PULSE2VIEW: SAME-SENSOR CROSS-VIEW PRE-TRAINING FOR PATIENT-LEVEL MULTI-DISEASE CLASSIFICATION FROM WEARABLE PPG

Anonymous authors
Paper under double-blind review

ABSTRACT

Photoplethysmography (PPG) is a widely available biosignal in consumer wearables, enabling scalable, low-burden cardiovascular screening. However, labeled PPG datasets are scarce, and single-view self-supervised pre-training can underexploit the complementary morphology, rhythm, and time-frequency structure of PPG. To address this, we propose Pulse2View (P2V), a same-sensor cross-view pre-training framework aligning two views of each 30-second recording without an additional sensing modality. For each of three consecutive 10-second segments, the continuous wavelet transform (CWT) branch constructs a three-channel Ricker-CWT image from the exported PPG and its first- and second-order differences, while a Conv1D-BiGRU encodes the full sequence as a temporal (time-domain) view. Symmetric InfoNCE aligns the two views without disease labels. Both encoders are transferred and fused end-to-end downstream, while inference remains PPG-only. Pre-trained on 99,701 unlabeled 30-second PPG windows, P2V was evaluated on the private hospital-derived Clinical Wearable PPG Cohort (CWPC; 14,811 patients and 100,657 windows). It achieved the highest three-seed mean on all four primary patient-level metrics among evaluated models, including a coronary heart disease (CHD) AUROC of 0.7859 ± 0.0044 and a Macro AUROC of 0.7150 ± 0.0012, exceeding a public image-pretrained ConvNeXtV2-Tiny initialization by 3.2 and 1.4 AUROC points and the strongest public baseline, MOMENT-small, by 2.6 and 0.8 AUROC points. At 25% and 50% label budgets, P2V improves over the image-pretrained initialization and a two-CWT-view control, while achieving the highest CHD and Macro AUPRC among these configurations. Paired bootstrap and ablations support the contributions of cross-view alignment and the pre-trained temporal branch. On the public Liu PPG arrhythmia six-class dataset, P2V also achieves a window-level Macro AUROC of 0.9697 ± 0.0011. Together, these results support label-efficient cross-view pre-training for PPG-only patient-level multi-label classification.

1 INTRODUCTION

Photoplethysmography (PPG) is now widely available in consumer wearables such as smartwatches and wristbands, and it is easier to acquire over long periods than ECG, which requires electrode contact, making it well suited to population-level screening and risk stratification (Pillai et al., 2025; Saha et al., 2025). We target a PPG-only deployment setting, where inference cannot assume an additional sensor or modality. However, extending this deployment constraint into a claim that pre-training must therefore use a single representation makes it difficult to model the morphology, rhythm, and time–frequency structure of a 30-second PPG recording together. **We argue that being restricted to a single sensor at inference time does not mean that pre-training must be restricted to a single representation.** The pronounced inter-subject variability of PPG further makes representation transfer under patient-disjoint evaluation a central requirement (Ghorbani et al., 2023).

Self-supervised learning on time series and physiological signals has constructed complementary supervision along several axes: PITS embeds temporal patches independently to model local structure, while FactorCL and FOCAL address redundancy and complementarity through view factorization

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.

and multimodal alignment, respectively (Lee et al., 2024; Liang et al., 2023; Liu et al., 2023). CoST learns disentangled seasonal–trend representations for forecasting (Woo et al., 2022), while TSPulse disentangles temporal, spectral, and semantic views for time-series diagnostics (Ekambaram et al., 2026). TF-C jointly learns time-domain and frequency-domain representations of the same series with a consistency objective (Zhang et al., 2022). This leaves open whether such a combination benefits PPG-only patient-level learning: two views of the same recording that differ in representation domain—a local multi-scale CWT image and the full waveform—modeled separately and then transferred jointly downstream.

We propose Pulse2View (P2V), a same-sensor cross-view pre-training framework that represents each 30-second exported PPG recording in two views. The first is an image-domain CWT time–frequency view, which we construct from three consecutive 10-second segments with the Ricker-CWT. The second is a full-waveform temporal view, which we obtain by encoding the complete sequence with a Conv1D-BiGRU. Both views are derived from the same PPG recording but introduce different inductive biases through different representation domains and dedicated encoders. Joint transfer lets the two pre-trained representations continue to participate in end-to-end adaptation on patient-level tasks, while inference itself remains PPG-only. We pre-train P2V on 99,701 unlabeled 30-second PPG windows. We evaluate the design on the Clinical Wearable PPG Cohort (CWPC; 14,811 patients and 100,657 windows) with patient-level main results, structural ablations, and label-efficiency experiments. We further evaluate transfer on a public PPG arrhythmia dataset under its window-level protocol.

We address how to exploit heterogeneous representations of the same signal under PPG-only inference, with the following contributions:

(1) Formulating heterogeneous representation learning under PPG-only deployment. For patient-level PPG multi-label classification, we study how to jointly exploit local multi-scale time–frequency structure and full-waveform dynamics from a single recording, without adding a sensing modality or a non-PPG input.

(2) P2V, a same-sensor heterogeneous cross-view framework. P2V pairs an image-domain Ricker-CWT view with a full-waveform PPG temporal view and models them with a vision encoder and a Conv1D-BiGRU. Both branches are pre-trained with a symmetric InfoNCE objective, then transferred and fused end-to-end for downstream prediction.

(3) Patient-level evidence and design conclusions. On CWPC, P2V attains the highest three-seed mean among evaluated models on all four primary patient-level metrics, including a coronary heart disease (CHD) AUROC of 0.7859. It also retains an advantage over both a public image-pretrained initialization and a two-CWT-view control at 25%/50% label budgets. Our ablations support the roles of heterogeneous view pairing and temporal-branch pre-training, and on the public arrhythmia task we reach a window-level Macro AUROC of 0.9697.

2 Related Work

PPG representation learning and wearable disease modeling. General time-series foundation models such as MOMENT (Goswami et al., 2024) and UniTS (Gao et al., 2024) build general-purpose representations for cross-task learning and transfer, whereas PPG-specific models draw on signal-informed objectives, data, and priors: PaPaGei separates participant-aware from morphology-aware objectives (Pillai et al., 2025), and Pulse-PPG and SIGMA-PPG improve PPG representations using field-collected data and statistical priors, respectively (Saha et al., 2025; Guo et al., 2026). Foundation models for multiple wearable physiological signals instead cover signal types such as PPG, ECG, and EEG, as in NormWear (Luo et al., 2026). Cross-subject transfer is a key dimension for assessing the value of self-supervised PPG representations (Ghorbani et al., 2023), and other work explores PPG–ECG association and cardiovascular risk modeling (Ding et al., 2026; Wei et al., 2025). These efforts improve general time-series objectives, PPG-specific priors, or multi-signal modeling; P2V instead constructs cross-view supervision from heterogeneous representations of the same PPG without adding a sensor.

Physiological time-series self-supervision and cross-view learning. Time-series self-supervision constructs label-free supervision through mechanisms such as masked reconstruction,

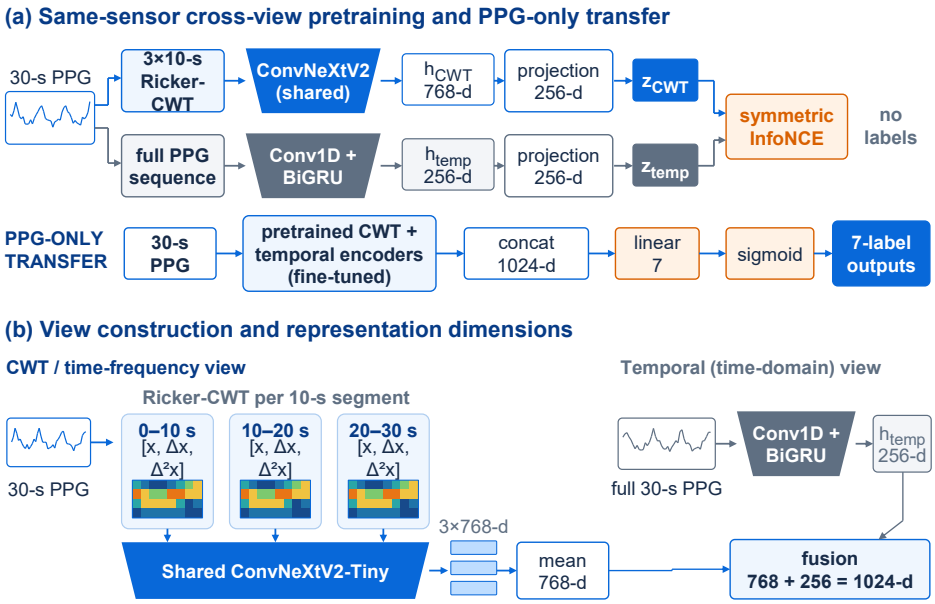

Figure 1: Overview of P2V. (a) The CWT and temporal views of the same exported PPG are pretrained with symmetric InfoNCE, and both encoders are transferred to PPG-only downstream fusion; (b) the three Ricker-CWT segments, the encoders, and the representation dimensions.

independent patch modeling, view factorization, and multimodal alignment (Lee et al., 2024; Liang et al., 2023; Liu et al., 2023). A second line of work draws its supervision from an additional modality or sensor: BioX-Bridge transfers knowledge across physiological modalities (Li et al., 2026), sleep2vec aligns heterogeneous nocturnal biosignals (Yuan et al., 2026), and SleepSMC targets sleep staging under a train-multimodal, infer-unimodal setting (Ma et al., 2025). P2V extends cross-view supervision to the single-sensor PPG setting: view diversity comes from an image-domain time–frequency representation and a full-waveform representation of the same exported PPG, requiring neither a paired second sensor nor a second physiological modality.

Time–frequency representation and same-sensor heterogeneous views. TF-C is the closest time–frequency self-supervised comparator to P2V: it jointly learns time-domain and frequency-domain representations of the same series and concatenates both representations downstream (Zhang et al., 2022). CoST learns disentangled representations along seasonal–trend dimensions (Woo et al., 2022), and TSPulse learns disentangled temporal, spectral, and semantic views (Ekambaram et al., 2026). In a supervised setting, XSleepNet likewise models physiological sequences together with their time–frequency images for sleep staging (Phan et al., 2022). These works show that time-domain and frequency-domain information can be aligned or disentangled. P2V combines both ingredients, pairing a local multi-scale image-domain Ricker-CWT view with a full-waveform temporal view of the same recording and introducing complementary inductive biases through a vision encoder and a sequence encoder.

3 Method

3.1 Problem Definition and Overall Pipeline

Let $\mathcal{D} = \{x^{(n)}\}_{n=1}^{N}$ be a set of 30-second PPG windows without disease labels. P2V aims to construct, from each window alone and without disease labels, two views that live in different representation domains, and to learn transferable encoders through cross-view alignment. The two views are complementary: an image-domain time–frequency view that characterizes local multi-scale pulse morphology and frequency content, and a full-waveform temporal view that preserves continuous temporal relations and rhythm. After pre-training, both encoders are transferred to a downstream model that predicts seven diseases per window and aggregates these predictions into patient-level

disease probabilities. Figure 1(a) summarizes the pre-training and transfer pipeline, and Figure 1(b) gives the view construction, tensor dimensions, and fusion path.

We use "exported PPG" for the PPG time series that a device or backend emits after its own internal processing, to distinguish it from the unprocessed raw optical signal. Input windows are sampled at 100 Hz over 30 seconds; length alignment and value normalization follow the input contract of Section 4.1, yielding $x \in \mathbb{R}^{3000}$. The CWT view partitions x into three consecutive 1,000-point segments and constructs time–frequency images; the temporal view uses the full 3,000-point sequence directly.

3.2 CWT View Encoder

For the i-th consecutive 1,000-point segment x_i, we compute the Ricker-wavelet CWT (Ricker, 1944; Torrence & Compo, 1998) of the exported PPG and of its first-order difference Δx_i and second-order difference $\Delta^2 x_i$ separately. We then stack the three transforms along the channel dimension into a three-channel time–frequency image

$$c_i \in \mathbb{R}^{3\times S\times L}, \tag{1}$$

where S is the number of scales and L is the segment length. The three channels provide complementary descriptions of local variation: the exported PPG waveform characterizes pulse morphology, the first-order difference emphasizes rising and falling edges, and the second-order difference highlights turning points and local curvature. Before entering the vision encoder, the time–frequency image passes through image preprocessing and resizing.

This image lies in the image domain and is therefore modeled by a vision backbone, while the exported one-dimensional series is handled by the temporal branch. The three segments share a single ConvNeXtV2-Tiny encoder E_v (Woo et al., 2023), each segment producing a 768-dimensional vector; we take their arithmetic mean to obtain the window-level CWT representation

$$h_{\mathrm{CWT}} = \frac{1}{3}\sum_{i=1}^{3} E_v(c_i) \in \mathbb{R}^{768}. \tag{2}$$

Sharing weights lets the encoder learn one set of local multi-scale structures across the three segments, and mean pooling gathers the evidence of all three into a single window-level representation. Temporal order within and across segments, in turn, is modeled by the temporal branch on the full waveform. Image size, wavelet scales, and difference boundaries are given in the supplementary material.

3.3 PPG Temporal View and Cross-View Objective

The PPG temporal (time-domain) branch is a Conv1D-BiGRU encoder E_t: one-dimensional convolutions extract local sequence features over multiple receptive fields, and a bidirectional GRU integrates context across the full window, giving

$$h_{\mathrm{temp}} = E_t(x) \in \mathbb{R}^{256}. \tag{3}$$

Per-layer channel counts, kernel sizes, and normalization settings are given in the supplementary material. The two representations h_{CWT} and h_{temp} are those transferred downstream; cross-view alignment is performed by two independent 256-dimensional projection heads,

$$z_{\mathrm{CWT}} = g_v(h_{\mathrm{CWT}}), \qquad z_{\mathrm{temp}} = g_t(h_{\mathrm{temp}}). \tag{4}$$

The projection heads take part in alignment only during pre-training and are not transferred to the downstream classifier.

During pre-training we apply a source-consistent weak augmentation to each of the two views separately. The CWT branch applies gain, Gaussian noise, and low-frequency baseline drift to the waveform before generating the time–frequency image, while the temporal branch encodes the full waveform after a second, independent augmentation. Pre-training builds batches from patient-group tokens, sampling at most one window per group per batch. Let a batch contain B windows. After ℓ_2 normalization, the two views' in-batch representations are compared by cosine similarity at

temperature τ, and the objective from CWT to temporal is

$$\ell_{v\to t}(k) = -\log \frac{\exp\Big(\mathrm{sim}(z^{(k)}_{\mathrm{CWT}}, z^{(k)}_{\mathrm{temp}})/\tau\Big)}{\sum_{j=1}^{B} \exp\Big(\mathrm{sim}(z^{(k)}_{\mathrm{CWT}}, z^{(j)}_{\mathrm{temp}})/\tau\Big)}, \tag{5}$$

For the anchor $z^{(k)}_{\mathrm{CWT}}$, the denominator contains all temporal-view representations in the batch, including the matched representation $z^{(k)}_{\mathrm{temp}}$. The remaining temporal representations serve as negatives because each batch contains at most one window per patient group. The symmetric objective averages the two directions:

$$\mathcal{L}_{\mathrm{P2V}} = \frac{1}{2}\left[\frac{1}{B}\sum_{k=1}^{B}\ell_{v\to t}(k) + \frac{1}{B}\sum_{k=1}^{B}\ell_{t\to v}(k)\right]. \tag{6}$$

This form matches symmetric in-batch matching in contrastive representation learning (Chen et al., 2020); its per-window form pulls the two views of the same window together and pushes them apart from the cross-view representations of other patient groups in the batch. Pre-training never uses disease labels.

3.4 Downstream PPG-only Fusion

P2V initializes the downstream model with the two pre-trained encoders and fine-tunes it end to end. For one window, concatenating the two representations gives a 1,024-dimensional feature,

$$h = [h_{\mathrm{CWT}}; h_{\mathrm{temp}}] \in \mathbb{R}^{1024}, \qquad \hat{y} = \sigma(Wh + b), \quad W \in \mathbb{R}^{7\times 1024}, \tag{7}$$

and the output $\hat{y} \in [0,1]^7$ gives the predicted probability of each of the seven diseases. Training uses a binary cross-entropy with positive-class weights,

$$\mathcal{L}_{\mathrm{down}} = -\frac{1}{7}\sum_{c=1}^{7}\left[w_c\, y_c \log \hat{y}_c + (1-y_c)\log(1-\hat{y}_c)\right], \tag{8}$$

where the positive-class weight w_c is computed from training-set label frequencies and clipped to $[1, 20]$; it multiplies the positive term only and compensates for disease-label imbalance. This expression gives the loss for a single window, and training averages it over the windows in a batch.

Window-level predictions are aggregated at the patient level: all valid windows of a patient inherit that patient's labels, and their predicted probabilities are averaged arithmetically into the patient-level probability,

$$y_{ij} = y_i, \qquad \hat{p}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{p}_{ij}, \tag{9}$$

where n_i is the number of valid windows of patient i in the current split. Windows are therefore the unit of training and encoding, and patients the unit of evaluation. P2V takes PPG as its only input, and joint fine-tuning of the two encoders together with window-prediction aggregation yields patient-level probabilities for the seven diseases.

4 Experiments

4.1 Experimental Setup

Dataset and task. CWPC is a private wearable-PPG cohort linked to hospital clinical records, and we fix a patient-level seven-disease multi-label classification task with the label order hypertension, hyperlipidemia, carotid plaque, arrhythmia, diabetes, CHD, and stroke. The source label set also contains lower-extremity arteriosclerosis obliterans, which enters neither the target vector nor training, model selection, Test, or result statistics.

CWPC comprises 100,657 30-second PPG windows from 14,811 patients. The data team excluded all patients in the upstream pre-training split before constructing CWPC; its Train, Validation, and

Table 1: Data overview. Counting units differ and the counts are not additive: pre-training is counted in patient-group tokens, the grouping key that limits each batch to at most one window per group, and each CWPC split in patients.

Corpus / split	Count unit	Count	Windows	Role
Unlabeled pre-training	Patient-group tokens	26,357	99,701	Self-supervised pre-training
CWPC Train	Patients	10,452	71,676	Supervised training
CWPC Validation	Patients	2,225	13,829	Model selection
CWPC Test	Patients	2,134	15,152	Patient-level evaluation

Test splits are also patient-disjoint. Table 1 summarizes the pre-training corpus and the role of each split. All inputs are 100 Hz exported PPG with 30-second windows. A CWPC audit confirms an effective exported length of 2,990 points for all records; the series is extended to 3,000 points by 22nd-order autoregression with a decaying junction correction, while a per-point mask records measured versus extrapolated samples but is not provided to the model. The sequence is then normalized per window and clipped to $[-10, 10]$. The CWT view uses three consecutive 1,000-point segments, and the temporal view the full 3,000-point sequence. Supplementary Section 2 summarizes the length audit; Table S1 gives the data contract. During training each patient contributes at most $K = 8$ sampled windows; Validation and Test retain all valid windows and compute metrics at the patient level, with aggregation defined in Section 3.

The acquisition system also supports PPG, ACC, and single-lead ECG, but every stage of this work uses PPG only, and ACC and ECG are neither model inputs nor used for artifact screening or quality correction. Disease labels are patient-level, with positive counts and rates for the seven target diseases in supplementary Table S2. Upstream pre-training draws on the separately registered unlabeled corpus summarized in Table 1.

Models and training protocol. P0 initializes the CWT branch with a publicly image-pretrained ConvNeXtV2-Tiny and performs no cross-view pre-training. P2V performs PPG CWT–temporal cross-view pre-training followed by downstream fusion of both branches; CWT-only keeps the pre-trained vision encoder, CWT-CWT aligns same-type time–frequency views upstream, and the randomly initialized temporal branch loads no pre-trained temporal weights. Public baselines are MOMENT-small, NormWear, PaPaGei-S, Pulse-PPG, SIGMA-PPG, UniTS x128, and TF-C with its native PPG+FFT interface, with Logistic Regression, Random Forest, and XGBoost as traditional controls. All models use random seeds 42, 43, and 44.

Each deep model selects its best epoch by patient-level Validation CHD AUROC, with Macro AUROC as a predefined fallback metric. The selected checkpoint is frozen before Test evaluation. P2V trains downstream with AdamW and BCEWithLogitsLoss with positive-class weights computed from Train label statistics and clipped, and pre-trains with a symmetric InfoNCE objective at temperature 0.2; optimizer, batch-size, augmentation, and encoder-layer details are in the supplementary material.

Public baselines are verified cell by cell against a unified patience=3 early-stopping target via direct re-runs or documented equivalence checks. Public deep baselines use effective batch 128 and NormWear 32, while TF-C keeps its native PPG+FFT interface and SIGMA-PPG its native pipeline; per-seed epochs, patience, best epoch, and the batch/accumulation split are in supplementary Table S12.

Label budgets and evaluation metrics. The label-efficiency experiments keep the original patient splits and restrict labeled Train patients to 25% or 50% with three seeds per setting, testing cross-view pre-training transfer under constrained annotation budgets.

We also evaluate external transfer on the public PPG arrhythmia six-class dataset of Liu et al. (2022): patient-independent window-level classification over normal sinus rhythm (SR) and PVC, PAC, VT, SVT, and AF, under its own protocol.

Primary discriminative metrics are CHD AUROC, Macro AUROC, CHD AUPRC, and Macro AUPRC; the supplementary material adds per-disease AUROC/AUPRC, Macro F1, Balanced Accuracy, Brier, ECE, validation-fitted calibration (Guo et al., 2017), and ROC/PR protocols.

Table 2: CWPC patient-level Test main results (three-seed mean ± sample standard deviation). The best value in each column is bold; Logistic Regression is deterministic, so its standard deviation is 0.

Model	CHD AUROC	Macro AUROC	CHD AUPRC	Macro AUPRC
P2V (CWT + pre-trained temporal)	**0.7859 ± 0.0044**	**0.7150 ± 0.0012**	**0.1831 ± 0.0010**	**0.2351 ± 0.0020**
P0 (public initialization)	0.7536 ± 0.0051	0.7012 ± 0.0016	0.1598 ± 0.0068	0.2236 ± 0.0060
P2V (CWT-only)	0.7606 ± 0.0237	0.6935 ± 0.0217	0.1609 ± 0.0208	0.2149 ± 0.0203
CWT-CWT	0.7469 ± 0.0010	0.6893 ± 0.0022	0.1480 ± 0.0031	0.2068 ± 0.0009
MOMENT-small	0.7599 ± 0.0080	0.7067 ± 0.0038	0.1648 ± 0.0048	0.2262 ± 0.0062
SIGMA-PPG	0.7505 ± 0.0050	0.6915 ± 0.0044	0.1489 ± 0.0049	0.2195 ± 0.0057
NormWear	0.7268 ± 0.0045	0.6715 ± 0.0074	0.1379 ± 0.0007	0.1979 ± 0.0040
PaPaGei-S	0.7114 ± 0.0092	0.6698 ± 0.0030	0.1317 ± 0.0020	0.1986 ± 0.0009
Pulse-PPG	0.6054 ± 0.0075	0.5825 ± 0.0107	0.0875 ± 0.0045	0.1557 ± 0.0070
UniTS x128	0.7155 ± 0.0008	0.6725 ± 0.0043	0.1375 ± 0.0023	0.1999 ± 0.0030
TF-C (native PPG+FFT)	0.6881 ± 0.0168	0.5980 ± 0.0209	0.1167 ± 0.0057	0.1502 ± 0.0060
Logistic Regression	0.7210 ± 0.0000	0.6740 ± 0.0000	0.1262 ± 0.0000	0.2036 ± 0.0000
Random Forest	0.6891 ± 0.0031	0.6339 ± 0.0001	0.1460 ± 0.0036	0.1945 ± 0.0009
XGBoost	0.6858 ± 0.0045	0.6483 ± 0.0014	0.1456 ± 0.0077	0.2016 ± 0.0026

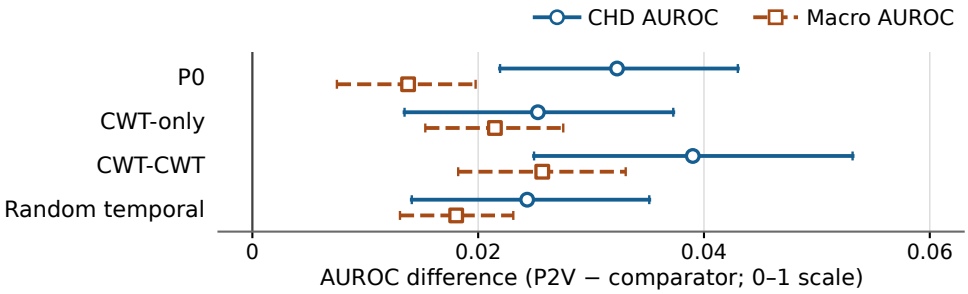

Figure 2: Patient-level paired Test differences on CWPC. Points are three-seed means of the original Test-set P2V-minus-comparator AUROC differences; horizontal bars show 95% patient-level paired-bootstrap confidence intervals, conditional on the three completed training seeds and quantifying patient-resampling uncertainty rather than between-seed variation.

4.2 Main Results

Table 2 summarizes CHD and macro-average metrics under the unified CWPC Test protocol. P2V reaches the highest values in the table on all four primary metrics (CHD AUROC 0.7859 ± 0.0044, Macro AUROC 0.7150 ± 0.0012, CHD AUPRC 0.1831 ± 0.0010, Macro AUPRC 0.2351 ± 0.0020), improving over P0 by 0.0323, 0.0138, 0.0233, and 0.0115 in that order, and over MOMENT-small, the strongest public three-seed baseline, by 0.0260 and 0.0083 in the two AUROCs.

CHD risk ranking and positive retrieval. Supplementary Figure S4 shows CHD patient-level probability ranking and retrieval, and Figure S2 shows the nine deep models included in the main baseline comparison without between-seed shading, with numerical values in Table 2; under class imbalance, P2V's advantage appears in both.

4.3 Structural Ablation and Paired Confidence Intervals

Table 3 gives the structural and downstream-input ablation, and supplementary Figure S5 the complete view. P2V improves Test CHD AUROC by 0.0253 over CWT-only, by 0.0390 over CWT-CWT, and by 0.0243 over the randomly initialized temporal branch; within this protocol the three controls support the design value of introducing and fusing the temporal branch, of heterogeneous view pairing, and of the pre-trained temporal initialization. A frozen-readout condition that fixes the pre-trained CWT encoder and trains a linear layer on its 768-dimensional representation alone reaches a Macro AUROC of 0.6845 ± 0.0152, whereas the full two-branch end-to-end model reaches 0.7150 ± 0.0012; the two conditions differ in branch set, fusion, and update rule together, and all rows are in supplementary Table S7.

Table 3: CWPC structural and downstream-input ablation (patient-level Test, three-seed mean ± sample standard deviation). The randomly initialized temporal branch row changes temporal initialization only, and the frozen-readout conditions are in supplementary Table S7. The best value in each column is bold.

Model or condition	CHD AUROC	Macro AUROC	CHD AUPRC	Macro AUPRC
P2V (CWT + pre-trained temporal)	$\mathbf{0.7859 \pm 0.0044}$	$\mathbf{0.7150 \pm 0.0012}$	$\mathbf{0.1831 \pm 0.0010}$	$\mathbf{0.2351 \pm 0.0020}$
P2V (CWT-only)	0.7606 ± 0.0237	0.6935 ± 0.0217	0.1609 ± 0.0208	0.2149 ± 0.0203
CWT-CWT	0.7469 ± 0.0010	0.6893 ± 0.0022	0.1480 ± 0.0031	0.2068 ± 0.0009
P2V + random temporal	0.7616 ± 0.0199	0.6969 ± 0.0148	0.1649 ± 0.0168	0.2196 ± 0.0138

Paired patient bootstrap further quantifies the differences among our own models. Figure 2 reports the CHD AUROC and Macro AUROC differences of P2V against P0, CWT-only, CWT-CWT, and the randomly initialized temporal branch with their 95% patient-level paired-bootstrap intervals, and all eight lie above zero; the resampling procedure and the Holm-corrected empirical p-values are in the supplementary material.

4.4 Label-Efficient Learning

Figure 3 compares P0, P2V CWT-only, CWT-CWT, and full P2V, with complete four-metric values and per-seed variation in supplementary Table S8. At 25% and 50% labeled patients, full P2V improves CHD AUROC/Macro AUROC over P0 by 0.0225/0.0137 and 0.0267/0.0138, and over CWT-CWT by 0.0352/0.0274 and 0.0271/0.0156.

On CHD AUROC, full P2V and CWT-only are numerically close, with mean differences of -0.0003 and -0.0033 at the 25% and 50% budgets, both smaller than the corresponding between-seed variation, while full P2V reaches the highest CHD AUPRC and Macro AUPRC among the four configurations at both budgets. This comparison is a three-seed descriptive result and does not represent a paired significance test.

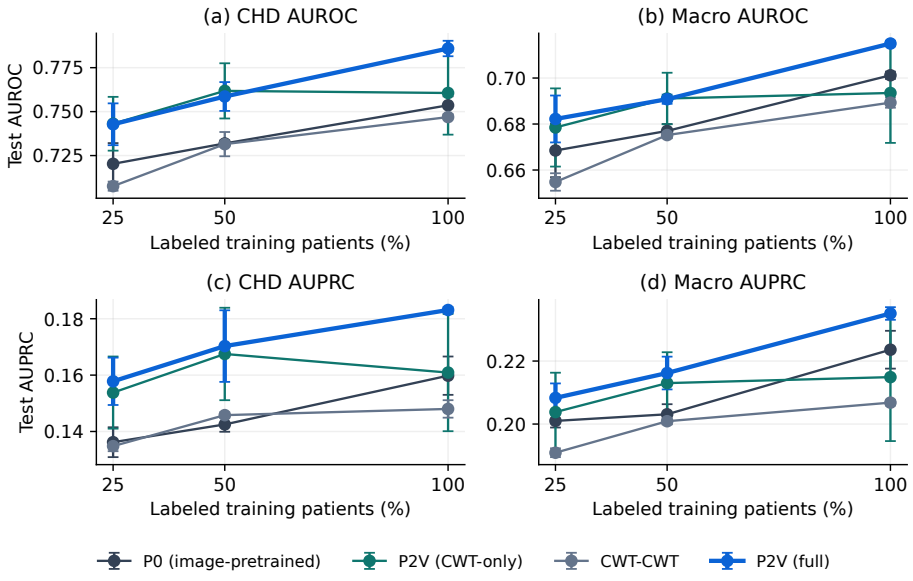

Figure 3: Label-efficient transfer. Points are patient-level Test three-seed means and error bars are sample standard deviations.

4.5 Transfer to Public PPG Arrhythmia Classification

On the PPG arrhythmia six-class dataset of Liu et al. (2022), P2V reaches an end-to-end three-seed Test Macro AUROC of 0.9697 ± 0.0011, Macro AUPRC of 0.8088 ± 0.0025, and Macro F1 of 0.7644 ± 0.0011; supplementary Figure S6 and Tables S9–S11 give the per-class values. This task

uses a different input duration and window-level evaluation, providing transfer evidence independent of the CWPC disease task.

5 Discussion and Conclusion

P2V's experimental chain supports a design attribution within the protocol. Results relative to CWT-only, CWT-CWT, and the randomly initialized temporal branch support, respectively, introducing and fusing the temporal branch, using heterogeneous view pairing, and the pre-trained temporal initialization. Results relative to P0 support the overall benefit of the P2V pre-training-and-transfer pipeline over public image initialization. The image-domain CWT time–frequency view exposes local multi-scale pulse morphology and frequency content to a vision backbone, while the full-waveform temporal view exposes continuous temporal relations and rhythm to a sequence encoder. Both views derive from the same exported PPG recording yet introduce different inductive biases through different representation domains and dedicated encoders. Cross-view alignment encourages agreement between image-domain and time-domain representations of the same PPG window, beyond the within-view invariance encouraged by augmentation. At 25% and 50% label budgets, full P2V and CWT-only achieve similar CHD AUROC, while full P2V attains the highest CHD and Macro AUPRC among the four configurations, indicating a metric-dependent benefit under limited supervision.

P2V targets multi-disease screening and prioritization from consumer-grade PPG, expanding the learnable representation space under a PPG-only constraint; its output is a disease risk ranking and does not constitute a clinical diagnosis. CWPC provides patient-level seven-disease multi-label evidence and the Liu task cross-dataset rhythm-transfer evidence, covering complementary disease and rhythm scenarios; our conclusions rest on the evaluated cohorts and task protocols. CWPC was constructed after excluding upstream pre-training patients, and its Train, Validation, and Test splits are patient-disjoint. Prospective evaluation across devices and populations will further test the applicability of this principle.

We presented P2V and validated a principle for wearable learning: a single-sensor deployment constraint does not require a single representation. P2V attains the highest three-seed means among evaluated models on all four primary patient-level metrics of the CWPC task, with a CHD AUROC of 0.7859 ± 0.0044, and our ablations support the roles of heterogeneous view pairing, temporal-branch fusion, and pre-trained temporal initialization. On the public PPG arrhythmia six-class task, it transfers under a window-level protocol.

Reproducibility Statement

The private CWPC cohort is not publicly released. Under patient-privacy and data-governance requirements, patient-level raw signals, direct identifiers, and individual clinical records are not disclosed, and this work reports aggregate statistics only. Input preprocessing, training sampling, and patient-level aggregation are given in Section 3 and Section 4.1; random seeds, model selection, encoder layers, augmentation, training hyperparameters, and curve coordinates are given in Section 4.1 and the supplementary material. The pre-training and downstream training code of P2V will be released upon paper acceptance.

Ethics Statement

This study uses hospital wearable-device data collected under institutional ethics approval and informed consent, and involves no new data collection or intervention. Although the acquisition system also supports PPG, ACC, and single-lead ECG, every step of this work uses PPG only. The model outputs independent probabilities for seven diseases and is intended for risk ranking and screening research; it does not constitute a clinical diagnosis and should not be used as the sole basis for medical decisions. Patient-level data are not publicly released.

Generative AI Usage Statement

This study used OpenAI ChatGPT and Codex to assist with literature search and organization, feedback on experimental design, drafting of code and data-processing scripts, figure preparation, interpretation of results, and the writing, translation, and polishing of the manuscript. The authors ran or tested all AI-assisted code, and every experimental number is traceable to saved results and experiment records. Generative AI was not used to generate or modify PPG measurements, disease labels, or experimental results. The authors reviewed all AI-assisted content and take responsibility for the final content, claims, and artifacts of the paper.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Zhengyao Ding, Yujian Hu, Ziyu Li, Yiheng Mao, Haitao Li, et al. 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://www.nature.com/articles/s41746-025-02228-3.

Vijay Ekambaram, Subodh Kumar, Arindam Jati, Sumanta Mukherjee, Tomoya Sakai, Pankaj Dayama, Wesley M. Gifford, and Jayant Kalagnanam. TSPulse: Tiny pre-trained models with disentangled representations for rapid time-series analysis. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6572721b15834cb37599abec23b5784f-Abstract-Conference.html.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. UniTS: A unified multi-task time series model. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fe248e22b241ae5a9adf11493c8c12bc-Paper-Conference.pdf.

Ramin Ghorbani, Marcel J. T. Reinders, and David M. J. Tax. Self-supervised PPG representation learning shows high inter-subject variability. In *Proceedings of the 8th International Conference on Machine Learning Technologies*, pp. 127–132, 2023. doi: 10.1145/3589883.3589902. URL https://doi.org/10.1145/3589883.3589902.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.03885.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

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. *arXiv preprint arXiv:2601.21031*, 2026. URL https://arxiv.org/abs/2601.21031.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Learning to embed time series patches independently. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/74b8ccb5ad21871fd436867df19c35ac-Abstract-Conference.html.

Chenqi Li, Yu Liu, Timothy Denison, and Tingting Zhu. BioX-Bridge: Model bridging for unsupervised cross-modal knowledge transfer across biosignals. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.02276.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1430. URL https://papers.nips.cc/paper_files/paper/2023/hash/6818dcc65fdf3cbd4b05770fb957803e-Abstract-Conference.html.

Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. FOCAL: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/93e98ddf39a9beb0a97fbbe56a986c80-Abstract-Conference.html.

Zengding Liu, Bin Zhou, Zhiming Jiang, Xi Chen, Ye Li, Min Tang, and Fen Miao. Multiclass arrhythmia detection and classification from photoplethysmography signals using a deep convolutional neural network. *Journal of the American Heart Association*, 11(7):e023555, 2022. doi: 10.1161/JAHA.121.023555. URL https://doi.org/10.1161/JAHA.121.023555.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals. *ACM Transactions on Computing for Healthcare*, 7(3):43:1–43:43, 2026. doi: 10.1145/3803808. URL https://doi.org/10.1145/3803808.

Shuo Ma, Yingwei Zhang, Yiqiang Chen, Hualei Wang, Yuan Jin, Wei Zhang, and Ziyu Jia. SleepSMC: Ubiquitous sleep staging via supervised multimodal coordination. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=B5VEi5d3p2.

Huy Phan, Oliver Y. Chén, Minh C. Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2022. doi: 10.1109/TPAMI.2021.3070057. URL https://arxiv.org/abs/2007.05492.

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.

Norman Ricker. Wavelet functions and their polynomials. *Geophysics*, 9(3):314–323, 1944. doi: 10.1190/1.1445082.

Mithun Saha, Maxwell A. Xu, Wanting Mao, Sameer Neupane, James M. Rehg, and Santosh Kumar. Pulse-PPG: An open-source field-trained PPG foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):126:1–126:35, 2025. doi: 10.1145/3749494. URL https://arxiv.org/abs/2502.01108.

Christopher Torrence and Gilbert P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998. doi: 10.1175/1520-0477(1998)079⟨0061:APGTWA⟩2.0.CO;2.

Riling Wei, Hanjie Chen, Kelu Yao, Chuanguang Yang, Jun Wang, and Chao Li. ECG-guided individual identification via PPG. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025. doi: 10.1109/ICASSP49660.2025.10890435. URL https://arxiv.org/abs/2501.01983.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.01575.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Woo_ConvNeXt_V2_Co-Designing_and_Scaling_ConvNets_With_Masked_Autoencoders_CVPR_2023_paper.html.

Weixuan Yuan, Zengrui Jin, Yichen Wang, Donglin Xie, Ziyi Ye, Chao Zhang, and Xuesong Chen. sleep2vec: Unified cross-modal alignment for heterogeneous nocturnal biosignals. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.13857.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.08496.