Synergy-Aware Contrastive Pretraining for Co-recorded Physiological Signals

Lei Chen [1] **Junetae Kim** [1]

Abstract

Co-recorded electrocardiography (ECG) and photoplethysmography (PPG) waveforms encode complementary electrical and volumetric perspectives of cardiovascular function. Prevailing contrastive pretraining frameworks adopt pairwise or leave-one-out alignment between per-channel views, capturing shared information but failing to recover the synergistic interactions that emerge only when channels are processed jointly. We present SyneCo, a synergy-aware contrastive pretraining framework that integrates per-channel embeddings through an attention-based fusion block before applying the contrastive objective on the fused representation. A dual InfoNCE loss supervises the fused representation and per-channel components jointly, and channel dropout confers robustness against incomplete recordings. Pretrained on 3.92 million co-recorded segments from MIMIC-III, MIMIC-IV, and VitalDB, SyneCo consistently surpasses pairwise and leave-one-out baselines on six ECG, four PPG, and three multi-channel benchmarks, with the largest gains on patient attribute prediction tasks that depending on cross-channel synergistic correlates.

1. Introduction

ECG and PPG record cardiovascular dynamics from complementary physical perspectives, the former through time-varying potential differences reflecting myocardial depolarization and repolarization, the latter through optical absorption tracking peripheral blood volume fluctuations (Bayoumy et al., 2021). In intensive care and intraoperative settings these two waveform classes are routinely acquired in synchrony, producing co-recorded data whose joint information content exceeds the sum of channel-wise contributions (Johnson et al., 2016; Johnson et al., 2023; Lee et al., 2022).

Recent self-supervised pretraining has produced strong unimodal foundation models for ECG (Li et al., 2025; Shu et al., 2025) and PPG (Pillai et al., 2025), yet these models inherently overlook synergistic interdependencies by design. Multi-channel pretraining frameworks have emerged to leverage such joint structure. SleepFM (Thapa et al., 2024; 2026) applied pairwise and leave-one-out InfoNCE objectives to brain activity, ECG, and respiratory waveforms. A methodological tension persists across these approaches. View-based contrastive learning aligns per-channel embeddings directly, treating each channel as an independent view of the same latent state. Recent analyses (Dufumier et al., 2025; Chen et al., 2025) have shown that this alignment paradigm captures shared information across channels but suppresses channel-specific structure and, more critically, fails to capture synergistic interactions that emerge only after channels are jointly processed.

This work moves beyond pairwise alignment by adopting a synergy-aware contrastive paradigm. Three per-channel encoders produce embeddings that are integrated through an attention-based fusion block into a single representation prior to the contrastive objective, explicitly targeting synergistic physiological coordination across simultaneously acquired electrical and volumetric signals rather than forcing per-channel embeddings to converge. We refer to the resulting framework as SyneCo, denoting synergy-aware contrastive learning over co-recorded physiological channels.

This paper makes three contributions. First, the pretraining corpus is expanded to encompass MIMIC-III, MIMIC-IV, and VitalDB, producing approximately 3.92 million co-recorded ten-second trimodal segments spanning 10,884 hours. Second, a dual InfoNCE objective is coupled with attention-based synergy-aware fusion, supervising the fused representation while preserving channel-aware discrimination through a complementary per-channel term computed before channel dropout. Third, a channel dropout mechanism stochastically zeros individual channel embeddings prior to fusion, training the model to remain stable under incomplete channel availability characteristic of clinical deployment.

[1]Department of Public Health and AI, National Cancer Center Graduate School of Cancer Science and Policy, Goyang-si, Gyeonggi-do, Republic of Korea. Correspondence to: Junetae Kim <lyjune0070@gmail.com>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

2. Method

2.1. Synergy-Aware Fusion of Per-channel Embeddings

As illustrated in Figure 1, each ten-second segment $s = (s_{\text{ECG-II}}, s_{\text{ECG-V}}, s_{\text{PPG}})$ is encoded by three independent ResNeXt1D backbones with non-shared weights, each producing a 512-dimensional L_2-normalized embedding. ECG signals are resampled to 500 Hz ($T = 5,000$ samples) and PPG to 125 Hz ($T = 1,250$ samples) prior to encoding (Shu et al., 2025; Pillai et al., 2025).

Rather than concatenating the three channel embeddings or averaging them, which would force premature commitment to a fixed interaction pattern, a Fusion Transformer module is employed to dynamically weight cross-channel interactions at the embedding level, instantiating the synergy-aware fusion mechanism that distinguishes SyneCo from alignment-based pretraining. The three L_2-normalized embeddings are reshaped into single-token sequences of shape $(B, 1, 512)$ and concatenated along the token axis to form a $(B, 3, 512)$ tensor. A single transformer encoder layer with four attention heads and a feed-forward expansion ratio of four operates on this three-token sequence. The attended tokens are subsequently mean-pooled along the token axis, and a final LayerNorm produces the fused embedding of dimension 512.

2.2. Channel Dropout

Within each training iteration, channel dropout is applied to the set of three L_2-normalized embeddings prior to fusion. Each embedding is independently zeroed with probability 0.4, and with probability 0.3 a complementary event simultaneously zeros two of the three channels, subject to a hard constraint that at least one channel remains active. Validation and inference forward all channels intact.

By presenting the fusion module with incomplete channel sets during training, it compels the downstream contrastive objective to learn representations that remain discriminative under partial channel availability, a condition frequently encountered in real-world clinical deployments where electrodes detach, optical sensors saturate, or individual channels fall below usable quality thresholds.

2.3. Dual InfoNCE Objective

For each segment, two augmented views $s^{(1)}$ and $s^{(2)}$ are generated independently through a two-step augmentation that first injects zero-mean Gaussian noise with standard deviation 0.05 into each waveform, then multiplies the perturbed waveform by an amplitude-scaling factor of 0.9. The same augmentation operator is applied consistently across ECG-II, ECG-V, and PPG within a view, preserving the temporal synchronization that conveys the physiological coupling between electrical and peripheral signals.

Encoding, channel dropout, and fusion produce, for each view, a 512-dimensional fused embedding $z^{(v)}$ and three per-channel embeddings $e_m^{(v)}$ for $m \in \{\text{ECG-II}, \text{ECG-V}, \text{PPG}\}$. Within a mini-batch of size $N = 128$, positive pairs are formed by the two augmented views derived from the same source segment, with all other in-batch pairings serving as negatives. The fused contrastive term $\mathcal{L}_{\text{fused}}$ adopts the symmetric InfoNCE formulation, with same-view similarity blocks aggregated into the negative pool to enrich the contrastive signal. The per-channel term is computed before channel dropout so that every encoder receives gradient supervision on every batch, and is given by:

$$\mathcal{L}_{\text{mod}} = \frac{1}{3} \sum_{m} \mathcal{L}_{\text{InfoNCE}} \left(e_m^{(1)}, e_m^{(2)} \right) \tag{1}$$

The two terms are combined as:

$$\mathcal{L} = \mathcal{L}_{\text{fused}} + \alpha \mathcal{L}_{\text{mod}} \tag{2}$$

with $\alpha = 0.5$. The dual objective supervises the fused representation at two complementary levels, with $\mathcal{L}_{\text{fused}}$ targeting the synergy-bearing post-fusion representation and $\mathcal{L}_{\text{mod}}$ preserving the channel-specific discriminative structure that synergistic fusion alone cannot recover.

2.4. Pretraining Corpus and Preprocessing

The pretraining corpus aggregates three large-scale clinical databases comprising MIMIC-III, MIMIC-IV, and VitalDB (Johnson et al., 2016; Johnson et al., 2023; Lee et al., 2022). Because subject counts differ markedly across sources, per-subject segment counts are capped to prevent a small number of long recordings from dominating the training distribution, with caps set to 50 segments per record for MIMIC-III, 1,000 per record for MIMIC-IV, and 500 per subject for VitalDB. Each recording is filtered by a minimum continuous duration of ten minutes, segmented into non-overlapping ten-second windows in which all three channels are simultaneously valid, and subjected to flatline detection that discards segments containing more than 25% constant samples in any channel (Pillai et al., 2025). ECG signals undergo a fifth-order Butterworth bandpass filter with cutoffs of 0.67 Hz and 40 Hz (Shu et al., 2025); PPG signals undergo a fourth-order Chebyshev bandpass filter with cutoffs of 0.5 Hz and 8 Hz (Pillai et al., 2025). All segments are individually z-score normalized. The combined pretraining set comprises approximately 3.92 million synchronized trimodal segments from 5,453 unique subjects spanning 10,884 hours of waveform data, with a held-out VitalDB test split of 646 subjects reserved for downstream evaluation. Detailed dataset statistics are provided in Appendix B.

3. Experiments

SyneCo is evaluated through linear probing on a frozen pretrained encoder. For each downstream segment a 512-dimensional embedding is extracted; for subjects with multiple segments a single subject-level representation is obtained by averaging valid segment embeddings, following the protocol of PaPaGei (Pillai et al., 2025). Frozen embeddings are evaluated under three probes, comprising a fully connected linear head (FC) trained with cross-entropy for classification, logistic regression (LR) for classification, and random forest (RF) for both classification and regression. Standardized embeddings are used for the LR and RF probes. All experiments are repeated with three seeds (0, 42, 1234) and the mean is reported.

The downstream evaluation spans six ECG benchmarks (PTB-XL super5 / sub23 / sex (Wagner et al., 2020); Chapman-Shaoxing rhythm and sex (Zheng et al., 2020); Georgia diagnosis (Alday et al., 2020)), four PPG benchmarks (Dalia activity and HR (Reiss et al., 2019); PPGBP hypertension and HR (Liang et al., 2018)), and three multi-channel benchmarks (VTaC false alarm detection (Lehman et al., 2023); VitalDB sex and emergency operation classification (Lee et al., 2022)). Baselines comprise BYOL, SimCLR, the SleepFM pairwise (SleepFM Pair) and leave-one-out (SleepFM LOO) objectives, a two-channel ECG-only variant of SyneCo (SyneCo 2mod), and a two-channel ECG-only SleepFM (SleepFM 2mod), all trained from scratch on the same pretraining corpus and budget. A naive concatenation baseline (Stack) replaces the attention-based fusion block with simple feature concatenation, isolating the contribution of attention-based fusion. (Macro) AUROC is reported for classification and mean absolute error (MAE) for regression. Detailed downstream dataset descriptions and the full numerical tables are provided in Appendix D.

4. Results

4.1. ECG Downstream Tasks

Table 1 summarizes representative ECG results under logistic regression probing. On diagnostic classification, SyneCo reaches AUROC 0.842 on PTB-XL super5 and 0.851 on sub23, matching BYOL (0.840, 0.851) and substantially exceeding SleepFM LOO (0.739, 0.724) and SleepFM Pair (0.721, 0.697). On Georgia diagnosis, SyneCo attains 0.835, comparable to BYOL (0.837) and well above the SleepFM variants. The contrast between SyneCo and SyneCo 2mod isolates the contribution of PPG, with PPG inclusion producing consistent gains across all six tasks, raising super5 from 0.802 to 0.842, sub23 from 0.796 to 0.851, and Georgia diagnosis from 0.775 to 0.835. The parallel SleepFM LOO versus SleepFM 2mod comparison produces markedly smaller improvements (super5 0.732 to 0.739; sub23 0.733 to 0.724), indicating that the multi-channel benefit conferred by the additional PPG channel is more effectively captured by synergy-aware fusion than by leave-one-out alignment.

4.2. PPG Downstream Tasks

PPG performance is summarized in Table 2 (regression in MAE under random forest, classification in AUROC under LR). On Dalia HR regression, SyneCo achieves MAE 4.87, comparable to BYOL (4.93) and SyneCo 2mod (4.84) and substantially better than SleepFM LOO (5.65), SleepFM Pair (6.19), and SleepFM 2mod (7.74). On Dalia activity classification, SyneCo achieves AUROC 0.844, exceeding SleepFM LOO (0.804), SleepFM Pair (0.815), and BYOL (0.809). Switching to classification, SyneCo achieves AUROC 0.844 on Dalia activity, exceeding SleepFM LOO (0.804), SleepFM Pair (0.815), and BYOL (0.809). On PPGBP hypertension, however, SyneCo (LR AUROC 0.637) is outperformed by SyneCo 2mod (0.717), suggesting that the three-channel alignment may redistribute representational capacity away from blood pressure-related morphological features in PPG.

4.3. Multi-channel Downstream Tasks

The multi-channel evaluation directly tests synergistic representation quality. On VitalDB sex prediction, SyneCo achieves AUROC 0.769 under logistic regression, exceeding SleepFM LOO by 7.88 percentage points and SleepFM Pair by 9.44 percentage points (Table 3). The pronounced advantage on patient attribute prediction, which depends on subtle cross-channel correlates spanning ventricular repolarization in ECG and peripheral pulse waveform shape in PPG, indicates that synergy-aware fusion preserves patient-level physiological signatures more effectively than alignment-based strategies when all channels are available during pretraining. On VitalDB emergency operation classification, SyneCo attains 0.674, marginally above SleepFM Pair (0.665) and SleepFM LOO (0.655). On VTaC false alarm detection, the two-channel ECG-only SleepFM 2mod marginally leads at AUROC 0.755, with SyneCo at 0.736, consistent with the observation that ECG-dominant temporal features are sufficient for arrhythmia false alarm detection and that additional PPG alignment constraints provide no clear advantage on this task.

4.4. Ablation

Two ablations isolate the contribution of attention-based fusion and the contribution of PPG inclusion. The Stack variant replaces attention fusion with naive concatenation, retaining all other components. On ECG (Table 1), Stack reaches LR AUROC 0.801, 0.764, and 0.703 on PTB-XL super5, sub23, and Georgia diagnosis, compared to 0.842, 0.851, and 0.835 for SyneCo. The degradation under Stack

Table 1. AUROC on ECG benchmarks under logistic regression probing on ECG-lead II and V inputs. Best per column in bold.

Method (LR)	PTB super5	PTB sub23	PTB sex	Chap rhy	Chap sex	GA dx
SyneCo	**0.842**	**0.851**	**0.755**	0.917	**0.748**	0.835
SleepFM LOO	0.739	0.724	0.661	**0.958**	0.662	0.768
SleepFM Pair	0.721	0.697	0.628	0.922	0.621	0.739
SyneCo 2mod	0.802	0.796	0.718	0.914	0.704	0.775
Stack	0.801	0.764	0.699	0.871	0.676	0.703
BYOL	0.840	**0.851**	0.727	0.935	0.637	**0.837**
SimCLR	0.733	0.714	0.615	0.940	0.623	0.750

Table 2. PPG benchmark results. Regression in MAE under random forest (lower is better); classification in AUROC under logistic regression.

Method Metric	Dalia HR MAE ↓	PPGBP HR MAE ↓	Dalia act. AUROC ↑	PPGBP hyp. AUROC ↑
SyneCo	4.867	5.185	**0.844**	0.637
SleepFM LOO	5.648	5.013	0.804	0.505
SleepFM Pair	6.191	5.392	0.815	0.520
SyneCo 2mod	4.835	5.346	0.833	**0.717**
Stack	10.086	6.417	0.777	0.656
BYOL	4.930	5.012	0.809	0.611
SimCLR	**4.470**	**4.933**	0.812	0.649

Table 3. Multi-channel benchmark AUROC under logistic regression probing. VitalDB tasks are restricted to three-channel pretrained models.

Method (LR)	VTaC	VitalDB sex	VitalDB em.
SyneCo	0.736	**0.769**	**0.674**
SleepFM LOO	0.730	0.690	0.655
SleepFM Pair	0.709	0.675	0.665
SleepFM 2mod	**0.755**	–	–
SyneCo 2mod	0.724	–	–
Stack	0.746	0.667	0.601

indicates that the discriminative structure of the learned representation depends critically on the dynamic cross-channel weighting induced by attention fusion rather than on the increased parameter capacity afforded by concatenation. On PPG regression (Table 2), the SyneCo advantage over Stack is most pronounced on Dalia HR, where Stack reaches MAE 10.09 compared to 4.87 for SyneCo, more than a twofold degradation. On the multi-channel evaluation (Table 3), Stack reaches LR AUROC 0.667 on VitalDB sex compared to 0.769 for SyneCo, a gap of 10.21 percentage points. These margins confirm that synergy-aware representation requires an explicit fusion mechanism that adaptively weights cross-channel contributions; static concatenation cannot recover synergistic interactions even when given the same per-channel embeddings. The SyneCo versus SyneCo 2mod comparison reported in Sections 4.1 and 4.2 further confirms that PPG inclusion enriches the learned representations in a manner not achievable through two-lead ECG alignment alone.

5. Conclusion

SyneCo operationalizes a synergy-aware contrastive paradigm by integrating per-channel embeddings through attention-based fusion before the contrastive objective, moving beyond the pairwise alignment strategies that dominate prior multi-channel pretraining. A dual InfoNCE loss preserves channel-specific discriminability while supervising the fused representation, and channel dropout confers robustness against the incomplete channel availability characteristic of clinical deployment. Across six ECG, four PPG, and three multi-channel benchmarks SyneCo consistently surpasses pairwise and leave-one-out baselines, with the largest gains on patient attribute prediction tasks that depend on cross-channel synergistic correlates. Future work will extend the framework to ambulatory cohorts, accommodate asynchronous channel configurations, and explore beat-level cross-channel correspondence.

Impact Statement

Pretrained on retrospective ICU and intraoperative recordings, the released representations may reflect cohort-specific biases and require task-specific validation before any clinical use.

References

Alday, E. A. P., Gu, A., Shah, A. J., Robichaux, C., Wong, A. K. I., Liu, C., Liu, F., Rad, A. B., Elola, A., Seyedi, S., Li, Q., Sharma, A., Clifford, G. D., and Reyna, M. A. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41(12):124003, December 2020. ISSN 0967-3334. doi: 10.1088/1361-6579/ABC960.

Bayoumy, K., Gaber, M., Elshafeey, A., Mhaimeed, O., Dineen, E. H., Marvel, F. A., Martin, S. S., Muse, E. D., Turakhia, M. P., Tarakji, K. G., and Elshazly, M. B. Smart wearable devices in cardiovascular care: where we are and how to move forward. *Nature Reviews Cardiology*, 18(8):581–599, August 2021. ISSN 1759-5010. doi: 10.1038/s41569-021-00522-7.

Chen, L., Park, K., and Kim, J. Multimodal Contrastive Learning with Early Fusion for Robust Medical Signal Representation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 4649–4653, New York, NY, USA, November 2025. Association for Computing Machinery. ISBN 979-8-4007-2040-6. doi: 10.1145/3746252.3760883.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations, July 2020. URL http://arxiv.org/abs/2002.05709. arXiv:2002.05709 [cs].

Dufumier, B., Castillo Navarro, J., Tuia, D., and Thiran, J.-P. What to align in multimodal contrastive learning? In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R. (eds.), *International Conference on Learning Representations*, volume 2025, pp. 5408–5432, 2025.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised Learning, September 2020. URL http://arxiv.org/abs/2006.07733. arXiv:2006.07733 [cs].

Johnson, Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–1, January 2023. doi: https://doi.org/10.1038/s41597-022-01899-x.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016. doi: https://doi.org/10.1038/sdata.2016.35.

Lee, H. C., Park, Y., Yoon, S. B., Yang, S. M., Park, D., and Jung, C. W. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data 2022 9:1*, 9(1):279–, June 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01411-5. URL https://www.nature.com/articles/s41597-022-01411-5.

Lehman, L.-w. H., Moody, B., Deep, H., Wu, F., Saeed, H., McCullum, L., Perry, D., Struja, T., Li, Q., Clifford, G., and Mark, R. G. VTaC: a benchmark dataset of ventricular tachycardia alarms from ICU monitors. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 38827–38843, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Li, J., Aguirre, A. D., Junior, V. M., Jin, J., Liu, C., Zhong, L., Sun, C., Clifford, G., Westover, M. B., and Hong, S. An Electrocardiogram Foundation Model Built on over 10 Million Recordings. *NEJM AI*, 2 (7), June 2025. ISSN 2836-9386. doi: 10.1056/AIOA2401033. URL https://ai.nejm.org/doi/full/10.1056/AIoa2401033.

Liang, Y., Chen, Z., Liu, G., and Elgendi, M. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China. *Scientific Data*, 5(1): 180020, February 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.20. URL https://www.nature.com/articles/sdata201820.

Pillai, A., Spathis, D., Kawsar, F., and Malekzadeh, M. PaPaGei: Open Foundation Models for Optical Physiological Signals. In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R. (eds.), *International Conference on Learning Representations*, volume 2025, pp. 48230–48261, 2025. doi: 10.48550/arXiv.2410.20542.

Reiss, A., Indlekofer, I., Schmidt, P., and Van Laerhoven, K. Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. *Sensors*, 19(14):3079, January 2019. ISSN 1424-8220. doi: 10.3390/s19143079. URL https://www.mdpi.com/1424-8220/19/14/3079.

Shu, Y., Charlton, P. H., Kawsar, F., Hernesniemi, J., and Malekzadeh, M. CLEF: Clinically-Guided Contrastive Learning for Electrocardiogram Foundation Models, December 2025. URL http://arxiv.org/abs/2512.02180. arXiv:2512.02180 [cs].

Thapa, R., He, B., Kjær, M. R., Moore, H., Ganjoo, G., Mignot, E., and Zou, J. SleepFM: Multi-modal Representation Learning for Sleep Across Brain Activity, ECG and Respiratory Signals. *Proceedings of Machine Learning*

Research, 235:48019–48037, May 2024. ISSN 26403498. URL https://arxiv.org/pdf/2405.17766.

Thapa, R., Kjaer, M. R., He, B., Covert, I., Moore IV, H., Hanif, U., Ganjoo, G., Westover, M. B., Jennum, P., Brink-Kjaer, A., Mignot, E., and Zou, J. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, 32(2):752–762, February 2026. ISSN 1546-170X. doi: 10.1038/s41591-025-04133-4. URL https://www.nature.com/articles/s41591-025-04133-4.

Tian, Y., Li, Z., Jin, Y., Wang, M., Wei, X., Zhao, L., Liu, Y., Liu, J., and Liu, C. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Reports Medicine*, 5(12), December 2024. ISSN 2666-3791. doi: 10.1016/j.xcrm.2024.101875. URL https://www.cell.com/cell-reports-medicine/abstract/S2666-3791(24)00646-3.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154–154, May 2020. doi: https://doi.org/10.1038/s41597-020-0495-6.

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., and Rakovski, C. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48, February 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0386-x. URL https://www.nature.com/articles/s41597-020-0386-x.

A. Overall Framework

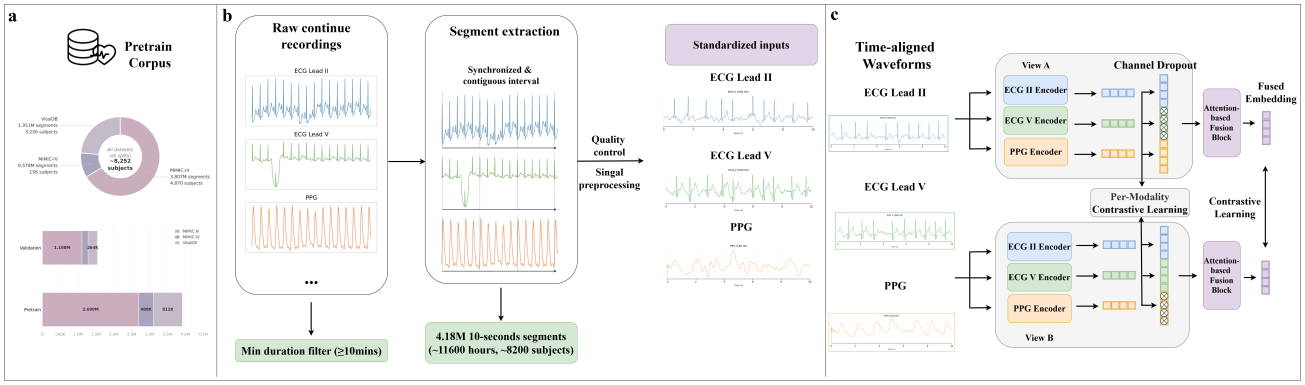

Figure 1. Overview of the SyneCo pretraining pipeline. **(a)** Composition of the pretraining corpus, comprising subject and segment counts across MIMIC-III, MIMIC-IV, and VitalDB. **(b)** Data processing pipeline, in which continuous recordings of ECG Lead II, ECG Lead V, and PPG are filtered by a minimum duration of ten minutes, segmented into co-recorded non-overlapping ten-second windows, and passed through quality control and channel-specific bandpass filtering to yield standardized inputs. **(c)** Pretraining architecture, in which two augmented views of the same trimodal segment are independently encoded by per-channel ResNeXt1D backbones, subjected to channel dropout, and aggregated through an attention-based fusion block to produce a fused embedding. A dual InfoNCE objective jointly supervises the fused representation through a cross-view contrastive term and the per-channel embeddings through a complementary per-channel term.

B. Pretraining Corpus Statistics

Section B reports the per-database splits after preprocessing. After subject-level partitioning, the combined pretraining set encompasses approximately 3.92 million ten-second synchronized trimodal segments totaling 10,884 hours of waveform data from 5,453 unique subjects. The validation set comprises approximately 1.54 million segments. The VitalDB test split (646 subjects) is held out exclusively for downstream evaluation.

Table B. Pretraining dataset statistics and splits after preprocessing.

Database	Split	Subjects	Records	Segments	Hours	Cap / record
MIMIC-III	Pretrain	3,409	54,639	2,699,233	7,498.0	50
	Validation	1,461	22,409	1,108,160	3,078.3	50
MIMIC-IV	Pretrain	109	640	407,932	1,133.1	1,000
	Validation	47	254	169,948	472.1	1,000
VitalDB	Pretrain	1,935	1,935	810,996	2,252.8	500
	Validation	645	645	264,381	734.4	500
	Test (held out)	646	646	275,156	764.3	–
Total	**Pretrain**	**5,453**	**57,214**	**3,918,161**	**10,884**	–

C. Implementation Details

C.1. Encoder Architecture

Each ten-second segment $\mathbf{s} = (\mathbf{s}_{\text{ECG-II}}, \mathbf{s}_{\text{ECG-V}}, \mathbf{s}_{\text{PPG}})$ is encoded by three independent ResNeXt1D backbones with non-shared weights, characterized by grouped bottleneck convolutions and Squeeze-and-Excitation channel attention. Each backbone primarily consists of an initial stride-2 convolution with 64 base filters and kernel size 16, followed by seven sequential stages containing $[2, 2, 2, 3, 3, 4, 4]$ bottleneck blocks with progressively expanded channel widths

$[64, 160, 160, 400, 400, 1024, 1024]$. Each backbone produces a 512-dimensional L_2-normalized embedding. ECG signals are resampled to 500 Hz ($T = 5{,}000$ samples) and PPG to 125 Hz ($T = 1{,}250$ samples) prior to encoding.

C.2. Pretraining

SyneCo is pretrained with the AdamW optimizer at learning rate 1×10^{-4}, weight decay 5×10^{-5}, and batch size 128, under a cosine annealing schedule for up to 300,000 gradient steps. Given the scale of the corpus, a cycle-based sampling strategy independently draws 50% of segments from each data source at the start of each cycle (approximately 1.95 million segments per cycle) and reseeds the sampler at the start of every new cycle. Validation loss is evaluated on a fixed 115,500-segment subset every 4,000 steps, with early stopping after five consecutive evaluations without improvement.

C.3. Downstream Probing

Frozen embeddings are evaluated under three probes, comprising a fully connected linear head (FC), logistic regression (LR), and random forest (RF). For classification tasks, FC is trained with cross-entropy loss, and LR and RF are fitted as classifiers on standardized embeddings. For regression tasks, RF is fitted as a regressor. FC is optimized for up to 100 epochs with early stopping on validation performance. LR and RF use the scikit-learn defaults, with L2 regularization for LR. For datasets in which a subject contributes multiple segments, segment-level embeddings are mean-pooled into a subject-level representation prior to probing. All experiments are repeated with three random seeds $\{0, 42, 1234\}$ and the mean is reported.

D. Downstream Datasets and Baselines

ECG benchmarks comprise PTB-XL (Wagner et al., 2020) (21,837 ten-second 12-lead recordings sampled at 500 Hz from 18,885 patients with hierarchical SNOMED-CT annotations including five superclasses and 23 subclasses), Chapman-Shaoxing (Zheng et al., 2020) (10,646 recordings with 11-class rhythm and demographic annotations), and Georgia from the PhysioNet Computing in Cardiology Challenge 2020 (Alday et al., 2020)) (10,344 recordings with 22-class multi-label SNOMED-CT annotations). All ECG evaluation uses Lead II and Lead V only, matching the pretraining configuration.

PPG benchmarks comprise PPG-DaLiA (Reiss et al., 2019)(15 subjects, 9-class daily activity with synchronized chest-ECG-derived heart rate as regression target, 8-second windows zero-padded to 10 seconds) and PPG-BP (Liang et al., 2018) (continuous finger PPG with simultaneous blood pressure measurements, hypertension defined as SBP ≥ 140 mmHg or DBP ≥ 90 mmHg, with SBP, DBP, and HR regression targets, individual pulse segments of approximately 2.1 seconds zero-padded to 10 seconds).

Multi-channel benchmarks comprise VTaC (Lehman et al., 2023) (5,037 annotated VT alarm events from ICU bedside monitors at three US hospitals, 6-minute records segmented into ten-second windows, official train/test split) and VitalDB (Lee et al., 2022) (held-out test split of 646 surgical cases excluded from pretraining, with sex and emergency operation classification at the per-case level using up to 50 segments mean-pooled into a single representation).

Multi-channel self-supervised baselines comprise BYOL (Grill et al., 2020) (online plus momentum target encoder, no negative pairs), SimCLR (Chen et al., 2020) (NT-Xent loss with in-batch negatives), SleepFM Pair (Thapa et al., 2024) (three pairwise NT-Xent objectives summed across modality pairs), SleepFM LOO (three leave-one-out objectives where the concatenation of two channels serves as positive counterpart for the third), and Stack (per-channel embeddings concatenated without any learned fusion). As BYOL and SimCLR are inherently two-view objectives, they are instantiated separately for each downstream modality, using the ECG-II and ECG-V pair for ECG tasks and the ECG-II and PPG pair for PPG tasks. All baselines share encoder architecture, augmentation strategy, dataset, and training budget with SyneCo to ensure controlled comparison; only the contrastive objective and the presence or absence of fusion differ.

E. Comparison with Open-Source Foundation Models

SyneCo is compared against publicly available foundation models trained on domain-specific large-scale corpora, comprising ECGFM-KED and ECGFounder for ECG representation, and PaPaGei-P and PaPaGei-S for PPG representation.

ECGFounder(Li et al., 2025) is an ECG foundation model pretrained on over 10 million recordings with a ResNeXt1D architecture from the same Net1D family used in this work. ECGFounder-1lead applies a 1-channel encoder independently to ECG-II and ECG-V and concatenates the resulting features into a 2,048-dimensional embedding. ECGFounder-12lead

Table E.1. AUROC results for ECG classification benchmarks under ogistic regression probing, comparing SyneCo against ECGFM-KED, ECGFounder-1lead, and ECGFounder-12lead across seven tasks.

Model	PTB super5	PTB sub23	PTB sex	Chap rhy	Chap sex	GA dX
SyneCo	0.842	0.851	0.755	0.917	0.748	0.835
ECGFM-KED	0.841	0.855	0.764	0.933	0.746	0.798
ECGFounder-1lead	0.868	0.874	0.813	0.984	0.812	0.876
ECGFounder-12lead	0.877	0.890	0.819	0.982	0.814	0.884

Table E.2. PPG benchmark results comparing SyneCo against PaPaGei-P and PaPaGei-S. Regression in MAE under random forest regressor (lower is better); classification in AUROC under linear probing (higher is better).

Method Metric	Dalia HR MAE ↓	PPGBP HR MAE ↓	Dalia act. AUROC ↑	PPGBP hyp. AUROC ↑
SyneCo	**4.866**	**5.185**	**0.844**	0.637
PaPaGei-P	9.149	6.381	0.800	**0.657**
PaPaGei-S	9.560	6.535	0.767	0.631

employs a 12-channel encoder pretrained on full 12-lead ECGs, with Lead II and Lead V placed at their standard channel indices and the remaining channels zero-filled, yielding a 1,024-dimensional embedding. In both variants, the classification head is discarded and the penultimate feature vector serves as the representation.

KED ECG Encoder(Tian et al., 2024) is a multimodal ECG foundation model that aligns ECG signals with cardiologist-verified diagnostic reports under a CLIP-style contrastive framework. Only the ECG encoder is retained for downstream evaluation, producing a 768-dimensional pooled representation.

PaPaGei(Pillai et al., 2025) is an open-source PPG foundation model pretrained on over 57,000 hours of wearable optical physiological signals. PaPaGei-P uses a ResNet1D backbone, whereas PaPaGei-S augments the same backbone with a 3-expert Mixture-of-Experts layer; both produce 512-dimensional embeddings.