WHAT TO KEEP, WHAT TO REPLACE: CROSS-MODAL DEPENDENCY-AWARE AUGMENTATION FOR WEARABLE CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning from multimodal wearable data requires augmented views that introduce sufficient variation while preserving both intra-modal semantics and cross-modal relationships. Cross-modal generation offers a way to exploit these relationships, but faithful reconstruction may provide little variation, while generation from other sensors alone can discard information specific to the target-modality signal. We introduce ReWeave, a plug-and-play augmentation framework that learns which target-modal-specific information to retain and which to replace using synchronized companion modalities. A learned gate preserves selected components in a frozen, variance–covariance-regularized latent space, while temporal cross-attention synthesizes the residual from helper signals. A sparsity penalty limits direct target retention, and the self-supervised objective encourages agreement between original and augmented views. To encourage preserving the sample-wise cross-modal relationships, we compare matched attention with a control that substitutes helper keys from another instance while retaining the original helper values. Decoding the edited embeddings produces signal-space augmentations compatible with different multimodal encoder architectures and self-supervised objectives. We pretrain on 3,565 hours of wearable recordings and evaluate frozen representations on four disjoint downstream datasets spanning emotion recognition, sleep staging, activity recognition, and oxygen-consumption estimation. Across two fusion architectures and two self-supervised objectives, ReWeave generally improves performance over alternative positive-view construction baselines, with gains of up to 9.7 macro-F1 points and an 11.4% reduction in oxygen-consumption MAE over the strongest baselines in the corresponding settings.

1 INTRODUCTION

Wearable devices are becoming increasingly ubiquitous, and they collect continuous and synchronized measurements from multiple sensors, including accelerometers (ACC), photoplethysmography (PPG), electrodermal activity (EDA), temperature, and electrocardiography. Together, these modalities provide complementary observations of human behavior and physiology, supporting applications such as emotion recognition, sleep staging, activity recognition, and cardiorespiratory monitoring. Obtaining labels for such data, however, is costly: annotation may require controlled protocols, specialized equipment, expert interpretation, or substantial participant effort. Consequently, self-supervised learning (SSL) has garnered significant attention as it offers a promising way to learn generalizable representations from abundant unlabeled wearable data (Narayanswamy et al., 2025; Erturk et al., 2025; Pillai et al., 2025; Xu et al., 2025; Zhang et al., 2025).

As a powerful SSL paradigm widely proven in computer vision and natural language processing, contrastive learning (CL) (Chen et al., 2020) learns by pulling positive views of the same instance together while separating negative instances in an embedding space. The effectiveness of contrastive learning depends critically on augmentation, which determines how positive views are constructed (Zheng et al., 2024). In the context of wearable health, a good positive view (Tian et al., 2020) should therefore differ sufficiently from the original input to prevent trivial matching, while remaining compatible with the same underlying physiological or behavioral state.

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.

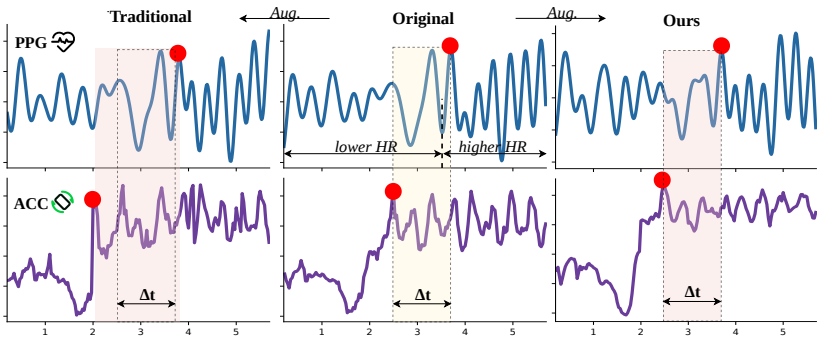


Figure 1: **Effect of augmentation on cross-modal temporal dependencies.** In a real-world sample (center), Δ t denotes the latency between activity onset in the accelerometer (ACC) signal and the subsequent heart-rate increase inferred from photoplethysmography (PPG). Applied to the same sample, traditional augmentation (left) shifts this latency, whereas our method (right) yields a closer one to that of the original while introducing obvious variation.

Existing time-series CL methods construct positive views either by retrieving related sequences (Franceschi et al., 2019; Tonekaboni et al., 2021; Xu et al., 2024) or by applying handcrafted or learned transformations such as cropping, jittering, masking, warping, and parametric perturbations (Zheng et al., 2024; Demirel & Holz, 2023). These methods provide useful variation, but most are designed around single-modality signals and do not explicitly model the relationships among synchronized modalities. Even applying the same crop or mask across modalities preserves alignment only mechanically, which cannot be adapted to different possible temporal dependencies (e.g., delays) across modalities. For example, as shown in Figure 1, an ACC-detected exercise onset may be followed by a characteristic PPG-derived heart-rate response, and the delay between them reflects physical activity, training status, and exercise performance (Bellenger et al., 2016; Nelson et al., 2014). Similarly, coordinated relationships among EDA, temperature, and heart rate change around epileptic seizures (Vieluf et al., 2021). Generic transformations can therefore produce signals that remain plausible on a single-modality basis but become jointly inconsistent in their event ordering, response latency, or coordinated variation.

Cross-modal reconstruction provides a useful starting point for constructing more coherent views. By predicting one signal from another, a model can exploit dependencies between synchronized streams; for example, xMAE (Zhou et al., 2026) uses masked cross-modal reconstruction to encourage representations that capture physiological timing structure. However, reconstruction and augmentation have different objectives. A highly accurate reconstruction may remain too similar to the original sample to provide useful view diversity, while synthesizing a view entirely from helper modalities may discard information unique to the target modality (Liu et al., 2023). Allowing the generator to retain part of the target can address the latter problem, but excessive retention makes direct copying an easy solution. These considerations expose a central design question **for a good positive view: how much of the target signal should be preserved, and which parts can be replaced using information from other modalities?**

We propose ReWeave, a plug-and-play data augmentation framework that addresses this question through selective cross-modal replacement. ReWeave consists of two core components: a VC-MAE that learns a VC-regularized latent space with reduced cross-dimensional redundancy to facilitate subsequent augmentation. Based on this space, a cross-modal replacement module explores the boundaries of irreplaceable information within each modality and learns to replace the remaining information using other modalities. The edited latent representation is then decoded back to the signal domain, providing augmented positive samples for contrastive learning. During training, a curriculum progressively reduces the retained target information to generate increasingly challenging positive views, while the contrastive objective encourages semantic consistency. A matched-versus-shuffled helper objective further discourages reliance on generic modality priors and promotes sample-specific cross-modal temporal dependency learning.

We evaluate ReWeave across diverse health-related classification and regression tasks, including emotion detection, activity recognition, sleep staging, and VO_2 prediction, and it outperforms existing augmentations in most comparisons. Additionally, ReWeave also consistently outperforms the baseline in positive-only SSL settings. Temporal-lag analyses and component ablations further

examine the behavior of the generated views and the roles of selective retention and cross-modal synthesis. Our key contributions are as follows:

- **Selective Cross-modal Augmentation.** We introduce a framework that learns which latent dimensions to retain from each wearable modality and which to synthesize from synchronized companions, with a retention penalty controlling direct reuse of the target signal.
- **Joint Optimization of Positive Sample Construction and Contrastive Learning.** We increase the difficulty of positive samples by progressively suppressing the upper bound of their mutual information with the original sample. Meanwhile, the contrastive objective encourages semantic alignment between the original and augmented views. Together, these objectives operationalize the InfoMin trade-off.
- **Temporal Dependency Learning via Shuffled Baseline.** To prevent the model from shortcutting and ignoring the cross-modal replacement components, we formulate a shuffled sample and a positive-vs-shuffled gap loss. By encouraging the interaction gain of a matched positive view that exceeds that of a counterfactual shuffled baseline, the gap objective promotes learning sample-wise cross-modal temporal dependencies.

2 METHOD

2.1 PROBLEM FORMULATION AND METHOD OVERVIEW

Multi-modal Wearable Data and Temporal Dependency. Let $\mathcal{M}$ denote the set of M synchronized wearable modalities. For sample i, let $\mathbf{x}_i^m \in \mathbb{R}^{T \times C_m}$ denote the complete length-T time window of modality $m \in \mathcal{M}$, and define the corresponding multimodal sample as $\mathbf{X}_i = \left(\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^M\right)$. These synchronized modality windows are drawn from a joint distribution that varies across subjects and physiological or behavioral regimes. This joint distribution captures the temporal relationships among modalities, including cross-modal delays, coordinated responses, and higher-order interactions over time. For example, when exercise begins, the movement change captured by the accelerometer can serve as an early marker of exercise onset, preceding the slower adjustment of PPG-derived heart rate, whose response kinetics have been associated with physical activity, training status, and exercise performance.

Good View via Cross-Modal Reconstruction. Following the InfoMin (Tian et al., 2020) principle, a good view for CL should minimize redundant input information while retaining the essential task-relevant semantics of the original sample. Mathematically, the classical InfoMin principle can be written as: $\max I(V^+; Y)$, s.t. $I(V^+; X) \leq \epsilon$, where X is the original sample and Y denotes its semantic embedding. In the context of cross-modal reconstruction that explicitly models inter-modal temporal dependencies, we formulate this InfoMin objective by conceptually partitioning the modality semantics into two functional components: an irreplaceable source-preserved component p_i (e.g., modal-private features (Liu et al., 2023)), and a residual component n_i. An optimal augmented view is constructed by retaining the minimal p_i and replacing n_i with $\hat{n}$ synthesized from a companion helper modality, yielding the positive view $V_i^+ = [p_i, \hat{n}]$. However, in our cross-modal generation setting, satisfying only these two conditions leaves a critical problem: a model could theoretically satisfy both sufficiency and minimality by retaining an overly informative p_i of the target (original) modality while ignoring the cross-modal helper, allowing $\hat{n}$ to degenerate into meaningless noise or generic modality priors. To prevent this trivial shortcut and force the capture of fine-grained cross-modal dependencies, we introduce a lower bound on the conditional information gain. The complete objective for generating a good cross-modal view is thus formulated as:

$$\max I(V^+; Y), \text{s.t. } I(V^+; X) \leq \epsilon,\ I(Y; \hat{n} \mid p) \geq \delta. \tag{1}$$

This third constraint explicitly dictates that the synthesized component $\hat{n}$ must contribute a meaningful margin of cross-modal information arising from temporal interactions.

Pipeline Overview. As shown in Figure. 2 ReWeave first encodes the synchronized modalities using pretrained and frozen autoencoders. For each target modality, a learned gate retains selected latent dimensions as queries, while cross-modal attention, translation, and helper fusion synthesize the replacement component from the other modalities. The retained and synthesized components are combined and decoded to construct the multimodal positive view, which is jointly optimized with the SSL encoder.

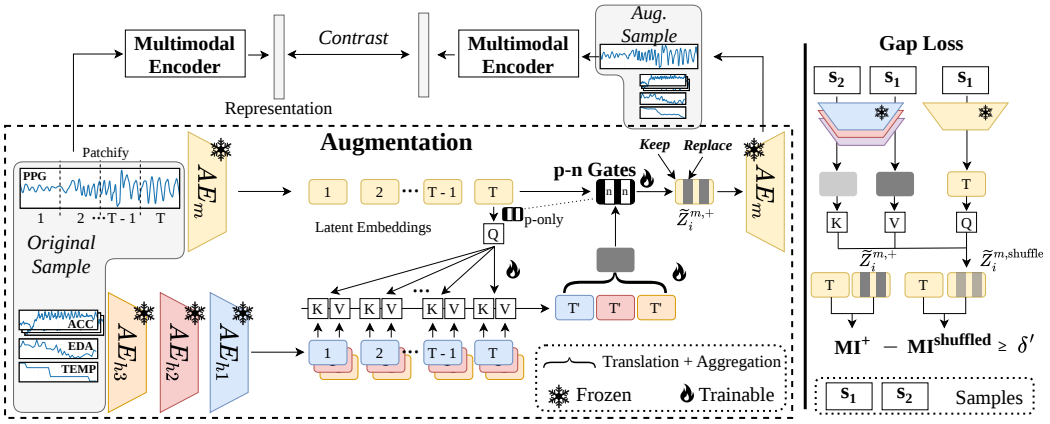


Figure 2: **Overview of ReWeave.** Frozen AEs encode all modalities into temporal tokens. For each target modality m, cross-modal attention over time aggregates the translated outputs of all helpers ($h \neq m$). A learned p-n gate replaces the n dimensions within the aggregate, which the frozen decoder maps back to raw space. Repeating this process for all modalities produces an augmented positive view for jointly training the augmentation module and multimodal encoder. The gap loss constructs a shuffled view using $\mathbf{Q}$ from m in sample s_1, $\mathbf{V}$ from h in s_1, and $\mathbf{K}$ from h in a random sample s_2. This view retains population-level modal-pair relationships , providing a lower-bound baseline that encourages fine-grained temporal dependency preservation in augmented positive samples.

2.2 Stage I: Pretrained VC-Regularized Autoencoder

To support dimension-wise partitioning and editing, we pretrain a variance-covariance-regularized autoencoder that learns a latent space with reduced cross-dimensional redundancy. For each modality (m), we independently pretrain an autoencoder consisting of an encoder (E^m_{AE}) and a decoder (D^m_{AE}): $Z^m_i = E^m_{\mathrm{AE}}(x^m_i), \quad \hat{x}^m_i = D^m_{\mathrm{AE}}(Z^m_i)$, where x^m_i is the raw signal and $Z^m_i \in \mathbb{R}^{T\times d_m}$ is its latent-token matrix, with T denoting the number of temporal tokens/patches. The reconstruction objective for modality m is $\mathcal{L}^m_{\mathrm{rec}} = \frac{1}{N}\sum_{i=1}^{N} \left| x^m_i - D^m_{\mathrm{AE}}\big(E^m_{\mathrm{AE}}(x^m_i)\big)\right|_2^2$, which encourages the latent representation to retain the information needed to recover the input. Reconstruction alone may concentrate information in a small subset of latent dimensions or produce highly correlated dimensions, leading the subsequent separation to exploit redundant shortcuts. We therefore apply variance–covariance regularization (Bardes et al., 2022) separately to the latent space of each modality. Let $\mathbf{Z}^m \in \mathbb{R}^{B\times d}$ denote a batch of d-dimensional latent vectors for modality m, where the batch and temporal-token axes are merged when necessary. The variance and covariance regularization are

$$\mathcal{L}^m_{\mathrm{var}} = \frac{1}{d}\sum_{j=1}^{d}\left[\gamma - \sqrt{\mathrm{Var}\left(\mathbf{Z}^m_{:,j}\right)+\epsilon}\right]_+, \quad \mathcal{L}^m_{\mathrm{cov}} = \frac{1}{d}\sum_{j\neq k}\left(\mathbf{C}^m_{jk}\right)^2, \tag{2}$$

where $[a]_+ = \max(0,a)$, and $\mathbf{C}^m = \frac{1}{B-1}\left(\overline{\mathbf{Z}}^m\right)^\top \overline{\mathbf{Z}}^m$, is the covariance matrix of the centered latent representations. The variance term discourages inactive dimensions, whereas the covariance term penalizes linear redundancy between dimensions. To improve the robustness of the latent space to dimension-wise recombination, we generate two independently masked views $\widetilde{x}^{m,(1)}_i$ and $\widetilde{x}^{m,(2)}_i$ of the same sample, encode them as $Z^{m,(1)}_i = E^m_{\mathrm{AE}}\left(\widetilde{x}^{m,(1)}_i\right)$ and $Z^{m,(2)}_i = E^m_{\mathrm{AE}}\left(\widetilde{x}^{m,(2)}_i\right)$, and construct a mixed representation $Z^{m,\mathrm{mix}}_i = b^m_i \odot Z^{m,(1)}_i + (\mathbf{1}-b^m_i)\odot Z^{m,(2)}_i$, where $b^m_i \in \{0,1\}^{d_m}$ is a randomly sampled binary mask. The mixed representation is decoded and compared with the original unmasked sample, yielding $\mathcal{L}^m_{\mathrm{robust}} = \frac{1}{N}\sum_{i=1}^{N}\left\|x^m_i - D^m_{\mathrm{AE}}\left(Z^{m,\mathrm{mix}}_i\right)\right\|_2^2$, which encourages the latent space and decoder to remain stable under such dimension-wise recombinations. The autoencoder objective for modality m is

$$\mathcal{L}^m_{\mathrm{AE}} = \mathcal{L}^m_{\mathrm{rec}} + \lambda_{\mathrm{v}}\mathcal{L}^m_{\mathrm{var}} + \lambda_{\mathrm{c}}\mathcal{L}^m_{\mathrm{cov}} + \lambda_{\mathrm{r}}\mathcal{L}^m_{\mathrm{robust}}. \tag{3}$$

The overall objective is obtained by summing the objectives of all modalities: $\mathcal{L}_{\mathrm{AE}} = \sum_{m=1}^{M}\mathcal{L}^m_{\mathrm{AE}}$.

After pretraining, all modality-wise encoders and decoders are frozen to provide fixed latent spaces. Given an edited latent-token matrix $\widetilde{Z}^m_i \in \mathbb{R}^{L\times d_m}$, the corresponding frozen decoder maps it back

to the raw-signal domain as $\widetilde{x}_i^m = D_{\mathrm{AE}}^m\left(\widetilde{Z}_i^m\right)$ to build the augmented sample. It is crucial to explicitly distinguish this AE from the main target encoder. The main encoder is the primary subject of our contrastive learning optimization, whereas the AE serves as a frozen, plug-and-play module placed ahead of it.

2.3 Stage II: Cross-Modal Augmentation Module

Dimension-wise Gating. For each modality, we maintain one trainable logit per latent dimension, $\lambda^m \in \mathbb{R}^{d_m}$, and compute the corresponding gate probabilities as $\pi_j^m = \sigma(\lambda_j^m)$. In the forward pass, the probabilities are converted into binary gates using a fixed threshold: $g_{j,\mathrm{hard}}^m = \mathbb{I}[\pi_j^m > 0.5]$. We use the straight-through estimator $\widetilde{g}_j^m = \mathrm{sg}(g_{j,\mathrm{hard}}^m - \pi_j^m) + \pi_j^m$ so that the forward pass remains binary while gradients are propagated through π_j^m. The normalized L_0 cardinality of the gates $r_m = \|g_{\mathrm{hard}}^m\|_0/d_m$ represents the fraction of target-modal dimensions retained by modality m, and is penalized when it exceeds the target retention level ρ. The gate is shared across patches and samples and partitions the target latent representation into a source-preserved component and a component designated for cross-modal synthesis: $P_i^m = Z_i^m \,\mathrm{Diag}(\widetilde{g}^m), N_i^m = Z_i^m \,\mathrm{Diag}(1-\widetilde{g}^m)$.

Route-wise Cross-Modal Synthesis. For a directed route $h \to m$, where m is the target modality and $h \neq m$ is a helper modality, the target-preserved component provides the query while the helper latent provides keys and values: $Q_i^{m\leftarrow h} = P_i^m W_Q^{m\leftarrow h}, \quad K_i^{m\leftarrow h} = Z_i^h W_K^{m\leftarrow h}, \quad V_i^h = Z_i^h W_V^h$. The attention output $Z_i^{m\leftarrow h}$ remains in the helper coordinate system. Although all modality-specific latent spaces have the same dimensionality, independently pretrained autoencoders do not align their latent coordinates. A route-specific translator therefore maps $Z_i^{m\leftarrow h}$ to the target latent space: $U_i^{m\leftarrow h} = T^{m\leftarrow h}(Z_i^{m\leftarrow h}) \in \mathbb{R}^{T\times d_m}$. Each directed route also has a scorer intended to reflect the helper's quality and its importance to the target modality: Given the helper latent patch $z_{i,\ell}^h$, the scorer produces a logit that is converted into a non-negative fusion weight, denoted as $w_i^{m\leftarrow h}$.

Helper Fusion and Decoding. Let $\mathcal{H}_m = \mathcal{M} \setminus \{m\}$ denote the helper modalities used for target m. Their translated outputs are fused dimension-wise and then masked by the target's p-n gate:

$$\widehat{N}_i^m = \sum_{h\in\mathcal{H}_m} w_i^{m\leftarrow h} \odot U_i^{m\leftarrow h}, \quad \widetilde{N}_i^m = (\mathbf{1} - g^m) \odot \widehat{N}_i^m, \quad \widetilde{Z}_i^{m,+} = P_i^m + \widetilde{N}_i^m. \tag{4}$$

This operation preserves the original p dimensions and replaces only the n dimensions while keeping the original latent channel order. The frozen target decoder then produces the augmented target modality: $\widetilde{x}_i^{m,+} = D_{\mathrm{AE}}^m(\widetilde{Z}_i^{m,+}), \quad \widetilde{X}_i^+ = \left(\widetilde{x}_i^{1,+}, \ldots, \widetilde{x}_i^{m,+}, \ldots, \widetilde{x}_i^{M,+}\right)$.

2.4 Training Objectives and Information-Theoretic Interpretation

Optimization Target. We next derive practical objectives corresponding to the three conditions in Eq. 1. Firstly, we jointly optimize the augmentation module and the final encoder, using the contrastive learning loss to optimize the first term.

For the second term, we apply the chain rule to the augmented view $V = [p^m, \hat{n}^{m\leftarrow h}]$, therefore

$$I(V; X^m) = I(p^m; X^m) + I(\hat{n}^{m\leftarrow h}; X^m | p^m). \tag{5}$$

Because p^m is selected from a VC-regularized latent space, we use its active dimensionality d_p as a tractable proxy for the amount of target information retained. Under the simplifying assumption that active dimensions carry comparable and weakly redundant information, $I(p^m; X^m)$ is expected to increase approximately with d_p, yielding $I(p^m; X^m) \approx c \cdot d_p$. Meanwhile, conditioned on p^m, the synthesized component $\hat{n}^{m\leftarrow h}$ is obtained through a transformation of the helper modality X^h. The conditional data-processing inequality therefore gives $I(\hat{n}^{m\leftarrow h}; X^m \mid p^m) \leq I(X^h; X^m \mid p^m)$. Since p^m is a deterministic representation of X^m, the chain rule gives $I(X^h; X^m \mid p^m) = I(X^h; X^m) - I(X^h; p^m) \leq I(X^h; X^m)$. Let $C_{h,m}$ denote the upper bound on the shared information between modalities h and m: $I(X^h; X^m) \leq C_{h,m}$. Combining these results:

$$I(V; X^m) \leq I(p^m; X^m) + C_{h,m} \approx c \cdot d_p + C_{h,m}. \tag{6}$$

Consequently, minimizing this term motivates reducing this tractable upper-bound surrogate by limiting d_p. In practice, for each modality m, we impose the normalized L_0 constraint $\|g^m\|_0/d_m \leq \rho$, where ρ denotes the target fraction of target-modal dimensions that may be preserved.

Finally, for the third term, directly optimizing the conditional mutual information $I(Y;\hat{n}|p)$ is difficult as it lacks straightforward proxy tasks. Therefore, we propose an optimization strategy that transforms this conditional objective into an unconditional mutual information problem. Specifically, we construct a counterfactual shuffled view $V_i^{\text{shuf}} = [p_i, \hat{n}_{i\leftarrow j}]$, where the helper keys are sampled from a different instance j: $\widetilde{Z}_i^{m,\text{shuf}} \leftarrow \text{Attn}\left(Q_i^{m\leftarrow h}, K_j^{m\leftarrow h}, V_i^h\right), j \neq i$. We then apply the chain rule to the unconditional mutual information of both the matched positive view $V_i^+ = [p_i, \hat{n}_{i\leftarrow i}]$ and the shuffled view:

$$I(Y_i; V_i^+) = I(Y_i; p_i) + I(Y_i; \hat{n}_{i\leftarrow i}|p_i); \quad I(Y_i; V_i^{\text{shuf}}) = I(Y_i; p_i) + I(Y_i; \hat{n}_{i\leftarrow j}|p_i). \tag{7}$$

By computing the difference between these two unconditional terms, the base information $I(Y_i; p_i)$ cancels out, yielding an interaction gap:

$$\text{Gap}_{\text{MI}} = I(Y_i; V_i^+) - I(Y_i; V_i^{\text{shuf}}) = I(Y_i; \hat{n}_{i\leftarrow i}|p_i) - I(Y_i; \hat{n}_{i\leftarrow j}|p_i). \tag{8}$$

Here, $I(Y_i; \hat{n}_{i\leftarrow j}|p_i)$, with $\hat{n}_{i\leftarrow j}$ denoting that sample j participates in the cross-modal temporal interaction of $\hat{n}$, serves as a counterfactual baseline for information available without matched sample-level temporal-dependency, which provides a reference for instantiating the lower bound δ defined in Eq. 1, yielding the adaptive bound $\delta = I(Y_i; \hat{n}_{i\leftarrow j}|p_i) + \delta'$. Through this derivation, we translate the complex conditional lower bound constraint into ensuring a margined unconditional mutual information gap $\text{Gap}_{\text{MI}} \geq \delta'$ between the positive and shuffled views.

Final Optimization Objective. To consolidate the theoretical constraints into a differentiable framework, we translate the three conditions from Eq. 1 into empirical loss terms. First, we use the contrastive loss on matched positive views as a practical surrogate for semantic sufficiency (maximizing $I(V^+; Y)$). Second, to control the minimality upper bound in Eq. 6, we constrain the normalized L_0 cardinality of the binary gates: $\frac{\|g^m_{\text{hard}}\|_0}{d_m} \leq \rho, \forall m \in \mathcal{M}$. We optimize a straight-through penalty relaxation of this constraint: $\mathcal{L}_{\text{sparse}} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\left[\frac{1}{d_m}\sum_{j=1}^{d_m}\widetilde{g}_j^m - \rho\right]_+^2$. Starting from an all-retained initialization, this penalty progressively reduces the retained fraction during joint training, inducing a curriculum from augmented views to increasingly replacement-heavy positive views. Finally, we operationalize the theoretical interaction gap Gap_{MI} using the empirical loss $\mathcal{L}_{\text{gap}}$ in the pretrained AE latent space to encourage the conditional-information lower bound (the third condition in Eq. 1). Inspired by the cross-correlation formulation of Barlow Twins, we use diagonal cross-correlation as a proxy for shared information. Let $\mathbf{Z}^{m,\text{orig}}$, $\widetilde{\mathbf{Z}}^{m,+}$, and $\widetilde{\mathbf{Z}}^{m,\text{shuf}}$ denote the batch-standardized AE latent embeddings of the original, augmented-positive, and shuffled views of modality m, respectively. The correlation matrices are: $\mathbf{C}^{m,+} = \frac{1}{B}\left(\mathbf{Z}^{m,\text{orig}}\right)^\top \widetilde{\mathbf{Z}}^{m,+}$, $\mathbf{C}^{m,\text{shuf}} = \frac{1}{B}\left(\mathbf{Z}^{m,\text{orig}}\right)^\top \widetilde{\mathbf{Z}}^{m,\text{shuf}}$. We define the gap loss as:

$$\mathcal{L}_{\text{gap}} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\frac{1}{d_m}\sum_{j=1}^{d_m}\left[\delta' - \left([\mathbf{C}^{m,+}]_{jj} - [\mathbf{C}^{m,\text{shuf}}]_{jj}\right)\right]_+, \tag{9}$$

where d_m is the latent dimensionality of the AE latent space. Notably, the shuffled branch is stop-gradient and serves only as a reference. Integrating these derivations, the final joint loss function for the end-to-end framework is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SSL}} + \alpha\mathcal{L}_{\text{gap}} + \beta\mathcal{L}_{\text{sparse}}, \tag{10}$$

where α and β are hyperparameters. Note that during joint training, gradients from L_{SSL} propagate through the frozen decoder to update the cross-modal augmentation module.

3 Experiments

3.1 Experiment Setup

Datasets and Tasks. All experiments utilize publicly accessible datasets, strictly ensuring zero dataset or subject overlap between the pretraining and downstream stages. Our pretraining corpus comprises 3,565 hours of daily-living data from 56 participants across three sources: BIG

Table 1: Comparison under early fusion and InfoNCE objective. Best results are marked in **bold**. Second-best results are underlined.

	WESAD		AAUWSS		DaLiA		WEEE	
Method	↑F_1	↑BACC	↑F_1	↑BACC	↑F_1	↑BACC	↓MAE	↑R^2
Ours (LP)	**.544**±.010	**.572**±.014	**.276**±.031	.259±.035	**.729**±.013	**.715**±.006	**3.003**±.211	**.554**±.047
SimCLR	.436±.006	.449±.002	.271±.040	**.307**±.085	.695±.004	.691±.013	3.806±.493	.277±.324
Chaos	.447±.003	.461±.006	.247±.020	.266±.034	.704±.008	.698±.013	3.543±.439	.424±.073
AutoTCL	.445±.014	.464±.013	.213±.008	.220±.012	.684±.009	.673±.011	3.534±.279	.420±.076
TF-C	.410±.014	.428±.020	.271±.024	.305±.031	.696±.010	.687±.021	3.391±.213	.418±.046
xMAE	.421±.107	.456±.060	.200±.005	.207±.026	.606±.022	.592±.020	3.670±.192	.366±.041
REBAR	.404±.083	.426±.078	.232±.019	.243±.020	.560±.032	.550±.021	3.810±.267	.306±.112
TS2Vec	.385±.112	.432±.104	.215±.010	.231±.013	.665±.008	.650±.001	3.535±.460	.422±.133
TNC	.410±.048	.453±.045	.219±.024	.236±.020	.646±.025	.629±.017	3.865±.122	.323±.135

Table 2: Comparison under early fusion and Barlow Twins objective. Best results are marked in **bold**. Second-best results are underlined.

	WESAD		AAUWSS		DaLiA		WEEE	
Method	↑F_1	↑BACC	↑F_1	↑BACC	↑F_1	↑BACC	↓MAE	↑R^2
Ours (LP)	**.486**±.063	**.522**±.067	**.257**±.009	.240±.012	**.738**±.005	**.729**±.010	**2.913**±.324	**.578**±.086
Barlow Twins	.419±.010	.437±.006	.250±.015	.265±.019	.702±.009	.690±.006	3.603±.092	.297±.064
Chaos	.430±.025	.450±.025	.247±.020	.256±.017	.686±.013	.675±.019	3.423±.234	.411±.103
AutoTCL	.430±.010	.444±.009	.237±.016	.253±.005	.704±.007	.686±.005	3.799±.423	.254±.291
TF-C	.447±.017	.459±.014	.254±.012	**.293**±.059	.699±.026	.686±.022	3.128±.361	.506±.176
xMAE	.424±.107	.451±.109	.249±.041	.238±.024	.698±.007	.683±.006	3.083±.097	.536±.034
REBAR	.404±.020	.424±.020	.235±.015	.227±.031	.700±.014	.687±.011	3.278±.154	.423±.101
TS2Vec	.482±.006	.491±.001	.212±.002	.229±.013	.665±.029	.649±.017	3.305±.142	.444±.060
TNC	.453±.029	.474±.036	.218±.010	.241±.024	.692±.007	.675±.013	3.471±.288	.404±.150

IDEAs (Bent et al., 2021), UE4W (Hinkle et al., 2023), and the non-cannabis-user split of CANSTRESS (Azghan et al., 2025). Downstream performance is assessed on four benchmarks: WESAD (Schmidt et al., 2018) for 3-class emotion recognition (15 subjects), DaLiA (Reiss et al., 2019) for 8-class activity recognition (15 subjects), AAUWSS (Djanian et al., 2025) for 5-class sleep staging (13 subjects), and WEEE (Gashi et al., 2022) for VO_2 estimation (17 subjects). We employ a subject-wise 5-fold cross-validation to ensure that there is no overlap or leakage.

Baselines and Metrics. In addition to the classic SimCLR-type data augmentations (e.g., scaling, shifting, and jittering), our baseline also incorporates three common data augmentation paradigms from time series contrastive learning: rule-based transformations, positive sample selection from real data, and generation-based positive sample construction. Specifically, TS2Vec (Yue et al., 2022) and TF-C (Zhang et al., 2022) employ fixed non-parametric transformations based on overlapping crops with timestamp masking, and time–frequency perturbations, respectively; TNC (Tonekaboni et al., 2021) and REBAR (Xu et al., 2024) construct positive pairs from observed subsequences using temporal-neighborhood membership, and learned reconstruction-based similarity, respectively; AutoTCL (Zheng et al., 2024) learns a parametric augmentation network, Finding Order in Chaos (Demirel & Holz, 2023) adaptively performs phase–amplitude mixup using VAE-derived latent similarity, and we use xMAE's (Zhou et al., 2026) directional cross-modal reconstructions as positive views designed to capture cross-modal temporal structure. For classification tasks, we evaluate using Macro-F1 and BAcc, while for regression tasks, we report MAE and R^2.

Evaluation Protocol. All baselines are evaluated using the same encoder and contrastive objective, differing only in their positive-view construction strategies. We evaluate two common strategies for multimodal fusion: early fusion and late fusion. In addition to the traditional InfoNCE-type contrastive learning, we also evaluate on positive-sample-only pipeline of Barlow Twins (Zbontar et al., 2021), which is also important because defining negative samples can be challenging in wearable health scenarios. All methods follow the same evaluation protocol: freezing the encoder after pretraining, training only a linear probe to adapt to downstream tasks, and ensuring no overlap between the pretraining dataset and the downstream task datasets used for evaluation.

3.2 Experimental Results

Performance on Early Fusion Encoder. We evaluated both the InfoNCE-type contrastive learning objective and the Barlow Twins-type positive-sample-only objective on the early fusion architecture, as shown in Table 1 and Table 2. Among the 16 listed metrics, ReWeave ranks first in 14 of them. Specially, ReWeave outperforms all baselines on the macro-F1 scores across three classification tasks under both objectives, with the largest improvement over the second-best method reaching approximately 0.1. For VO_2 prediction, ReWeave reduces the MAE by up to 0.388 (over

Table 3: Comparison under late fusion and InfoNCE objective. Best results are marked in **bold**. Second-best results are underlined.

	WESAD		AAUWSS		DaLiA		WEEE	
Method	↑F_1	↑BACC	↑F_1	↑BACC	↑F_1	↑BACC	↓MAE	↑R^2
Ours (LP)	**.608**±.010	**.622**±.024	**.288**±.016	**.315**±.041	**.736**±.011	**.732**±.025	**3.262**±.439	**.454**±.083
SimCLR	.427±.005	.439±.002	.259±.031	.280±.052	.691±.003	.688±.009	3.801±.442	.299±.242
Chaos	.519±.024	.526±.028	.256±.006	.270±.013	.731±.013	.720±.013	3.551±.308	.402±.079
AutoTCL	.600±.017	.619±.018	.279±.017	.286±.045	.715±.025	.705±.026	3.637±.409	.410±.130
TF-C	.560±.017	.558±.028	.272±.017	.309±.009	.720±.018	.708±.010	3.785±.303	.337±.102
xMAE	.549±.028	.564±.042	.218±.013	.226±.031	.713±.003	.704±.011	3.647±.490	.402±.098
REBAR	.553±.085	.563±.091	.263±.015	.283±.019	.721±.024	.719±.024	3.773±.438	.339±.175
TS2Vec	.579±.018	.590±.023	.270±.042	.278±.041	.698±.014	.688±.021	3.694±.526	.397±.168
TNC	.571±.025	.580±.030	.279±.013	.286±.023	.722±.018	.715±.021	3.456±.192	.438±.149

Table 4: Comparison under late fusion and Barlow Twins objective. Best results are marked in **bold**. Second-best results are underlined.

	WESAD		AAUWSS		DaLiA		WEEE	
Method	↑F_1	↑BACC	↑F_1	↑BACC	↑F_1	↑BACC	↓MAE	↑R^2
Ours (LP)	**.652**±.017	**.654**±.024	**.301**±.002	**.325**±.024	**.720**±.013	**.718**±.018	**3.270**±.407	.442±.122
Barlow Twins	.448±.029	.462±.035	.285±.006	.307±.012	.636±.016	.623±.032	3.714±.486	.261±.170
Chaos	.457±.021	.474±.042	.270±.015	.262±.027	**.720**±.013	.704±.026	3.342±.233	**.473**±.067
AutoTCL	.571±.024	.565±.017	.270±.013	.302±.023	.675±.011	.662±.014	3.668±.670	.377±.191
TF-C	.598±.010	.591±.013	.264±.012	.290±.027	.702±.018	.692±.021	3.332±.291	.464±.126
xMAE	.589±.014	.602±.013	.245±.018	.269±.024	.647±.016	.637±.017	3.508±.130	.436±.020
REBAR	.569±.033	.574±.045	.275±.012	.294±.020	.696±.025	.682±.040	3.758±.917	.359±.318
TS2Vec	.519±.012	.530±.057	.283±.017	.296±.019	.649±.007	.635±.018	3.564±.456	.424±.080
TNC	.600±.056	.588±.051	.273±.013	.288±.023	.694±.021	.677±.014	3.334±.452	.458±.142

11%) compared to the second-best method. This task heavily relies on the interaction between PPG and ACC signals, highlighting the importance of the temporal dependency preserved by our method.

Performance on Late Fusion Encoder. As shown in Table 3 and Table 4, under the late fusion encoder, ReWeave achieves first place in 15 out of the 16 metrics. The results indicate that the late fusion encoder is more effective for emotion detection and sleep staging tasks. Based on this, ReWeave is the only data augmentation method that achieves an F1 score above 0.65 for emotion detection (WESAD) and above 0.3 for the more challenging sleep staging task (AAUWSS).

3.3 Temporal Dependency Preservation

Previous work quantifies temporal dependency using modality-specific metrics, such as the temporal offset between the ECG R-peak and the PPG onset valley (Zhou et al., 2026). However, in our multimodal setting, not every modal pair has such specific combinations. Therefore, we adopt a more general approach to approximate temporal dependency.

Specifically, we define temporal dependency preservation error in order to quantify augmentation-induced changes in cross-modal temporal dependencies. Because changes in other modalities are generally reflected in PPG pulse frequency rather than amplitude (e.g., accelerometer-detected motion may be accompanied by an increased heart rate), we first preprocess PPG into heart-rate (HR). For each modality pair (e.g., PPG-ACC), we construct a common temporal grid at the lower sampling rate f_p and average the signal values within each bin. We shift the second modality in one-bin increments within the shared duration and compute the Pearson correlation between the overlapping portions at each shift. This produces a correlation sequence with a temporal resolution of $1/f_p$ seconds, and the temporal offset with the largest absolute correlation is taken as the modality-pair lag for that sample. We compute the delay separately for the original and augmented signals, define their absolute difference as the error for each modality pair.

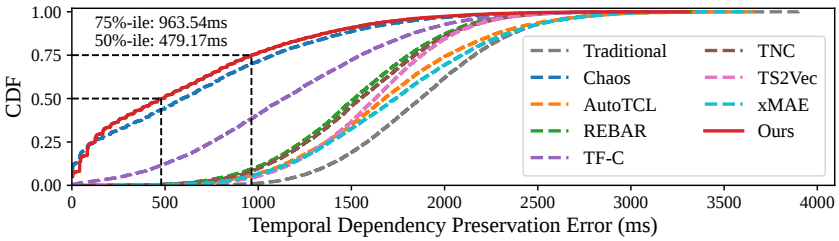


Figure 3: CDF of temporal dependency preservation error across augmentation methods.

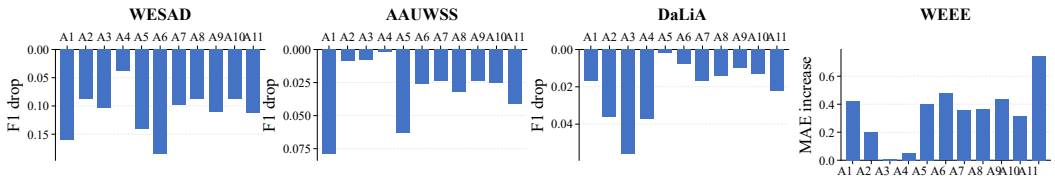


Figure 4: **Performance drop of each ablation setting compared to the full ReWeave. For WESAD, AAUWSS, and DaLiA classification tasks, the Macro-F1 drop is reported; for WEEE regression task, the MAE increase is shown.** A1: AE w/o VC-Reg. A2: Random Gate. A3: Fixed Gate Ratio (25%). A4: Fixed Gate Ratio (50%). A5: Fixed Gate Ratio (75%). A6: w/o Cross-Attention. A7: w/o Gap Loss. A8: w/o Helper Scorer. A9: w/o Translator. A10: KV-Shuffled Sample. A11: AE-only Augmentation.

We first compute the average temporal dependency error across all modal-pairs for each original-augmented sample pair, and then obtain its distribution. The cumulative distribution function (CDF) of this distribution is plotted in Figure 3. It can be observed that the augmented samples produced by ReWeave significantly better preserve the temporal dependency of the original samples compared to the baseline. While Finding Order in Chaos (Demirel & Holz, 2023) also demonstrates strong dependency preservation, this is largely because it does not explicitly encourage the introduction of variation during augmented sample construction like ReWeave. Empirical results also show that this could lead to poorer downstream task performance.

3.4 Ablation Study

To validate the necessity of each component in ReWeave and evaluate the performance drop when they are removed, we conducted 11 different ablation studies, as shown in Figure 4. Specifically, we conduct the following ablations: (A1) replace the pretrained autoencoder originally trained with VC-reg in Section 2.2, with a standard reconstruction-based autoencoder; (A2) for Dimension-wise Gating, replace the learnable gates with randomly selected gates while keeping the gate ratio consistent with ReWeave; (A3-A5) keep the gates learnable but fix the gate ratio throughout training to 25%, 50%, or 75%, respectively; (A6) remove the cross-temporal cross-attention between modalities, forcing all target modality helpers to come only from the same timestamp in other modalities, which disables learning temporal dependencies across modalities; (A7) remove $\mathcal{L}_{gap}$, eliminating the construction of shuffled samples as counterfactual baselines to encourage temporal dependency learning; (A8) remove the helper scorer, using equal weights for all helper modalities rather than using learned weights; (A9) exclude the translator used to map between different modality autoencoder latent spaces; (A10) replace the values $\mathbf{v}$ in cross-attention for the shuffled samples with $\mathbf{v}$ from other samples, similar to the keys $\mathbf{k}$, instead of using the current sample's $\mathbf{v}$; and (A11) use only the VC-Reg AE to generate reconstructed samples, which are then used as augmented samples. As shown in the results, removing any of these components results in a significant performance drop across all tasks, demonstrating the necessity of these key design choices.

4 Conclusion

We introduced ReWeave, a plug-and-play augmentation framework for multimodal wearable contrastive learning that learns what to retain from each modality and what to replace using synchronized companion signals. ReWeave utilizes a VC-reg latent space and learnable dimension-wise gates to preserve target-modal information that is difficult to recover from the available helpers while synthesizing the remaining content from temporally structured cross-modal context, thereby producing diverse yet semantically valid positive views. Together with a progressive sparsity curriculum and a matched-vs-shuffled gap objective, ReWeave prevents shortcutting through excessive target retention or generic modality priors and explicitly promotes sample-wise cross-modal temporal dependency learning. Experiments across four health-related tasks, two fusion architectures, and both contrastive and positive-only objectives show that ReWeave generally improves downstream performance over alternative positive-view construction strategies. Together, these findings support selective cross-modal replacement as a practical approach to constructing useful positive views, highlighting synchronized companion signals as a source of structured variation for wearable representation learning.

AI USE STATEMENT

In this work, we used generative AI tools to clean and reformat dataset. We have not used generative AI tools for any other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to edit a research paper to improve readability, identify relevant literature, and create or modify scientific figures or images. We have reviewed all AI-assisted work. All AI-assisted text was reviewed by two authors to ensure that it preserved the intended technical meaning. All references identified with the assistance of generative AI were verified by two authors against the original sources for their existence and relevance. The AI-assisted dataset processing was validated against the original data by two authors. The resulting figures were also checked against the underlying data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses only publicly available wearable-sensor datasets and involves no new data collection or interaction with human participants. We followed the applicable data-use terms and relied on the original dataset providers for informed consent and ethical approval.

REPRODUCIBILITY STATEMENT

The implementation details including model architecture, data preprocessing configuration, hyperparameter configuration, and optimization settings are detailed in Appendix F. We will make the code publicly available upon acceptance to support reproducibility and future research.

REFERENCES

Reza Rahimi Azghan, Nicholas C Glodosky, Ramesh Kumar Sah, Carrie Cuttler, Ryan McLaughlin, Michael J Cleveland, and Hassan Ghasemzadeh. Can-stress: A real-world multimodal dataset for understanding cannabis use, stress, and physiological responses. *arXiv preprint arXiv:2503.19935*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Clint R Bellenger, Joel T Fuller, Rebecca L Thomson, Kade Davison, Eileen Y Robertson, and Jonathan D Buckley. Monitoring athletic training status through autonomic heart rate regulation: a systematic review and meta-analysis. *Sports medicine*, 46(10):1461–1486, 2016.

Brinnae Bent, Peter J Cho, Maria Henriquez, April Wittmann, Connie Thacker, Mark Feinglos, Matthew J Crowley, and Jessilyn P Dunn. Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches. *NPJ Digital Medicine*, 4(1):89, 2021.

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

Berken Utku Demirel and Christian Holz. Finding order in chaos: A novel data augmentation method for time series in contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:30750–30783, 2023.

Shagen Djanian, Thomas Dyhre Nielsen, Søren H. Nielsen, and Anders Bruun. Aalborg University Wearable Sleep Study (AAUWSS), August 2025. URL `https://doi.org/10.5281/zenodo.16919071`. Version 1.0.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*,

pp. 2352–2359. International Joint Conferences on Artificial Intelligence Organization (IJCAI), 8 2021. doi: 10.24963/ijcai.2021/324. Main Track.

Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J Foti, and Joseph Futoma. Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. In *Forty-second International Conference on Machine Learning (ICML)*, 2025.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. *Advances in neural information processing systems (NeurIPS)*, 32, 2019.

Shkurta Gashi, Chulhong Min, Alessandro Montanari, Silvia Santini, and Fahim Kawsar. A multidevice and multimodal dataset for human energy expenditure estimation using wearable devices. *Scientific Data*, 9(1):537, 2022.

Lee B Hinkle, Gentry Atkinson, and Vangelis Metsis. Fusion of learned representations for multimodal sensor data classification. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 404–415. Springer, 2023.

Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. Focal: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:47309–47338, 2023.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jacob Sunshine, Yun Liu, et al. Scaling wearable foundation models. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 59382–59413, 2025.

Maximillian J Nelson, Rebecca L Thomson, Daniel K Rogers, Peter RC Howe, and Jonathan D Buckley. Maximal rate of increase in heart rate during the rest-exercise transition tracks reductions in exercise performance when training load is increased. *Journal of science and medicine in sport*, 17(1):129–133, 2014.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 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.

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.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems (NeurIPS)*, 33:6827–6839, 2020.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations (ICLR)*, 2021.

Solveig Vieluf, Tanuj Hasija, Peter J Schreier, Rima El Atrache, Sarah Hammond, Fatemeh Mohammadpour Touserkani, Rani A Sarkis, Tobias Loddenkemper, and Claus Reinsberger. Generalized tonic-clonic seizures are accompanied by changes of interrelations within the autonomic nervous system. *Epilepsy & Behavior*, 124:108321, 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 (ICLR)*, volume 2024, pp. 15440–15464, 2024.

Maxwell A Xu, Jaya Narain, Gregory Darnell, Haraldur T Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Andres Fineman, Karthik Jayaraman Raghuram, James Matthew Rehg, and Shirley You Ren. Relcon: Relative contrastive learning for a motion foundation model for wearable data. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

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 (AAAI)*, volume 36, pp. 8980–8987, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning (ICML)*, pp. 12310–12320. PMLR, 2021.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems (NeurIPS)*, 35:3988–4003, 2022.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed Metwally, Jinhua Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the language of wearable sensors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Xu Zheng, Tianchun Wang, Wei Cheng, Aitian Ma, Haifeng Chen, Mo Sha, and Dongsheng Luo. Parametric augmentation for time series contrastive learning. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 46906–46931, 2024.

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 (ICML)*, 2026.

A Related Work

Building on instance-discrimination frameworks such as SimCLR (Chen et al., 2020), time-series contrastive learning adapts the definition of positive and negative pairs to the temporal structure of sequential data. Early methods primarily exploit temporal locality: T-Loss learns sequence-level representations through a triplet objective over nested subsequences and time-based negative sampling (Franceschi et al., 2019), whereas TNC determines pair membership according to temporal neighborhoods (Tonekaboni et al., 2021). Subsequent approaches introduce more structured contrastive objectives. TS2Vec performs hierarchical timestamp- and instance-level contrasting to learn representations at multiple temporal scales (Yue et al., 2022); TS-TCC combines cross-view temporal prediction with contextual contrasting (Eldele et al., 2021); and TF-C enforces consistency between time- and frequency-domain representations (Zhang et al., 2022). For wearable and multimodal signals, FOCAL factorizes representations into shared and modality-private components while imposing a temporal-locality constraint (Liu et al., 2023), whereas RelCon learns motif-based relative similarities for large-scale accelerometry pretraining (Xu et al., 2025). Collectively, these studies demonstrate the value of incorporating temporal and modality-specific inductive biases.

Existing positive-view construction strategies for time series can be broadly divided into three categories. Rule-based transformations apply predefined operations such as jittering, scaling, permutation, cropping, masking, and time–frequency perturbation. For example, TS-TCC combines weak and strong transformations (Eldele et al., 2021), TS2Vec constructs contextual views through overlapping crops and timestamp masking (Yue et al., 2022), and TF-C perturbs signals in both the temporal and frequency domains (Zhang et al., 2022). Positive sample selection from real data instead identifies naturally occurring related subsequences: T-Loss samples nested subsequences from the same series (Franceschi et al., 2019), TNC uses temporal-neighborhood membership (Tonekaboni et al., 2021), and REBAR retrieves positive pairs using a learned reconstruction-based similarity measure (Xu et al., 2024). Finally, generation-based positive sample construction learns or derives

new views rather than selecting fixed transformations or existing samples. AutoTCL learns dataset- and instance-adaptive parametric perturbations (Zheng et al., 2024), Finding Order in Chaos performs phase–amplitude mixup using latent similarity between quasi-periodic signals (Demirel & Holz, 2023). However, these approaches do not explicitly optimize positive-view construction to preserve the source instance's cross-modal temporal dependency, such as event ordering, response latency, and coordinated variation across synchronized modalities.

B Downstream Dataset Details.

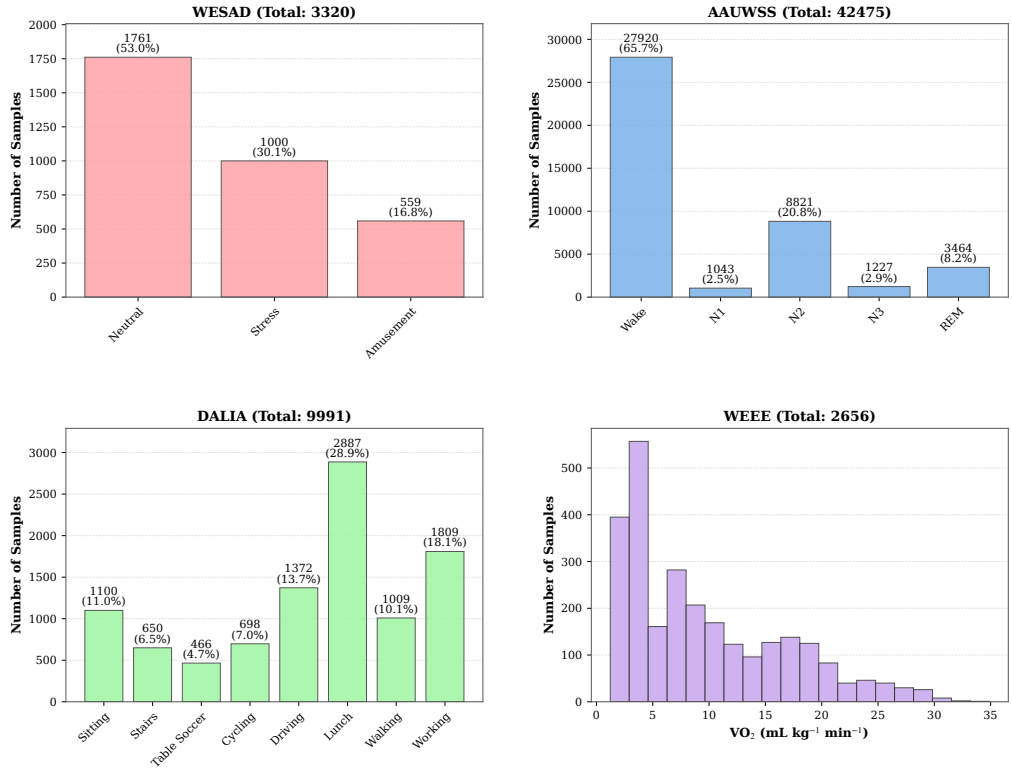


Figure 5: Distributions of Four Downstream Tasks (WESAD, AAUWSS, DaLiA, and WEEE).

C Hyperparameter Ablation

Training ReWeave involves four hyperparameters: α, β, ρ, and δ'. The hyperparameters α and β control the relative contributions of the auxiliary losses $\mathcal{L}_{\text{gap}}$ and $\mathcal{L}_{\text{sparse}}$ to the main contrastive objective $\mathcal{L}_{\text{SSL}}$. We choose them such that the weighted auxiliary terms are approximately one order of magnitude smaller than $\mathcal{L}_{\text{SSL}}$ in the total objective. Based on the observed scales of $\mathcal{L}_{\text{gap}}$ and $\mathcal{L}_{\text{sparse}}$, we set $\alpha = 1$ and $\beta = 10$. We further conduct sensitivity analyses over order-of-magnitude variations in α and β on WESAD and DaLiA. As shown in Figure 6, the performance of ReWeave remains stable within the selected ranges.

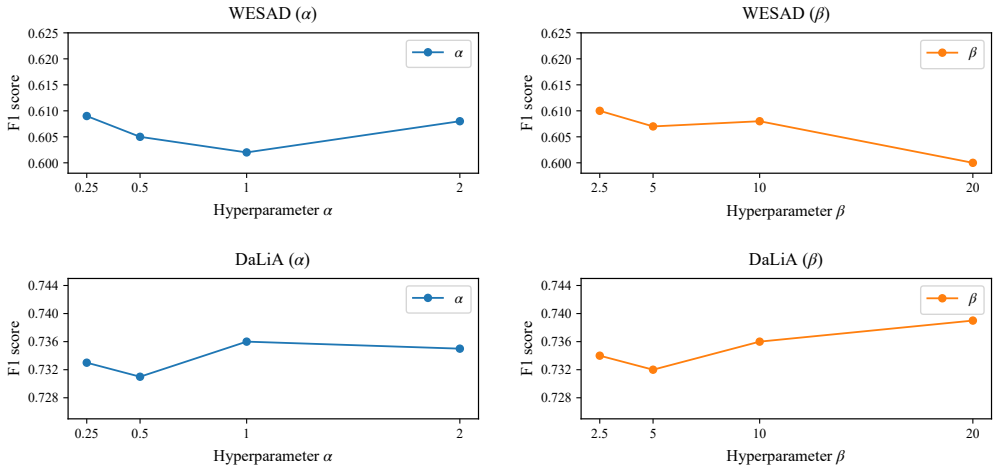


Figure 6: Sensitivity Analysis of α and β.

For δ', its value is bounded within $[0, 1]$. A larger δ' places greater emphasis on separating the shuffled and positive views, thereby encouraging the model to capture temporal dependencies. We use the intermediate value $\delta' = 0.5$ in our experiments. As shown in Figure 7, the model is relatively insensitive to δ' over a broad intermediate range. The hyperparameter ρ directly controls the strength of positive construction: a smaller ρ permits more aggressive cross-modal replacement. Figure 7 shows that different health-related tasks favor different levels of augmentation difficulty, with some benefiting from harder positive views and others from milder perturbations. Consequently, values that are too large or too small may be suboptimal for particular tasks. However, the curriculum induced by our joint training strategy makes the optimization robust to ρ, preventing performance collapse even when ρ is not ideally set.

Beyond validating sensitivity to ρ, these results reveal task-dependent granularity in the learned representations. Affective states in WESAD are largely characterized by coarse, slowly varying physiological summaries, such as heart-rate level, EDA level, and temperature trend. Stronger augmentation (smaller ρ) can therefore suppress fine-grained nuisance variations without removing task semantics. Conversely, DaLiA relies more on fine-grained motion dynamics: walking and stairing may have similar mean gravity-aligned acceleration but differ in their within-step vertical impulses, making weaker augmentation (larger ρ) preferable.

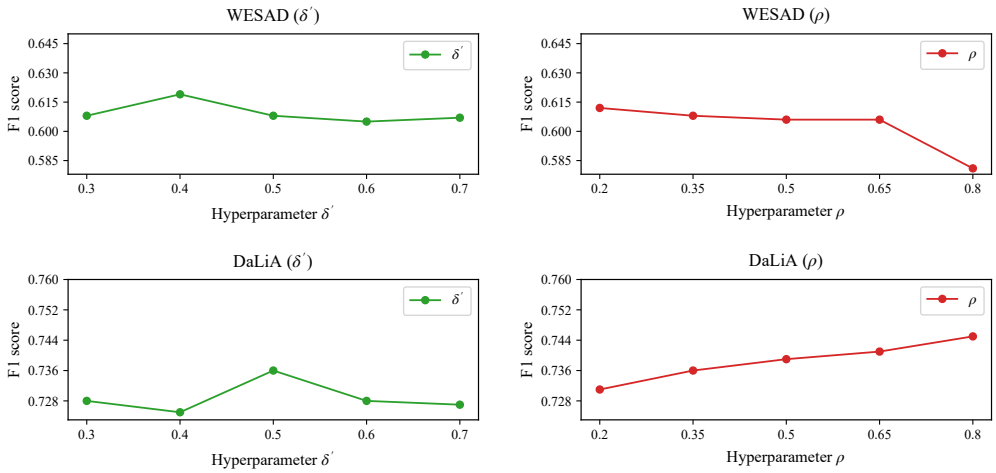


Figure 7: Sensitivity Analysis of δ' and ρ.

D VISUALIZATION OF P-RATIO

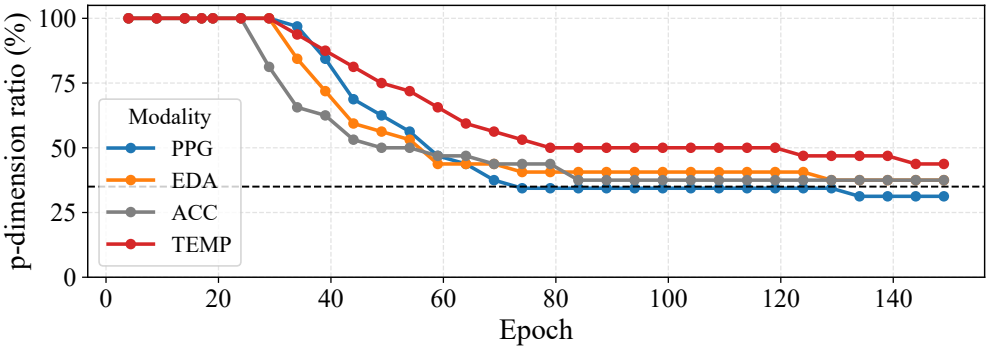


Figure 8: Evolution of the p-ratio across training epochs for different modalities with ρ set to 0.35.

E GENERALIZATION TO MULTIMODAL CONTRASTIVE LEARNING FRAMEWORK

Beyond the standard early- and late-fusion contrastive learning frameworks, we further evaluate ReWeave and the baselines under FOCAL (Liu et al., 2023), a contrastive learning framework specifically designed for multimodal time series. FOCAL disentangles each modality into modality-private and modality-shared representations, where data augmentation plays a critical role in helping the model identify modality-private information. As shown in Table 5, FOCAL generally outperforms conventional fusion structures and contrastive objectives on these multimodal tasks. ReWeave further improves FOCAL on most tasks, achieving the best performance on 6 of the 8 metrics and the second-best performance on the remaining 2 when compared with the baselines. These results further demonstrate the generalizability of ReWeave across different contrastive learning frameworks.

Table 5: FOCAL comparison under linear probing (LP). Best results are marked in **bold**. Second-best results are underlined.

	WESAD		AAUWSS		DaLiA		WEEE	
Method	↑F_1	↑BACC	↑F_1	↑BACC	↑F_1	↑BACC	↓MAE	↑R^2
Ours (LP)	**.577**±.031	**.588**±.046	.265±.023	**.308**±.036	**.742**±.031	.736±.039	**3.279**±.487	**.471**±.124
FOCAL	.513±.050	.533±.067	**.270**±.011	.285±.011	.737±.039	.728±.043	3.347±.561	.468±.132
Chaos	.546±.043	.560±.066	.251±.029	.250±.018	.740±.015	.733±.023	3.528±.458	.439±.124
AutoTCL	.558±.046	.568±.085	.254±.036	.261±.029	.736±.019	.735±.019	3.449±.973	.416±.396
REBAR	.558±.062	.570±.081	.259±.040	.265±.034	.733±.012	.732±.008	3.563±1.039	.409±.305
TF-C	.530±.012	.539±.012	.252±.014	.272±.014	.735±.035	.734±.035	3.589±.540	.416±.186
TNC	.566±.050	.583±.058	.254±.011	.271±.020	.733±.031	.736±.039	3.410±.544	.467±.128
TS2Vec	.571±.008	.573±.027	.247±.040	.261±.036	**.742**±.019	**.740**±.027	3.528±.548	.441±.177
XMAE	.562±.063	.576±.081	.260±.036	.267±.037	.732±.010	.731±.015	3.630±.965	.385±.272

F IMPLEMENTATION DETAILS

Table 6: Implementation details and hyperparameters of the ReWeave augmentation module.

Hyperparameters		Values
Input Configuration	Sampling Rates (Hz)	{PPG: 64, ACC: 32, EDA: 4, TEMP: 4}
	Input Channels	{1, 3, 1, 1}
	Window Duration	10 s
	Patch Size	1 s
Frozen VC-Reg AE	Architecture	Modality-specific, patch-wise 1-D convolutional autoencoder
	Number of Parameters	5,746,342 (5.75M)
	Latent Dimension (d)	256
	Loss Weights ($\lambda_v/\lambda_c/\lambda_r$)	$25.0/1.0/1.0$
	Pre-training Setup	AdamW, 100 epochs, 10^{-4} learning rate, 10^{-5} weight decay
	Status during Augmentation Training	Frozen
Augmentation Module	Number of Trainable Parameters	9,782,212 (excluding the frozen AE)
	Cross-Attention Heads	4
	Attention Head Dimension	64
	Translator Hidden Dimension	512
	Translator Dropout	0.1
	Helper-Scorer Hidden Dimension	256
	Gate-Logit Initialization	1.0
Augmentation Objective	Retention Upper Bound (ρ)	0.35
	Gap-Loss Weight (α)	1.0
	Sparsity-Loss Weight (β)	10.0
	Gap Margin (δ')	0.5
Augmentation Optimization	Batch Size	1024
	Optimizer	AdamW
	Base Learning Rate	1×10^{-4}
	Weight Decay	1×10^{-5}
	Learning-Rate Schedule	5-epoch linear warm-up, then cosine decay to 1×10^{-6}
	Training Epochs	150
	Gradient-Clipping Norm	1.0