PhysioDyT: Physiologically Directed Dynamic Transport for ECG–PPG Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

ECG precedes the peripheral pulse measured by PPG, with a delay that varies across beats and recordings. We present PhysioDyT, a joint-embedding model that predicts future PPG tokens from ECG through a positive-delay distribution, monotone soft transport, and an unmatched-token state. An auxiliary PPG-to-ECG objective supports representation learning, while deployment uses PPG alone. An archived three-pre-training-seed summary on waveform-de-duplicated Clinical Wearable PPG Cohort (CWPC) data reports patient-level coronary-heart-disease (CHD) validation AUROC of 0.7572 ± 0.0050. In a separate clean frozen-feature evaluation with a shared PPG-only MIL head, PhysioDyT performs comparably to leading public encoders. Reversing PPG, permitting negative delays, and disrupting token order increase alignment loss across seeds; injected teacher-token shifts elicit a graded response (slope 0.654; Spearman $\rho = 0.878$). Together, the transfer and intervention results characterize directed temporal correspondence as a testable representation-learning bias.

1 Introduction

Photoplethysmography is inexpensive, wearable, and suitable for repeated monitoring, but it observes cardiovascular function indirectly through optical changes in peripheral blood volume. Its morphology is affected by vascular state, sensor placement, motion, contact pressure, and device filtering (Charlton et al., 2023). ECG provides a complementary view: ventricular electrical activation precedes mechanical ejection and the arrival of the pulse at a peripheral sensor. Synchronized ECG can therefore serve as privileged information during pre-training even when deployment must use PPG alone. We use a separately catalogued unlabeled PPG corpus for pre-training and a labeled Clinical Wearable PPG Cohort (CWPC) for downstream evaluation; the two data sources are not interchangeable.

The modalities have a directed, variable-delay relation whose apparent correspondence also depends on acquisition pipelines, filtering, tokenization, missed beats, motion, and window boundaries. A useful representation learner should encode direction and ordering while allowing uncertain or unmatched tokens. Its internal alignment can then be examined through controlled timing interventions.

General time-series foundation models such as MOMENT (Goswami et al., 2024) and UniTS (Gao et al., 2024) learn transferable representations across heterogeneous domains. PPG-focused models including PaPaGei (Pillai et al., 2025) and Pulse-PPG (Saha et al., 2025), together with wearable models such as NormWear (Luo et al., 2024), exploit large optical or multimodal corpora. These approaches provide strong transfer baselines but do not explicitly constrain ECG-to-PPG temporal geometry. Likewise, ordered cross-modal reconstruction can exploit physiological sequence structure, so the observation that ECG precedes PPG is not itself a sufficient contribution. The open question is whether cross-modal correspondence can be represented by a probabilistic monotone operator whose direction, pairing, and shift response can be falsified.

PhysioDyT addresses this question with a JEPA-style latent prediction objective (Assran et al., 2023). For each ECG token, the model predicts a distribution over future PPG offsets

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

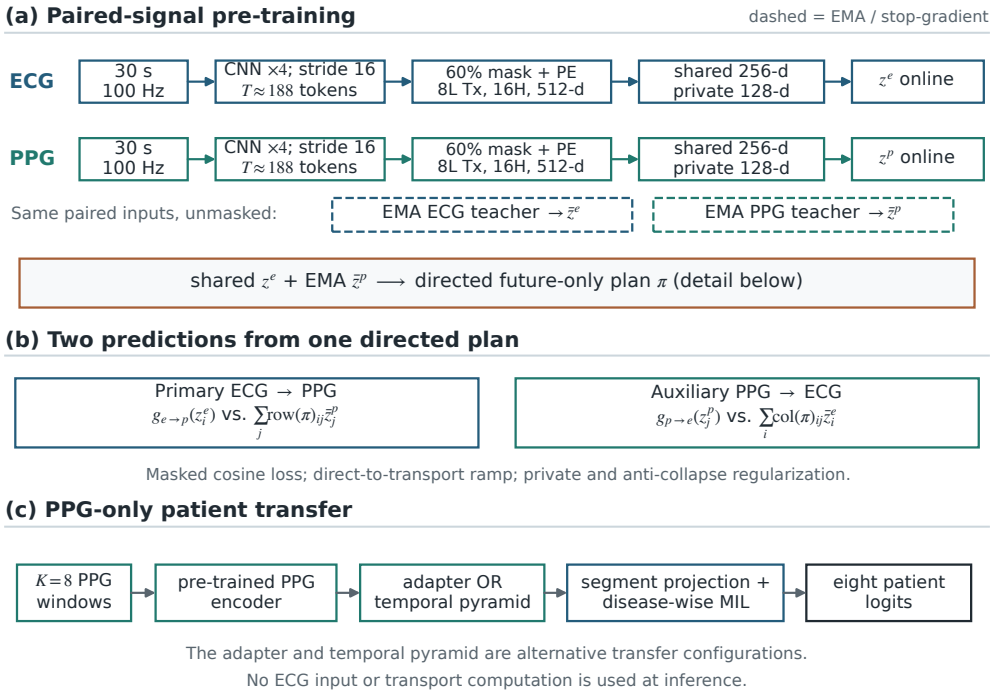


Figure 1: PhysioDyT computation and transfer. Paired 30-s ECG and PPG windows are drawn from the separately catalogued unlabeled PPG corpus. They pass through modality-specific CNN–Transformer encoders: online tokens are block-masked after the CNN stem, while EMA teachers see unmasked inputs. Only shared online ECG and EMA PPG tokens construct the future-only plan π. Forward and auxiliary reverse prediction use row- and column-normalized views of this same plan. At CWPC transfer, a configuration-selected adapter or temporal-pyramid path precedes PPG-only disease-conditioned patient MIL. PE denotes positional encoding; Tx denotes Transformer.

and an unmatched state. Expected target positions are regularized for monotonicity and local smoothness, while the dustbin prevents low-quality or boundary tokens from being forced into a match. An auxiliary reverse predictor regularizes representation learning, but the physiological transport itself remains ECG-to-PPG. At transfer time, the ECG branch and transport head are discarded. Figure 1 shows the pre-training and PPG-only transfer pathways; Figure 2 expands the transport operator.

Our contributions are:

- a physiology-directed latent transport objective with positive-delay support, probabilistic matching, soft monotonicity, local smoothness, and unmatched-token handling;
- a paired-pre-training, PPG-only-deployment design evaluated on CWPC after deduplication against the upstream waveforms; and
- an intervention-based audit of direction, pairing, temporal order, and shift response across independent pre-training seeds.

2 Related Work

Time-series and physiological foundation models. MOMENT uses masked time-series modeling to support broad transfer (Goswami et al., 2024), while UniTS unifies forecasting, classification, and related tasks (Gao et al., 2024). PaPaGei and Pulse-PPG focus on optical

physiological signals (Pillai et al., 2025; Saha et al., 2025); NormWear targets multivariate wearable sensing (Luo et al., 2024). These models motivate a controlled representation-quality comparison in which the input channel, patient split, encoder status, aggregation head, and model-selection rule are held fixed.

Joint-embedding and cross-modal prediction. Joint-embedding predictive architectures estimate target representations instead of reconstructing every input sample, reducing pressure to model nuisance details (Assran et al., 2023). Cross-modal masked reconstruction can also capture ordered physiological relations. Latent prediction alone leaves the correspondence between ECG and PPG tokens implicit. PhysioDyT makes this correspondence explicit through soft weights over future PPG tokens and an unmatched state. The resulting plan can be constrained for temporal order and probed by reversing, mismatching, or shifting PPG.

Differentiable temporal alignment. Soft-DTW (Cuturi & Blondel, 2017) and entropically regularized optimal transport (Peyre & Cuturi, 2019) provide differentiable correspondences between sequences. Unconstrained alignment may connect implausible temporal locations or force corrupted tokens to match. PhysioDyT restricts support to future PPG positions, regularizes order, and includes a dustbin to accommodate uncertain matches.

3 Method

3.1 Problem formulation

Let x^e and x^p denote synchronized ECG and PPG windows. Modality-specific CNN stems (total stride 16) and eight-layer Transformers produce token sequences

$$z^e_{1:T_e} = f^e_\theta(x^e), \qquad z^p_{1:T_p} = f^p_\theta(x^p), \tag{1}$$

while exponential-moving-average (EMA) target encoders produce stop-gradient targets $\bar{z}^e$ and $\bar{z}^p$. The online Transformer receives a 60% block-masked token sequence after CNN tokenization; EMA encoders see unmasked inputs. A shared/private projection separates the 256-dimensional tokens used for cross-modal prediction from private features regularized by reconstruction and orthogonality.

3.2 Positive-delay probabilistic transport

For ECG token i and future offset d, the transport head combines a 128-dimensional projected content similarity $c_{i,i+d}$, global delay logits G_d from pooled ECG/PPG features, and local ECG-conditioned delay logits $L_{i,d}$:

$$s_{i,d} = \big(w_c c_{i,i+d} + w_g G_d + w_l L_{i,d}\big)/\tau, \quad d \in \mathcal{D}^+. \tag{2}$$

The local head also predicts an unmatched logit. The five supported offsets correspond to approximately $\{160, 320, 480, 640, 800\}$ ms; all other ECG–PPG pairs are masked. A dustbin row and column augment the score matrix, and 20 unbalanced entropic Sinkhorn steps produce a soft token-alignment plan π.

Let T^f and T^r be the row- and column-normalized matched portion of the same π. The EMA targets transported to the two online predictors are

$$\tilde{z}^p_i = \sum_j T^f_{ij} \bar{z}^p_j, \qquad \tilde{z}^e_j = \sum_i T^r_{ij} \bar{z}^e_i. \tag{3}$$

For masked tokens with nonzero matched mass, the forward loss is

$$\mathcal{L}_{e\to p} = \frac{1}{|\Omega_e|} \sum_{i\in\Omega_e} \left[1 - \cos(g_{e\to p}(z^e_i), \tilde{z}^p_i)\right]. \tag{4}$$

The teacher tokens are stop-gradient, but the weighted targets retain gradients through the transport plan. The reverse term analogously compares $g_{p\to e}(z^p_j)$ with $\tilde{z}^e_j$; no independently learned reverse transport head is introduced.

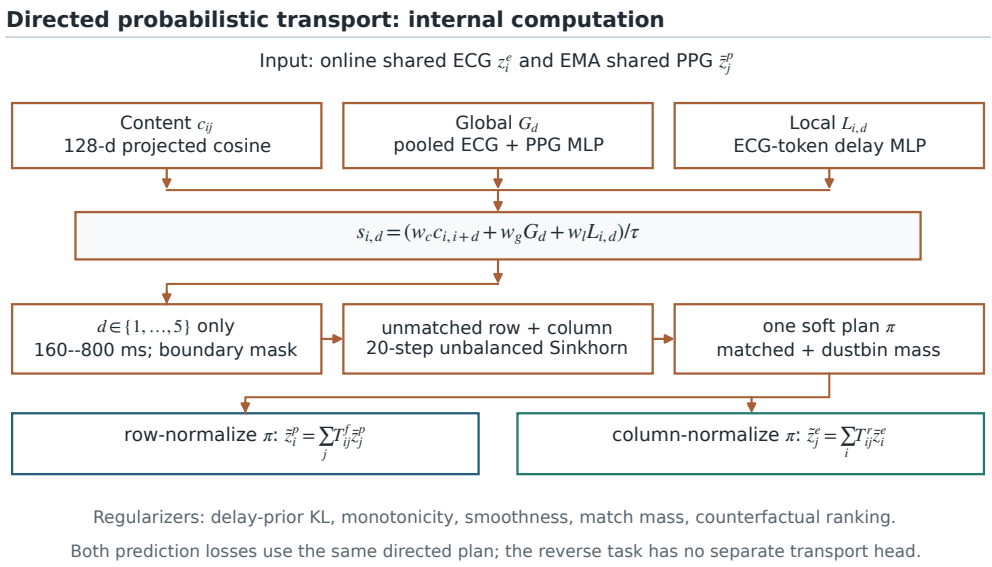


Figure 2: Transport head, expanded. A 128-dimensional projected content score, record-level global delay, and ECG-token local delay define scores at positive offsets. Unmatched row/column states and unbalanced Sinkhorn produce one soft plan. Its matched mass is row-normalized for EMA PPG aggregation and column-normalized for EMA ECG aggregation.

3.3 Monotonicity, smoothness, and auxiliary prediction

Let δ_i be the matched conditional expected offset and $\mu_i = i + \delta_i$. We penalize reversals with the mean $[\mu_i - \mu_{i+1}]_+$ over valid adjacent rows, and abrupt offset changes with a smooth-ℓ_1 loss on normalized δ_i. Other transport terms are a positive-delay prior KL, a matched-mass target, and a counterfactual pair-ranking margin. Variance/covariance and private reconstruction/orthogonality terms regularize the representations. Training ramps from direct masked prediction to transported masked prediction:

$$\mathcal{L}_{\text{token}} = (1 - r)\mathcal{L}_{\text{direct}} + r\beta \frac{\mathcal{L}_{e\to p} + \alpha \mathcal{L}_{p\to e}}{1 + \alpha}, \tag{5}$$

with scheduled regularization added to $\mathcal{L}_{\text{token}}$. The reverse task supports representation learning without changing the ECG-to-PPG direction of the transport plan. An independent three-pre-training-seed study found no reliable advantage for asymmetric weighting, so the canonical model uses $\alpha = 1$.

3.4 PPG-only patient transfer

Each patient contributes multiple short PPG windows. The pre-trained PPG encoder yields token or window features. In the current implementation, a shared/private feature adapter and temporal-pyramid pooling are alternative, configuration-selected transfer paths, not sequential modules. A segment projection and disease-conditioned attention MIL module (Ilse et al., 2018) aggregate $K = 8$ sampled windows into patient-level logits. Paired ECG is required only for pre-training; neither ECG nor the transport head is used at inference.

4 Experimental Design

4.1 Data sources, leakage audit, and protocols

The upstream source is a separately catalogued, unlabeled PPG corpus. Its five-channel recordings also contain synchronized ECG; we extract ECG (channel 0) and PPG (channel

4) for self-supervised cross-modal pre-training, without using disease labels. The downstream labeled resource is CWPC, which supplies patient-level disease outcomes for PPG-only transfer. Thus, "PPG corpus" names the upstream data source and does not mean that pre-training observes PPG alone.

An audit identified patient or waveform overlap risk between the upstream corpus and the original CWPC data. We retained historical pre-training checkpoints but removed waveform-matched patients from the clean CWPC cohort. Candidate matches were identified by record time and verified by pointwise PPG equality before patient-level removal. The primary clean CWPC split (V1-A) contains 10,452/2,225/2,134 train/validation/test patients. The test partition remains sealed.

We report three related evaluations with distinct provenance. Clean CWPC transfer uses V1-A and an archived summary over three independent pre-training seeds. Historical public-model comparison uses an earlier development cohort, frozen PPG encoders, one patient-level MIL head, and three downstream seeds; some run-level artifacts remain pending. Clean frozen-feature replication repeats the shared-head comparison on V1-A using one PhysioDyT pre-training checkpoint and aligned patient-level predictions. We report these protocols separately because their AUROCs answer different questions.

4.2 Evaluation and baselines

The primary endpoint is patient-level CHD AUROC. CHD AUPRC and macro AUROC are secondary. Public encoders include MOMENT-small, UniTS-x128, NormWear, Pulse-PPG, and PaPaGei-S. For the controlled comparison, every encoder is frozen and only the same lightweight MIL head is trained. We report mean and standard deviation over three downstream seeds.

Mechanism experiments intervene on the frozen alignment pathway. Controls replace dynamic delay with a segment-static delay or fixed prior, shuffle delay tokens, permit negative delays, reverse PPG order, or mismatch ECG and PPG patients. Shift recovery injects known PPG shifts and measures the compensation selected by the loss profile. Patient-level bootstrap intervals are used for mechanism summaries.

5 Results

5.1 De-duplicated CWPC transfer

An archived summary of PPG-only transfer on de-duplicated CWPC V1-A reports validation CHD AUROC 0.7572 ± 0.0050 across three independent PhysioDyT pre-training seeds. Its run-level patient predictions still require source-level verification. The 2,134-patient test partition remains sealed.

5.2 Controlled comparison with public encoders

Table 1: Historical development-cohort frozen PPG comparison. All encoders use PPG-only input, the same patient split, and the same MIL head. Values are mean ± SD over three downstream seeds. The later clean CWPC replication is reported separately in the text and Appendix B.

Model	CHD AUROC	CHD AUPRC	Macro AUROC
PhysioDyT	0.7695±0.0012	0.3370±0.0048	0.7290±0.0024
MOMENT-small	0.7217±0.0005	0.2988±0.0003	0.7022±0.0009
UniTS-x128	0.7207±0.0011	0.2797±0.0027	0.6981±0.0036
NormWear	0.7185±0.0018	0.2826±0.0010	0.7123±0.0002
Pulse-PPG	0.7028±0.0022	0.2785±0.0061	0.6959±0.0059
PaPaGei-S	0.7018±0.0039	0.2594±0.0034	0.6801±0.0092

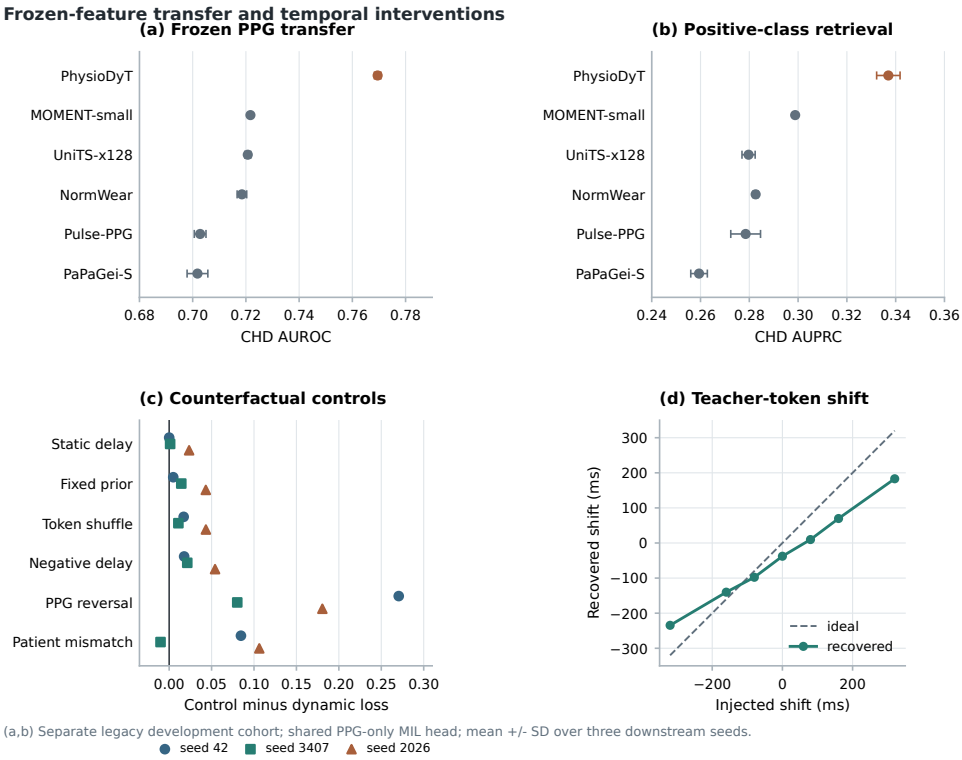


Figure 3: Transfer and mechanism results. (a–b) Public-encoder comparison under the separate frozen-PPG development protocol, not the clean CWPC V1-A protocol; points and error bars show mean and SD over three downstream seeds. (c) Control-minus-dynamic alignment loss for each independent pre-training seed; positive values mean higher loss under intervention. (d) Recovered compensation after injecting shifts into teacher PPG tokens. The dashed line is ideal recovery.

Table 1 and Figures 3a–b hold the downstream head fixed on the historical development cohort. PhysioDyT exceeds MOMENT-small there by 0.0478 CHD AUROC and 0.0382 CHD AUPRC. A three-seed probability ensemble yields an exploratory patient-level AUROC difference of 0.0487 (95% bootstrap CI $[0.0070, 0.0919]$; unadjusted $p = 0.0176$). In the later clean CWPC V1-A replication, the PhysioDyT frozen-feature ensemble reaches 0.7187 CHD AUROC, compared with 0.7193 for MOMENT-small and 0.7200 for NormWear; paired differences versus these two encoders have confidence intervals spanning zero (Appendix B). The historical advantage therefore does not replicate as a clear frozen-feature lead on the clean cohort. These comparisons evaluate transferable PPG representations; the within-model interventions below probe the proposed alignment mechanism.

5.3 Counterfactual mechanism tests

We test whether the learned alignment uses the temporal structure imposed during training. PPG reversal and negative delays challenge direction; token shuffling challenges order; fixed priors and segment-static delays challenge adaptation; cross-patient mismatch challenges pairing. Figure 3c shows that reversal, negative delay, fixed priors, and shuffling increase alignment loss for all three pre-training seeds. Segment-static replacement is also adverse, though its magnitude varies substantially across checkpoints.

Cross-patient mismatch increases loss for seeds 42 and 2026 but decreases it for seed 3407. Pairing sensitivity is thus less stable than direction and order sensitivity across these checkpoints.

5.4 Shift recovery and physiological interpretation

The loss controls above establish sensitivity, while the shift experiment tests whether the response changes with the size of a known temporal perturbation. Injecting offsets into teacher PPG tokens produces graded loss-minimizing compensation with slope 0.654 (95% CI $[0.629, 0.679]$), Spearman $\rho = 0.878$, and MAE 118 ms (Figure 3d). The corresponding waveform-domain response is weaker (slope 0.076, Spearman 0.315, MAE 206 ms), localizing the stronger effect to the learned token space.

Agreement between the learned delay and waveform-derived pulse arrival time is weak (Spearman approximately 0.022; MAE 256 ms). This comparison helps distinguish latent correspondence from physical timing.

6 Discussion

Transfer and mechanism evidence. The archived clean CWPC V1-A summary suggests reproducible PPG-only CHD discrimination after waveform-level de-duplication, pending its run-level audit. The historical frozen-feature comparison favored PhysioDyT, whereas the clean replication places it near MOMENT-small and NormWear. The intervention suite identifies more consistent sensitivity to temporal direction and order than to patient-specific pairing. Together, these studies motivate the transport hypothesis while locating its most reproducible evidence in the alignment behavior.

Interpretation and scope. The transport plan is a learned token correspondence shaped by physiological timing, tokenization, noise, and optimization. The weaker waveform response locates the clearest timing effect in representation space. Loss-level interventions establish that the model responds to timing changes; controlled downstream ablations are still needed to attribute CHD discrimination to individual transport constraints. External 25-Hz prospective evaluation also faces substantial device and cohort shift.

Why transport may help. Forcing exact synchronous correspondence would make ECG and PPG encoders explain acquisition offsets and beat-level timing variability as if they were content. A positive-delay distribution instead permits multiple plausible future targets; monotonicity preserves sequence order; and the dustbin avoids contaminating the objective with low-quality or boundary tokens. These mechanisms provide a coherent explanation for the observed intervention sensitivity, although clean downstream component ablations are still required to attribute predictive gains to individual terms.

Confirmatory evaluation. The clean public-encoder replication now provides a shared-head reference. The next confirmatory step is to verify the independent-seed transfer records, evaluate transport-off, fixed-delay, no-monotonicity, and no-dustbin variants on the same clean manifest, then lock model selection before one-time sealed testing. Patient-level AUROC, AUPRC, paired bootstrap intervals, ROC/precision-recall curves, and calibration will be reported.

7 Conclusion

PhysioDyT models ECG-to-PPG correspondence with a positive-delay, probabilistic, monotone transport plan. An archived clean-cohort transfer summary reports CHD discrimination across independent pre-training seeds; a clean shared-head comparison finds frozen PPG features comparable to leading public encoders. Controlled interventions reveal a graded, direction- and order-sensitive latent response, while pairing effects vary by seed. The contribution is a structured and testable temporal inductive bias for asynchronously coupled physiological signals.

Ethics Statement

The study uses de-identified physiological recordings and patient-level disease labels under the applicable data-governance procedures. The model is intended for representation-learning research and retrospective risk stratification, not autonomous diagnosis. We report data-overlap risk, domain shift, and calibration limitations because unqualified deployment could disproportionately harm underrepresented device, age, or disease groups. No patient data will be released through the supplementary material.

Reproducibility Statement

Sections 4–5 and the appendix document data provenance, splits, endpoints, and the planned sealed-test procedure. Supplementary materials will provide anonymized configurations, split and checkpoint hashes, intervention and bootstrap scripts, and permissible patient-level predictions. We distinguish variation across pre-training seeds from variation across downstream-head seeds.

AI Use Statement

Generative AI assisted with hypothesis and experiment refinement, method implementation and data processing, result interpretation, figure preparation, and manuscript editing and translation. The authors checked AI-assisted content against experimental records and cited literature, made the research and reporting decisions, and take responsibility for the text, code, and results.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, et al. Self-supervised learning from images with a joint-embedding predictive architecture. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

Peter H. Charlton et al. The 2023 wearable photoplethysmography roadmap. Physiological Measurement, 2023.

Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. In International Conference on Machine Learning, 2017.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. UniTS: A unified multi-task time series model. arXiv preprint arXiv:2403.00131, 2024.

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.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In International Conference on Machine Learning, 2018.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation models for multivariate wearable sensing of physiological signals. arXiv preprint arXiv:2412.09758, 2024.

Gabriel Peyre and Marco Cuturi. Computational Optimal Transport. Now Publishers, 2019.

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

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. arXiv preprint arXiv:2502.01108, 2025.

A Evidence ledger and evaluation boundary

- Clean CWPC transfer: an archived three-pre-training-seed summary reports CHD validation AUROC 0.7572 ± 0.0050; run-level verification is pending.
- Frozen-feature comparison: the historical development cohort favors PhysioDyT, while the clean V1-A replication places it near MOMENT-small and NormWear.
- Mechanism: direction and order interventions increase alignment loss across three seeds; cross-patient pairing responses vary by seed.
- Generalization: clean component ablations and the sealed test remain to be completed.

B Clean CWPC frozen-feature replication

All six encoders use PPG-only input, the V1-A validation cohort, frozen encoders, and the same patient-level MIL head. The PhysioDyT encoder is one pre-training seed (42); variation in the table is over three downstream head seeds. The ensemble ROC uses aligned predictions from 2,225 patients, including 134 CHD-positive patients.

Table 2: Clean CWPC V1-A CHD results. Mean ± SD varies downstream head seeds; ensemble AUROC intervals use patient bootstrap resampling.

Model	CHD AUROC	Ensemble AUROC [95% CI]	Ensemble AUPRC
PhysioDyT	0.7156±0.0020	0.7187 [0.6739, 0.7593]	0.1325
MOMENT-small	0.7189±0.0012	0.7193 [0.6739, 0.7606]	0.1476
NormWear	0.7202±0.0002	0.7200 [0.6766, 0.7606]	0.1362
UniTS-x128	0.7088±0.0021	0.7090 [0.6652, 0.7499]	0.1274
PaPaGei-S	0.6955±0.0039	0.6978 [0.6510, 0.7406]	0.1240
Pulse-PPG	0.5146±0.0225	0.5169 [0.4660, 0.5659]	0.0690

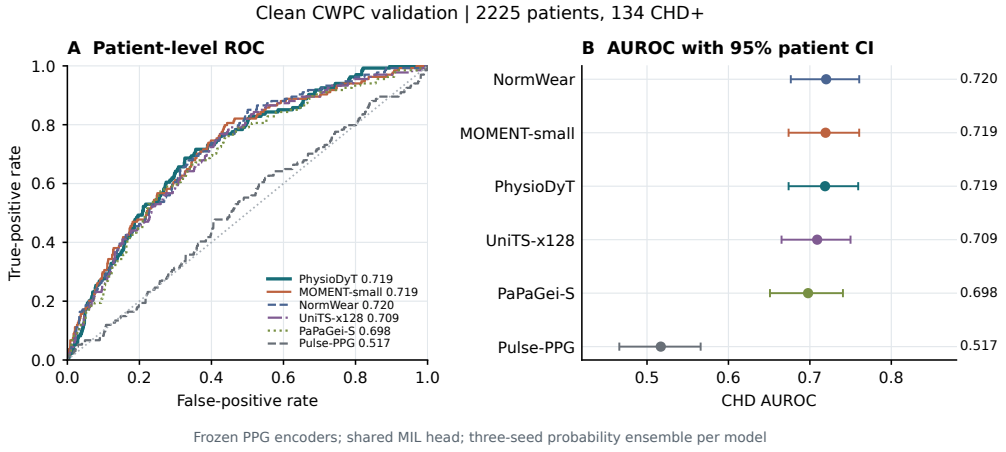


Figure 4: Patient-level CHD ROC curves for the clean V1-A validation cohort. Curves and AUROCs use aligned ensemble probabilities from three downstream head seeds. Pulse-PPG remains an adapter diagnostic pending protocol verification.

C Prespecified sealed-test analysis

The final CWPC evaluation will use one immutable canonical manifest. CHD AUROC is the primary endpoint; CHD AUPRC and macro AUROC are secondary. Model selection uses validation only. Thresholds and calibration are fixed before test access. Patient-level paired bootstrap resampling will compare PhysioDyT with the strongest clean public baseline and

transport-off model. Exact two-sided paired p values and 95% confidence intervals will be reported, with multiplicity correction when more than one public baseline is treated as confirmatory.