. The resulting beat from modality $r \in \{\mathrm{ECG}, \mathrm{PPG}\}$ is denoted by $x_k^{(r)} \in \mathbb{R}^L$. The complete procedure, including signal preprocessing, R-peak detection, beat-boundary determination, resampling, and within-beat standardization, is described in Appendix B.

Modality-specific phase encoders. Before any cross-modal interaction, each modality r uses its own 1D CNN encoder to map the normalized beat $x_k^{(r)}$ to $P = 8$ ordered content tokens $C_{k,p}^{(r)} \in \mathbb{R}^d$, where $d = 768$ is the shared token dimension. Because each beat has been resampled and standardized, these tokens no longer carry the absolute timing, duration, or amplitude of the original beat. We restore this information through two additive terms. The first is a normalized phase coordinate $\phi_p = (p - \frac{1}{2})/P$, which indexes the same relative position within every beat. The second is a modality-specific beat-context descriptor $\rho_{k,p}^{(r)} \in \mathbb{R}^6$ that summarizes local rhythm, beat duration, and

pre-normalization signal statistics. The resulting phase token is

$$Z_{k,p}^{(r)} = C_{k,p}^{(r)} + E_{\text{phase}}^{(r)}(\phi_p) + W_\rho^{(r)}\rho_{k,p}^{(r)},$$

where $E_{\text{phase}}^{(r)}$ embeds the phase coordinate and $W_\rho^{(r)}$ projects the descriptor into the token space (Appendix C.1).

2.1.2 DELAY-AWARE CROSS-MODAL FUSION.

Beat-wise delay posterior. Because ECG electrical activation and the corresponding peripheral pulse in PPG are separated by a beat-dependent delay related to pulse arrival time (PAT), we estimate a delay posterior before cross-modal fusion. We consider the admissible nonnegative phase-lag set $\mathcal{L}_{\text{lag}} = \{0, 1, \ldots, P-1\}$. For each $\ell \in \mathcal{L}_{\text{lag}}$, the ECG token at phase position p is compared with the PPG token at position $p+\ell$ over all valid overlapping positions using waveform-content tokens only. Let $\Psi_{k,p,q}$ denote the similarity between the ECG and PPG content tokens at phase indices p and q, respectively (Appendix C.2). The resulting delay posterior is

$$\pi_k(\ell) = \text{softmax}_{\ell\in\mathcal{L}_{\text{lag}}}\left(\frac{\sum_{p:\, p+\ell\le P}\Psi_{k,p,p+\ell}}{\eta}\right), \qquad T_k = \frac{e_k - s_k}{f_s},$$
$$\widehat{\delta}_k = \text{Mean}_{\pi_k}\left(\frac{T_k\ell}{P}\right), \qquad \gamma_k = \exp\left[-\frac{\text{Var}_{\pi_k}(T_k\ell/P)}{2s_\delta^2}\right].$$

Here, Mean_{π_k} and Var_{π_k} denote the posterior-weighted mean and variance over $\ell \in \mathcal{L}_{\text{lag}}$. Here, f_s denotes the sampling frequency of the corresponding recording, measured in samples per second (Hz). The quantity $T_k = (e_k - s_k)/f_s$ is the physical duration of beat k in seconds. Therefore, $T_k\ell/P$ is the physical delay corresponding to candidate phase lag ℓ. The posterior mean $\widehat{\delta}_k$ estimates the physical displacement of the PPG event relative to its ECG counterpart. Its dispersion yields the confidence γ_k, which is passed to the cross-modal attention to control how strongly the estimated delay is enforced. The posterior temperature $\eta > 0$ controls the concentration of π_k. The confidence scale $s_\delta > 0$ controls the sensitivity of γ_k to posterior dispersion. Both are learnable scalar parameters optimized jointly with the model.

Delay-conditioned information exchange. We use the estimated beat-specific delay to guide bidirectional cross-modal attention. ECG serves as the reference coordinate. The PPG coordinate is shifted by the normalized delay, and the coordinate mismatch defines an attention bias for query modality r and source modality s,

$$\chi_{k,p}^{\text{ECG}} = \phi_p, \qquad \chi_{k,q}^{\text{PPG}} = \phi_q - \frac{\widehat{\delta}_k}{T_k}, \qquad [B_k^{r\leftarrow s}]_{p,q} = -\gamma_k\big(\chi_{k,p}^{(r)} - \chi_{k,q}^{(s)}\big)^2,$$

so that phase positions describing the same physiological event share a coordinate, and mismatched pairs are penalized in proportion to the posterior confidence γ_k. Writing $Z_k^{(r)} \in \mathbb{R}^{P\times d}$ for the P phase tokens of beat k, each modality is updated from the other by cross-modal attention whose logits are biased toward delay-consistent pairs,

$$\widetilde{Z}_k^{(r)} = Z_k^{(r)} + \lambda_{\text{delay}}\,\text{softmax}\left(\frac{\big(Z_k^{(r)}W_Q^r\big)\big(Z_k^{(s)}W_K^s\big)^{\mathsf{T}}}{\sqrt{d}} + B_k^{r\leftarrow s}\right)Z_k^{(s)}W_V^s,$$

where W_Q^r, W_K^s, and W_V^s are learned modality-specific projections, and λ_{delay} is a learnable scalar. The dot-product term measures content compatibility, whereas B favors pairs consistent with the estimated delay: a concentrated posterior (large γ_k) enforces the delay strongly, whereas a diffuse one makes content dominate.

After bidirectional exchange, a learnable beat-level gate scores phase-pooled summaries from ECG and PPG. It normalizes the scores over available modalities and applies the resulting weights at every phase position: $H_{k,p} = \sum_{r:a_r=1}\omega_{k,r}\widetilde{Z}_{k,p}^{(r)}$. Here, $\omega_{k,r}$ denotes the weight produced by the gate for modality r at beat k, and $a_r \in \{0,1\}$ indicates whether modality r is available. Finally, we add a modality-shared embedding of each beat's position within the window,

$$H_{k,p}^{[0]} = H_{k,p} + E_{\text{time}}(t_k), \qquad t_k = \frac{R_k - R_1}{f_s},$$

where R_1 is the first retained R peak of the window and E_{time} is a sinusoidal-plus-linear embedding (Appendix C.3).

2.1.3 ORTHOGONAL BEAT–PHASE TRANSFORMER

Within a beat, the phase axis describes how different waveform portions jointly define cycle morphology. Across beats, the beat axis describes how a matched cycle portion changes over time. Applying a single attention operation to all KP tokens would impose the same interaction rule on every pair and conflate these two relations. We therefore factorize each block into within-beat phase attention followed by across-beat attention, and equip each with a physiological relative-position bias.

Physiological relative-position biases. Let $\tau_{k,p}$ denote the physical time of phase position p in beat k relative to its R peak, and RR_k its preceding R–R interval. Two learned functions g_{phase} and g_{beat} produce a $P \times P$ bias for each beat and a $K \times K$ bias for each phase position,

$$\left[B_k^{\text{phase}}\right]_{p,q} = g_{\text{phase}}\left(\phi_p - \phi_q,\, \tau_{k,p} - \tau_{k,q},\, \log T_k\right),$$
$$\left[B_p^{\text{beat}}\right]_{i,j} = g_{\text{beat}}\left(t_i - t_j,\, \log RR_i - \log RR_j,\, \phi_p\right).$$

The phase bias takes as input the normalized phase offset, the corresponding physical time offset in seconds, and the beat duration, so that the same relative offset can be weighted differently in long and short beats. The beat bias sees the physical separation between two beats, their local rhythm change, and the phase position at which they are compared, so that, for example, QRS-region and diastolic tokens can follow different across-beat interaction rules.

Block structure. Cross-modal fusion produces the fused beat–phase grid $H^{[0]} \in \mathbb{R}^{K\times P\times d}$. Write $H_{k,\cdot} \in \mathbb{R}^{P\times d}$ for the phase tokens of beat k and $H_{\cdot,p} \in \mathbb{R}^{K\times d}$ for the beat tokens at phase p. We proposed two biased self-attention operators over these slices,

$$\texttt{PhaseAttn}^{(h)}(H_{k,\cdot}) = \text{softmax}_q\left(\frac{\left(H_{k,\cdot}W_Q^h\right)\left(H_{k,\cdot}W_K^h\right)^\top}{\sqrt{d_h}} + \lambda_{\text{phase}}B_k^{\text{phase}}\right)H_{k,\cdot}W_V^h,$$
$$\texttt{BeatAttn}^{(h)}(H_{\cdot,p}) = \text{softmax}_j\left(\frac{\left(H_{\cdot,p}W_Q^h\right)\left(H_{\cdot,p}W_K^h\right)^\top}{\sqrt{d_h}} + \lambda_{\text{beat}}B_p^{\text{beat}} + M\right)H_{\cdot,p}W_V^h.$$

where $h = 1, \ldots, N_h$ indexes the attention head, $d_h = d/N_h$, and $M_{i,j} = 0$ for a valid beat slot j and $-\infty$ for a padded one. The coefficients λ_{phase} and λ_{beat} are learnable scalar parameters jointly optimized with the model. One orthogonal block applies them in sequence with pre-norm residuals and a position-wise feed-forward network (FFN),

$$\begin{aligned}
H'_{k,\cdot} &= H^{[n]}_{k,\cdot} + \texttt{MultiHeadPhaseAttn}\big(\text{LN}(H^{[n]}_{k,\cdot})\big), && k = 1, \ldots, K,\\
H''_{\cdot,p} &= H'_{\cdot,p} + \texttt{MultiHeadBeatAttn}\big(\text{LN}(H'_{\cdot,p})\big), && p = 1, \ldots, P,\\
H^{[n+1]} &= H'' + \text{FFN}\big(\text{LN}(H'')\big), && n = 0, \ldots, N-1,
\end{aligned}$$

where LN denotes layer normalization. We stack $N = 9$ blocks. A two-level readout then pools the final grid over phase to obtain one representation per beat, H_k^{beat}, and applies reliability-aware pooling over valid beats to obtain the window-level representation H^{cls} (Appendix C.5).

2.2 PRE-TRAINING OBJECTIVES

Pre-training supervises BEATFORMER at three physiological scales: waveform samples, individual beats, and analysis windows. Each objective reads one of the representations defined above and regresses a physiologically defined target (Appendix D).

Masked reconstruction. A reconstruction head restores the masked ECG and PPG samples of beat k from H_k^{beat}, preserving fine-grained morphology that the regression targets below do not capture.

Pulse arrival time. A PAT head predicts pulse arrival time from H_k^{beat}, using as the target the interval from the ECG R peak to the point of maximum first derivative on the filtered PPG upstroke.

Beat morphology. A morphology head predicts six descriptors of the pre-normalization PPG pulse from H_k^{beat}: amplitude, half-amplitude width, rise time, normalized area, peak position, and maximum upstroke slope.

Heart-rate variability (HRV). An HRV head predicts the standard deviation of normal-to-normal intervals (SDNN), the root mean square of successive differences (RMSSD), and mean heart rate from the pooled window-level representation H^{cls}.

All four objectives use the masked mean squared error $\mathcal{L}_u = \mathrm{mean}_{i:m_i^{(u)}=1}[(\widehat{y}_i^{(u)} - y_i^{(u)})^2]$, where $u \in \{\mathrm{recon}, \mathrm{PAT}, \mathrm{morph}, \mathrm{HRV}\}$ and i ranges over the entries supervised by task u (masked samples, beats, or windows). The pre-training objective is their weighted sum,

$$\mathcal{L}_{\mathrm{pre}} = \mathcal{L}_{\mathrm{recon}} + \lambda_{\mathrm{PAT}}\mathcal{L}_{\mathrm{PAT}} + \lambda_{\mathrm{morph}}\mathcal{L}_{\mathrm{morph}} + \lambda_{\mathrm{HRV}}\mathcal{L}_{\mathrm{HRV}},$$

with $\lambda_{\mathrm{PAT}} = 0.5$ and $\lambda_{\mathrm{morph}} = \lambda_{\mathrm{HRV}} = 0.2$ throughout our experiments.

3 EXPERIMENTS

Training datasets. We pretrain on paired ECG–PPG recordings from five source cohorts (Table 1). These include intensive-care MIMIC-III and intraoperative VitalDB recordings in PulseDB (Johnson et al., 2016; Lee et al., 2022; Wang et al., 2023), intensive-care MIMIC-II recordings in UCI (Saeed et al., 2011; UCI Machine Learning Repository, 2015), and two in-house datasets collected by our team, SYSU and TPTCOM, spanning hospital-based and real-world patient-monitoring settings. Together, the corpus contains approximately 5.26 million paired segments and 84.76 million valid cardiac cycles.

Table 1: Pretraining datasets.

Dataset	Participants	Segments	Beats
PulseDB–MIMIC	2,423	3,791,004	59,773,387
PulseDB–Vital	2,938	1,454,450	24,060,392
UCI	942	12,000	660,908
SYSU	169	2,237	142,317
TPTCOM	114	2,005	124,088

Experimental setup. BEATFORMER contains approximately 73.8 million parameters. We pretrain a single checkpoint for approximately 40 hours using two NVIDIA A800 GPUs (80GB each). All downstream prediction tasks use the same pretrained backbone. For classification and regression, we freeze the backbone and train a task-specific predictor. For waveform reconstruction and denoising, we jointly fine-tune the pretrained backbone and a trainable recovery head. .

Downstream tasks. We evaluate BEATFORMER on classification and regression, masked waveform reconstruction, and signal recovery under controlled noise. For all tasks, adaptation and test sets are subject-disjoint, and test participants are excluded from pretraining and task-specific adaptation.

The downstream prediction benchmark comprises 15 classification and regression tasks across nine datasets. Table 2 summarizes their prediction targets and evaluation metrics. Dataset details, evaluation sizes, and split protocols are provided in Appendix E.1.

Table 2: Downstream classification and regression tasks.

Dataset	Task	Metric
PTT-PPG	Heart rate	MAE (bpm)
PTT-PPG	Activity	Accuracy
Simultaneous	Heart rate	MAE (bpm)
Simultaneous	Load	Accuracy
Li ECG–PPG	PAT	Subject-macro MAE (ms)
Li ECG–PPG	Systolic BP	MAE (mmHg)
Li ECG–PPG	Diastolic BP	MAE (mmHg)
BUT PPG	Signal quality	AUROC
WESAD	Stress	AUROC
PPG-DaLiA	Activity	Accuracy
PPG-DaLiA	Age	MAE (years)
PPG-DaLiA	BMI	MAE (kg/m^2)
ScientISST	Activity	Accuracy
SensSmartTech	Before/after activity	AUROC
VTaC	VT false alarm	AUROC

Reconstruction and denoising share 2,654 paired 10-second ECG–PPG test windows from 214 VTaC participants (Lehman et al., 2024). Test signals are segmented into beats and normalized before masking or corruption. Reconstruction has 11 conditions: one to seven contiguous phase bins are masked at corresponding ECG and PPG positions within a target beat, or gaps of 0.5, 1, 2, or 4 s without regard to beat boundaries, allowing them to span adjacent beats. Root mean square error (RMSE) is computed only at missing positions. The reconstruction masking procedures are detailed in Appendix E.2.

For denoising, the target beat is corrupted with Gaussian noise, baseline wander, narrowband interference, motion bursts, or mixed noise at mild, moderate, and severe levels, while neighboring beats remain clean. We evaluate window-level signal-to-noise ratio (SNR) improvement and beat-level peak-to-peak amplitude error on the target beat, and separately measure distortion introduced by the recovery model on uncorrupted inputs. Noise types and severity parameters are summarized in Appendix Table E.3.

Baseline models. For downstream prediction, we compare with six pretrained baselines: AnyPPG (Nie et al., 2025), Pulse-PPG (Saha et al., 2025), PaPaGei-P, PaPaGei-S (Pillai et al.,

2025), SIGMA-PPG (Guo et al., 2026), and CSFM (Gu et al., 2026). AnyPPG uses synchronized ECG to guide PPG representation learning. Pulse-PPG applies relative contrastive learning to PPG collected in everyday settings. PaPaGei-P treats segments from the same participant as positive pairs, while PaPaGei-S incorporates PPG morphology and signal-quality objectives. SIGMA-PPG combines prior-guided adversarial masking with vector-quantized semantic consistency. CSFM uses masked generative pretraining and supports ECG, PPG, and paired ECG–PPG inputs. In our downstream prediction evaluation, BEATFORMER and CSFM use paired ECG–PPG, whereas the other baselines use PPG. The corresponding model and predictor settings for these classification and regression tasks are provided in Appendix F.1.

For reconstruction and denoising, we compare with CSFM, VCT-M2AE (Hu et al., 2025), which uses cross-modal masked reconstruction, MOMENT-1-small (Goswami et al., 2024), a general-purpose time-series model, and Small Transformer trained from random initialization, used as a non-pretrained reference. We evaluate only foundation models with publicly available pretrained weights, and thus exclude the previously mentioned xMAE, GPT-PPG, and CardioFM because their weights were unavailable. All models use the same data splits and test windows. The corresponding recovery input and adaptation settings are provided in Appendix F.2.

3.1 RESULTS

Table 3: Comparison of BEATFORMER and six pretrained models on 15 physiological-signal classification and regression tasks. Results are reported as mean ± standard deviation across outer test folds. Training and test sets are subject-disjoint for all tasks. Boldface and underlining indicate the best and second-best mean results for each task, respectively.

Task	Metric	BEATFORMER	AnyPPG	Pulse-PPG	CSFM	PaPaGei-S	PaPaGei-P	SIGMA-PPG
Physiological signal assessment								
PTT PPG HR	MAE (bpm)↓	**4.025±3.177**	5.623±4.652	6.798±4.190	4.209±5.241	9.702±3.602	7.424±4.281	8.197±3.955
Simultaneous HR	MAE (bpm)↓	**9.861±4.114**	15.711±5.848	19.448±11.182	11.050±3.892	20.735±5.226	20.235±11.342	19.698±6.601
Li ECG–PPG PAT	MAE (ms)↓	**11.475±2.185**	16.418±2.503	16.929±2.682	15.220±2.924	20.334±2.388	22.263±4.465	19.616±2.477
BUT PPG Quality	AUROC↑	**0.839±0.033**	0.801±0.040	0.798±0.022	0.757±0.044	0.642±0.053	0.731±0.035	0.707±0.035
State and activity recognition								
WESAD Stress	AUROC↑	0.919±0.027	**0.922±0.026**	0.882±0.033	0.913±0.046	0.768±0.048	0.841±0.037	0.838±0.030
PTT PPG Activity	Acc.↑	0.680±0.145	0.550±0.129	0.537±0.131	**0.732±0.133**	0.510±0.127	0.550±0.107	0.535±0.147
PPG-DaLiA Activity	Acc.↑	**0.517±0.108**	0.457±0.067	0.413±0.084	0.492±0.099	0.356±0.096	0.380±0.085	0.360±0.091
ScientISST Activity	Acc.↑	0.656±0.245	**0.674±0.207**	0.666±0.210	0.653±0.238	0.668±0.206	0.660±0.187	0.670±0.206
Simultaneous Load	Acc.↑	**0.569±0.098**	0.452±0.082	0.466±0.079	0.568±0.074	0.454±0.060	0.464±0.089	0.434±0.051
SensSmartTech Before/after activity	AUROC↑	0.871±0.049	0.787±0.085	0.770±0.072	**0.907±0.046**	0.620±0.055	0.810±0.045	0.766±0.104
VTaC VT-FalseAlarm	AUROC↑	0.843±0.035	0.780±0.043	0.712±0.018	**0.862±0.052**	0.665±0.037	0.695±0.021	0.596±0.074
Demographic regression								
PPG-DaLiA Age	MAE (years)↓	**6.449±3.096**	6.810±2.177	8.921±1.967	8.707±1.278	8.936±3.686	7.592±2.855	7.882±2.655
PPG-DaLiA BMI	MAE (kg/m^2)↓	1.459±0.523	**1.309±0.522**	1.605±0.397	1.628±0.538	1.972±0.425	1.588±0.347	1.348±0.608
Blood pressure regression								
Li ECG–PPG SBP	MAE (mmHg)↓	**10.752±1.474**	10.975±1.528	11.826±2.140	11.376±1.283	11.071±1.908	11.242±2.096	10.887±1.969
Li ECG–PPG DBP	MAE (mmHg)↓	**7.114±0.505**	7.206±0.896	7.526±0.979	7.708±0.718	7.802±1.021	7.704±1.078	7.140±0.880

Physiological-signal classification and regression. With the backbone frozen, BEATFORMER achieves the best mean result on 9 of 15 tasks and ranks among the top two on 13 (Table 3). On Simultaneous heart-rate estimation, its MAE is 9.861 bpm, 10.8% lower than that of the strongest baseline, CSFM. On Li ECG–PPG PAT prediction, its subject-macro MAE is 11.475 ms, 24.6% lower than that of CSFM. On BUT PPG signal-quality assessment, its AUROC is 0.839, compared with 0.801 for the second-best model, AnyPPG. On three additional PAT datasets, BEATFORMER achieved the lowest MAE on PTT-PPG and UQ (18.945 and 87.500 ms, respectively). On ErgoBP, its MAE was 93.089 ms, higher than the 78.868 ms achieved by the best-performing model, Pulse-PPG (Appendix Table G.3). The Bland–Altman plots show small mean biases on PTT-PPG and Li ECG–PPG, whereas errors are more dispersed on UQ and ErgoBP (Appendix Figure G.6). Representation aggregation and task metric definitions are provided in Appendix Sections F.3 and F.4.

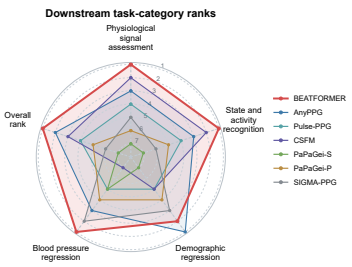


Figure 3: Ranks across four downstream task categories and overall, based on mean task ranks. Outer vertices indicate better ranks.

The rank radar chart (Figure 3) ranks models by their mean task rank within each category. At the category level, BEATFORMER ranks first in physiological signal assessment, state and activity recognition, and blood pressure regression, and second in demographic regression. Its mean rank across all 15 tasks is 1.73, ahead of the second-ranked AnyPPG (2.77). These results show that its learned ECG–PPG representations transfer across diverse physiological-signal tasks.

Masked waveform reconstruction. BEATFORMER achieved the lowest RMSE in 10 of 11 reconstruction conditions. Its equally weighted mean RMSE across conditions was 0.510, 11.3% lower than that of the strongest baseline, VCT-M2AE (0.575). Under within-beat contiguous masking, it had the lowest RMSE with one to six missing phase bins (Figure 4(a)). Its mean across all seven settings was 0.433, 15.3% lower than VCT-M2AE (0.511). With seven bins masked and only one visible, VCT-M2AE had lower RMSE (0.544 versus 0.561). For 0.5-, 1-, 2-, and 4-second gaps placed independently of beat boundaries, BEATFORMER had the lowest RMSE at every length, with a mean of 0.644, 6.0% lower than CSFM (0.685; Figure 4(b)). Figure 4(c) shows three missingness patterns in the same test window. Across the seven within-beat settings, its mean ECG and PPG RMSEs were 0.542 and 0.280, below the strongest channel-specific baselines, VCT-M2AE (0.591) and MOMENT-1-small (0.362), respectively (Figure 4(d)). Window-level error distributions and whole-beat masking results appear in Figure G.2 and Table G.2, respectively. Evaluation metrics for waveform reconstruction and denoising are defined in Appendix F.5.

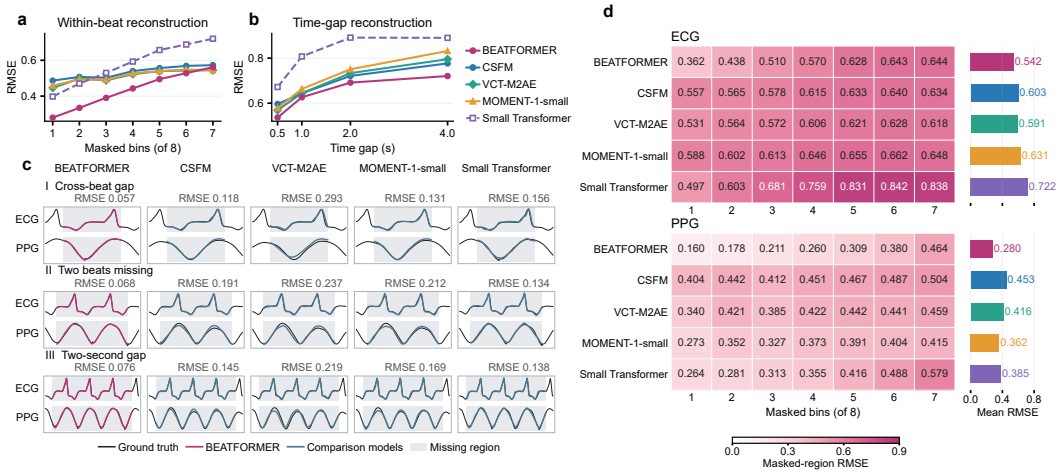


Figure 4: ECG–PPG waveform reconstruction under different missingness patterns. (a) Reconstruction errors of five models with one to seven contiguous within-beat phase bins masked. Each beat is divided into eight phase bins. (b) Reconstruction errors for time gaps of 0.5, 1, 2, and 4 seconds. (c) Fixed-window ECG and PPG examples for three missingness patterns; gray regions mark missing samples. All RMSE values are computed over missing regions; lower values indicate smaller reconstruction errors. (d) Separate ECG and PPG results for the seven within-beat missingness settings. The bars on the right report each model's average RMSE across these settings.

Signal recovery under controlled noise. ECG and PPG recordings are susceptible to artifacts during acquisition, motivating evaluation across controlled noise types and severities. Denoising performance is summarized as median sample-level SNR improvement, measured in decibels (dB). Its window-level definition and condition-level aggregation are given in Appendix F.5. BEATFORMER achieved a positive median SNR improvement in all 15 conditions and the highest median among the five models in 13 (Table 4). It ranked first in all nine conditions involving baseline wander, motion bursts, or mixed noise. Under severe baseline wander, its improvement was 8.83 dB, compared with 6.74 dB for Small Transformer and 3.02 dB for VCT-M2AE; under severe motion bursts, the respective values were 13.77, 12.24, and 4.81 dB. In addition, BEATFORMER achieves the highest median window-level SNR improvement across all aggregated noise-type and severity groups. The median gain is largest for motion bursts and smallest for narrowband interference (Figure 5(a)), while the full distribution shifts upward from mild to severe corruption (Figure 5(b)). Condition-wise distributions appear in Appendix Figure G.3.

Across the five severe noise types, we compared ECG and PPG peak-to-peak amplitude errors on the target beat (Appendix Figure G.4). BEATFORMER had the highest median SNR improvement

Severity	BEATFORMER	CSFM	VCT-M2AE	MOMENT-1-small	Small Transformer
Gaussian noise					
Mild	**6.92**	−8.60	−0.48	−1.62	6.87
Moderate	**8.36**	−4.55	2.65	1.32	8.14
Severe	**9.58**	−1.13	5.43	3.72	9.25
Baseline wander					
Mild	**4.16**	−4.74	0.61	0.25	2.73
Moderate	**6.58**	−1.39	2.06	1.78	4.83
Severe	**8.83**	1.48	3.02	2.78	6.74
Narrowband interference					
Mild	1.65	−14.66	−5.95	−7.00	**2.24**
Moderate	3.78	−10.49	−2.39	−3.34	**4.03**
Severe	**5.98**	−6.25	1.00	−0.14	5.92
Motion bursts					
Mild	**9.25**	−6.08	0.85	0.02	7.95
Moderate	**11.69**	−2.69	2.94	2.28	10.08
Severe	**13.77**	0.84	4.81	4.03	12.24
Mixed noise					
Mild	**5.00**	−6.51	0.53	−0.15	4.13
Moderate	**6.15**	−2.67	2.51	1.84	5.26
Severe	**7.34**	0.85	3.94	3.37	6.48

Table 4: Denoising results across five noise types and three severity levels. Each entry is the median SNR improvement over 2,654 test windows, in decibels (dB); higher is better, and bold indicates the highest value for that condition. Models are evaluated using their native input processing and scoring protocols.

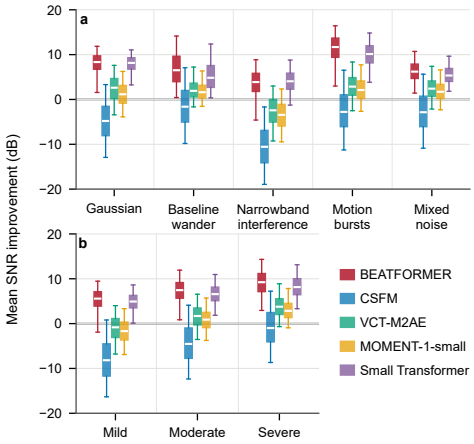


Figure 5: SNR improvement summarized by noise type and severity. (a) For each window, values are averaged across the three severity levels within each noise type. (b) For each window, values are averaged across the five noise types at each severity level. Each group contains 2,654 windows.

among the five models under each noise type and reduced the mean absolute amplitude error for both ECG and PPG in all five. Under severe Gaussian noise, its mean absolute amplitude error decreased from 0.435 to 0.130 for ECG and from 0.725 to 0.102 for PPG, in normalized amplitude units. These results show that the SNR gains were accompanied by smaller peak-to-peak amplitude errors in both signals. In the fixed-window examples (Appendix Figure G.5), BEATFORMER recovered the main ECG peaks and PPG contours under severe baseline wander, motion bursts, and mixed noise.

Ablation analysis. Removing any component increased MAE on all four tasks (Table 5). Morphology supervision had the largest effect on the two heart-rate tasks, increasing their MAEs by 45.0% and 55.9%, respectively. Removing cumulative-time, PAT, or HRV supervision also increased the heart-rate MAEs by 15.2%–37.4%. For SBP and DBP, all ablations increased MAE by 2.7%–14.5% and 7.7%–12.4%, respectively. Additional ablation results and sensitivity analyses are provided in Appendices G.6 and G.7, respectively.

Table 5: Percentage change in mean MAE after removing each component. Positive values indicate higher error.

Configuration	Simultaneous HR	PTT PPG HR	Li ECG–PPG SBP	Li ECG–PPG DBP
Without within-beat phase attention	+8.2%	+13.6%	+3.1%	+12.4%
Without across-beat attention	+3.2%	+14.9%	+2.7%	+7.8%
Without cumulative-time encoding	+19.1%	+22.3%	+8.4%	+8.2%
Without PAT supervision	+17.9%	+30.0%	+5.1%	+9.8%
Without morphology supervision	+45.0%	+55.9%	+14.5%	+12.1%
Without HRV supervision	+15.2%	+37.4%	+10.3%	+7.7%

4 Conclusion

We introduced BEATFORMER, a beat-resolved ECG–PPG foundation model that represents synchronized signals as paired heartbeats anchored on ECG R peaks. It estimates beat-specific ECG–PPG delay, uses beat-level PAT supervision during pretraining, and gates cross-modal fusion according to delay confidence. An orthogonal beat–phase Transformer then separates within-beat morphology from across-beat dynamics using physical-time positional biases. Pretrained on approximately 84.76 million valid cardiac cycles from paired ECG–PPG recordings, BEATFORMER achieved the best mean result on 9 of 15 prediction tasks, including a PAT mean absolute error of 11.48 ms, 24.6% lower than the strongest baseline. After adaptation, it achieved the lowest RMSE in 10 of 11 reconstruction conditions and the highest median SNR improvement in 13 of 15 denoising conditions. Together with the ablation results, these findings support incorporating beat structure, within-beat position, and cross-modal delay relationships into ECG–PPG representation learning for physiological prediction and waveform recovery.

AI USE STATEMENT

In this work, large language models (LLMs) were used primarily to support language editing and polishing, the retrieval and identification of related work, and limited assistance with editing, documenting, and organizing the public software release. AI-assisted literature suggestions and citations were checked against the original sources, and AI-assisted code was reviewed and tested by the authors. LLMs were not used to analyze or interpret the reported experimental results, generate synthetic datasets, or formulate or prove mathematical claims.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility of this work. A detailed description of the BEATFORMER architecture, pre-training objectives, experimental settings, data-processing procedures, and evaluation metrics is provided in Sections 2–3, the appendices, and the supplementary material. We release anonymized code and configuration files at `https://anonymous.4open.science/r/anonymous-model-release-6541/`. The repository contains the model definition, ECG–PPG preprocessing and beat-segmentation code, pre-training and training scripts, and downstream task-head implementations. The publicly available datasets used in this work are identified and properly cited in the paper. The SYSU and TPTCOM datasets contain non-public human physiological data and are therefore not redistributed. Trained model checkpoints will be released after acceptance.

ETHICS STATEMENT

The SYSU and TPTCOM datasets consist of recordings collected from human participants and used in this study. All participants provided informed consent before data collection. The recordings were de-identified and processed to protect participant privacy before analysis. Because these datasets contain private human physiological recordings, the raw data are not publicly released. The other datasets used in this study are publicly available and were used in accordance with their respective terms of use, with the original sources properly cited.

REFERENCES

Bella Eszter Ajtay, Szabolcs Béres, and László Hejjel. The oscillating pulse arrival time as a physiological explanation regarding the difference between ECG- and photoplethysmogram-derived heart rate variability parameters. *Biomedical Signal Processing and Control*, 79:104033, 2023. doi: 10.1016/j.bspc.2022.104033. URL https://doi.org/10.1016/j.bspc.2022.104033.

João Areias Saraiva, Mariana Abreu, Ana Sofia Carmo, Hugo Plácido da Silva, and Ana Fred. ScientISST MOVE: Annotated wearable multimodal biosignals recorded during everyday life activities in naturalistic environments. PhysioNet, version 1.0.1, 2024. URL https://doi.org/10.13026/hyxq-r919.

Joseph Azar, Abdallah Makhoul, Raphaël Couturier, and Jacques Demerjian. Deep recurrent neural network-based autoencoder for photoplethysmogram artifacts filtering. *Computers and Electrical Engineering*, 92:107065, 2021. doi: 10.1016/j.compeleceng.2021.107065. URL https://doi.org/10.1016/j.compeleceng.2021.107065.

Karim Bayoumy, Mohammed Gaber, Abdallah Elshafeey, Omar Mhaimeed, Elizabeth H. Dineen, Francoise A. Marvel, Seth S. Martin, Evan D. Muse, Mintu P. Turakhia, Khaldoun G. Tarakji, and Mohamed B. Elshazly. Smart wearable devices in cardiovascular care: Where we are and how to move forward. *Nature Reviews Cardiology*, 18(8):581–599, 2021. doi: 10.1038/s41569-021-00522-7.

J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986. doi: 10.1016/S0140-6736(86)90837-8.

Xiangmao Chang, Gangkai Li, Linlin Tu, Guoliang Xing, and Tian Hao. DeepHeart: Accurate heart rate estimation from PPG signals based on deep learning. In *2019 IEEE 16th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, pp. 371–379, 2019. doi: 10.1109/MASS.2019.00051. URL https://doi.org/10.1109/MASS.2019.00051.

Peter H. Charlton, Panicos A. Kyriacou, Jonathan Mant, Vaidotas Marozas, Phil Chowienczyk, and Jordi Alastruey. Wearable photoplethysmography for cardiovascular monitoring. *Proceedings of the IEEE*, 110(3):355–381, 2022. doi: 10.1109/JPROC.2022.3149785.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Zehua Chen, Yuyang Miao, Liyuan Wang, Luyun Fan, Danilo P. Mandic, and Jun Zhu. Versatile cardiovascular signal generation with a unified diffusion transformer. *Nature Machine Intelligence*, 8(1):6–19, 2026. doi: 10.1038/s42256-025-01147-y. URL https://doi.org/10.1038/s42256-025-01147-y.

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. URL https://doi.org/10.1088/1361-6579/add988.

Hong-Yu Chiu, Hong-Han Shuai, and Paul C.-P. Chao. Reconstructing qrs complex from ppg by transformed attentional neural networks. *IEEE Sensors Journal*, 20(20):12374–12383, 2020. doi: 10.1109/JSEN.2020.3000344. URL https://doi.org/10.1109/JSEN.2020.3000344.

I Chiu, Yu-Tung Liu, Kuan-Chen Wang, Hung-Yu Wei, and Yu Tsao. Robust photoplethysmography signal denoising via mamba networks, 2025. URL https://arxiv.org/abs/2510.11058. Preprint accepted to EMBC 2026.

Zhengyao Ding, Yujian Hu, Ziyu Li, Yiheng Mao, Haitao Li, Dongchen Zhou, Xuesen Chu, Long Yu, Ziyi Liu, Fei Wu, Hongkun Zhang, Qingbo Xu, Ting Chen, and Zhengxing Huang. Ai modeling photoplethysmography to electrocardiography useful for predicting cardiovascular disease. *npj Digital Medicine*, 9:61, 2026. doi: 10.1038/s41746-025-02228-3. URL https://doi.org/10.1038/s41746-025-02228-3.

Eoin Finnegan, Shaun Davidson, Mirae Harford, João Jorge, Peter Watkinson, Duncan Young, Lionel Tarassenko, and Mauricio Villarroel. Pulse arrival time as a surrogate of blood pressure. *Scientific Reports*, 11(1):22767, 2021. doi: 10.1038/s41598-021-01358-4. URL https://doi.org/10.1038/s41598-021-01358-4.

Daniel Franklin, Angeliki Tzavelis, J. Y. Lee, et al. Synchronized wearables for the detection of haemodynamic states via electrocardiography and multispectral photoplethysmography. *Nature Biomedical Engineering*, 7:1229–1241, 2023. doi: 10.1038/s41551-023-01098-y.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Xiao Gu, Wei Tang, Jinpei Han, Veer Sangha, Fenglin Liu, Shreyank N. Gowda, Antonio H. Ribeiro, Patrick Schwab, Kim Branson, Lei Clifton, Antonio Luiz P. Ribeiro, Zhangdaihong Liu, and David A. Clifton. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, 8(2):220–233, 2026. doi: 10.1038/s42256-026-01180-5. URL https://doi.org/10.1038/s42256-026-01180-5.

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, 2026. URL https://arxiv.org/abs/2601.21031.

Md Hassanuzzaman, Tilendra Choudhary, Alasdair Gent, Mihai V. Podgoreanu, Suresh Mitu Agarwal, Vijay Krishnamoorthy, Sivasubramanium Bhavani, Philip Yang, Annette Esper, and Rishikesan Kamaleswaran. Cardiofm: A multimodal foundation model for joint ecg and ppg representation learning. Research Square preprint, 2026. URL https://doi.org/10.21203/rs.3.rs-9652631/v1.

Qihan Hu, Daomiao Wang, Hong Wu, Jian Liu, and Cuiwei Yang. Efficient multi-view fusion and flexible adaptation to view missing in cardiovascular system signals. *Neural Networks*, 181:106760, 2025. doi: 10.1016/j.neunet.2024.106760. URL https://doi.org/10.1016/j.neunet.2024.106760.

Pranay Jain, Cheng Ding, Cynthia Rudin, and Xiao Hu. A self-supervised algorithm for denoising photoplethysmography signals for heart rate estimation from wearables. *Harvard Data Science Review*, 6(3), 2024. doi: 10.1162/99608f92.8636cb81. URL https://doi.org/10.1162/99608f92.8636cb81.

Alistair E. W. Johnson et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3: 160035, 2016. doi: 10.1038/sdata.2016.35.

Ella Lan. Performer: A novel ppg-to-ecg reconstruction transformer for a digital biomarker of cardiovascular disease detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1991–1999, 2023. doi: 10.1109/WACV56688.2023.00203. URL https://openaccess.thecvf.com/content/WACV2023/html/Lan_Performer_A_Novel_PPG-to-ECG_Reconstruction_Transformer_for_a_Digital_Biomarker_WACV_2023_paper.html.

Aleksandar Lazović, Predrag Tadić, et al. SensSmartTech database of cardiovascular signals synchronously recorded by an electrocardiograph, phonocardiograph, photoplethysmograph and accelerometer. PhysioNet, version 1.0.0, 2024. URL https://doi.org/10.13026/fy9p-n277.

Hyung-Chul Lee et al. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9:279, 2022. doi: 10.1038/s41597-022-01411-5.

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

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.

David Liu, Matthias Görges, and Simon A. Jenkins. University of queensland vital signs dataset: Development of an accessible repository of anesthesia patient monitoring data for research. *Anesthesia & Analgesia*, 114(3):584–589, 2012. doi: 10.1213/ANE.0b013e318241f7c0.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *JAMIA Open*, 8(5):ooaf122, 2025. doi: 10.1093/jamiaopen/ooaf122.

Philipp Mehrgardt, Matloob Khushi, Stephen Poon, and Anusha Withana. Pulse transit time PPG dataset. PhysioNet, version 1.0.0, 2022. URL https://doi.org/10.13026/g3me-rt62.

Fahimeh Mohagheghian, Dong Han, Om Ghetia, Andrew Peitzsch, Nishat Nishita, Mahdi Pirayesh Shirazi Nejad, Eric Y. Ding, Kamran Noorishirazi, Alexander Hamel, Edith Mensah Otabil, Danielle DiMezza, Emily L. Dickson, Khanh-Van Tran, David D. McManus, and Ki H. Chon. Noise reduction in photoplethysmography signals using a convolutional denoising autoencoder with unconventional training scheme. *IEEE Transactions on Biomedical Engineering*, 71(2): 456–466, 2024. doi: 10.1109/TBME.2023.3307400. URL https://doi.org/10.1109/TBME.2023.3307400.

Ramakrishna Mukkamala, Jin-Oh Hahn, Omer T. Inan, Lalit K. Mestha, Chang-Sei Kim, Hakan Töreyin, and Survi Kyal. Toward ubiquitous blood pressure monitoring via pulse transit time: Theory and practice. *IEEE Transactions on Biomedical Engineering*, 62(8):1879–1901, 2015. doi: 10.1109/TBME.2015.2441951. URL https://doi.org/10.1109/TBME.2015.2441951.

Andrea Nemcova, Eniko Vargova, Radovan Smisek, Lucie Marsanova, Lukas Smital, and Martin Vitek. Brno university of technology smartphone PPG database (BUT PPG): Annotated dataset for PPG quality assessment and heart rate estimation. *BioMed Research International*, 2021:3453007, 2021. doi: 10.1155/2021/3453007. URL https://doi.org/10.1155/2021/3453007.

Andrea Nemcova, Radovan Smisek, Eniko Vargova, et al. Brno university of technology smartphone PPG database (BUT PPG). PhysioNet, version 2.0.0, 2024. URL https://doi.org/10.13026/tn53-8153.

Guangkun Nie, Gongzheng Tang, Yujie Xiao, Jun Li, Shun Huang, Deyun Zhang, Qinghao Zhao, and Shenda Hong. Anyppg: An ecg-guided ppg foundation model trained on over 100,000 hours of recordings for holistic health profiling, 2025. URL https://arxiv.org/abs/2511.01747.

Jiapu Pan and Willis J. Tompkins. A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985. doi: 10.1109/TBME.1985.325532. URL https://doi.org/10.1109/TBME.1985.325532.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kYwTmlq6Vn.

Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019. doi: 10.3390/s19143079.

Mohammed Saeed et al. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39(5):952–960, 2011. doi: 10.1097/CCM.0b013e31820a92c6.

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, 2025. URL https://arxiv.org/abs/2502.01108.

Pritam Sarkar and Ali Etemad. Cardiogan: Attentive generative adversarial network with dual discriminators for synthesis of ecg from ppg. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 488–496, 2021. doi: 10.1609/aaai.v35i1.16126. URL https://doi.org/10.1609/aaai.v35i1.16126.

Filip Emil Schjerven, Torjus Lines Steffensen, Frank Lindseth, Hans Martin Flade, and Idar Kirkeby-Garstad. Descriptor: Invasive blood pressure with multimodal physiological signals during ergometer exercise dataset (ErgoBP). *IEEE Data Descriptions*, 3:100–108, 2026. doi: 10.1109/IEEEDATA.2025.3650442.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 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. URL https://arxiv.org/abs/2608.12944.

Monalisa Singha Roy, Rajarshi Gupta, Jayanta K. Chandra, Kaushik Das Sharma, and Arunansu Talukdar. Improving photoplethysmographic measurements under motion artifacts using artificial neural network for personal healthcare. *IEEE Transactions on Instrumentation and Measurement*, 67(12):2820–2829, 2018. doi: 10.1109/TIM.2018.2829488. URL https://doi.org/10.1109/TIM.2018.2829488.

Ananya Tripathi, Yurie Obata, Pavel Ruzankin, Narwan Askaryar, Dan E. Berkowitz, Jochen Steppan, and Viachaslau Barodka. A pulse wave velocity based method to assess the mean arterial blood pressure limits of autoregulation in peripheral arteries. *Frontiers in Physiology*, 8:855, 2017. doi: 10.3389/fphys.2017.00855. URL https://doi.org/10.3389/fphys.2017.00855.

UCI Machine Learning Repository. Cuff-less blood pressure estimation dataset, 2015. URL https://doi.org/10.24432/C5B602.

Muthiah Vaduganathan, George A. Mensah, Justine Varieur Turco, Valentin Fuster, and Gregory A. Roth. The global burden of cardiovascular diseases and risk: A compass for future health. *Journal of the American College of Cardiology*, 80(25):2361–2371, 2022. doi: 10.1016/j.jacc.2022.11.005. URL https://doi.org/10.1016/j.jacc.2022.11.005.

Marcus Vollmer, Dominic Bläsing, Julian Elias Reiser, Maria Nisser, and Anja Buder. Simultaneous physiological measurements with five devices at different cognitive and physical loads. PhysioNet, version 1.0.0, 2020. URL https://doi.org/10.13026/chd5-t946.

W. Wang, P. Mohseni, K. L. Kilgore, and L. Najafizadeh. PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023. doi: 10.3389/fdgth.2022.1090854.

Qirui Zhang, Qingsong Xie, Min Wang, and Guoxing Wang. Motion artifact removal for PPG signals based on accurate fundamental frequency estimation and notch filtering. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2965–2968, 2018. doi: 10.1109/EMBC.2018.8512878. URL https://doi.org/10.1109/EMBC.2018.8512878.

Hao Zhou, Simon A. Lee, Cyrus Tanade, Keum San Chun, Juhyeon Lee, Migyeong Gwak, Megha Thukral, Justin Sung, Eugene Hwang, Mehrab Bin Morshed, Li Zhu, Viswam Nathan, Md Mahbubur Rahman, Subramaniam Venkatraman, and Sharanya Arcot Desai. Physiology-aware masked cross-modal reconstruction for biosignal representation learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Qiang Zhu, Xin Tian, Chau-Wai Wong, and Min Wu. Learning your heart actions from pulse: Ecg waveform reconstruction from ppg. *IEEE Internet of Things Journal*, 8(23):16734–16748, 2021. doi: 10.1109/JIOT.2021.3097946. URL https://doi.org/10.1109/JIOT.2021.3097946.

A Related Work

This section surveys PPG-only and multimodal foundation models for cardiovascular signals, together with methods for ECG–PPG waveform reconstruction and signal denoising.

PPG-only foundation models. PPG-only foundation models pretrain single-modality encoders on large unlabelled PPG corpora with self-supervised objectives. PaPaGei (Pillai et al., 2025), including its PaPaGei-S and PaPaGei-P variants, uses morphology-aware contrastive learning. Pulse-PPG (Saha et al., 2025) pretrains on raw PPG collected in field settings using a relative contrastive objective with a learned motif-distance function. GPT-PPG (Chen et al., 2025) adapts autoregressive GPT pretraining to continuous PPG. Public code for GPT-PPG is **not available**. SIGMA-PPG (Guo et al., 2026) combines prior-guided adversarial masking with vector-quantized semantic consistency to learn transferable representations. These models capture PPG morphology and temporal patterns, but lack an ECG reference for directly representing ECG-referenced timing quantities such as PAT. They therefore cannot explicitly organize the delayed peripheral response around its corresponding electrical event.

ECG-guided and multimodal foundation models. ECG-guided and multimodal models combine complementary physiological signals through cross-modal supervision, shared representations, joint encoding, or signal translation. VCT-M2AE (Hu et al., 2025) couples modality-specific encoders through a shared bottleneck and combines masked reconstruction with cross-modal contrast learning. AnyPPG (Nie et al., 2025) uses ECG as a supervisory signal for PPG pretraining. UniCardio (Chen et al., 2026) uses a diffusion Transformer for denoising, imputation, and translation among PPG, ECG, and blood-pressure waveforms. CSFM (Gu et al., 2026) extends multimodal pretraining across heterogeneous cohorts and devices. CardioFM (Hassanuzzaman et al., 2026) jointly encodes ECG and PPG with bidirectional cross-modal attention and a residual vector-quantized codebook. Public code for CardioFM is **not available**. CardioState-JEPA (Shafiq et al., 2026) learns a shared ECG–PPG–PCG representation with a Transformer and a physiology-supervised delay aligner, using PAT as the ECG–PPG delay anchor. Despite these advances, most models still operate on windows or uniformly sampled patches, leaving shared beat correspondence and the separation of within-beat morphology from across-beat dynamics implicit.

ECG–PPG waveform reconstruction. ECG–PPG reconstruction methods learn mappings from PPG to ECG at different output scales, from local electrical complexes to complete waveforms, using supervised regression, adversarial synthesis, or sequence models. Transformed attentional neural networks (Chiu et al., 2020) reconstruct the QRS complex from PPG. CardioGAN (Sarkar & Etemad, 2021) uses adversarial synthesis. The cardiovascular signal model-based transform (Zhu et al., 2021) maps the discrete cosine transform (DCT) coefficients of each PPG cycle to those of its corresponding ECG cycle. Performer (Lan, 2023) applies a Transformer to PPG-to-ECG reconstruction and evaluates the reconstructed signal as a digital biomarker for cardiovascular-disease detection. CardioPPG (Ding et al., 2026) evaluates reconstructed ECG for cardiovascular-disease prediction. These methods focus on waveform translation or downstream prediction and treat the ECG–PPG delay implicitly.

Physiological signal denoising. Earlier PPG denoising approaches include frequency-aware notch filtering (Zhang et al., 2018), artificial neural networks to correct motion-corrupted beats (Singha Roy et al., 2018), DnCNN-based denoising (Chang et al., 2019), and recurrent autoencoder filtering (Azar et al., 2021). A convolutional denoising autoencoder (Mohagheghian et al., 2024) was trained using an unconventional scheme. SPEAR (Jain et al., 2024) uses self-supervised training to restore corrupted PPG segments while preserving clean regions. DPNet (Chiu et al., 2025) combines Mamba-based temporal modelling with waveform-fidelity and auxiliary heart-rate objectives. These methods focus on PPG restoration and do not explicitly model ECG–PPG beat correspondence.

B Data Preprocessing and Beat Construction

The pre-training cache stores each paired ECG–PPG record as a variable-length sequence of beats. The following subsections describe beat construction, retained timing metadata, and conversion to the fixed-size packed input used by BEATFORMER.

Missing-value repair and filtering. Before beat detection, missing samples were linearly interpolated. The resulting ECG and PPG signals were then filtered separately using third-order, zero-phase Butterworth band-pass filters, with passbands of 0.5–40 Hz and 0.3–8 Hz, respectively.

Cardiac anchors. For paired ECG–PPG input, R-like cardiac centers are detected from the filtered ECG. The detector uses the derivative–squaring–moving-average processing common to classical QRS detection (Pan & Tompkins, 1985), followed by robust thresholding and local refinement.

Let x_t denote the filtered ECG sampled at f_s. Then we form the smoothed derivative-energy signal m_t as

$$m_t = \mathrm{MA}_{\lfloor 0.10 f_s \rfloor}\big((x_t - x_{t-1})^2\big),$$

where MA_w denotes a centered moving average over w samples. To suppress weak peaks caused by noise, motion artifacts, or other waveform fluctuations, we retain a sample c_i as a candidate center only if it is a local maximum of m_t and satisfies $m_{c_i} \geq \tau$. The record-wise robust threshold τ is computed from the median and median absolute deviation (MAD) of m_t:

$$\widetilde{m} = \mathrm{median}_t(m_t), \qquad \mathrm{MAD}_m = \mathrm{median}_t\,|m_t - \widetilde{m}|\,, \qquad \tau = \widetilde{m} + 3 \times 1.4826 \times \mathrm{MAD}_m\,.$$

Additionally, candidate centers are required to be separated by at least $d_{\min} = \lfloor 60 f_s / 220 \rfloor$ samples, where 220 bpm is defined as an upper heart-rate bound. Each candidate c_i is refined by selecting the maximum-amplitude sample of the filtered ECG within $\lfloor 0.05 f_s \rfloor$ samples on either side. We denote the resulting centers, ordered in time, by $R_1, \ldots, R_N$; these centers serve as shared cardiac anchors for segmenting both ECG and PPG.

Beat intervals, resampling, and normalization. After defining the beat interval $[s_i, e_i)$, the same interval is applied to the ECG and PPG channels. Each variable-length segment is linearly resampled to $L = 128$ samples. For modality r, let $\mu_k^{(r)}$ and $v_k^{(r)}$ denote the within-beat mean and variance of the resampled signal before normalization. Each beat and channel is then standardized independently, and $(\mu_k^{(r)}, v_k^{(r)})$ are retained for constructing the subsequent beat-context descriptor.

Beat context. Using the retained R-peak centers, beat boundaries, and normalized phase coordinates defined above, we construct a phase-token context vector for each phase position p within an interior beat k:

$$\phi_p = \frac{p - \frac{1}{2}}{P}, \qquad RR_k^{\mathrm{pre}} = \frac{R_k - R_{k-1}}{f_s}, \qquad RR_k^{\mathrm{post}} = \frac{R_{k+1} - R_k}{f_s}, \qquad T_k = \frac{e_k - s_k}{f_s},$$

$$\tau_{k,p} = \frac{s_k + \phi_p(e_k - s_k) - R_k}{f_s}, \qquad \rho_{k,p}^{(r)} = \Big[RR_k^{\mathrm{pre}}, RR_k^{\mathrm{post}}, T_k, \tau_{k,p}, \mu_k^{(r)}, v_k^{(r)}\Big]^{\mathsf{T}} \in \mathbb{R}^6.$$

Here, RR_k^{pre} and RR_k^{post} denote the physical R–R intervals immediately before and after beat k, respectively. T_k is the physical duration of the extracted beat, and $\tau_{k,p}$ is the physical timing offset of phase position p relative to R_k. The quantities $\mu_k^{(r)}$ and $v_k^{(r)}$ are the pre-normalization within-beat mean and variance defined above. Thus, $\rho_{k,p}^{(r)}$ combines local rhythm, physical timing, and modality-specific amplitude-scale information. During denoising and reconstruction, we inject $\rho_{k,p}^{(r)}$ only when the corresponding waveform segment is available in the model input and the complete descriptor can be computed from that input. Otherwise, we omit it to prevent target leakage.

Fixed-size packing. Each packed window contains $K = 25$ beat slots. Windows with fewer than K valid beats are zero-padded, and a beat-validity mask identifies the occupied slots. If more than K beats are available, the loader selects a subset of K beats. For paired input, each window is represented by a waveform tensor of shape $\mathbb{R}^{25 \times 2 \times 128}$. The beat-validity mask is constructed during loading and passed to BEATFORMER, while the modality-availability mask indicates whether each input modality is present.

C Implementation Details

This section details the beat-centered encoders, modality-aware fusion, cumulative beat-time encoding, orthogonal beat–phase Transformer blocks, two-level readout, and pre-training heads used by BEATFORMER.

C.1 TOKEN INITIALIZATION

For each modality, the normalized beat is processed by an independent one-dimensional convolutional encoder. The encoder consists of three convolutional blocks with channel widths of 32, 64, and 128, kernel sizes of 9, 7, and 5, and stride 2 in the first convolution of each block. Each block uses batch normalization and GELU activation. Adaptive average pooling then produces $P = 8$ ordered phase features, which are linearly projected from 128 channels to the model dimension $d = 768$, yielding $C_k^{(r)} \in \mathbb{R}^{P\times d}$. The initial phase token combines encoded waveform content with its phase identity and beat context:

$$C_k^{(r)} = \mathrm{Enc}^{(r)}\left(x_k^{(r)}\right), \quad Z_{k,p}^{(r)} = C_{k,p}^{(r)} + E_{\mathrm{phase}}^{(r)}(\phi_p) + W_\rho^{(r)}\rho_{k,p}^{(r)}.$$

where $E_{\mathrm{phase}}^{(r)}$ embeds the normalized phase coordinate and $W_\rho^{(r)} \in \mathbb{R}^{d\times 6}$ projects the beat-context descriptor into the token space. The resulting modality-specific tokens are passed to the delay-aware fusion module.

C.2 DELAY-CONDITIONED CROSS-MODAL FUSION

The delay posterior is estimated from the encoder outputs before adding the phase and beat-context embeddings. For ECG and PPG content tokens at phase positions p and q, respectively, we compute their cosine similarity as

$$\Psi_{k,p,q} = \frac{\left\langle C_{k,p}^{(\mathrm{ECG})}, C_{k,q}^{(\mathrm{PPG})}\right\rangle}{\left\|C_{k,p}^{(\mathrm{ECG})}\right\|_2 \left\|C_{k,q}^{(\mathrm{PPG})}\right\|_2 + \epsilon}.$$

Here, $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product, $\|\cdot\|_2$ denotes the Euclidean norm, and $\epsilon = 10^{-6}$ is a fixed constant for numerical stability.

For each candidate phase lag $\ell \in \mathcal{L}_{\mathrm{lag}}$, where $\mathcal{L}_{\mathrm{lag}} = \{0, 1, \ldots, P-1\}$, we compare the ECG token at phase position p with the PPG token at position $p + \ell$ over all valid overlapping positions for which both indices lie in $\{1, \ldots, P\}$. The delay posterior, its physical mean, and its confidence are computed as

$$\pi_k(\ell) = \mathrm{softmax}_{\ell\in\mathcal{L}_{\mathrm{lag}}}\left(\frac{\sum_{p:\, p+\ell\le P}\Psi_{k,p,p+\ell}}{\eta}\right), \qquad T_k = \frac{e_k - s_k}{f_s},$$
$$\widehat{\delta}_k = \mathrm{Mean}_{\pi_k}\left(\frac{T_k\ell}{P}\right), \qquad \gamma_k = \exp\left[-\frac{\mathrm{Var}_{\pi_k}(T_k\ell/P)}{2s_\delta^2}\right].$$

Here, f_s is the sampling frequency, and T_k is the physical duration of beat k. The positive scalars η and s_δ are learned jointly with the model. The posterior mean $\widehat{\delta}_k$ provides the estimated physical delay, whereas γ_k measures the confidence of this estimate.

The estimated delay is converted into a coordinate shift for bidirectional cross-modal attention. Taking ECG as the reference coordinate, we define

$$\chi_{k,p}^{(\mathrm{ECG})} = \phi_p, \qquad \chi_{k,q}^{(\mathrm{PPG})} = \phi_q - \frac{\widehat{\delta}_k}{T_k}, \qquad [B_k^{r\leftarrow s}]_{p,q} = -\gamma_k\left(\chi_{k,p}^{(r)} - \chi_{k,q}^{(s)}\right)^2,$$

where $r, s \in \{\mathrm{ECG}, \mathrm{PPG}\}$ and $r \neq s$. The resulting bias favors cross-modal pairs with delay-consistent coordinates. Let $Z_k^{(r)} \in \mathbb{R}^{P\times d}$ denote the phase tokens after token initialization. For each direction $r \leftarrow s$, the cross-modal update is

$$\widetilde{Z}_k^{(r)} = Z_k^{(r)} + \lambda_{\mathrm{delay}}\,\mathrm{softmax}_q\left(\frac{\left(Z_k^{(r)}W_Q^r\right)\left(Z_k^{(s)}W_K^s\right)^{\mathsf{T}}}{\sqrt{d}} + B_k^{r\leftarrow s}\right)Z_k^{(s)}W_V^s, \qquad r \neq s.$$

The query, key, and value projections are modality-specific learned single-head matrices $W_Q^r, W_K^s, W_V^s \in \mathbb{R}^{d\times d}$. The learned scalar λ_{delay} controls the strength of the delay-conditioned exchange. If one modality is unavailable, the corresponding cross-modal update is skipped. The updated streams $\widetilde{Z}_k^{(r)}$ are then passed to the modality gate.

C.3 MODALITY GATE AND BEAT-TIME ENCODING

Following the delay-aware bidirectional exchange, the updated phase-token streams $\widetilde{Z}^{(r)}$ are fused by a beat-level modality gate. For each beat, phase tokens within each modality are mean-pooled and mapped to a scalar logit by a two-layer gate MLP f_{gate} with hidden width $d/2$ and GELU activation. Let $a_r \in \{0, 1\}$ indicate whether modality r is available. The gate normalizes the logits over available modalities and fuses the modality streams at every phase position:

$$g_{k,r} = f_{\text{gate}}\left(\frac{1}{P}\sum_{p=1}^{P}\widetilde{Z}^{(r)}_{k,p}\right), \qquad \omega_{k,r} = \frac{a_r \exp(g_{k,r})}{\sum_{r'} a_{r'} \exp(g_{k,r'})}. \qquad H_{k,p} = \sum_{r:\, a_r=1} \omega_{k,r}\widetilde{Z}^{(r)}_{k,p}.$$

The gate parameters are learned jointly with the rest of the model.

After modality fusion, the model adds a shared beat-time embedding of each valid beat's cumulative position within the input window. Let $t_k = (R_k - R_1)/f_s$ denote the position of beat k in seconds, where R_1 is the first retained R peak in the window. For $j = 0, \ldots, M-1$, the beat-time embedding and its injection into every phase token $p = 1, \ldots, P$ are defined as

$$\nu_j = \exp\left(-\frac{j \log \Lambda}{M-1}\right), \qquad u_k = \left[t_k,\, \{\sin(\nu_j t_k)\}_{j=0}^{M-1},\, \{\cos(\nu_j t_k)\}_{j=0}^{M-1}\right]^{\mathsf{T}},$$

$$E_{\text{time}}(t_k) = W_t u_k + b_t \in \mathbb{R}^d, \qquad H^{[0]}_{k,p} = H_{k,p} + E_{\text{time}}(t_k).$$

Here, $W_t \in \mathbb{R}^{d\times(2M+1)}$ and $b_t \in \mathbb{R}^d$ are learnable parameters. In our implementation, $M = 32$ and $\Lambda = 10{,}000$. The same beat-time embedding is broadcast over all phase tokens within beat k, and the resulting beat–phase grid $H^{[0]}$ is passed to the orthogonal beat–phase Transformer.

C.4 ORTHOGONAL BEAT–PHASE TRANSFORMER

For the implementation described here, each Transformer block sequentially applies within-beat phase attention, across-beat attention at each fixed phase position, and a position-wise feed-forward network, with residual connections around these updates. We stack $N = 9$ such blocks with model dimension $d = 768$. The other Transformer and pre-training settings are listed in Table C.1.

Table C.1: Transformer configuration and pre-training optimization settings.

Hyperparameter	Default value
Number of attention heads N_h	12
Attention-head dimension d_h	64
Feed-forward dimension d_{ff}	2048
Bias-generating MLP hidden width h_g	32
Bias-generating MLP dropout rate	0.1
FFN dropout rate	0.1
Time-encoding bands M	32
Time-encoding scale Λ	10,000
Masking ratio	0.60
Batch size	64
Learning rate	2×10^{-5}
Weight decay	1×10^{-4}
Activation function	GELU
Optimizer	AdamW
Mixed precision	bfloat16
Gradient-norm clipping	1.0

Physiological relative-position biases. The physiological relative-position biases are generated from the following beat- and phase-dependent inputs:

$$\left[B^{\text{phase}}_k\right]_{p,q} = g_{\text{phase}}\left(\phi_p - \phi_q,\, \tau_{k,p} - \tau_{k,q},\, \log T_k\right),$$
$$\left[B^{\text{beat}}_p\right]_{i,j} = g_{\text{beat}}\left(t_i - t_j,\, \log RR_i - \log RR_j,\, \phi_p\right).$$

Here, $t_i = (R_i - R_1)/f_s$ denotes the cumulative physical time of beat i within the window, and $RR_i = (R_i - R_{i-1})/f_s$ denotes its preceding R–R interval. The quantities ϕ_p, $\tau_{k,p}$, and T_k are defined above. The functions g_{phase} and g_{beat} independently map these three-dimensional inputs to scalar bias values. Each function is implemented as a two-layer MLP with layer normalization, GELU activation, dropout, and a scalar linear output. The two functions use separate learnable parameters.

Within-beat phase attention. Let $H^{[n]} \in \mathbb{R}^{K\times P\times d}$ denote the input to block n, and let $\bar{H}^{[n]}_{k,\cdot} = \mathrm{LN}(H^{[n]}_{k,\cdot})$. For each beat $k = 1, \ldots, K$ and attention head $h = 1, \ldots, N_h$, let $Q^{\text{phase}}_{k,h}$, $K^{\text{phase}}_{k,h}$, and $V^{\text{phase}}_{k,h}$ denote the d_h-dimensional outputs of learned query, key, and value projections applied to $\bar{H}^{[n]}_{k,\cdot}$. The phase-attention output of head h is

$$\texttt{PhaseAttn}^{(h)}\left(\bar{H}^{[n]}_{k,\cdot}\right) = \mathrm{softmax}_q\left(\frac{Q^{\text{phase}}_{k,h}\left(K^{\text{phase}}_{k,h}\right)^{\mathsf{T}}}{\sqrt{d_h}} + \lambda_{\text{phase}} B^{\text{phase}}_k\right) V^{\text{phase}}_{k,h}.$$

The head outputs are concatenated and added to the input through a residual connection:

$$H'_{k,\cdot} = H^{[n]}_{k,\cdot} + \mathrm{Concat}_{h=1}^{N_h}\left[\texttt{PhaseAttn}^{(h)}\left(\bar{H}^{[n]}_{k,\cdot}\right)\right] W_{O,\text{phase}}.$$

The query, key, and value projections, the output projection $W_{O,\text{phase}}$, and the scalar coefficient λ_{phase} are learned jointly with the model. Here, $d_h = d/N_h$.

Across-beat attention. For each fixed phase position p, beat attention operates on the K beat slots of the phase-refined representation. Let $\bar{H}'_{\cdot,p} = \mathrm{LN}(H'_{\cdot,p})$. For each phase position $p = 1, \ldots, P$ and attention head $h = 1, \ldots, N_h$, let $Q^{\text{beat}}_{p,h}$, $K^{\text{beat}}_{p,h}$, and $V^{\text{beat}}_{p,h}$ denote the d_h-dimensional outputs of learned query, key, and value projections applied to $\bar{H}'_{\cdot,p}$. The additive padding mask satisfies $\mathcal{M}^{\text{pad}}_{i,j} = 0$ for a valid beat key and $\mathcal{M}^{\text{pad}}_{i,j} = -\infty$ for a padded beat key . The beat-attention output of head h is

$$\texttt{BeatAttn}^{(h)}\left(\bar{H}'_{\cdot,p}\right) = \mathrm{softmax}_j\left(\frac{Q^{\text{beat}}_{p,h}\left(K^{\text{beat}}_{p,h}\right)^{\mathsf{T}}}{\sqrt{d_h}} + \lambda_{\text{beat}} B^{\text{beat}}_p + \mathcal{M}^{\text{pad}}\right) V^{\text{beat}}_{p,h}.$$

The head outputs are concatenated and added through a residual connection:

$$H''_{\cdot,p} = H'_{\cdot,p} + \mathrm{Concat}_{h=1}^{N_h}\left[\texttt{BeatAttn}^{(h)}\left(\bar{H}'_{\cdot,p}\right)\right] W_{O,\text{beat}}.$$

The query, key, and value projections, the output projection $W_{O,\text{beat}}$, and the scalar coefficient λ_{beat} are learned jointly with the model.

Feed-forward update and repetition. A position-wise feed-forward network completes each Transformer block. The FFN consists of two linear layers with dimensions $d \to d_{\text{ff}} \to d$, GELU activation, and dropout:

$$H^{[n+1]} = H'' + \mathrm{FFN}\left(\mathrm{LN}(H'')\right), \qquad n = 0, \ldots, N-1.$$

The FFN parameters and the affine parameters of layer normalization are learned jointly with the rest of the model. Each block applies phase attention, then beat attention, and then the feed-forward update.

C.5 Readout

After the Transformer stack, the final beat–phase grid is layer-normalized and mean-pooled over the P phase tokens within each beat. Let $m_k \in \{0, 1\}$ denote the beat-validity mask, where $m_k = 1$ indicates a valid beat and $m_k = 0$ indicates a padded beat slot. The reliability score is computed by a two-layer MLP that maps each d-dimensional beat representation to a scalar logit. The MLP consists of layer normalization, a linear layer with hidden width $h_{\text{rel}} = \max(d/2, 32)$, GELU activation, dropout rate 0.1, and a final linear layer with one output.

$$H_k^{\text{beat}} = \frac{1}{P}\sum_{p=1}^{P} \text{LN}\Big(H_{k,p}^{[N]}\Big), \qquad s_k^{\text{rel}} = f_{\text{rel}}\big(H_k^{\text{beat}}\big),$$

$$w_k = \frac{m_k \exp\big(s_k^{\text{rel}}\big)}{\sum_{j=1}^{K} m_j \exp\big(s_j^{\text{rel}}\big)}, \qquad H^{\text{cls}} = \sum_{k=1}^{K} w_k H_k^{\text{beat}}.$$

The beat-level representations H_k^{beat} are used for the masked waveform reconstruction, PAT prediction, and PPG morphology prediction objectives, whereas the window-level representation H^{cls} is used for the HRV prediction objective. The representation H^{cls} is obtained by reliability-weighted pooling rather than by introducing an additional CLS token.

D Pre-training Target Construction

The reconstruction head applies LayerNorm followed by a two-layer MLP with hidden width d and output dimension $2L$, producing $\widehat{\mathbf{x}}_k \in \mathbb{R}^{2\times L}$. The three physiological regression heads are independent two-layer MLPs with hidden width $d/2$. They produce the PAT prediction $\widehat{y}_k^{\text{PAT}} \in \mathbb{R}$, the morphology prediction $\widehat{\mathbf{y}}_k^{\text{morph}} \in \mathbb{R}^6$, and the HRV prediction $\widehat{\mathbf{y}}^{\text{HRV}} \in \mathbb{R}^3$ respectively.

Masked beat reconstruction. During each pre-training step, masked beat slots are sampled uniformly without replacement from the valid slots according to the masking ratio; padded slots are never selected. Let $m_k^{\text{mask}} \in \{0,1\}$ indicate whether beat slot k is selected for reconstruction. For each selected beat, all $P = 8$ phase tokens entering the Transformer stack are replaced by a learned mask embedding $e_{\text{mask}} \in \mathbb{R}^d$:

$$\widetilde{H}_{k,p}^{[0]} = \begin{cases} e_{\text{mask}}, & m_k^{\text{mask}} = 1, \\ H_{k,p}^{[0]}, & m_k^{\text{mask}} = 0, \end{cases} \qquad p = 1, \ldots, P.$$

The reconstruction head maps the resulting beat-level representation to a two-channel waveform, $\widehat{\mathbf{x}}_k = \text{Head}_{\text{recon}}(H_k^{\text{beat}}) \in \mathbb{R}^{2\times L}$. The target waveform $\mathbf{x}_k \in \mathbb{R}^{2\times L}$ has ECG and PPG in its two rows and is used only for computing the reconstruction loss, not as a model input. Here, $L = 128$ is the number of resampled samples per beat, as defined in Appendix B. The reconstruction loss is the masked mean squared error over valid waveform entries:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{I}_{\text{recon}}|} \sum_{(k,r,n)\in\mathcal{I}_{\text{recon}}} \Big(\widehat{x}_{k,n}^{(r)} - x_{k,n}^{(r)}\Big)^2,$$

where $r \in \{\text{ECG}, \text{PPG}\}$ indexes the modality, n indexes the resampled waveform samples, and $\mathcal{I}_{\text{recon}}$ contains only masked, valid waveform entries. Batch indices are suppressed.

Pulse arrival time. For each beat, we locate the maximum PPG upstroke slope within 0.08–$0.55\,\text{s}$ after the ECG R peak. Let f_s denote the sampling frequency and g_n the preprocessed PPG sample at index n. The candidate samples, maximum-slope location, and PAT target are

$$\mathcal{S}_k^{\text{PAT}} = \left\{ n : 0.08 \le \frac{n - R_k}{f_s} \le 0.55 \right\},$$

$$n_k^\star = \arg\max_{n\in\mathcal{S}_k^{\text{PAT}}} f_s(g_n - g_{n-1}), \qquad y_k^{\text{PAT}} = \frac{n_k^\star - R_k}{f_s}.$$

Thus, y_k^{PAT} denotes the elapsed time from the ECG R peak to the maximum-slope point of the corresponding PPG upstroke. The PAT loss is the masked mean squared error $\mathcal{L}_{\text{PAT}} = |\mathcal{I}_{\text{PAT}}|^{-1} \sum_{i\in\mathcal{I}_{\text{PAT}}} \big(\widehat{y}_i^{\text{PAT}} - y_i^{\text{PAT}}\big)^2$, where $\mathcal{I}_{\text{PAT}}$ contains valid beat entries with available PAT targets.

PPG morphology. For each beat, let $x_{k,n}$ denote the pre-normalization PPG value at sample n of beat k, where $n = 1, \ldots, L$. Let $t_{k,n}$ denote the physical time associated with sample n. The morphology targets and their definitions are summarized in Table D.1.

Table D.1: Beat-level PPG morphology targets.

Target	Definition
Pulse amplitude A_k	$A_k = \max_n x_{k,n} - \min_n x_{k,n}$; the peak-to-trough PPG amplitude within beat k.
Half-amplitude width W_k^{50}	$W_k^{50} = t_{k,\mathrm{R}}^{50} - t_{k,\mathrm{L}}^{50}$; here $b_k = \min_n x_{k,n}$, $h_k = b_k + A_k/2$, and $n_k^{\mathrm{peak}} = \arg\max_n x_{k,n}$. The times $t_{k,\mathrm{L}}^{50}$ and $t_{k,\mathrm{R}}^{50}$ are the first and last sample times reaching h_k around the pulse peak.
Rise time T_k^{rise}	$T_k^{\mathrm{rise}} = t_{k,n_k^{\mathrm{peak}}} - t_k^{\mathrm{foot}}$; t_k^{foot} denotes the local PPG foot preceding the pulse peak.
Normalized area A_k^{norm}	$A_k^{\mathrm{norm}} = \dfrac{1}{LA_k} \sum_{n=2}^{L} \left(\dfrac{x_{k,n-1} + x_{k,n}}{2} - b_k \right)$; the trapezoidal area above the beat minimum, normalized by pulse amplitude and sample count.
Peak position ϕ_k^{peak}	$\phi_k^{\mathrm{peak}} = (n_k^{\mathrm{peak}} - 1)/(L-1)$; the pulse-peak location within the normalized beat.
Maximum upstroke slope S_k^{up}	$S_k^{\mathrm{up}} = \dfrac{1}{A_k} \max\limits_{2 \le n \le L} \dfrac{x_{k,n} - x_{k,n-1}}{t_{k,n} - t_{k,n-1}}$; the steepest PPG increase normalized by pulse amplitude.

The six values, ordered as in the table, form $\mathbf{y}_k^{\mathrm{morph}} \in \mathbb{R}^6$. The morphology loss is the masked mean squared error $\mathcal{L}_{\mathrm{morph}} = |\mathcal{I}_{\mathrm{morph}}|^{-1} \sum_{i \in \mathcal{I}_{\mathrm{morph}}} \left(\widehat{y}_i^{\mathrm{morph}} - y_i^{\mathrm{morph}} \right)^2$, where $\mathcal{I}_{\mathrm{morph}}$ contains valid beat–descriptor entries with available morphology targets.

Heart-rate variability. For each window, let $RR_i = (R_{i+1} - R_i)/f_s$ denote the interval between two adjacent retained R peaks. Let N_{RR} denote the number of valid intervals, and let $\overline{RR} = \frac{1}{N_{\mathrm{RR}}} \sum_{i=1}^{N_{\mathrm{RR}}} RR_i$ denote their mean. The window-level HRV descriptors are

$$\mathrm{SDNN} = \sqrt{\frac{1}{N_{\mathrm{RR}} - 1} \sum_{i=1}^{N_{\mathrm{RR}}} (RR_i - \overline{RR})^2},$$

$$\mathrm{RMSSD} = \sqrt{\frac{1}{N_{\mathrm{RR}} - 1} \sum_{i=1}^{N_{\mathrm{RR}}-1} (RR_{i+1} - RR_i)^2}, \qquad \overline{HR} = \frac{1}{N_{\mathrm{RR}}} \sum_{i=1}^{N_{\mathrm{RR}}} \frac{60}{RR_i}.$$

The three descriptors form the window-level target vector $\mathbf{y}^{\mathrm{HRV}} \in \mathbb{R}^3$. The HRV loss is the masked mean squared error $\mathcal{L}_{\mathrm{HRV}} = |\mathcal{I}_{\mathrm{HRV}}|^{-1} \sum_{i \in \mathcal{I}_{\mathrm{HRV}}} \left(\widehat{y}_i^{\mathrm{HRV}} - y_i^{\mathrm{HRV}} \right)^2$, where $\mathcal{I}_{\mathrm{HRV}}$ contains valid window–descriptor entries with sufficient consecutive R–R intervals. Windows without sufficient valid consecutive R–R intervals do not contribute to this loss.

E Datasets and Task Construction

E.1 Classification and Physiological Regression

Window-level tasks. Activity recognition, load classification, stress recognition, signal-quality classification and heart-rate regression use 10-second windows. The source datasets are ScientISST MOVE (Areias Saraiva et al., 2024), Simultaneous (Vollmer et al., 2020), SensSmartTech (Lazović et al., 2024), PTT PPG (Mehrgardt et al., 2022), BUT PPG (Nemcova et al., 2021; 2024), WESAD (Schmidt et al., 2018), and PPG-DaLiA (Reiss et al., 2019). Table E.1 lists their evaluation sizes and outer splits. ScientISST activity, Simultaneous load and heart rate, PTT PPG activity and PPG-DaLiA activity use leave-one-subject-out (LOSO) cross-validation. All windows from one subject are held out for testing in each fold. The other window-level tasks use fixed five-fold group splits. Validation subjects are drawn from the outer training set to select the predictor epoch. Training, validation and test subjects are disjoint within each fold. PTT PPG, BUT PPG and WESAD contain 22, 50 and 15 subjects, respectively. Each PTT PPG activity fold has 16 training, five validation and one test subject; the corresponding counts for PPG-DaLiA activity are 11, three and one.

Table E.1: Evaluation sizes and outer splits for window-level classification and heart-rate regression. Window counts are summed over disjoint outer test folds. The number of LOSO folds equals the number of evaluated subjects.

Task	Test windows	Folds	Split
ScientISST activity	3652	16	LOSO
Simultaneous load	2411	13	LOSO
Simultaneous heart rate	2411	13	LOSO
SensSmartTech before/after activity	338	5	Grouped five-fold
PTT PPG heart rate	3207	5	Subject-grouped five-fold
BUT PPG signal quality	3888	5	Subject-grouped five-fold
WESAD stress	3320	5	Subject-grouped five-fold
PTT PPG activity	3207	22	LOSO
PPG-DaLiA activity	47196	15	LOSO

Classification labels. Each dataset retains its own label space. ScientISST has 11 activity classes: `gesticulate`, `greetings`, `jumps`, `lift`, `run`, `sprint`, `walk_after`, `walk_before`, `walk_before_downstairs`, `walk_before_elevatordown` and `walk_before_elevatorup`. Simultaneous load has four classes: 2back, baseline, uphill walking and walking. SensSmartTech retains the A/B codes for before/after activity. PTT PPG activity distinguishes sitting, walking and running. PPG-DaLiA distinguishes sitting, stairs, soccer, cycling, driving, lunch, walking and working. WESAD retains original labels 1–3: stress (2) is positive, whereas baseline (1) and amusement (3) form the negative class. BUT PPG retains the binary encoding of its original Quality field. BUT PPG quality, WESAD stress and SensSmartTech before/after classification are scored using AUROC from continuous predictions. Multiclass activity and load tasks use accuracy.

Heart-rate regression. PTT PPG and Simultaneous use continuous window-level heart-rate targets and report MAE in bpm. They use fixed five-fold and LOSO splits, respectively. Errors are computed on held-out windows, separately from demographic and blood-pressure regression using subject-level representations.

Age and body mass index. PPG-DaLiA age and body mass index (BMI) regression uses questionnaire labels from 15 subjects. Age is taken from the questionnaire; BMI is weight divided by squared height. Five-fold, sex-stratified evaluation assigns 10 subjects to training, two to validation and three to testing per fold. Validation includes one woman and one man, sampled with seed $42 + 1000 + k$ for fold index k. A two-layer MLP is selected by validation MAE and scored at the subject level.

Window representations from each subject are aggregated by concatenating their feature-wise mean and standard deviation. Feature and target standardization parameters are fitted only on training subjects within each outer fold. Predictions are returned to the original units before scoring: years for age and kg/m^2 for BMI. Multiple windows represent the same person and are not independent demographic test samples; the five outer test folds together contain 15 subjects.

Ventricular tachycardia false-alarm detection. VTaC prediction operates on alarm events and is evaluated over five folds. ECG and PLETH from seconds 240–300 of each record provide the minute preceding the alarm at second 300. This interval is divided into six 10-second windows. The mean and standard deviation of their encoder representations are concatenated into an event representation. The dataset contains 4,637 events and 27,822 windows in total. The split illustrated below contains 3,734 training, 460 validation and 443 test events. Evaluation is event-level.

The original `decision=true` label is encoded as 1 and all other values as 0. XGBoost predicts a continuous true-alarm score from frozen representations. A decision threshold is selected from 181 equally spaced values between 0.05 and 0.95 by validation F1. Ties are resolved by balanced accuracy and then proximity to 0.5. This threshold is fixed for testing; AUROC uses continuous scores.

Blood-pressure regression. Li ECG–PPG blood-pressure prediction uses resting red-light PPG and synchronized ECG, with subject-level cuff systolic and diastolic blood pressure as references. Feature-wise means, standard deviations, minima and maxima of window representations are concatenated for

Table E.2: PAT evaluation sizes, pooled over the five held-out test folds.

Dataset	Test windows	Groups	Grouping unit
PTT-PPG	3189	22	Subject
Li ECG–PPG	7860	148	Subject
UQ	18553	32	Case
ErgoBP	1028	23	Subject

each subject. Evaluation uses fixed five-fold splits. The predictor has hidden dimension 64, dropout 0.3, learning rate 3×10^{-4}, weight decay 10^{-3} and batch size 64. Systolic and diastolic MAE are 10.752 ± 1.474 and 7.114 ± 0.505 mmHg, respectively. Blood pressure uses cuff measurements, whereas PAT uses intervals extracted from synchronized ECG–PPG. The targets and errors are constructed separately. Blood-pressure scoring remains subject-level despite using multiple windows for aggregation.

PAT reference construction. PAT regression is evaluated on PTT-PPG, Li ECG–PPG, UQ (Liu et al., 2012) and ErgoBP (Schjerven et al., 2026). Beat-level PAT is the interval from the ECG R peak to the corresponding PPG foot. The target is the median over valid paired beats in each window:

$$y_i^{\mathrm{PAT}} = \mathrm{median}_{b \in \mathcal{B}_i} \left(t_{i,b}^{\mathrm{foot}} - t_{i,b}^{R} \right).$$

Here $\mathcal{B}_i$ contains valid pairs in window i, and both times are in milliseconds. The downstream reference starts at the ECG R peak and ends at the PPG foot. The pre-training timing target in Appendix D instead ends at the point of maximum PPG upslope; these targets are constructed separately.

Before label extraction, ECG and PPG are band-pass filtered at 5–35 Hz and 0.5–8 Hz, respectively. R peaks have a minimum separation of 0.30 seconds. The PPG foot is the intersection of the upslope tangent with the local baseline. The search interval is 40–700 ms after the R peak and ends no later than 40 ms before the next R peak. Candidate upslopes must be positive and have a slope-quality ratio of at least 1.5. PAT values outside 60–600 ms are discarded. A window must contain at least five valid paired beats.

PAT grouped splits. All four datasets use five-fold group splits, grouped by subject for PTT-PPG, Li ECG–PPG and ErgoBP and by case for UQ. Groups are shuffled with seed 42 and assigned to five folds. For $k \in \{0, 1, 2, 3, 4\}$, fold k is the test set, fold $(k + 1) \bmod 5$ is validation, and the remaining three folds form the training set. All windows of a group remain in the same fold. Models share window identifiers, groups, test folds and reference PAT values within each dataset.

E.2 Missing-Signal and Noise Construction

Reconstruction and denoising use 2,654 paired 10-second windows from 214 VTaC subjects. These waveform targets differ from the alarm-event targets in Appendix E.1. Conditions reuse fixed test windows; valid counts for time gaps and whole-beat masking depend on the condition. Complete reference signals are segmented into beats and normalized before artificial masking or corruption. The evaluation therefore measures controlled recovery given beat segmentation and normalization, rather than end-to-end recovery when raw signals, beat boundaries and complete-signal statistics are unavailable.

Within-beat masking. Each beat is resampled to 128 points and divided into eight phase bins of 16 points. We mask one to seven consecutive bins at matching ECG and PPG positions. Each condition uses all 2,654 windows.

Whole-beat masking. We remove one to four consecutive complete beats from both channels. No visible samples remain within the masked beats. Valid window counts are reported in Table G.2.

Time-defined gaps. Gaps last 0.5, 1, 2 or 4 seconds and may cross beat boundaries. Start times are drawn uniformly using a generator fixed by sample and condition, retaining up to 0.25 seconds of

Table E.3: Noise applied to the BEATFORMER target beat. $U(a, b)$ denotes uniform sampling on $[a, b]$. Sinusoidal frequency is expressed as cycles per normalized beat; amplitudes are in per-beat normalized coordinates.

Noise	Parameters
Gaussian	Standard deviation $0.16s\,U(0.75, 1.25)$.
Baseline drift	Sinusoidal amplitude $0.30s\,U(0.70, 1.30)$; cycles per beat $U(0.20, 1.50)$; phase $U(0, 2\pi)$.
Narrowband	Sinusoidal amplitude $0.10s\,U(0.70, 1.30)$; uniformly sampled integer cycles from 4 to 18; phase $U(0, 2\pi)$.
Motion artifact	Gaussian envelope with center $U(0.10, 0.90)$, width $U(0.035, 0.155)$ and amplitude $0.65s\,U(0.60, 1.40)$; within the envelope, 0.55 times standard Gaussian noise plus an offset $U(-1, 1)$.
Mixed	Sum of the four components above, multiplied by 0.55.

context on either side where possible. A resampled point is scored if its center lies in the requested interval. All phase bins containing such points are masked, but only points inside the requested interval are scored. The four conditions contain 2,654, 2,653, 2,650 and 2,627 valid windows, respectively. Beat boundaries are determined from complete signals in advance, and evaluation uses fixed segmentation and masks.

Controlled noise. Gaussian noise, baseline drift, narrowband interference, motion artifacts and mixed noise each have mild, moderate and severe levels. For BEATFORMER, noise is applied to the target beat after per-beat normalization. ECG and PPG corruption parameters are sampled separately; other beats remain clean. Let $u \in [0, 1]$ denote normalized within-beat position. The dimensionless severity factor s is 0.65, 1 or 1.55. Table E.3 specifies the parameters.

The evaluation seed is 7142; target beats and corruption seeds are fixed by sample and condition. During training, clean input, Gaussian noise, baseline drift, narrowband interference, motion artifacts and mixed noise are sampled with probabilities 0.10, 0.20, 0.15, 0.15, 0.20 and 0.20, respectively. This setting tests recovery of a corrupted target beat with clean neighboring context.

F Downstream Training and Evaluation Details

F.1 Frozen Representations and Predictors

Representation extraction. Classification and regression use the epoch-14 pre-training checkpoint with the encoder frozen. For input window $\mathbf{x}_i$,

$$\mathbf{z}_i = f_{\theta^*}(\mathbf{x}_i), \qquad \hat{y}_i = g_\phi(\mathbf{z}_i),$$

where θ^* is fixed and only predictor parameters ϕ are updated. Window-level tasks use window representations directly; subject- and event-level tasks use the aggregation in Appendix E.1.

ScientISST activity, Simultaneous load and heart rate, SensSmartTech before/after activity, PTT PPG heart rate and activity, BUT PPG quality, WESAD stress and PPG-DaLiA activity use an MLP with hidden dimension 128 and dropout 0.1. Predictors train for 40 epochs with batch size 256, learning rate 10^{-3}, weight decay 10^{-4} and seed 42. Each fold retains the predictor with the best validation metric: balanced accuracy for ScientISST, MAE for PTT PPG heart rate, AUROC for BUT PPG and WESAD, and accuracy for PTT PPG and PPG-DaLiA activity.

PTT PPG, BUT PPG, WESAD and PPG-DaLiA activity predictors use AdamW. Feature means and standard deviations are estimated only from inner training subjects and applied to validation and test representations. Heart-rate targets are likewise standardized using training data only and predictions returned to bpm before scoring. Binary tasks use binary cross-entropy with training-set class weights. Validation selects the predictor epoch but does not contribute to these standardization estimates.

Demographic and alarm predictors. Age and BMI use an MLP with hidden dimension 128, dropout 0.1, AdamW and SmoothL1 loss. Training lasts 40 epochs with learning rate 10^{-3} and weight decay 10^{-4}. The configured batch size is 256; with only 10 training subjects per fold, each

Table F.1: Input modalities and parameter counts of models used for downstream prediction and waveform recovery. Counts (millions) refer to the pretrained model or representation encoder used in evaluation, excluding separately fitted downstream predictors and recovery heads. The Small Transformer count covers the complete randomly initialized model.

Model	Input modality	Parameters (M)
BEATFORMER	ECG–PPG	73.8
CSFM-Base	ECG–PPG	117
AnyPPG	PPG	5.8
Pulse-PPG	PPG	28.5
SIGMA-PPG-Pro	PPG	30
PaPaGei-S	PPG	5
PaPaGei-P	PPG	5
VCT-M2AE	ECG–PPG	25.9
MOMENT-1-small	ECG–PPG	40
Small Transformer	ECG–PPG	3.764

epoch uses one batch containing all training subjects. Predictor selection uses validation MAE. Standardization and conversion to original units follow Appendix E.1.

VTaC uses XGBoost (Chen & Guestrin, 2016) with 300 trees, maximum depth 4, learning rate 0.05, row and per-tree feature subsampling ratios of 0.9, a binary logistic objective, log-loss evaluation, histogram tree construction and seed 42. The positive-class weight is the training negative-to-positive count ratio. Within each fold, only training events fit the classifier; validation events select the decision threshold.

PAT linear probing. PAT uses Ridge regression on frozen window representations. BEATFORMER produces 768 features, giving 769 trainable predictor parameters including the intercept. Inputs are non-overlapping 10-second windows. Beat resampling, channel-wise standardization and fixed-length packing follow Appendix B.

For each BEATFORMER outer fold, feature means and standard deviations are fitted on training windows only. With standardized features $\widetilde{\mathbf{z}}_i$, the objective is

$$\min_{\mathbf{w},b_0} \sum_{i\in\mathcal{T}} \left(y_i^{\mathrm{PAT}} - \mathbf{w}^{\mathsf{T}}\widetilde{\mathbf{z}}_i - b_0\right)^2 + \lambda\|\mathbf{w}\|_2^2,$$

where $\mathcal{T}$ contains training windows. Validation-window MAE selects λ from $\{0.01, 0.1, 1, 10, 100, 1000\}$. After selecting λ, BEATFORMER and all six baselines refit the feature scaler and Ridge predictor on the combined training and validation sets. The test fold remains held out. All five Li ECG–PPG folds select $\lambda = 1000$ for BEATFORMER. BEATFORMER and CSFM use ECG–PPG, whereas AnyPPG, Pulse-PPG, SIGMA-PPG, PaPaGei-S and PaPaGei-P use PPG.

F.2 Adaptation for Waveform Recovery

Training settings. Reconstruction and denoising independently initialize from the pre-training checkpoint, randomly initialize a recovery head and jointly update the backbone and head. Both use 5,261,492 training samples for 10 epochs without a validation set; evaluation uses the fixed epoch-10 checkpoint. Backbone and head learning rates are 10^{-5} and 10^{-4}, respectively. Weight decay is 10^{-4}, warm-up lasts 1,000 steps, the minimum learning-rate ratio is 0.05 and gradient-norm clipping is 1.0. Training uses mixed precision and gradient checkpointing. Table F.2 gives the remaining settings.

Reconstruction training samples whole-beat, cross-beat contiguous, mixed and time-gap masks with probabilities 0.60, 0.15, 0.15 and 0.10. Requested whole-beat counts are uniform from one to eight. Cross-beat contiguous lengths are selected from 4, 8, 12, 16 and 24 phase tokens; time gaps are 0.5, 1, 2 or 4 seconds. ECG and PPG masks are synchronized. The fixed evaluation conditions in Appendix E.2 are defined separately from this training distribution.

Table F.2: BEATFORMER adaptation settings. Steps count parameter updates over the full run. Numbers following loss terms are weights. Reconstruction losses target missing regions; denoising losses target the corrupted beat.

Setting	Reconstruction	Denoising
GPUs	2	2
Per-GPU / global batch	32 / 64	64 / 128
Optimization steps	822,110	411,060
Seed	42	142
Head hidden dimension	256	256
Losses and weights	Masked SmoothL1: 1; masked derivative: 0.1; boundary continuity: 0.1; visible-context anchoring: 0.01	Target-beat SmoothL1: 1; target-beat derivative: 0.1; clean-context anchoring: 0.005

Table F.3: Shared denoising evaluation settings. All models are evaluated on the same 2,654 VTaC test windows under five noise types and three severity levels. Noise is applied to the target beat while surrounding context remains clean. For each condition, the reported result is the median window-level SNR improvement relative to the corrupted input.

Evaluation item	Setting
Test set	The same 2,654 VTaC test windows
Signals	Paired ECG and PPG
Corrupted region	Target beat, with clean surrounding context retained
Noise conditions	Five noise types at three severity levels each
Recovery target	ECG and PPG waveforms of the target beat
Primary measure	Window-level SNR improvement relative to the corrupted input; the median across 2,654 windows is reported for each condition
Clean-input test	Output RMSE is evaluated separately to measure distortion introduced by recovery

Input representations and scoring. BEATFORMER and Small Transformer use packed beats; CSFM, VCT-M2AE and MOMENT-1-small use continuous 10-second waveforms. Small Transformer reuses the same beat loader, per-beat normalization and missing/corrupted-condition construction as BEATFORMER. All paths corrupt the target beat and retain the remaining context. Table F.3 summarizes the shared denoising evaluation settings.

All models are evaluated on the same fixed set of 2,654 VTaC test windows under the predefined waveform-recovery conditions. Reconstruction targets the missing ECG and PPG waveforms. Under contiguous within-beat masking, predictions and references for all models are scored on a 128-point grid per beat, with the same number of masked points used to compute RMSE. Denoising targets the corrupted beat and reports SNR improvement relative to the corrupted input for five noise types at three severity levels. These settings evaluate the models on the same test windows, recovery tasks, and metric definitions.

VCT-M2AE and MOMENT-1-small reconstruction and denoising train for 10 epochs on 5,559,335 cached samples using four GPUs, batch size 368 per GPU and global batch size 1,472, for 37,770 updates. Encoder and decoder learning rates are 10^{-5} and 10^{-4}, warm-up is 1,000 steps, weight decay is 10^{-4} and gradient clipping is 1.0. Sample organization, batch sizes and update counts follow each model's training configuration.

CSFM initializes from the public Base encoder and jointly fine-tunes the encoder and channel-independent recovery heads. Both tasks use the fixed epoch-10 checkpoint after 37,770 updates on 5,559,335 samples, four GPUs, batch size 368 per GPU, global batch size 1,472 and seed 42. AdamW uses backbone and head learning rates of 10^{-5} and 10^{-4}; the reconstruction missing-position embedding uses 5×10^{-5}. After 1,000 linear warm-up steps, cosine decay reaches a minimum ratio of 0.05. There is no layer-wise learning-rate decay. Weight decay is 10^{-4} and gradient clipping is 1.0. The recovery heads preserve 100 temporal tokens for each channel. A $768 \to 256 \to 25$ mapping and convolutional upsampling residual produce 2,500 points per channel, without additional cross-channel mixing after the encoder. Reconstruction receives missing-position encodings; denoising receives

only corrupted waveforms. VTaC does not select either checkpoint. Reconstruction uses masked SmoothL1, derivative, boundary-continuity and visible-region anchoring weights of 1, 0.1, 0.1 and 0.01. Denoising uses target-beat SmoothL1, derivative and clean-context anchoring weights of 1, 0.1 and 0.005.

Small Transformer reconstruction trains from random weights with four layers, eight attention heads, hidden dimension 256, feed-forward dimension 1,024 and 3,763,842 parameters. It divides the packed waveform into 200 tokens of length 16 and does not use the BEATFORMER pre-trained backbone. Training uses 5,261,492 samples, one GPU, batch size 392, learning rate and weight decay 10^{-4}, gradient clipping 1.0, and 10 epochs totaling 134,230 steps. Data and initialization seeds are 42 and 4201. Mask sampling and the four loss weights match BEATFORMER reconstruction. No validation set is used; evaluation fixes epoch 10.

Small Transformer denoising uses the same architecture with independent random initialization and no reconstruction or pre-training weights. It uses 5,261,492 samples, one GPU, batch size 392 and 10 epochs totaling 134,230 steps. Learning rate and weight decay are 10^{-4} and gradient clipping is 1.0. The overall seed is 42; initialization, data-loading and corruption seeds are 5201, 7043 and 9243. Training uses bfloat16 mixed precision, no validation set and the predefined epoch-10 checkpoint. VTaC does not select the checkpoint. Inputs contain corrupted waveforms and padding-validity information, but no target-beat index, noise type, severity or noise mask.

F.3 Scoring Units and Representation Aggregation

For window representations $\mathbf{z}_i$ belonging to a subject or event s, let $\mathcal{W}_s$ be the corresponding window set. Mean and standard-deviation aggregation are

$$\boldsymbol{\mu}_s = \frac{1}{|\mathcal{W}_s|} \sum_{i \in \mathcal{W}_s} \mathbf{z}_i, \qquad \boldsymbol{\sigma}_s = \sqrt{\frac{1}{|\mathcal{W}_s|} \sum_{i \in \mathcal{W}_s} (\mathbf{z}_i - \boldsymbol{\mu}_s)^2}.$$

Operations are feature-wise. Age, BMI and VTaC use $[\boldsymbol{\mu}_s; \boldsymbol{\sigma}_s]$; Li ECG–PPG blood pressure uses $[\boldsymbol{\mu}_s; \boldsymbol{\sigma}_s; \min_{i \in \mathcal{W}_s} \mathbf{z}_i; \max_{i \in \mathcal{W}_s} \mathbf{z}_i]$. Window classification, heart rate and PAT use window representations directly. Aggregation never mixes subjects, cases or events.

Classification and heart rate are scored per window, age, BMI and blood pressure per subject, and VTaC per event. PAT first averages absolute window errors within each subject or case, then gives groups equal weight. Thus window counts, demographic subject counts and alarm-event counts refer to different evaluation units.

Across-fold summaries. For BEATFORMER's ScientISST activity, Simultaneous load and heart rate, SensSmartTech state, PTT PPG heart rate and activity, Li ECG-PPG PAT, BUT PPG quality, WESAD stress, PPG-DaLiA activity, age and BMI, and VTaC false-alarm detection, metrics are first computed in each outer test fold and then averaged with equal fold weights. Standard deviations use the number of folds as the denominator. They describe differences across test folds, not variation across independent pre-training seeds or confidence intervals for the mean.

F.4 Classification and Regression Metrics

For N scored units with reference values y_i and predictions $\hat{y}_i$,

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|.$$

Heart-rate, age, BMI and blood-pressure errors are in bpm, years, kg/m^2 and mmHg, respectively; N counts the actual scoring units.

PAT uses group-equal MAE. Let S be the number of test groups. For held-out windows $\mathcal{I}_s$ of group s, with $N_s = |\mathcal{I}_s|$,

$$\mathrm{MAE}_s = \frac{1}{N_s} \sum_{i \in \mathcal{I}_s} |\hat{y}_i - y_i|, \qquad \mathrm{MacroMAE} = \frac{1}{S} \sum_{s=1}^{S} \mathrm{MAE}_s \,.$$

Predictions from all five test folds are pooled before aggregation. Groups are subjects, or cases for UQ. The metric is in ms and weights groups equally despite different recording lengths.

Multiclass accuracy is the proportion of correct predictions; balanced accuracy is mean class recall. Binary AUROC uses continuous scores. VTaC sensitivity, specificity, precision and F1 quantify true-alarm detection, rejection of false alarms, the fraction of predicted true alarms that are correct, and the harmonic mean of precision and sensitivity, respectively. Validation alone selects the decision threshold.

F.5 Waveform Recovery Metrics

Error in missing regions. Let $x_{i,c,t}$ and $\hat{x}_{i,c,t}$ be the reference and reconstruction for window i, channel c and point t. Let $m_{i,c,t} \in \{0, 1\}$ indicate whether that point is scored, with at least one scored point per evaluated window and channel.

$$r_{i,c}^{\text{mask}} = \sqrt{\frac{\sum_t m_{i,c,t}(\hat{x}_{i,c,t} - x_{i,c,t})^2}{\sum_t m_{i,c,t}}}.$$

Condition-level RMSE pools valid scoring points across windows. For window RMSE r_i and point count n_i,

$$R_{\text{pooled}} = \sqrt{\frac{\sum_i n_i r_i^2}{\sum_i n_i}}.$$

Channel-specific scores pool ECG or PPG points separately; joint scores pool both channels. Condition-group means equally average the constituent pooled RMSE values. For the clean-input condition, let r_i^{clean} denote the joint-channel recovered RMSE reported for test window i. Clean-input distortion is the arithmetic mean $N^{-1} \sum_{i=1}^{N} r_i^{\text{clean}}$ over $N = 2{,}654$ windows.

SNR improvement. For window i, let Ω_i contain the scored points from both ECG and PPG in the target beat. The clean reference x_i, corrupted input u_i, and recovered output $\hat{x}_i$ are expressed in each model's native normalized coordinates. Let $r_i^{\text{in}} = \text{RMSE}_{\Omega_i}(u_i, x_i)$ and $r_i^{\text{out}} = \text{RMSE}_{\Omega_i}(\hat{x}_i, x_i)$, where each RMSE pools squared errors over the same reference points in both channels before taking the square root. For positive input and output RMSE, window-level improvement is

$$\Delta\,\text{SNR}_i = 20 \log_{10} \frac{r_i^{\text{in}}}{r_i^{\text{out}}}.$$

This error-ratio measure is expressed in decibels (dB): positive values indicate reduced error, zero indicates no change, and negative values indicate increased error. For each model and noise–severity condition, Table 4 reports the median of the 2,654 window-level dB values, not a dB value calculated from pooled condition-level RMSE. Figure G.3 shows the same window-level values: boxes span the 25th–75th percentiles, center lines mark medians, and whiskers span the 5th–95th percentiles. The common test conditions and reported measures are summarized in Table F.3. Clean inputs have zero input error and are evaluated separately using output RMSE.

Peak-to-peak amplitude error. For beat waveform x, define $A(x) = \max_t x_t - \min_t x_t$. The signed error of corrupted or recovered waveform u is

$$e_A(u, x) = A(u) - A(x).$$

Positive and negative values indicate an expanded and compressed amplitude range, respectively. ECG and PPG are evaluated in normalized amplitude coordinates. Joint-distribution contours enclose 90% probability mass.

G Supplementary Experimental Results

G.1 Additional Classification Results

G.1.1 Variation Across Held-Out Folds

ScientISST activity accuracy has a median of 0.719 and ranges from 0.037 to 0.923 across 16 LOSO folds. Simultaneous load accuracy has a median of 0.581 and ranges from 0.360 to 0.754 across

Table G.1: VTaC test-event counts for the illustrated split at its validation-selected threshold of 0.38. Rows are reference labels; label 1 corresponds to the original `decision=true`.

Reference	Predicted false alarm (0)	Predicted true alarm (1)	Total
False alarm (0)	260	63	323
True alarm (1)	28	92	120
Total	288	155	443

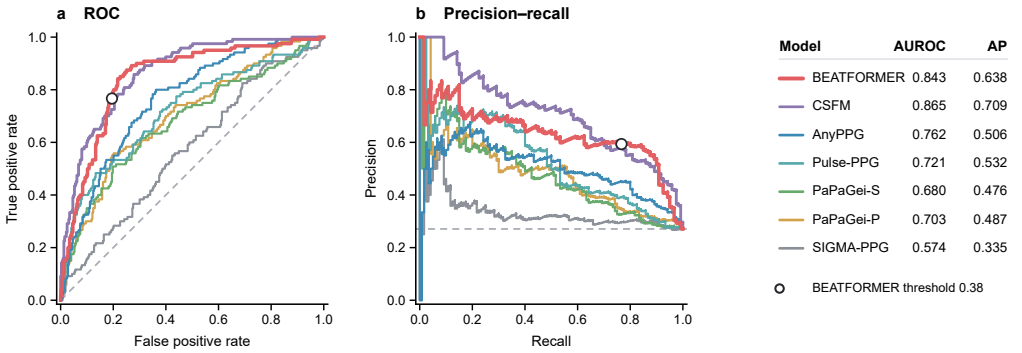


Figure G.1: VTaC ROC and precision–recall curves for seven models on the same 443 test events (120 true and 323 false alarms), excluding validation events. Open circles mark BEATFORMER's operating point at the validation-selected threshold of 0.38. Gray dashed lines denote random ranking and positive prevalence (0.271). The figure reports split-specific AUROC and AP; AP uses recall-increment weighting, not trapezoidal integration.

13 LOSO folds. Each fold tests a different subject; variation also reflects window counts and class composition. SensSmartTech AUROC ranges from 0.779 to 0.916, with median 0.892, across five grouped folds. These group-level metrics have a different statistical unit from single-subject LOSO scores.

Across five subject-grouped folds, BUT PPG quality AUROC ranges from 0.787 to 0.877, with median 0.841, and WESAD stress AUROC ranges from 0.883 to 0.960, with median 0.921. PTT PPG activity accuracy ranges from 0.389 to 0.924, with median 0.655, across 22 LOSO folds. PPG-DaLiA activity accuracy ranges from 0.301 to 0.739, with median 0.503, across 15 LOSO folds. These ranges show variation that mean performance alone does not describe.

G.1.2 Alarm-Event Operating Point

For the illustrated VTaC split, the validation-selected threshold is 0.38. On its 443 test events, it identifies 92 of 120 true alarms and rejects 260 of 323 false alarms (Table G.1). Sensitivity, specificity, precision and F1 are 0.767, 0.805, 0.594 and 0.669, respectively. These split-specific results describe the operating-point trade-off; they are not five-fold averages.

Figure G.1 compares score-ranking performance for seven models on the same 443 test events. BEATFORMER ranks second in both AUROC (0.843) and average precision (AP, 0.638), behind CSFM (0.865 and 0.709) and ahead of the other five baselines. Its AP exceeds the true-alarm prevalence of 0.271, the reference level for random ranking in the precision–recall panel.AP weights precision by recall increments and differs from precision at a fixed threshold.

G.2 Additional Heart-Rate and Demographic Results

Individual differences in heart-rate error. Across 13 Simultaneous LOSO folds, heart-rate MAE has median 9.022 bpm and range 5.406–20.237 bpm. Mean signed error, $\mathrm{ME} = N^{-1}\sum_i(\hat{y}_i - y_i)$, ranges from -17.194 to 8.689 bpm. Positive and negative values indicate over- and underestimation. A single aggregate obscures these subject-dependent directions. PTT PPG heart-rate MAE ranges from 2.105 to 10.348 bpm across five subject-grouped folds, with median 2.331 bpm and mean

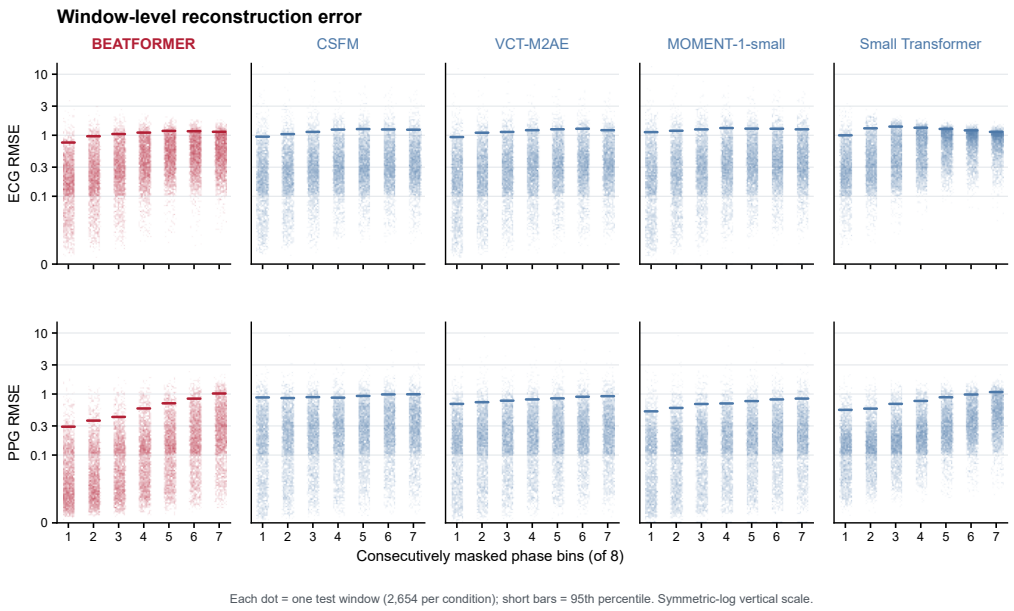


Figure G.2: Window-level ECG and PPG masked-region RMSE for five models under one to seven consecutively masked phase bins. Each dot represents one of the same 2,654 test windows per condition; horizontal jitter is for display only. Short horizontal bars mark the 95th percentile. The symmetric-logarithmic vertical scale retains all observations, including high-error windows. Models retain their native normalized amplitude and scoring coordinates. These distributions are not subject-level confidence intervals.

4.025 bpm. High-error folds raise the mean; these group-level results are interpreted separately from Simultaneous LOSO scores.

Age and BMI variation. Each of the five PPG-DaLiA test folds contains three subjects. Age MAE ranges from 1.071 to 10.602 years; BMI MAE ranges from 0.663 to 2.122 kg/m^2 on the same splits. Both tasks have positive and negative fold-level mean signed errors. With three subjects per fold, these results describe variation among the 15 evaluated subjects rather than performance in particular age or BMI subpopulations.

G.3 Masked Waveform Reconstruction

G.3.1 Window-Level Variation Under Within-Beat Masking

Figure G.2 shows the masked-region RMSE of every test window for each model and masking level. BEATFORMER has the numerically lowest 95th-percentile ECG error with one to six masked bins: 0.760 versus the next-lowest 0.934 at one bin, and 1.164 versus 1.202 at six bins. With seven bins, Small Transformer is marginally lower (1.137 versus 1.139). For PPG, BEATFORMER has the lowest 95th-percentile error with one to five bins (0.292 versus the next-lowest 0.520 at one bin; 0.705 versus 0.766 at five), but not with six or seven. Its median ECG window error is lowest only with one or two bins masked, showing that typical-window and high-error-tail rankings can differ. These window-level quantiles differ from the pooled condition RMSE defined in Appendix F.5.

G.3.2 Consecutive Whole-Beat Masking

Removing one to four complete beats leaves no visible samples within the masked beats. Recovery relies on inputs outside the missing interval. Table G.2 reports joint-channel RMSE. BEATFORMER error rises from 0.6419 to 0.7114 as the gap grows from one to four beats. Values retain native scoring coordinates and do not establish a common-amplitude-scale ranking.

Table G.2: Whole-beat reconstruction RMSE, pooled over ECG and PPG scoring points in missing regions. Each model retains its native input, normalization and scoring protocol. The last column gives valid windows per model.

Missing beats	BEATFORMER	CSFM	VCT-M2AE	MOMENT-1-small	Small Transformer	Windows
1	0.6419	0.5678	0.5363	0.5443	0.7970	2654
2	0.6963	0.6371	0.6227	0.6358	0.9052	2648
3	0.7078	0.6765	0.6764	0.6966	0.9269	2644
4	0.7114	0.7013	0.7059	0.7244	0.9184	2622

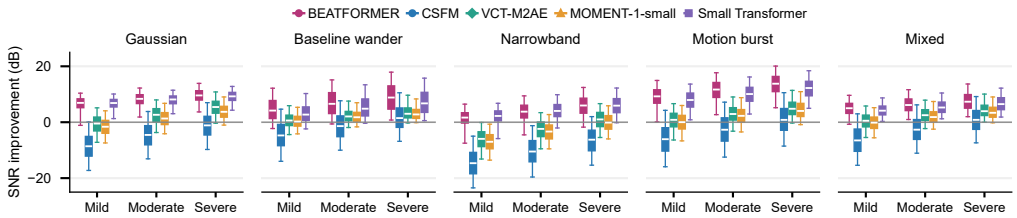


Figure G.3: SNR improvement across five noise types and three noise levels. Boxes show quartiles and medians; whiskers span the 5th–95th percentiles. Positive values indicate reduced error.

G.3.3 Reconstruction of a Fixed Test Window

Figure 4(c) shows cross-beat contiguous masking, two missing beats and a two-second gap in the same test window. Under cross-beat masking, BEATFORMER continues local ECG turns and the PPG trough between visible context. Under two-beat and two-second gaps, the output retains the arrangement of several neighboring peaks and troughs in this example. Local amplitudes and details do not coincide everywhere and should be considered alongside masked-region errors. The fixed example illustrates local morphology; Table G.2 provides whole-test-set evidence.

G.4 Denoising and Clean-Input Preservation

Figure G.3 shows the window-level distribution of SNR improvement for the 15 denoising conditions summarized in Table 4.

G.4.1 Clean-Input Distortion

For 2,654 uncorrupted test windows passed through the BEATFORMER denoising model, mean window output RMSE is 0.040986, median RMSE is 0.026616 and the interquartile range is [0.021358, 0.037093]. Thus the recovery model makes nonzero changes even without input corruption. The mean exceeding the median reflects higher-error windows. This control measures distortion in native normalized coordinates, separately from recovery gains on corrupted inputs.

G.4.2 Recovery of a Corrupted Beat

Figure G.5 shows a fixed target beat under severe baseline drift, motion artifacts and mixed noise. Baseline drift changes the slowly varying PPG contour; in this example, BEATFORMER brings the upslope and post-peak descent closer to the reference. Recovery attenuates rapid local fluctuations from motion artifacts. Under mixed noise, major ECG peaks and PPG contours remain visible, although small deviations persist. These local observations complement the clean-input control above.

G.5 PAT Regression and Error Distributions Across Four Datasets

The main text reports Li ECG–PPG PAT regression; Table G.3 adds PTT-PPG, UQ and ErgoBP. Frozen-representation predictors are trained and evaluated separately within each dataset using five group-disjoint folds and group-equal MAE. This is not cross-dataset training and testing. BEATFORMER achieves 18.945 ms on PTT-PPG and 87.500 ms on UQ, below the evaluated baselines;

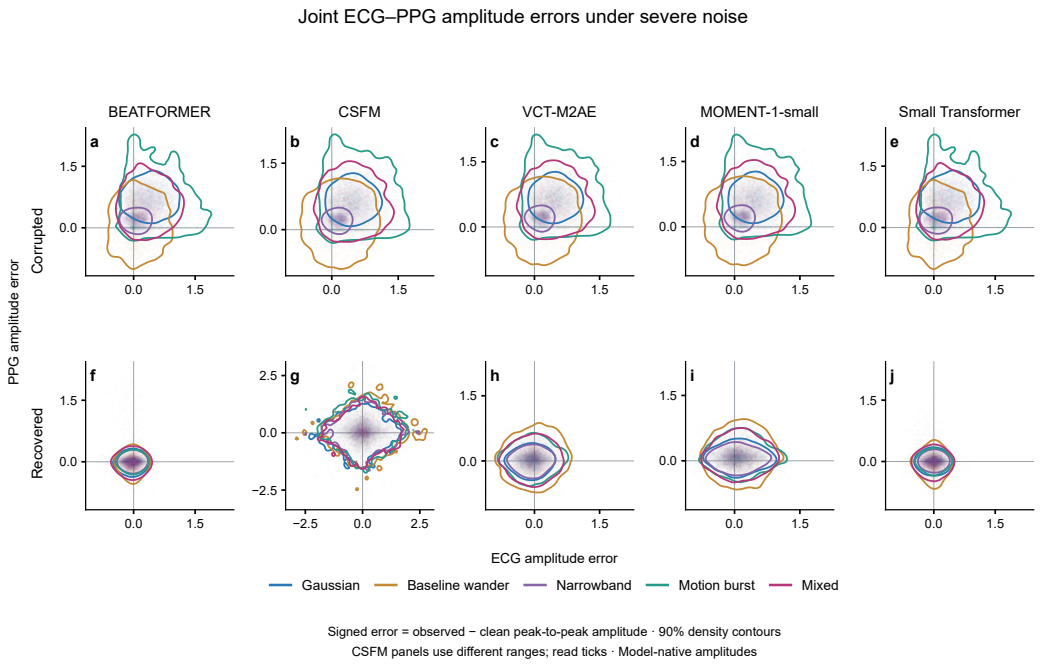


Figure G.4: Signed ECG–PPG peak-to-peak amplitude errors for five models under five severe noise types, before (top) and after (bottom) recovery. Errors are observed minus clean amplitudes; contours enclose 90% of estimated density. The CSFM panels use different axis ranges, as indicated by their ticks. Amplitude coordinates are model-native and do not support absolute cross-model error ranking.

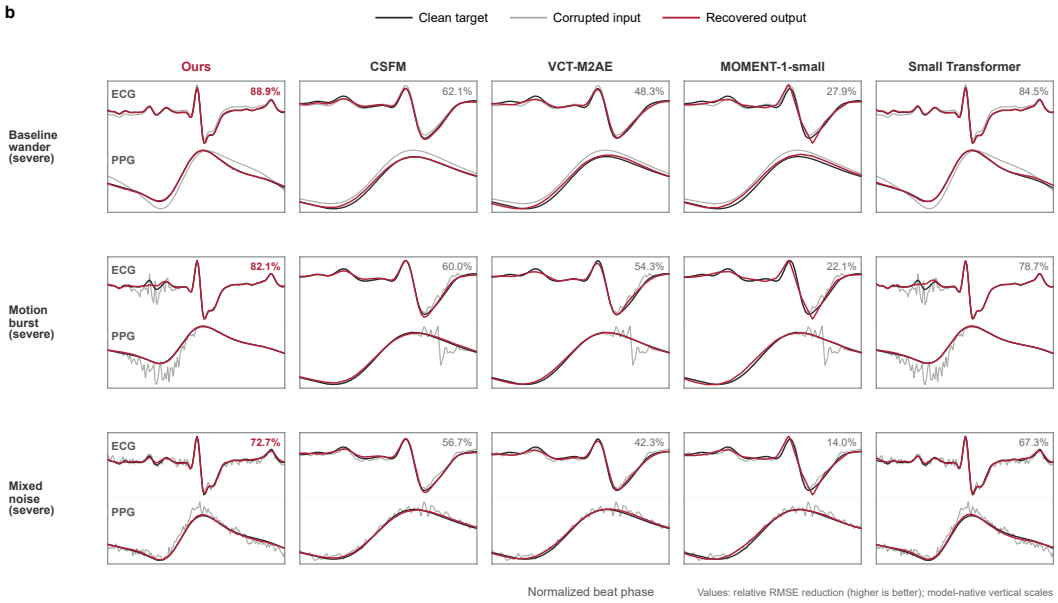


Figure G.5: Denoising the target beat of `vtac:004f0c_2475_w00`. Rows show severe baseline drift, motion artifacts and mixed noise; columns show the five recovery models. Black, gray and red curves denote clean references, corrupted inputs and recovered outputs. Axes use normalized within-beat phase and native normalized amplitudes. References and corruptions follow each model's protocol. Insets give this example's joint-channel relative RMSE reduction, $100(1 - r^{\text{out}}/r^{\text{in}})\%$.

Table G.3: PAT group-equal MAE (ms) from pooled five-fold held-out predictions. PTT-PPG and ErgoBP weight subjects equally; UQ weights cases equally. BEATFORMER and CSFM use ECG–PPG; other models use PPG. Predictor fitting and model-selection procedures are specified in Appendix F.1.

Model	PTT-PPG	UQ	ErgoBP
BEATFORMER	18.945	87.500	93.089
CSFM	28.543	93.863	87.035
AnyPPG	26.602	121.741	88.332
Pulse-PPG	25.708	116.715	78.868
SIGMA-PPG	24.804	116.070	85.602
PaPaGei-S	24.863	111.074	94.067
PaPaGei-P	29.592	118.003	89.790

the lowest-error baselines are SIGMA-PPG and CSFM, respectively. On ErgoBP, Pulse-PPG has the lowest error, 78.868 ms, compared with BEATFORMER's 93.089 ms.

G.5.1 Prediction Bias and Dispersion

Bland–Altman analysis. Following Bland and Altman (Bland & Altman, 1986), we define each window's difference and mean from five-fold held-out predictions as

$$d_i = \hat{y}_i - y_i, \qquad a_i = \frac{\hat{y}_i + y_i}{2}.$$

Positive differences indicate overestimation. Plotting d_i against a_i shows how errors vary with PAT level. Bias and sample standard deviation are

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i, \qquad s_d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2}.$$

The descriptive bounds are $\bar{d} \pm 1.96 s_d$. Subjects and cases contribute multiple windows, so these equally weighted pooled-window bounds describe test-error distributions rather than repeated-measures-adjusted agreement.

PTT-PPG and Li ECG–PPG have mean biases of 1.134 and -1.624 ms, but difference standard deviations of 33.336 and 25.424 ms. Small overall bias therefore coexists with window-level dispersion. UQ and ErgoBP have positive mean biases and standard deviations above 120 ms. These statistics complement group-equal MAE: the former describe window-level bias and spread, whereas the latter weights each subject or case equally when measuring absolute error.

G.6 Component Ablation Analysis

To examine the contribution of each component, we train full BEATFORMER and six ablation variants from the same joint-training initialization through the complete second stage of pretraining. All settings use the same 5,261,696 paired ECG–PPG samples, input processing, and training schedule. We fix the epoch-14 checkpoint for downstream evaluation without selecting a checkpoint based on downstream performance. The original schedule uses global batch sizes of 64 for epochs 1–5, 32 for epochs 6–10, and 64 for epochs 11–14; epoch 14 corresponds to 1,562,066 cumulative optimizer updates.

The architecture ablations separately disable within-beat phase attention, across-beat attention, or cumulative-time encoding. For each attention ablation, we zero the output of the corresponding branch in every layer while retaining the other branches and residual paths. For the time-encoding ablation, we zero the module's complete output, including its projection bias, while preserving the input timestamps and other processing steps. The objective ablations separately set the PAT, morphology, or HRV loss weight to zero. The remaining loss weights retain their default values and are not renormalized. Each change remains active throughout the variant's second-stage training and downstream evaluation.

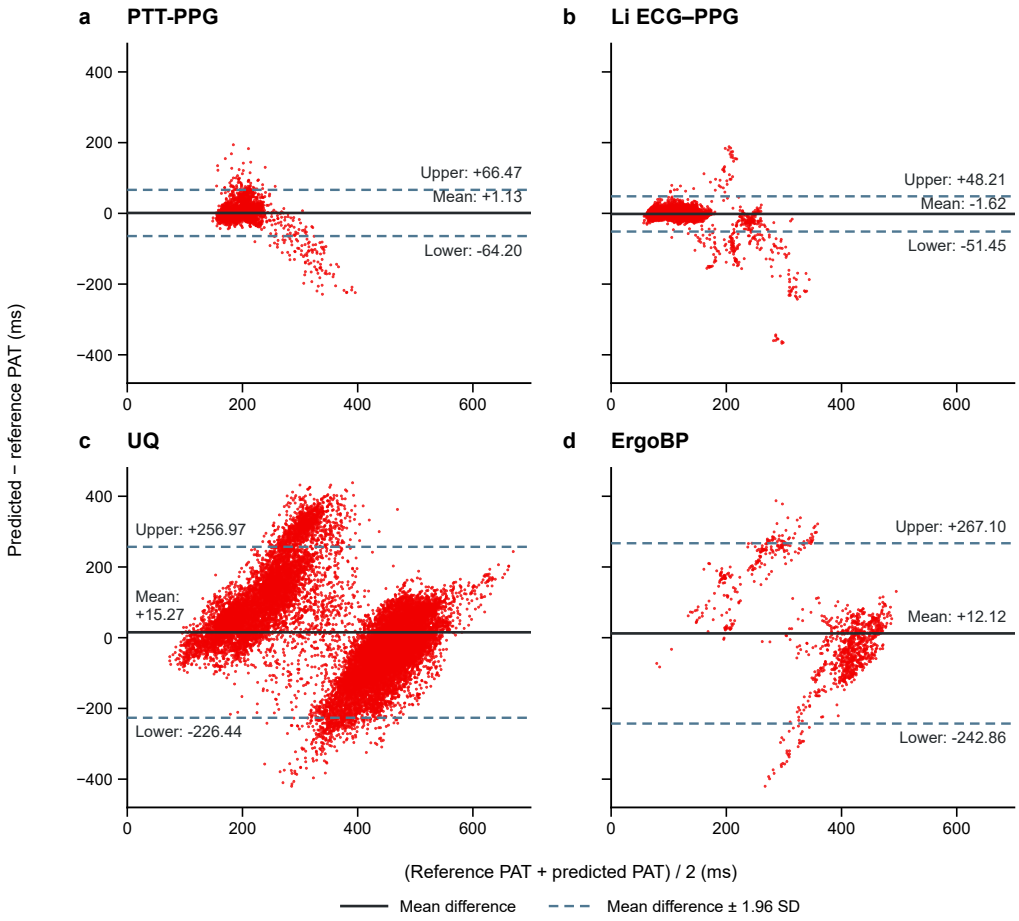


Figure G.6: BEATFORMER PAT Bland–Altman plots for PTT-PPG, Li ECG–PPG, UQ and ErgoBP. Each point is a held-out window from five group-disjoint folds with seed 42. Axes show the reference–prediction mean and prediction minus reference, both in ms. Solid lines mark mean differences; dashed lines mark mean differences ± 1.96 sample standard deviations. Panels share axis ranges and retain all 30,630 windows, including outliers: 3,189, 7,860, 18,553 and 1,028 per dataset. Group counts are in Table E.2. Bounds are descriptive, not repeated-measures-adjusted or clinical acceptance limits.

For each trained checkpoint, we freeze the backbone, extract that model's own downstream representations, and refit the prediction head using the same data splits. Table 5 reports percentage changes in mean MAE relative to full BEATFORMER for Simultaneous and PTT PPG heart-rate estimation and Li ECG–PPG systolic and diastolic blood-pressure estimation.

Every reported change is positive. PTT PPG heart-rate MAE increased more than Simultaneous heart-rate MAE for each variant. Removing morphology supervision produced the largest increases in both heart-rate tasks (45.0% and 55.9%); removing PAT supervision increased them by 17.9% and 30.0%, respectively. For systolic blood pressure, removing morphology or HRV supervision increased MAE by 14.5% and 10.3%. For diastolic blood pressure, removing within-beat phase attention and morphology supervision produced the largest increases, 12.4% and 12.1%. The relative effect of each removal therefore differed across the four tasks.

Table G.4: Pretraining hyperparameter values used in the sensitivity analysis. Each non-default setting changes only the specified parameter.

Parameter	Lower value	Default	Higher value
Masking ratio	0.15	0.60	0.30
Learning rate	1×10^{-5}	2×10^{-5}	4×10^{-5}
PAT loss weight	0.25	0.50	1.00
Morphology loss weight	0.10	0.20	0.40
HRV loss weight	0.10	0.20	0.40
Dropout	0.05	0.10	0.20

Table G.5: Downstream performance under different pretraining hyperparameter settings. MAE is reported for Simultaneous HR and Li ECG–PPG SBP and DBP (lower is better); ordinary accuracy is reported for Simultaneous Load (higher is better). The default and variant settings use the same downstream splits and evaluation procedure.

Setting	HR MAE	Load accuracy	SBP MAE	DBP MAE
Default	9.861	0.569	10.752	7.114
Mask ratio 0.15	9.817	0.554	10.477	7.101
Mask ratio 0.30	9.958	0.559	10.800	7.218
Learning rate 1×10^{-5}	9.819	0.559	10.882	7.206
Learning rate 4×10^{-5}	9.764	0.563	10.639	7.132
PAT weight 0.25	9.867	0.558	11.019	7.224
PAT weight 1.00	9.951	0.565	10.420	7.083
Morphology weight 0.10	10.346	0.563	10.877	7.161
Morphology weight 0.40	9.855	0.561	10.899	7.112
HRV weight 0.10	9.804	0.561	10.625	7.101
HRV weight 0.40	9.890	0.563	10.841	7.141
Dropout 0.05	10.166	0.562	10.887	7.212
Dropout 0.20	9.853	0.558	10.648	7.174

G.7 Hyperparameter Sensitivity

Using the same training and evaluation conditions, we vary one pretraining hyperparameter at a time and keep the others at their default values. We examine the masking ratio, learning rate, three auxiliary-loss weights, and dropout (Table G.4). The auxiliary-loss weights remain nonzero in this analysis; setting a weight to zero is part of the component ablation study above.

Across the tested values in Table G.5, Simultaneous HR MAE ranges from 9.764 to 10.346 bpm, and Load accuracy ranges from 0.554 to 0.569. SBP and DBP MAEs range from 10.420 to 11.019 mmHg and from 7.083 to 7.224 mmHg, respectively. Reducing the morphology loss weight from its default of 0.20 to 0.10 increases HR MAE by about 4.9%; increasing it to 0.40 gives a result close to the default. Raising the PAT loss weight from 0.25 to 1.00 reduces SBP and DBP MAEs by about 5.4% and 2.0%, respectively, while increasing HR MAE by about 0.9%. Responses to parameter changes differ across tasks, and no tested setting performs best on every task.

H Limitations

While BEATFORMER advances beat-resolved foundation modeling for synchronized ECG–PPG signals, it has several limitations. First, the current formulation is based on detected cardiac anchors and a fixed beat–phase representation. Although this design provides a consistent coordinate system for local morphology and rhythm, it may not fully capture variability across cardiac rhythms, recording conditions, and temporal scales. Future extensions could investigate adaptive beat construction, uncertainty-aware alignment, and longer temporal contexts. Second, the model is developed and evaluated primarily on paired ECG–PPG recordings and a defined set of downstream tasks. Extending the pretraining corpus and evaluation to additional devices, acquisition conditions, populations, longitudinal recordings, and complementary physiological modalities would help characterize its generalization and robustness. Developing unimodal variants that operate on ECG or PPG alone

could further reduce the dependence on synchronized multimodal acquisition and extend the model to settings in which only one signal is available. Finally, the current pretraining objectives focus on waveform reconstruction, PAT, PPG morphology, and HRV, while broader physiological states and clinically relevant tasks remain to be explored. Future work could incorporate additional physiological supervision and multi-task learning objectives to improve the interpretability and downstream utility of the learned representations. Despite these limitations, BEATFORMER provides a structured framework for modeling synchronized cardiac signals, and these directions may further improve its temporal modeling, generalization and practical utility.