TEX-MEX: SIMILARITY-GUIDED TEMPORAL CONTRASTIVE LEARNING FOR TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning for time series depends on deciding which pairs of inputs should be treated as similar and dissimilar. Because few augmentations reliably preserve meaning in time series, prior methods have relied on two disjoint notions of similarity between windows: structural similarity, which relates windows with similar morphology, and temporal similarity, which relates windows that occur close together in time. We introduce TEX-MEX, to our knowledge the first approach to unify the two into a single relationship. Given a random batch of windows drawn across the dataset, TEX-MEX uses the encoder to identify which windows are structurally similar or dissimilar to each other. It then pulls each window toward the temporally shifted versions of similar windows and pushes it away from those of dissimilar ones, learning what typically accompanies similar windows in time. Across PPG, EEG, ECG, accelerometry, and continuous glucose monitoring, TEX-MEX broadly achieves state-of-the-art performance against contrastive, reconstructive, JEPA-based, and modality-specific approaches, establishing it as a strong general SSL objective. We further train a 560K-parameter PPG foundation model that achieves state-of-the-art performance across a broad range of downstream benchmarks while being roughly 50$\times$ smaller than the next-best unimodally pretrained PPG foundation model in our evaluation. We will release this foundation model open-source and open-weight to the community to support future work and practical applications.

1 INTRODUCTION

Self-supervised learning (SSL) has emerged as the dominant paradigm for learning generalizable representations from large-scale unlabeled data (Chen et al., 2020; He et al., 2022). This approach is particularly attractive for modeling physiological signals, such as photoplethysmography (PPG) or electrocardiography (ECG), which are routinely captured as time series data by wearables (e.g. smart watches) or medical devices. Labeling clinically-important events in such data (such as atrial fibrillation in an ECG recording) typically requires expert interpretation, invasive assays, or controlled clinical settings (Xu et al., 2024), and is very difficult to perform at scale. At the same time, the increasing adoption of smart watches by the general public and the large-scale aggregation of waveform data within electronic health records has created an opportunity to develop foundation models for such physiological signals (Xu et al., 2025; Saha et al., 2025; Ansari et al., 2025). Such models could enable new data-driven approaches to modeling states of health and disease.

Time series representation learning is challenging due to the need to capture statistical dependencies across widely-varying temporal scales (Goldberger et al., 2002). For example, heart rate, which is measurable via PPG, is correlated at the level of minutes in responding to physiological demands such as exercise, but is also correlated at the level of hours due to the diurnal cycle. Weekly patterns of stress or physical activity, such as regular commute times or exercise routines, can induce even longer-range correlational structure. A key step in developing foundation models for such data is the identification of effective pretext tasks. Two common approaches are contrastive learning, which extracts windows of samples from a time series recording to construct positive and negative pairs, and masked autoencoding, which uses temporal context to reconstruct the samples within a masked-out window. Contrastive learning is a particularly useful lens for examining time series representation learning as it provides a way to explicitly reason about similarity and dissimilarity.

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.

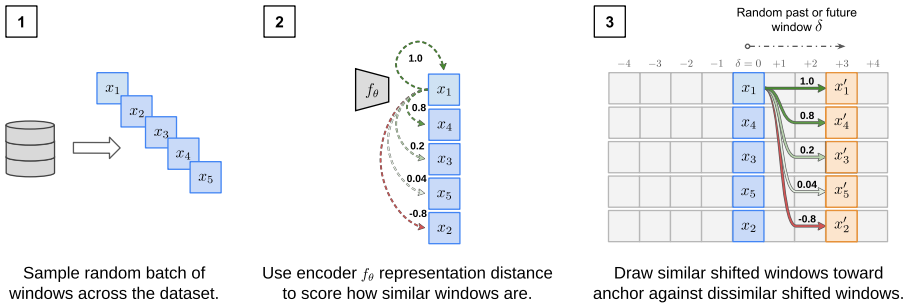

Figure 1: **Overview of TEX-MEX.** (1) Windows of sensor data are randomly sampled from the dataset; (2) The encoder scores each window's similarity to an anchor (here, x_1) using cosine similarity; (3) Each window is offset in time (forward or backward) by a random δ (here, $+3$ windows). The anchor is pulled toward the temporal shifts of similar windows and away from those of dissimilar ones, in proportion to the similarity, composing structural and temporal notions.

Unfortunately, time series recordings typically lack a rich set of well-defined invariances from which positive pairs can be constructed (Xu et al., 2024). As a consequence, prior contrastive learning works have adopted one of two disjoint approaches to defining similarity between windows: *structural similarity* approaches that index on signal morphology or frequency content (i.e., windows are similar if they contain similar waveform shapes, like two heart beats in an ECG waveform) (Xu et al., 2025; 2024), and *temporal similarity* measures that index on temporal proximity (i.e., two windows are similar if they occur close together in time, such as when a cool-down period follows a period of vigorous exercise) (Tonekaboni et al., 2021; Oord et al., 2018). In contrast, this paper proposes, to our knowledge, the first contrastive learning approach which unifies structural and temporal similarity measures into a single framework. We demonstrate that the standard approach of generating positive pairs by mining morphologically-similar patterns from time series recordings (Xu et al., 2024) can be extended by composing it with temporal similarity, drawing positive and negative pairs from the temporal context of similar and dissimilar windows. This unifies the two primary sources of similarity in time series modeling: windows that correspond to similar physiological processes (e.g., two heartbeats) will have similar morphology, while windows which map onto stable patterns of human behavior (e.g., cool-down following exercise) will be temporally co-occurring. An exercise window, for instance, can learn from the cool-downs that follow exercise in other recordings, capturing what typically follows exercise rather than what followed it once.

We introduce TEX-MEX: for each anchor window, it first identifies similar and dissimilar windows within a random batch drawn across the dataset. Every window is then shifted by the same random offset, forward or backward in time, redrawn in each batch to span many timescales. TEX-MEX pulls the anchor toward the shifted versions of similar windows and pushes it away from those of dissimilar windows. We find that the encoder's own representations suffice for measuring this similarity, eliminating the costly signal distance measures of prior methods (Xu et al., 2024). Unlike those fixed measures, this similarity improves as the encoder trains, ultimately surpassing them. Because it is graded rather than binary, TEX-MEX is naturally a soft contrastive method (Lee et al., 2024; Xu et al., 2025), weighting each shifted window by its source window's similarity to the anchor, which captures degrees of relatedness and adds robustness to false negatives. The main contributions of our work are as follows:

1. We introduce TEX-MEX, a contrastive objective that unifies structural and temporal similarity by contrasting the shifted versions of similar windows against those of dissimilar windows, with similarity derived from the encoder's own representations.
2. We show that this representation-space similarity improves throughout training, surpassing existing signal distance approaches as a predictor of mutual class membership.
3. We demonstrate the generality of TEX-MEX across PPG, ECG, accelerometry, EEG, and CGM, where it broadly achieves state-of-the-art performance against leading contrastive, reconstructive, JEPA-based, and modality-specific methods.
4. Leveraging its efficiency, we train and will release an open-source, open-weight PPG foundation model that achieves state-of-the-art performance across a range of downstream tasks

with just 560K parameters, roughly 50× smaller than the next-best unimodally pretrained PPG foundation model in our evaluation.

2 Related Work

Time-Series Self-Supervised Learning. Temporal contrastive methods use proximity in time as supervision: CPC (Oord et al., 2018) contrasts a window's future against other samples, and TNC (Tonekaboni et al., 2021) treats windows within a locally stationary neighborhood as positives, reweighting others as unlabeled. A single neighborhood, however, ties relatedness to one timescale and one recording. Structural methods instead relate windows by learned distances, such as REBAR's reconstruction-based motif similarity (Xu et al., 2024) and RelCon's learnable distance (Xu et al., 2025), which RelCon also uses to grade relationships for soft contrastive learning, as in our work. These, however, miss temporal relationships, relating only windows that look alike. TS2Vec (Yue et al., 2022) and SoftCLT (Lee et al., 2024) apply instance-wise and temporal contrast as losses independent of one another, and their temporal contrast stays within a single input. SoftCLT softens both, using DTW distance and a monotonically decaying timestamp distance, and reports that embedding-space similarity, as in NNCLR (Dwibedi et al., 2021), degrades performance; in our formulation, where it determines how temporal context is contrasted rather than serving directly as positives, it instead surpasses data-space distances. Masked modeling approaches such as TimeMAE (Cheng et al., 2026) and PatchTST (Nie et al., 2022) instead reconstruct masked portions of an input.

PPG Foundation Models. Recent PPG foundation models span contrastive, reconstructive, and autoregressive pretraining. Contrastive approaches differ mainly in how they relate windows: through shared subject identity (Abbaspourazad et al., 2024; Pillai et al., 2025), morphology-based similarity (Pillai et al., 2025; Saha et al., 2025), quality-aware temporal proximity (Ding et al., 2024), or synchronization with paired ECG (Nie et al., 2026). xMAE (Zhou et al., 2026) instead learns through cross-modal reconstruction of masked ECG, and GPT-PPG (Chen et al., 2025) models PPG autoregressively through next-patch prediction. We build on Pulse-PPG (Saha et al., 2025), adopting its pretraining corpus and downstream evaluation.

3 Methodology

Batch Construction At each training step, we randomly sample a batch of B windows, $\mathcal{X} = \{x_1, \ldots, x_B\}$, across the pretraining dataset. Then, for each batch, we sample a single temporal offset $\delta \sim \mathcal{N}(0, \sigma^2)$, measured in window lengths. Unlike in TNC, δ does not determine positives: it is shared across the batch, and each shifted window's role as positive or negative is set by encoder-estimated similarity, as described next. The offset thus only selects the timescale each batch probes, and the Gaussian acts as a simple, symmetric prior over where useful relationships are likely to be exposed, with σ setting their scale. We reject and resample draws outside $\delta_{\min} \leq |\delta| \leq \delta_{\max}$. The lower bound prevents shifted windows from overlapping their originals, which would permit trivial identity shortcuts; the upper bound limits shifts that cross gaps or recording boundaries. We set $\delta_{\min} = 1$ and $\delta_{\max} = 2.5\sigma$ by default, the latter truncating only $\approx$1.2% of the Gaussian's mass. Each window x_i is then paired with a shifted offset (a past or future window) $x'_i = T_\delta(x_i)$ from the same recording, where T_δ denotes displacement by δ along the time series. This produces paired unshifted and shifted sets, $\mathcal{X} = \{x_1, \ldots, x_B\}$ and $\mathcal{X}' = \{x'_1, \ldots, x'_B\}$, respectively.

Finding Similar and Dissimilar Windows: Encoder-Based Batch Grading A useful property of TEX-MEX is the availability of continually improving estimates of the degree of similarity between windows, derived directly from the encoder's representation space (as examined in Section 4.1). Let $z_i = f_\theta(x_i)$ denote the ℓ_2-normalized representation of window x_i. For each anchor x_i, we compute its cosine similarity with every member x_k of the unshifted batch as $c_{ik} = z_i^\top z_k$. We then transform these similarities into positive and negative weights,

$$w_{ik}^+ = \frac{\exp(c_{ik}/\tau_w)}{\sum_{j=1}^B \exp(c_{ij}/\tau_w)} \qquad w_{ik}^- = \frac{\exp(-c_{ik}/\tau_w)}{\sum_{j=1}^B \exp(-c_{ij}/\tau_w)} \tag{1}$$

where τ_w controls the sharpness of the weighting distributions. Thus, w_{ik}^+ places greater weight on batch members most similar to the anchor, while w_{ik}^- places greater weight on those least similar.

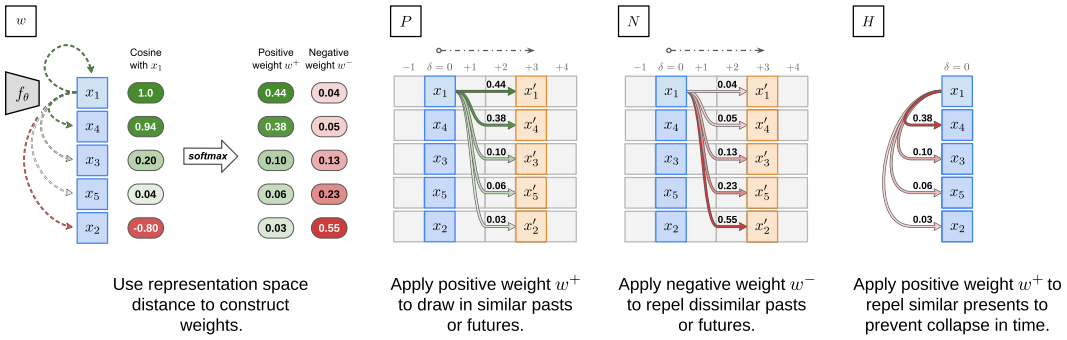

Figure 2: **Components of the TEX-MEX loss**, shown for anchor x_1 with a batch of five windows and $\delta = +3$. (w) The anchor's cosine similarity to each window is turned by two softmaxes into positive weights w^+, which favor similar windows, and negative weights w^-, which favor dissimilar ones. (P) The shifted windows, weighted by w^+, are attracted to the anchor. (N) The same shifted windows, weighted by w^-, are repelled. Every shifted window therefore counts as partly positive and partly negative, according to how related it is. (H) The unshifted versions of the other windows, weighted by w^+, are also repelled. This stops the encoder from ignoring time by mapping a window and its shift to the same point. The loss is $-\log\big(P/(P+N+H)\big)$.

Together, these provide complementary estimates of the degree of similarity between each anchor and all B members of the batch. When producing these weights, we apply a stop-gradient to c_{ik}, preventing the contrastive objective described below from directly optimizing the similarity estimates used to construct its own weights. However, because explicit negatives remain, removing the stop-gradient does not strictly result in collapse.

Soft Contrastive Terms on a Graded, Temporally-Shifted Batch Having estimated similarity among the unshifted windows, we apply these relationships to attract and repel shifted windows in time. Let $z'_k = f_\theta(x'_k)$ denote the normalized representation of shifted window x'_k. For each anchor x_i, we compute its similarity to every shifted batch member as $s'_{ik} = z_i^\top z'_k/\tau$, where τ is the contrastive temperature. We then use the weights derived above to construct positive and negative terms,

$$P_i = \sum_{k=1}^{B} w^+_{ik} \exp(s'_{ik}) \qquad N_i = \sum_{k=1}^{B} w^-_{ik} \exp(s'_{ik}) \tag{2}$$

Thus, the same set of shifted candidates participates in both terms, but their contributions differ according to their estimated relationship with the anchor: shifted windows of more similar members contribute more strongly to P_i, while those of more dissimilar members contribute more strongly to N_i. In this way, TEX-MEX contrasts the temporal shifts of batch members according to their graded cross-member relationships rather than temporal distance.

Encouraging Temporal Frame Separation The positive and negative terms above contrast relationships across batch members in time, but alone do not require representations to preserve features that distinguish a window from its temporal shift: an encoder could merely learn a collapsed invariance to time. For example, user identity shortcuts may be learned. To prevent this, we additionally contrast the temporal shifts of similar batch members against their corresponding unshifted windows. We define $s_{ik} = z_i^\top z_k/\tau$ and introduce the temporal frame-separation term

$$H_i = \sum_{k \neq i} w^+_{ik} \exp(s_{ik}) \tag{3}$$

Here, unshifted windows of more similar batch members receive greater weight as negatives. We exclude $k = i$, as a sample cannot serve as a negative against itself. Consequently, for a similar member k, the shifted representation z'_k contributes to the positive term P_i, while its unshifted representation z_k contributes with the same weight w^+ to the negative term H_i, encouraging separation between temporal frames.

Combining the three terms, the TEX-MEX objective is

$$\mathcal{L}_{\text{TEX-MEX}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{P_i}{P_i + N_i + H_i} \tag{4}$$

The resulting objective attracts an anchor toward the temporal shifts of similar batch members, repels it from the temporal shifts of dissimilar members, and distinguishes those shifted representations from the original unshifted frame: temporal cross-member exchange.

4 TEX-MEX AS A GENERAL TIME-SERIES SSL OBJECTIVE

To evaluate TEX-MEX as a general time-series SSL objective, we adopt two recently proposed evaluation settings spanning five signal domains: first, following PULSE (Chen et al., 2026) (PPG, EEG, ECG, and HAR) and second, following CGM-JEPA (Muhammad et al., 2026) (for CGM).

4.1 MULTI-DOMAIN SELF-SUPERVISED LEARNING: PPG, EEG, ECG, AND HAR

Setup The core evaluation setting of PULSE (Chen et al., 2026) utilizes four datasets spanning four time-series domains. For PPG, WESAD (Schmidt et al., 2018) is used for four-class affective-state classification across baseline, stress, amusement, and meditation states. For EEG, Sleep-EDF (Kemp et al., 2000) is used for five-class sleep-stage classification. For ECG, MIT-BIH Atrial Fibrillation (Moody, 1983) is used for binary classification of atrial fibrillation versus normal rhythm. Finally, the accelerometry dataset from Reyes-Ortiz et al. (2015) is used for six-class human activity recognition. We retain preprocessing, windowing, and predefined 70/15/15 inter-subject splits.

The PULSE evaluation protocol provides a standardized setup for introducing new SSL objectives while preserving the encoder architecture and downstream evaluation. We use the same 10-layer dilated-convolution encoder with 320-dimensional representations. Each method is tuned through 30 random-search trials over shared training hyperparameters, while allowing method-specific hyperparameters to vary independently. The best hyperparameter configuration is selected by validation performance, and the lowest-validation-loss checkpoint within each training run is retained. Frozen representations are evaluated using a fixed logistic-regression linear probe, with results averaged across five model initialization seeds (0–4) on the predefined split. We reuse results for existing baselines from Chen et al. (2026), which cover representative modern approaches spanning contrastive learning, masked autoencoding, and sequential variational autoencoding.

Alongside TEX-MEX, we implement TNC (Tonekaboni et al., 2021) and SoftCLT (Lee et al., 2024) as additional baselines representing prior temporal contrastive approaches. For each, we sweep method-specific hyperparameters following the search spaces of their original implementations. Further details on hyperparameter sweeps and final choices can be found in Appendix A.1.

TEX-MEX as a Distance Function is a Predictor of Mutual Class Membership The use of signal-distance functions to estimate similarity has been applied in several prior time-series SSL works. REBAR (Xu et al., 2024) learns a reconstruction-based distance that captures shared motifs between windows, while SoftCLT (Lee et al., 2024) uses DTW in the raw signal space to estimate graded relationships between instances. We derive the former from Chen et al. (2026). Figure 3 compares these approaches with the representation-space distance learned directly by TEX-MEX, as well as raw-signal L_2 distance. For each domain, we sample 100 batches of 64 windows from the held-out set and measure the proportion of each anchor's $k \in \{1, 3, 5\}$ nearest neighbors that share its ground-truth class. All distance measures are evaluated on identical batches, and we evaluate TEX-MEX at one epoch, halfway through training, and after training.

A few observations emerge from this analysis. First, by 50% of training, the distance learned by TEX-MEX is the strongest measure across all three values of k, exceeding REBAR by 4–5 percentage points on average. Second, its continuous improvement suggests that the representation space becomes an increasingly informative source of cross-member relationships as optimization proceeds. Unlike REBAR or DTW, for which the distance measure is frozen for contrastive learning, there is no fixed ceiling in supervision quality.

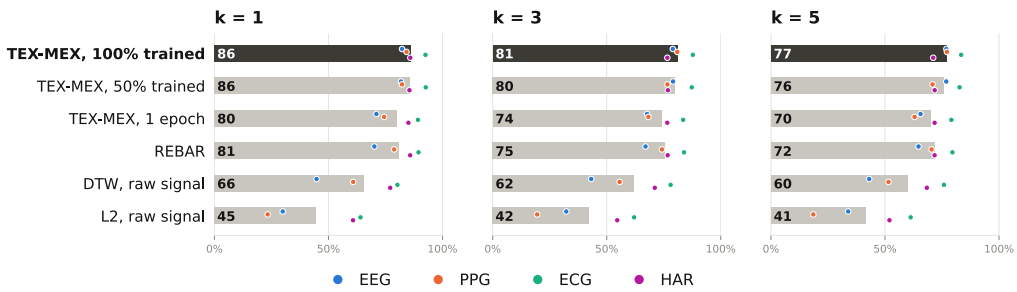

Figure 3: **Comparison of signal distance measures by mutual class membership.** On the held-out set, we sample 100 batches of 64 windows per domain and, for each anchor, retrieve its $k \in \{1, 3, 5\}$ nearest neighbors according to each distance measure, reporting the proportion sharing the anchor's ground-truth class. Bars show the unweighted mean across PPG, EEG, ECG, and HAR, with colored points indicating individual domains. TEX-MEX is evaluated at 1 epoch, 50% of training, and after training, and compared with REBAR and raw-signal DTW and L_2 distances. Higher values indicate that a distance measure more reliably identifies semantically similar signals.

Table 1: Linear probe performance across four sensor domains. **Bold** is best, underline second best. The bracketed block is reproduced from Chen et al. (2026); TNC and SoftCLT are run by us under the same protocol.

	Method	AUROC ↑				PRAUC ↑			
		PPG	ECG	HAR	EEG	PPG	ECG	HAR	EEG
Chen et al. (2026)	SimCLR (Chen et al., 2020)	.6119	.8254	.9938	.8545	.3608	.8063	.9763	.5095
	TS2Vec (Yue et al., 2022)	.6428	.8656	.9931	.9499	.3959	.8516	.9766	.7022
	REBAR (Xu et al., 2024)	.6977	.9146	.9965	.9508	.4457	.8985	.9891	.7077
	PatchTST (Nie et al., 2022)	.7108	.7096	.9744	.9355	.5291	.6816	.8990	.6737
	TimeMAE (Cheng et al., 2026)	.7808	.7661	.9914	.9509	.5674	.7662	.9705	.7322
	LFADS (Sussillo et al., 2016)	.7110	.7169	.9949	.9449	.4959	.6921	.9829	.6855
	DSVAE (Yingzhen & Mandt, 2018)	.7678	.8288	.9936	.9542	.5538	.8131	.9769	.7225
	PULSE (Chen et al., 2026)	.8029	.9493	.9942	.9617	.5989	**.9475**	.9810	.7382
	TNC (Tonekaboni et al., 2021)	.7213	.8723	.9923	.9628	.5016	.8421	.9734	.7459
	SoftCLT (Lee et al., 2024)	.7089	.8598	**.9969**	.9527	.4550	.8118	**.9893**	.7170
	TEX-MEX	**.8162**	**.9539**	.9944	**.9684**	**.6094**	**.9475**	.9811	**.7580**

Results Table 1 summarizes linear-probe performance across the four domains. For PPG, EEG, and ECG, TEX-MEX achieves the highest (or, in one case, tied-highest) performance on both AUROC and PRAUC, with PULSE generally providing the closest baseline performance. Further, TEX-MEX retains its computational efficiency: on a Quadro RTX 6000, a controlled runtime benchmark found that TEX-MEX reduced peak GPU memory utilization by 65.4% and pretraining wall-clock time by 49.6% on average relative to PULSE. On HAR, however, TEX-MEX is outperformed by SoftCLT, LFADS, and REBAR, ranking 4/11. Nevertheless, performance on this task is highly saturated: all but one baseline exceeds 99% AUROC, with TEX-MEX reaching 99.44% compared with 99.69% for SoftCLT, the best-performing method. A similar, though somewhat larger, gap is observed for PRAUC, where TEX-MEX reaches 98.11% compared with 98.93%.

The advantage of TEX-MEX is especially clear in comparison with contrastive methods that similarly exploit temporal structure. Across the benchmark, TEX-MEX is the only such approach that remains consistently competitive with the strongest SSL methods, with particularly large gains on PPG and ECG over TNC, SoftCLT, and the SimCLR implementation of Chen et al. (2026), which uses temporal shifting as an augmentation. Each of these methods prescribes similarity in advance, whether by temporal proximity (TNC, SimCLR) or by DTW and timestamp distance (SoftCLT). TEX-MEX instead learns similarity through the encoder itself, contrasting the temporal shifts of similar windows against those of dissimilar ones. These learned relationships carry most of the positive weight, with only 30.2% of w^+ (averaged across domains and epochs) falling on the anchor's own shift. Because similarity comes from representations already computed in the forward pass, TEX-MEX also avoids the highly costly dataset-wide preprocessing required by TNC (ADF

Table 2: Linear probe evaluation for CGM metabolic subphenotyping across three regimes and two endpoints. All results are mean over 20 seeds.

		AUROC ↑						PRAUC ↑					
		Home		Venous		Ven.→Home		Home		Venous		Ven.→Home	
	Method	β-cell	IR	β-cell	IR	β-cell	IR	β-cell	IR	β-cell	IR	β-cell	IR
Muhammad et al. (2026)	PCA (Shlens, 2014)	.925	.821	.790	.744	.927	.827	.909	.842	.784	.770	.911	.847
	GluFormer (Lutsker et al., 2025)	.857	.889	.742	.709	.801	.530	.809	.880	.703	.758	.756	.555
	TS2Vec (Yue et al., 2022)	.854	.848	.735	.733	.805	.713	.880	.868	.716	.741	.754	.720
	MOMENT_{S} (Goswami et al., 2024)	.849	.819	.712	.674	.736	.830	.834	.829	.668	.671	.751	.852
	MOMENT_{L} (Goswami et al., 2024)	.652	.605	.683	.630	.606	.713	.649	.651	.604	.579	.615	.750
	Mantis (Feofanov et al., 2025)	.758	.690	.674	.709	.821	.766	.777	.759	.634	.687	.826	.809
	CGM-JEPA (Muhammad et al., 2026)	.932	.842	.845	.801	.946	.866	.931	.872	.821	.824	.941	.892
	X-CGM-JEPA (Muhammad et al., 2026)	.946	.857	**.855**	.800	.949	.865	.941	.883	.815	.822	.943	.892
	TNC (Tonekaboni et al., 2021)	.914	.832	.831	**.805**	.912	.845	.909	.850	.831	.841	.882	.853
	SoftCLT (Lee et al., 2024)	.902	.811	.735	.741	.911	.831	.916	.857	.717	.774	.921	.872
	TEX-MEX	**.980**	**.952**	**.855**	.796	**.978**	**.952**	**.972**	**.946**	**.882**	**.848**	**.967**	**.943**

stationarity testing) and SoftCLT (DTW distances), making it substantially more amenable to large-scale pretraining. We exploit this advantage of TEX-MEX in Section 5.

4.2 TEX-MEX Applied to Continuous Glucose Monitoring

Setup We follow the experimental setting of CGM-JEPA (Muhammad et al., 2026), preserving its transformer-based encoder, pretraining data, downstream tasks, seeds, and evaluation protocol. Pretraining uses unlabeled continuous glucose monitoring (CGM) recordings from a Stanford study of metabolic health (Metwally et al., 2025) together with a public CGM dataset (Colás et al., 2019). Downstream evaluation uses two labeled groups from the Stanford study. One group contains glucose responses measured from venous blood draws during an in-clinic oral glucose tolerance test (OGTT), while a separate held-out group contains both venous OGTT measurements and wearable CGM collected at home. Each participant is clinically labeled for insulin resistance and β-cell dysfunction, yielding two binary prediction tasks. We evaluate frozen representations in three settings: training on venous measurements from the first group and testing on venous measurements from the held-out group, training and testing on home CGM within the held-out group, and training on venous measurements and testing on home CGM within the held-out group. The logistic-regression probe and average of 20 seeds of subject-level 2-fold cross validation are unchanged, and all baseline results (other than TNC and SoftCLT) are derived from Muhammad et al. (2026).

Incorporating TEX-MEX requires a few tweaks in this setup. First, the corpus does not support temporal shifts at CGM-JEPA's 24-hour (288-reading) window: 391 of its 413 recordings are exactly one day long, leaving only 22 long enough to hold both an anchor and a displaced partner. We therefore rewindow and pretrain on 39-reading (3.25-hour) windows, matching the 39-point curves that form every downstream input in their setup. Second, we z-score inputs using corpus statistics; this is applied to TS2Vec in their setup. We derive all TEX-MEX-specific hyperparameters without modification from defaults derived in Section 5, with an exception for $\sigma = 1$ chosen beforehand as a more plausible timescale for the domain. We also expand the baselines list by including TNC and SoftCLT. For the former, we use the TEX-MEX window size of 39, while for the latter, we preserve the window size of 288 as SoftCLT only performs within-window temporal contrast. For both, we use defaults for method-specific hyperparameters.

Results Table 2 summarizes linear-probe performance for metabolic subphenotyping. TEX-MEX achieves the best or tied-best performance on 11 of 12 metric-task combinations, spanning both β-cell dysfunction and insulin resistance across the three evaluation settings. The sole exception is venous insulin resistance AUROC, where TNC achieves the highest performance.

The largest gains emerge on at-home CGM, particularly for insulin resistance. TEX-MEX exceeds the strongest competing baseline by 6.3 AUROC points for in-domain home evaluation and by 8.6 points under venous→home transfer. The latter is the largest improvement across the evaluated settings and is particularly relevant for practical deployment, as it tests whether a probe learned from clinically collected venous OGTT curves transfers to CGM measurements obtained during an at-home glucose challenge. This setting reflects a plausible path toward scalable metabolic screening, where models can leverage high-quality clinical measurements during development while operating on substantially more accessible wearable measurements at deployment (Metwally et al., 2025).

Table 3: Ablation of TEX-MEX components on CGM metabolic subphenotyping. The top row reports absolute performance; all other rows report the change relative to it. Values are means over 20 seeds.

	AUROC ↑						PRAUC ↑					
	Home		Venous		Ven.→Home		Home		Venous		Ven.→Home	
Variant	β-cell	IR	β-cell	IR	β-cell	IR	β-cell	IR	β-cell	IR	β-cell	IR
TEX-MEX	.980	.952	.855	.796	.978	.952	.972	.946	.882	.848	.967	.943
w/o graded weighting	−.042	−.097	−.049	+.001	−.171	−.142	−.046	−.090	−.118	−.008	−.159	−.125
w/o frame separation H	−.018	−.066	+.018	+.004	−.012	−.057	−.020	−.044	−.010	−.015	−.010	−.033
w/o stop-gradient	−.037	−.079	+.019	−.010	−.025	−.069	−.034	−.052	−.009	−.019	−.025	−.047
EMA grader (m=0.999)	−.034	−.090	+.009	+.016	−.032	−.090	−.031	−.057	−.020	−.009	−.026	−.054
$\delta = 0$	−.227	−.198	−.398	−.326	−.220	−.146	−.265	−.203	−.492	−.404	−.232	−.138

Ablations of TEX-MEX We perform five ablations of TEX-MEX (Table 3). The most direct test of its core idea removes graded weighting: the anchor's own shift becomes the sole positive, while all other shifted and unshifted windows serve as uniformly weighted negatives. This effectively reduces TEX-MEX to more standard hard contrastive learning with temporal shifts as positives. Performance drops sharply, most severely under venous→home transfer (-17.1 and -14.2 AUROC on β-cell and IR). Temporal shifting alone therefore does not account for TEX-MEX's performance; the gains come from contrasting the shifts of similar windows against those of dissimilar ones. Only fixing $\delta = 0$, which removes temporal shifting entirely, is more damaging, reducing venous performance to chance level. The remaining ablations have effects concentrated on home and transfer evaluation and particularly on insulin resistance. Removing the frame-separation term H or the stop-gradient costs up to 6.6 and 7.9 AUROC points, respectively. Computing c_{ik} from an EMA copy of the encoder trails the live encoder by up to 9.0 points. Since the encoder's similarity estimates improve throughout training (Figure 3), this suggests that the EMA's lag degrades the weights by tying them to an earlier, weaker encoder. On venous evaluation, these three variants perform comparably to the full model.

5 Scaling TEX-MEX: A Bite-Sized, SOTA PPG Foundation Model

Setup We broadly derive our setup from Pulse-PPG (Saha et al., 2025), a prior state-of-the-art open PPG foundation model pretrained on field-collected wearable data. We inherit its Mobile Open Observation of Daily Stressors (MOODS) pretraining dataset (Neupane et al., 2024), a large-scale collection of raw smartwatch PPG recorded during participants' everyday lives, as well as its diverse downstream evaluation: PPG-BP (Liang et al., 2018), WESAD (Schmidt et al., 2018), PPG-DaLiA (Reiss et al., 2019), MOODS (Neupane et al., 2024), and SDB (Garde et al., 2014). The downstream tasks span wearable field, wearable lab, and clinical settings, cover both classification and regression problems such as stress, activity, heart rate, sleep disturbance, blood pressure, and hypertension, and use input windows at a variety of temporal scales. We strengthen the downstream evaluation protocol for all models. Rather than evaluating on a single subject split with one random seed, we use subject-level five-fold cross-validation repeated across 15 random seeds for every downstream task. We also expand the set of foundation-model baselines to include more recent general time-series and PPG-specific models. These include MOMENT (Goswami et al., 2024), Chronos 1 & 2 (Ansari et al., 2024; 2025), PaPaGei (both S and P variants) (Pillai et al., 2025), AnyPPG (Nie et al., 2026), and Pulse-PPG (Saha et al., 2025).

For the TEX-MEX PPG foundation model, we apply a pretraining window length of 1 minute. This contrasts with Saha et al. (2025), where longer, four-minute windows were specifically chosen to capture the multi-minute physiological dynamics associated with stress. TEX-MEX, however, explicitly learns temporal relationships across shifted signal windows during pretraining. We therefore use a shorter window intended to provide a more general representation across tasks operating at different temporal scales. Similar to Saha et al. (2025) we use a ResNet backbone, however, we substantially reduce the size of the encoder to 560K parameters, yielding a 96-dim representation. We manually explored a range of architectural and training hyperparameters for the TEX-MEX foundation model, including model depth and width, learning rate, and TEX-MEX-specific parameters. Full details of the TEX-MEX ResNet architecture and pretraining hyperparameters are provided in Appendix B.

Table 4: Downstream performance of the TEX-MEX PPG foundation model against general time-series and PPG-specific foundation models. All results are mean over 15 seeds of 5-fold cross-validation. **Bold** is best, underline second best.

			General TSFMs			*PPG FMs*				
			MOMENT	Chronos	Chronos-2	Pulse-PPG	PaPaGei-S	PaPaGei-P	AnyPPG	TEX-MEX
	Params (M)		341.25	201.37	119.48	28.50	5.79	4.99	4.04	**0.56**
	Embedding dim		1024	768	768	512	512	512	512	96
	Pretraining data		*Gen. TS*	*Gen. TS*	*Gen. TS*	*PPG*	*PPG*	*PPG*	*PPG+ECG*	*PPG*
PPG-BP	Hypertension	AUROC ↑	.6749	.6353	.6659	.6487	.6373	.6357	.6739	**.7347**
		PRAUC ↑	.3951	.3642	.4103	.3848	.3573	.3567	.4250	**.4960**
	Systolic BP	MAE ↓	17.487	16.979	16.195	16.622	15.960	15.470	18.429	**14.007**
	Diastolic BP	MAE ↓	10.610	10.203	10.103	9.649	10.096	9.148	11.164	**8.294**
	HR	MAE ↓	4.950	5.525	4.815	4.815	5.030	4.765	5.747	**4.433**
WESAD	Stress (2)	AUROC ↑	.8507	.8642	.9302	.9388	.8383	.8486	.9520	**.9630**
		PRAUC ↑	.7259	.7218	.8675	.8734	.6537	.7168	.9017	**.9387**
	Stress (4)	AUROC ↑	.6969	.7394	.7431	.7197	.7046	.6599	**.7562**	.7455
		PRAUC ↑	.4665	.5174	.5425	.5049	.4594	.4299	**.5588**	.5404
DaLiA	Activity	AUROC ↑	.7798	**.8069**	.7937	.7744	.7164	.7399	.7988	.7992
		PRAUC ↑	.3250	**.3773**	.3461	.3126	.2478	.2821	.3641	.3543
	HR	MAE ↓	8.460	8.830	9.628	10.105	14.099	12.232	**6.643**	9.790
MOODS	Activity	AUROC ↑	.7387	.8325	.8574	.8635	.6917	.7140	.8657	**.8835**
		PRAUC ↑	.3645	.5001	.5415	.5465	.2982	.3168	.5660	**.6058**
	Stress	AUROC ↑	.5163	.5384	.5360	.5438	.5316	.5526	.5556	**.5591**
		PRAUC ↑	.4193	.4378	.4292	.4363	.4282	.4391	.4464	**.4500**
SDB	Sleep	AUROC ↑	.5039	.5364	.5425	.4967	.5257	**.5507**	.4929	.4928
		PRAUC ↑	.3823	.3931	.4017	.3632	.3892	**.4118**	.3727	.3544

Results Table 4 summarizes linear-probe performance across 11 downstream classification and regression tasks, spanning five datasets and three settings: wearable field, wearable lab, and clinical. Overall, TEX-MEX is the strongest-performing approach, achieving the best performance on 7 of 11 tasks and outperforming all unimodally pretrained PPG foundation models on 10 of 11. The gains are particularly pronounced for hypertension prediction, where TEX-MEX exceeds the next-best models by roughly 6–7 percentage points in both AUROC and PRAUC.

While TEX-MEX is the best-performing model overall, it is also by far the smallest in our evaluation: the next-smallest, AnyPPG, is over 7× larger and additionally requires paired PPG and ECG recordings during pretraining. The comparison with Pulse-PPG, the strongest unimodally pretrained baseline (outperforming both PaPaGei variants on 6 of 11 tasks), is especially stark. It is roughly 50× larger and uses the same pretraining dataset and evaluation splits, yet TEX-MEX outperforms it on 10 of 11 tasks; on the sole exception, sleep, the gap is under one percentage point on both AUROC and PRAUC. Further, attempts at downscaling by Saha et al. (2025) show that reducing Pulse-PPG even to 4.73M parameters causes AUROC losses up to 0.139 in downstream tasks.

On the four tasks where TEX-MEX is not the top performer, two involve marginal differences from the best result in the table, with gaps to the best competing model averaging 1.5 percentage points. The other two cases (DaLiA heart-rate estimation and sleep classification on SDB) are more substantial. For the former, TEX-MEX nevertheless remains the best-performing unimodally pretrained PPG model, with only general time-series foundation models and multimodally pretrained AnyPPG performing better. On sleep, however, TEX-MEX underperforms by roughly six percentage points relative to the best performer, PaPaGei-P. We suspect this reflects a limitation of the MOODS pretraining dataset, which contains no sleep recordings. Consistent with this explanation, Pulse-PPG, a generally strong model trained on the same corpus, also performs poorly on SDB. PaPaGei, by contrast, is pretrained with sleep data and performs best in this setting though even it is hardly better than random. This limitation could be addressed in future work by expanding MOODS to include sleep recordings.

6 Conclusion

We introduce TEX-MEX, an efficient, scalable self-supervised objective for time series that unifies structural and temporal similarity by contrasting the temporal shifts of windows the encoder deems similar against those it deems dissimilar. Across PPG, ECG, EEG, accelerometry, and CGM, TEX-MEX broadly achieves state-of-the-art performance. We further demonstrate it at scale with a 560K-parameter PPG foundation model that achieves state-of-the-art downstream performance while having significantly fewer parameters than the PPG foundation models in our evaluation.

AI USE STATEMENT

In this work, we used generative AI tools to help write code, polish prose and grammar, and fix typos in the paper. We did not use them to develop ideas, to design the method or experiments, or to interpret results. The authors reviewed all AI-assisted work and take full responsibility for the contents of the paper.

ETHICS STATEMENT

Our paper leverages physiological signals to solve healthcare problems. Models trained on such signals raise concerns not only on data privacy, but also unequal performance across populations that are underrepresented in the training data. Our work uses de-identified data that are collected under IRB-approved protocols, and no subject identification information is included throughout the process. We believe that our work follows the ICLR Code of Ethics, and the ethical risks associated with the use of physiological data have been appropriately considered and mitigated. At the same time, our work may support more personalized and accessible health assessment, with the potential to improve health monitoring and inform future care.

REPRODUCIBILITY STATEMENT

All experiments in Section 4 follow publicly available benchmark settings, PULSE (Chen et al., 2026) and CGM-JEPA (Muhammad et al., 2026), whose datasets, preprocessing, splits, encoder architectures, and evaluation protocols we reuse, with any modifications described in Section 4.2. Hyperparameter search spaces and selected configurations for TEX-MEX and our reimplemented baselines (TNC (Tonekaboni et al., 2021) and SoftCLT (Lee et al., 2024)) are reported in Appendix A.1. The PPG foundation model in Section 5 builds on the Pulse-PPG setup (Saha et al., 2025), and full architecture and pretraining details are given in Appendix B. With the exception of MOODS, which serves as both the pretraining corpus and the source of two downstream tasks (activity and stress; Table 4), all downstream evaluation datasets are publicly available. MOODS cannot be released due to participant privacy restrictions. To support reproducibility and practical use, we will release upon acceptance the codebase developed for our PPG foundation model, including pretraining, preprocessing, and evaluation, along with the pretrained weights. The codebase implements TEX-MEX in a general-purpose form that can be readily applied to other encoders and time-series domains.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 3158–3181, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0d99a8c048befb6dd6e17d7684adacac-Paper-Conference.pdf.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Yenho Chen, Maxwell A Xu, James Matthew Rehg, and Christopher John Rozell. Self-supervised dynamical system representations for physiological time-series. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=cKsANUDt1m.

Zhaoliang Chen, Cheng Ding, Saurabh Kataria, Runze Yan, Minxiao Wang, Randall Lee, and Xiao Hu. Gpt-ppg: a gpt-based foundation model for photoplethysmography signals. *Physiological Measurement*, 46(5):055004, 2025.

Mingyue Cheng, Xiaoyu Tao, Zhiding Liu, Qi Liu, Hao Zhang, Rujiao Zhang, and Enhong Chen. Timemae: Self-supervised representations of time series with decoupled masked autoencoders. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pp. 498–508, 2026.

Ana Colás, Luis Vigil, Borja Vargas, David Cuesta-Frau, and Manuel Varela. Detrended fluctuation analysis in the prediction of type 2 diabetes mellitus in patients at risk: Model optimization and comparison with other metrics. *PloS one*, 14(12):e0225817, 2019.

Cheng Ding, Zhicheng Guo, Zhaoliang Chen, Randall J Lee, Cynthia Rudin, and Xiao Hu. Siamquality: a convnet-based foundation model for photoplethysmography signals. *Physiological Measurement*, 45(8):085004, aug 2024. doi: 10.1088/1361-6579/ad6747. URL https://doi.org/10.1088/1361-6579/ad6747.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9568–9577. IEEE, 2021.

Vasilii Feofanov, Marius Alonso, Songkang Wen, Romain Ilbert, Hongbo Guo, Malik Tiomoko, Lujia Pan, Jianfeng Zhang, and Ievgen Redko. Mantis: Lightweight calibrated foundation model for user-friendly time series classification. In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

Ainara Garde, Parastoo Dehkordi, Walter Karlen, David Wensley, J Mark Ansermino, and Guy A Dumont. Development of a screening tool for sleep disordered breathing in children using the phone oximeter™. *PloS one*, 9(11):e112959, 2014.

Ary L Goldberger, Luis AN Amaral, Jeffrey M Hausdorff, Plamen Ch Ivanov, C-K Peng, and H Eugene Stanley. Fractal dynamics in physiology: alterations with disease and aging. *Proceedings of the national academy of sciences*, 99(suppl_1):2466–2472, 2002.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Soft contrastive learning for time series. In *International Conference on Learning Representations*, volume 2024, pp. 46815–46839, 2024.

Yongbo Liang, Zhencheng Chen, Guiyong Liu, and Mohamed Elgendi. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in china. *Scientific data*, 5(1):180020, 2018.

Guy Lutsker, Gal Sapir, Smadar Shilo, Jordi Merino, Anastasia Godneva, Jerry R Greenfield, Dorit Samocha-Bonet, Raja Dhir, Francisco Gude, Shie Mannor, et al. Gluformer: Learning generalizable representations from continuous glucose monitoring data. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*, 2025.

Ahmed A Metwally, Dalia Perelman, Heyjun Park, Yue Wu, Alokkumar Jha, Seth Sharp, Alessandra Celli, Ekrem Ayhan, Fahim Abbasi, Anna L Gloyn, et al. Prediction of metabolic subphenotypes of type 2 diabetes via continuous glucose monitoring and machine learning. *Nature biomedical engineering*, 9(8):1222–1239, 2025.

George Moody. A new method for detecting atrial fibrillation using rr intervals. *Proc. Comput. Cardiol.*, 10:227–230, 1983.

Hada Melino Muhammad, Zechen Li, Flora Salim, and Ahmed A Metwally. Cgm-jepa: Learning consistent continuous glucose monitor representations via predictive self-supervised pretraining. *arXiv preprint arXiv:2605.00933*, 2026.

Sameer Neupane, Mithun Saha, Nasir Ali, Timothy Hnat, Shahin Alan Samiei, Anandatirtha Nandugudi, David M Almeida, and Santosh Kumar. Momentary stressor logging and reflective visualizations: Implications for stress management with wearables. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2024.

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

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations*, volume 2025, pp. 48230–48261, 2025.

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

Jorge L Reyes-Ortiz, D Anguita, L Oneto, and X Parra. Smartphone-based recognition of human activities and postural transitions data set. *UCI Machine Learning Repository. School Inf. Comput. Sci. Univ. California at Irvine, Irvine, CA, USA, available online: http://archive. ics. uci. edu/ml/datasets/Smartphone-Based+ Recognition+ of+ Human+ Activities+ and+ Postural+ Transitions*, 2015.

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

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 400–408, 2018.

Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

David Sussillo, Rafal Jozefowicz, LF Abbott, and Chethan Pandarinath. Lfads-latent factor analysis via dynamical systems. *arXiv preprint arXiv:1608.06315*, 2016.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.

Maxwell Xu, Alexander Moreno, Hui Wei, Benjamin M Marlin, and James Rehg. Rebar: Retrieval-based reconstruction for time-series contrastive learning. In *International Conference on Learning Representations*, volume 2024, pp. 15440–15464, 2024.

Maxwell Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik Raghuram, James Rehg, and Shirley Ren. Relcon: Relative contrastive learning for a motion foundation model for wearable data. In *International Conference on Learning Representations*, volume 2025, pp. 53188–53206, 2025.

Li Yingzhen and Stephan Mandt. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, pp. 5670–5679. PMLR, 2018.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8980–8987, 2022.

Hao Zhou, Simon A. Lee, Cyrus Tanade, Keum San Chun, Juhyeon Lee, Migyeong Gwak, Megha Thukral, Justin Sung, Eugene Hwang, Mehrab Bin Morshed, Li Zhu, Viswam Nathan, Md Mahbubur Rahman, Subramaniam Venkatraman, and Sharanya Arcot Desai. Physiology-aware masked cross-modal reconstruction for biosignal representation learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=NxbSbkLNMc.

A Multi-Domain Benchmark Details

A.1 Hyperparameter Sweeps

All methods share the training-hyperparameter search described in Section 4.1 (30 random-search trials), with method-specific hyperparameters varying independently within the same trials. The best configuration is selected by validation performance.

TEX-MEX. We sweep three method-specific hyperparameters: the offset scale $\sigma \in \{0.5, 1, 2, 4, 8, 16\}$ (in window lengths), the contrastive temperature $\tau \in \{0.03, 0.04, 0.05, 0.07, 0.10, 0.15\}$, and the weighting temperature $\tau_w \in \{0.04, 0.06, 0.10, 0.14, 0.20\}$. The offset is drawn as $\delta \sim \mathcal{N}(0, \sigma^2)$ and resampled until $\delta_{\min} \leq |\delta| \leq \delta_{\max}$, with $\delta_{\min} = 1$ and $\delta_{\max} = 2.5\sigma$.

SoftCLT. Following the classification search space of Lee et al. (2024), we sweep the instance-wise sharpness $\tau_I \in \{1, 2, 3, 4, 5, 10, 20\}$ and the temporal sharpness $\tau_T \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$, keeping the loss weight fixed at $\lambda = 0.5$ as in the original work.

TNC. We sweep the positive-unlabeled weight $w \in \{0, 0.05, 0.1\}$, where $w = 0$ disables the PU correction, the Monte Carlo sample size $\in \{10, 20, 40\}$, and the augmentation factor $\in \{5, 7\}$. These ranges span the per-dataset settings used in the official implementation of Tonekaboni et al. (2021).

Table 5: Final method-specific hyperparameters selected on the validation set for each domain.

Method	Hyperparameter	PPG	ECG	HAR	EEG
TEX-MEX	σ	1	8	8	2
	τ	0.03	0.04	0.03	0.04
	τ_w	0.06	0.14	0.10	0.10
SoftCLT	τ_I	10	4	20	5
	τ_T	0.5	1.0	2.0	2.5
TNC	w	0	0	0.1	0.05
	MC sample size	20	10	10	20
	Augmentation factor	5	5	5	7

Table 6: Selected shared training hyperparameters. Following Chen et al. (2026), all three methods draw from the same grid (epochs $\in \{50, 100, 200\}$, learning rate $\in \{10^{-3}, 5\times10^{-4}, 10^{-4}\}$, weight decay $\in \{10^{-3}, 10^{-4}, 10^{-5}\}$, batch size $\in \{64, 128\}$). A batch size of 64 was selected for all methods and domains, except TEX-MEX on HAR, which selected 128.

Method	Hyperparameter	PPG	ECG	HAR	EEG
TEX-MEX	Epochs	100	200	50	100
	Learning rate	10^{-4}	10^{-4}	10^{-4}	10^{-4}
	Weight decay	10^{-5}	10^{-5}	10^{-5}	10^{-4}
SoftCLT	Epochs	200	100	200	200
	Learning rate	5×10^{-4}	10^{-4}	5×10^{-4}	10^{-3}
	Weight decay	10^{-4}	10^{-4}	10^{-4}	10^{-4}
TNC	Epochs	100	50	200	100
	Learning rate	10^{-3}	10^{-3}	10^{-3}	5×10^{-4}
	Weight decay	10^{-5}	10^{-3}	10^{-3}	10^{-5}

Table 7: PPG foundation model architecture and pretraining hyperparameters.

Encoder	
Architecture	Stem conv + 6 residual blocks
Channels	$48 \rightarrow 96$
Kernel size	11
Downsampling	Stride 2 every second block ($8\times$ total)
Dropout	0.5
Pooling	Global average
Embedding dimension	96
Parameters	560K
Objective	
Window	60 s (3000 samples at 50 Hz)
σ	4
$[\delta_{\min}, \delta_{\max}]$	$[1, 10]$ windows
τ	0.03
τ_w	0.10
Optimization	
Optimizer	Adam, learning rate 10^{-4}, no weight decay
Batch size	32
Training	4 epochs $\times$ 40,000 steps
Checkpoint	Lowest validation loss (epoch 4)

B PPG Foundation Model Details

Following Saha et al. (2025), the encoder is a 1D ResNet; we reduce it to 560K parameters. Table 7 lists the final configuration.