VCR: Learning Valid Contextual Representation for Incomplete Wearable Signals

Yuxuan Weng, Wenhan Luo and Qijia Shao*
The Hong Kong University of Science and Technology (HKUST)
ywengac@connect.ust.hk, whluo@ust.hk, qijiashao@ust.hk

Abstract

Wearable devices enable continuous health monitoring from multimodal signals, but real-world deployment is hindered by limited labeled data and pervasive sensor incompleteness. While large-scale self-supervised pretraining reduces label dependence, most existing methods assume full modality availability during downstream inference. Current approaches for handling modality missingness often reconstruct entire absent signals, which can encourage hallucinating modality-specific details that are not inferable from the observed sensor signals and degrade robustness. We propose VCR, a self-supervised framework that learns to extract valid representations robust to modality missingness. VCR employs an orthogonal tokenizer to enforce strict orthogonal disentanglement by rectifying latent manifolds and applying a geometric projection, separating each modality into shared semantics and modality-specific residuals. This design preserves complete information integrity while serving as a structural foundation for robust learning under modality missingness. The resulting tokens are processed by a missing-aware mixture-of-experts backbone that adapts to varying patterns of modality availability. By constraining the objective to reconstruct only the inferable shared components of missing modalities, VCR effectively mitigates hallucination-like completion of non-inferable modality-specific details. Across multiple health monitoring tasks, VCR consistently improves performance and robustness under full, single-missing, and multiple-missing modality settings compared with strong supervised and self-supervised baselines. Our code is available at https://github.com/yuxuanweng/VCR.

1 Introduction

Wearable devices enable continuous and non-invasive monitoring by capturing synchronized motion and physiological signals such as photoplethysmography (PPG), electrodermal activity (EDA), accelerometer (ACC), and skin temperature (TEMP). Beyond basic signal processing (e.g., estimating heart rate from PPG), these modalities support a broad range of downstream health applications, including emotion recognition [1, 2], sleep staging [3, 4], and activity recognition [5, 6]. However, reliable generalization in real-world wearable deployments remains challenging due to two coupled factors: (i) label scarcity across diverse tasks [7], and (ii) pervasive modality missingness [1] caused by hardware constraints and user-dependent effects. For example, commercial wearable devices feature heterogeneous sensor suites (e.g., Oura Ring 4 lacks ECG, EDA, and an altimeter [8]; Fitbit Sense 2 includes all three [9]; and Apple Watch Series 9 lacks EDA [10]). Furthermore, these devices may experience complete modality loss due to hardware malfunction, battery conservation, or sensor detachment [11, 12, 13]. Empirical studies [14] also report that over three months of daily tracking, valid heart rate data was available on only 31.2%, 80.1%, and 49.3% of days for the Samsung Galaxy

*Corresponding author.

[1] In this paper, "missingness" refers to the complete absence of a modality throughout the given window.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

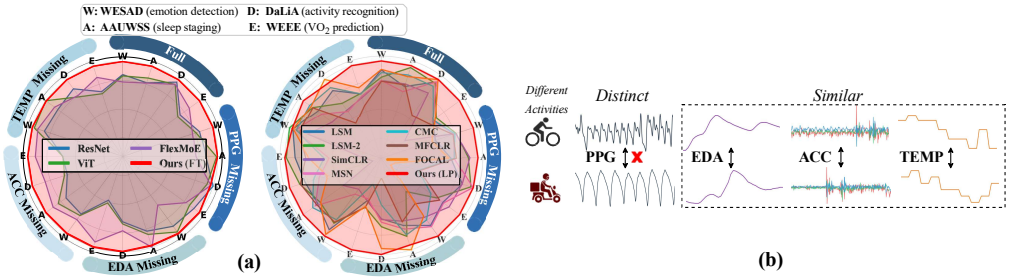

Figure 1: (a) **Performance under full and single-modal missingness** (see Experiments for multiple missingness). We report normalized Macro-F1 (value/max) and MAE (min/value) within each dataset to compare methods across tasks and datasets. Left: comparison to supervised baselines trained on labeled data only. Right: comparison to pretrained/self-supervised baselines. Higher is better. (b) **Modality-specific information can be non-identifiable from observed sensors.** Two activity windows may exhibit similar EDA, ACC, and TEMP signals while showing distinct PPG morphology.

Watch, Fitbit, and Apple Watch, respectively. Collectively, these examples demonstrate that complete modality loss is a common challenge for wearable devices in practical use.

Recent progress in self-supervised learning and foundation models [15, 16, 17] has made label-efficient wearable intelligence increasingly feasible: models can be pretrained on large-scale unlabeled data streams and transferred to downstream tasks with limited labeled data [18, 7]. However, most existing models implicitly assume full modality availability during downstream inference, thereby limiting their real-world utility under heterogeneous modality missingness. While contrastive learning approaches [19, 20, 21] can handle missing modalities by aligning shared representations, their emphasis on cross-modal commonality may suppress modality-specific cues. Conversely, missingness-aware reconstruction [11] often relies on masked reconstruction of the entire missing signal. When the absent modalities contain modality-specific details that are not identifiable from the observed sensors, such objectives can encourage hallucination-like completions and inject spurious information into the learned representation. Meanwhile, task-specific modality-routing methods [22, 23] may improve a particular downstream task but are not naturally reusable across multiple tasks. These limitations motivate a general-purpose pretrain model that remains valid and robust under train-on-full, test-on-missing scenarios. Achieving this goal requires addressing the following challenges:

- **Non-identifiability of Missing Modality-Specific Information.** When a modality is absent, its modality-specific details cannot be reliably inferred from the remaining sensors as illustrated in Figure 1(b). Nevertheless, reconstruction-based objectives can still incentivize generating them, leading to hallucination-like completions and injecting noise into representations.
- **Availability-Dependent Cross-Modal Dependencies.** Different missing-modality patterns determine which inter-modal correlations are observable, making it challenging for a unified dense architecture to model all modality combinations without interference.
- **Information Loss Caused by Imbalanced Learning.** Wearable modalities differ substantially in sampling rate, variance, and semantic content. As a result, masked reconstruction can be dominated by high-variance modalities and underutilize other modalities, while contrastive alignment can overemphasize shared semantics and underrepresent modality-specific cues.

To address these obstacles, we propose VCR, a missingness-robust general-purpose framework for multimodal wearable sensor data, pre-trained on raw PPG, EDA, ACC, TEMP signals. VCR consists of two stages: an Orthogonal Tokenizer that separates each modality into shared and specific components, and a Missing-Aware MoE Backbone that learns contextual representations under diverse modality-availability patterns. Together, these components enable VCR to preserve information under full observations while avoiding unsupported reconstruction when modalities are missing.

Our key contributions are as follows:

- **Validity-Aware Learning:** To address the non-identifiability of missing modality-specific information, we propose an orthogonal tokenizer that rectifies encoder representations, applies whitening, and uses a learned orthogonal projection to decompose each modality into modality-shared semantics and modality-specific residuals. During structural missingness, our validity-aware objective reconstructs only inferable shared components of absent modalities, thereby avoiding unsupported modality-specific reconstruction and mitigating hallucination-like completions.

- **Missing-Aware MoE Backbone:** To handle availability-dependent cross-modal dependencies, we introduce a sparse MoE backbone with learnable partition embeddings. This preserves modality/component identity and enables experts to specialize under different observed modality sets.
- **Two-Stage Information Preservation:** To mitigate imbalanced and conflicting SSL objectives, we first train the tokenizer to preserve shared semantics via cross-modal contrast and specific details via intra-modal reconstruction. We then train the backbone with masked latent reconstruction, reducing domination by high-frequency or high-variance raw signals.

We evaluate VCR on a comprehensive benchmark spanning three classification tasks (i.e., emotion detection, activity recognition, sleep staging), and one regression task (i.e., VO_2 prediction). Across these tasks as shown in Figure 1(a), VCR achieves highly competitive performance under both full-modality and all missing-modality settings, demonstrating its effectiveness in preserving information integrity and ensuring robustness against modality missingness.

2 Method

This section introduces VCR. We first formalize the problem and motivate disentanglement (Sec. 2.1), then describe the overall framework (Sec. 2.2), the orthogonal tokenizer (Sec. 2.3), and the missing-aware MoE backbone with training objectives (Sec. 2.4).

2.1 Preliminary

Problem Formalization. We consider multimodal wearable time-series data collected from M sensing modalities, including PPG, EDA, ACC, and TEMP. After synchronization and non-overlapping windowing, each sample is denoted as $X = \{\mathbf{x}^m\}_m^M$. Here, $\mathbf{x}^m \in \mathbb{R}^{T_m \times C_m}$ denotes the sequence for modality m, where T_m is the number of time steps (the product of window duration and sampling rate) and C_m is the number of channels. In actual deployments, not all modalities are always available due to heterogeneous device configurations, sensor failures, or user non-compliance. In this paper, a missing modality refers to the complete absence of an entire modality throughout the current window. We formalize this by defining the observed and missing modality sets as $\mathcal{M}_{\text{obs}}$ and $\mathcal{M}_{\text{miss}}$, respectively, where $\mathcal{M}_{\text{obs}}, \mathcal{M}_{\text{miss}} \subseteq \mathcal{M} = \{1, \ldots, M\}$. Accordingly, we denote the set of observed modalities as $\mathcal{X}_{obs} = \{\mathbf{x}^k | k \in \mathcal{M}_{obs}\}$ and the set of missing modalities as $\mathcal{X}_{miss} = \{\mathbf{x}^k | k \in \mathcal{M}_{miss}\}$. When handling missingness, the model receives only $\mathcal{X}_{\text{obs}}$ from the current window.

Information Disentanglement Hypothesis. We assume the latent information of each modality $\mathbf{x}^m$ can be structurally disentangled into two components [24, 25]: a **modal-shared** component $\mathbf{s}^m$ that captures semantics correlated with other modalities (e.g., heart-rate-related patterns in PPG may correlate with ACC micro-vibrations), and a **modal-specific** component $\mathbf{p}^m$ that is unique to m (e.g., skin impedance drift in EDA). By definition, $\mathbf{p}^m$ contains the information that is not *reliably identifiable* from other modalities. Accordingly, our framework encourages the disentangled components to be statistically independent, i.e., $I(\mathbf{s}^m; \mathbf{p}^m) = 0$.

Hallucination under structural modality missingness. Reconstruction is well-posed only when the target information is identifiable from the available context. Under our decomposition, a modality signal is represented by $(\mathbf{s}^m, \mathbf{p}^m)$. Standard masked autoencoding approaches implicitly maximize the likelihood of reconstructing the full modality signal: $\log p(\mathbf{x}^m \mid \mathcal{X}_{obs}) = \log p(\mathbf{s}^m, \mathbf{p}^m \mid \mathcal{X}_{obs})$. When modality m is structurally missing, the modal-specific component $\mathbf{p}^m$ is not identifiable from the observed modalities. Thus, reconstructing the full signal forces the model to generate plausible but ungrounded modal-specific patterns, which we term *hallucination-like completion*.

2.2 Overview of the Framework

Figure 2 illustrates the proposed two-stage VCR framework for learning robust multimodal representations from wearable time-series under heterogeneous modality missingness. Given a windowed sample with an availability pattern, VCR consists of two components: (i) an *Orthogonal Tokenizer* that is pretrained on fully observed windows and decomposes each modality into modality-shared and modality-specific factors, and (ii) a *Missing-Aware MoE Backbone* that receives only the tokens from observed modalities and adapts computation to different observed modality sets. During backbone pretraining, we use both random patch masking and structural masking, where 1–3 entire modalities are removed at the sample level.

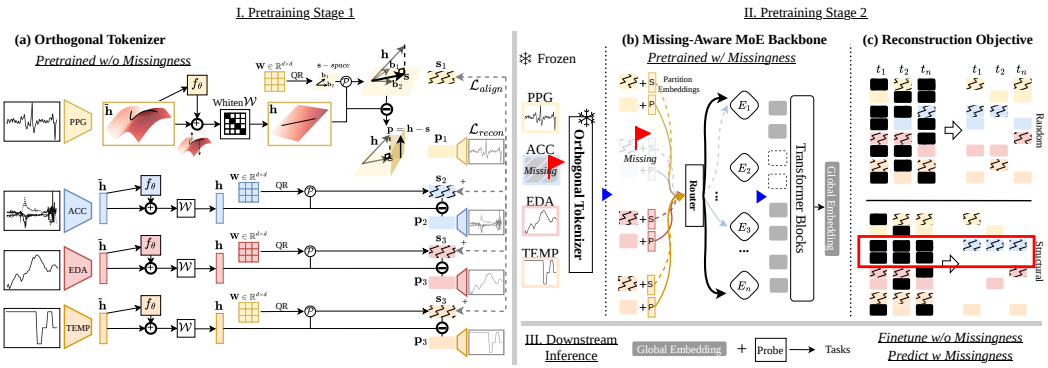

Figure 2: Overview of VCR. **(a) Orthogonal Tokenizer**: The Orthogonal Tokenizer is pretrained on the complete modality set. During training, encoders map raw signals to embeddings $\tilde{\mathbf{h}}$, which are rectified and projected by $\mathcal{P}$ into orthogonal shared $\mathbf{s}$ and specific $\mathbf{p}$ subspaces. Optimization combines cross-modal contrastive learning on $\mathbf{s}$ and intra-modal reconstruction on $\mathbf{p}$ in a full modality setting. During inference, our tokenizer uses a zero-padded placeholder with a "missingness flag" for missing modality to ensure tensor-shape consistency. **(b) Missing-Aware MoE Backbone:** After tokenization using a frozen tokenizer, learnable partition embeddings are added to the resulting tokens, while modalities flagged as missing are completely excluded at the input level.**(c) Masked Token Reconstruction:** Latent reconstruction is used only during backbone pretraining. For a single missing modality, only its shared component s is reconstructed; reconstruction of missing modalities is skipped when two or three modalities are missing.

2.3 Orthogonal Tokenizer

Disentangling modal-shared and modal-specific features is a prerequisite for explicitly preserving information integrity [24] and enabling selective reconstruction. Existing disentanglement methods typically impose soft constraints (e.g., orthogonality penalties in loss functions) between the shared and specific components [26]. Such an approach can lead to information leakage between the two components, causing $\mathbf{s}$ to attempt to reconstruct $\mathbf{p}$-related information.

Geometric Construction. We define a learnable parameter matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$, where d denotes the embedding dimension. We then perform QR decomposition on $\mathbf{W}$ during the forward pass: $\mathbf{W} = \mathbf{QR}$. Each column of the orthonormal matrix $\mathbf{Q}$ serves as a candidate basis vector and is associated with a learnable binary gate. We employ hard binary gating with a straight-through estimator (STE). The active columns selected by these gates form the basis matrix $\mathbf{B} \in \mathbb{R}^{d\times r}$, where r is the number of active basis vectors. Since the selected columns remain orthonormal, $\mathbf{B}^\top\mathbf{B} = \mathbf{I}_r$. Using this basis, we construct an orthogonal projection operator $\mathcal{P} = \mathbf{B}\mathbf{B}^\top \in \mathbb{R}^{d\times d}$. By construction, $\mathcal{P}$ is symmetric ($\mathcal{P}^\top = \mathcal{P}$) and idempotent ($\mathcal{P}^2 = \mathcal{P}$), satisfying the necessary and sufficient conditions for an orthogonal projection operator.

Structural Disentanglement and Independence. Since the raw latent representations from the encoder may reside on a highly non-linear manifold, directly applying a linear projection $\mathcal{P}$ might be insufficient to disentangle complex dependencies. To address this, we introduce a lightweight residual adapter f_θ, which predicts an additive correction $f_\theta(\tilde{\mathbf{h}})$, effectively rectifying the representation to better align with the target subspace. The rectified and whitened embedding $\mathbf{h}$ is obtained via:

$$\mathbf{h} = \mathcal{W}(\tilde{\mathbf{h}} + f_\theta(\tilde{\mathbf{h}})), \tag{1}$$

where $\mathcal{W}$ is the whitening process [27] and $\tilde{\mathbf{h}}$ is the original encoder embedding. Then we employ $\mathcal{P}$ to project $\mathbf{h}$ onto the modal-shared subspace, yielding the shared representation:

$$\mathbf{s} = \mathcal{P}\mathbf{h}. \tag{2}$$

Accordingly, the modality-specific representation is defined as the residual:

$$\mathbf{p} = \mathbf{h} - \mathbf{s} = (\mathbf{I} - \mathcal{P})\mathbf{h}, \tag{3}$$

which constitutes the strict geometric orthogonal complement of $\mathbf{s}$. Crucially, this geometric orthogonality translates to statistical independence under the condition that the input embeddings are

whitened. Assuming the covariance of the whitened input approximates identity ($\text{Cov}(h) \approx \mathbf{I}$), we can derive that $\mathbf{s}$ and $\mathbf{p}$ are uncorrelated:

$$\text{Cov}(\mathbf{s}, \mathbf{p}) = \mathbb{E}[\mathbf{s}\mathbf{p}^\top] = \mathbb{E}[\mathcal{P}\mathbf{h}\mathbf{h}^\top(\mathbf{I} - \mathcal{P})^\top] = \mathcal{P}\mathbb{E}[\mathbf{h}\mathbf{h}^\top](\mathbf{I} - \mathcal{P}) \approx \mathcal{P}\mathbf{I}(\mathbf{I} - \mathcal{P}) = \mathcal{P} - \mathcal{P}^2 = \mathbf{0}.$$

Under the standard assumption that latent representations follow a multivariate Gaussian distribution, this zero covariance implies statistical independence, i.e., $I(\mathbf{s}; \mathbf{p}) \approx 0$. Under the stated Gaussian assumption, this supports approximate statistical independence between $\mathbf{s}$ and $\mathbf{p}$, effectively eliminating feature redundancy by construction.

Hybrid Optimization Objectives. With the architecture ensuring structural separation, we need to populate $\mathbf{s}$ and $\mathbf{p}$ with semantic correctness and align the shared spaces across modalities. A single objective is insufficient for this dual purpose. Solely relying on masked reconstruction would still force the model to generate specific details using other modalities. Conversely, applying contrastive learning to $\mathbf{p}$ is detrimental, as standard data augmentations (e.g., scaling ACC data or adding noise to TEMP) can destroy the physical semantics of motion and physiological signals. Therefore, we adopt a hybrid strategy with three specific objectives.

Cross-Modal Alignment. To force the shared subspace to capture inter-modal semantics, we apply contrastive learning [28] exclusively to $\mathbf{s}$. We employ a masked contrastive loss to handle the temporal structure of wearable signals. For a given anchor token $\mathbf{s}_i^m$ (from modality m in sample i), we define the positive set $\mathcal{S}_i^+$ as the synchronized tokens from other modalities $\{\mathbf{s}_i^n\}_{n \neq m}$, the negative set $\mathcal{S}_i^-$ as the tokens from other modalities at different time steps, and the full candidate set as $\mathcal{A}_i = \mathcal{S}_i^+ \cup \mathcal{S}_i^-$. For brevity, let $\phi(\mathbf{u}, \mathbf{v}) = \exp(sim(\mathbf{u}, \mathbf{v})/\tau)$. The loss for modality m is defined as:

$$\mathcal{L}_{\text{align}}(s_i^m) = -\log \frac{\sum_{s^+ \in S_i^+} \phi(s_i^m, s^+)}{\sum_{s \in A_i} \phi(s_i^m, s)}. \tag{4}$$

Intra-Modal Reconstruction. To ensure the specific component $\mathbf{p}$ captures the unique details of each modality (e.g., high-frequency noise in ACC), we apply a reconstruction objective. A modality-specific decoder D^m attempts to recover the original input $\mathbf{x}^m$ solely from $\mathbf{p}^m$. With $\mathcal{L}_{recon}$, $\mathbf{p}$ retains all information not captured by $\mathbf{s}$:

$$\mathcal{L}_{recon}(\mathbf{p}^m) = \|\mathbf{x}^m - D^m(\mathbf{p}^m)\|_2^2. \tag{5}$$

Whitening Regularization. To satisfy the independence derivation condition ($\text{Cov}(\mathbf{h}) \approx \mathbf{I}$), we impose a whitening penalty on the encoder outputs. We minimize the squared Frobenius norm of the off-diagonal elements of the covariance matrix of $\mathbf{h}^m$:

$$\mathcal{L}_{\text{white}}(\mathbf{h}^m) = \frac{\|\text{Cov}(\mathbf{h}^m) - \text{diag}(\text{Cov}(\mathbf{h}^m))\|_F^2}{d(d-1)} + \frac{1}{d}\sum_{j=1}^{d} \max\left(0,\, 1 - \sqrt{\text{Var}(\mathbf{h}_j^m)}\right), \tag{6}$$

Total Tokenizer Loss. The final training objective for the tokenizer aggregates these components across all modalities:

$$\mathcal{L}_{tok} = \sum_{m \in \mathcal{M}} \Big(\mathcal{L}_{align}(\mathbf{s}^m) + \lambda_{recon}\mathcal{L}_{recon}(\mathbf{p}^m) + \lambda_{white}\mathcal{L}_{white}(\mathbf{h}^m) \Big), \tag{7}$$

where λ_{recon} and λ_{white} balance the specific reconstruction and whitening constraints.

2.4 Missing-Aware MoE Backbone

Architecture. Wearable sensors exhibit intrinsic physical and semantic entanglements. For example, interpreting EDA activation often requires temperature signals to distinguish between physical activity and stress. However, when the model learns to handle modality missingness, it tends to weaken cross-modal dependencies. In traditional dense architectures, using a shared set of parameters to learn both objectives leads to modality conflict, preventing the model from effectively capturing cross-modal entanglements. To address this, we incorporate a MoE [29] block at the early layer of the backbone. Each expert autonomously learns specialized combinations of modalities. When

certain modalities are missing, unaffected experts naturally gain greater influence, ensuring globally optimal representations under the current conditions without compromising for the missing modality patterns. Meanwhile, conventional approaches to modality missingness often rely on zero-padding or special tokens, which introduce noise and degrade model performance. Instead, we discard missing modalities to avoid such interference. To ensure the router is aware of each token's modality origin and semantic role, we define 8 distinct learnable partition embeddings corresponding to all combinations of the $M = 4$ modalities and the 2 disentangled component types ($\mathbf{s}$ vs. $\mathbf{p}$). Specifically, for each observed token representation $\mathbf{x}_i \in \{\mathbf{s}_t^m, \mathbf{p}_t^m\}$ ($m \in \mathcal{M}_{obs}$) and its corresponding partition embedding $\mathbf{e}_i$, the input to the MoE backbone is formed via element-wise addition: $\mathbf{x}_i + \mathbf{e}_i$. Unlike mask tokens in MAE that act as shared placeholders to fill in missing positions for reconstruction, our partition embeddings are injected strictly into the observed tokens to explicitly inform the MoE router of their modality identities and semantic roles. By inheriting the structural inductive bias from the Orthogonal Tokenizer, this design simplifies the decision boundary for the router, significantly reducing optimization difficulty and allowing the backbone to focus on robust inference.

Optimization Objective. The backbone is tasked with integrating within-modality temporal dependencies and cross-modal interactions across the observed tokens to produce a global contextual representation. Given the difficulty of augmenting human physiological signals, we employ masked reconstruction for self-supervised learning. However, reconstructing the raw signals of missing modalities forces the model to predict modal-specific details that are statistically unpredictable from the remaining context. To avoid this and mitigate modality competition arising from variations in frequency and complexity, we perform masked reconstruction on the latent embeddings [30, 31] ($\mathbf{s}$ and $\mathbf{p}$) with two complementary masking strategies as the optimization objectives.

Random Masking ($\mathcal{L}_{rand}$). We randomly mask [32] a proportion p_r of temporal patches for all available modalities. The model is required to reconstruct both the shared ($\mathbf{s}$) and specific ($\mathbf{p}$) embeddings for these masked patches, as the full information is theoretically recoverable from the context. Let $\mathcal{M}$ denote the set of modalities, and $\mathbb{M}_t^m \in \{0, 1\}$ be the binary mask indicator for the patch at time t of modality m. The loss is defined as:

$$\mathcal{L}_{rand} = \sum_{m \in \mathcal{M}} \sum_{t=1}^{T_m} \mathbb{M}_t^m \cdot \left(\|\hat{\mathbf{s}}_t^m - \mathbf{s}_t^m\|_2^2 + \|\hat{\mathbf{p}}_t^m - \mathbf{p}_t^m\|_2^2 \right). \tag{8}$$

Structural Masking ($\mathcal{L}_{struct}$). To simulate missing modalities, we randomly drop 1-3 entire modalities at the sample level. The frozen tokenizer first provides target tokens for all modalities, after which the selected missing-modality tokens are excluded from the backbone input. To prevent the model from over-focusing on absent sensors, we also apply random patch masking to the remaining observed modalities. When a single modality is missing ($|\mathcal{M}_{miss}| = 1$), we reconstruct only its modal-shared component ($\mathbf{s}^m$), while explicitly excluding the non-identifiable modal-specific component ($\mathbf{p}^m$). However, when 2 or 3 modalities are missing, we skip their reconstruction entirely (neither $\mathbf{s}^m$ nor $\mathbf{p}^m$), as even their shared components may be uninferable from the limited observed context. For the observed modalities ($m \in \mathcal{M}_{obs}$), the standard random masking loss applies. The structural objective is formulated as:

$$\mathcal{L}_{struct} = \mathbb{I}(|\mathcal{M}_{miss}| = 1) \cdot \left(\sum_{m \in \mathcal{M}_{miss}} \sum_{t=1}^{T_m} \|\hat{\mathbf{s}}_t^m - \mathbf{s}_t^m\|_2^2 \right) + \sum_{m \in \mathcal{M}_{obs}} \sum_{t=1}^{T_m} \mathbb{M}_t^m \cdot \left(\|\hat{\mathbf{s}}_t^m - \mathbf{s}_t^m\|_2^2 + \|\hat{\mathbf{p}}_t^m - \mathbf{p}_t^m\|_2^2 \right), \tag{9}$$

where $\mathbb{I}(\cdot)$is the indicator function. Note that although the $\mathbf{s}$ of missing modality can be inferred from other modalities, we still reconstruct $\mathbf{s}$ in our objective. This encourages spatial consistency and enables the model to learn to handle missing modalities effectively.

Total Backbone Loss. The final optimization objective is the weighted sum of these two components:

$$\mathcal{L}_{backbone} = \mathcal{L}_{rand} + \lambda_{struct} \mathcal{L}_{struct}, \tag{10}$$

where $\hat{s}$ and $\hat{p}$ denote the reconstructed embeddings predicted by the backbone, and s, p are the ground-truth target embeddings frozen from the Orthogonal Tokenizer.

3 Experiments

3.1 Experiment Setup

Model Configuration. VCR uses a two-stage architecture (detailed in Appendix I) with 50.7M parameters at inference: 16.94M in the orthogonal tokenizer and 33.74M in the missing-aware MoE

Table 1: **Full Modality Results.** Best results are marked in **bold**. Second best results are underlined.

	WESAD			AAUWSS			DaLiA			WEEE	
Method	↑F_1	↑BAcc	↑ κ	↑F_1	↑BAcc	↑ κ	↑F_1	↑BAcc	↑ κ	↓MAE	↑R^2
Supervised											
ResNet	.610±.069	.635±.058	.561±.054	.236±.021	.254±.019	.110±.044	.733±.004	.723±.006	.656±.008	3.529±.123	.384±.049
ViT	.600±.041	.636±.020	.588±.029	.251±.095	.268±.088	.169±.170	.711±.016	.708±.012	.643±.010	3.677±.154	.307±.062
FlexMoE	.555±.030	.612±.038	.609±.070	.235±.022	.246±.030	.098±.050	.711±.006	.697±.008	.632±.007	3.039±.203	.470±.073
MAESTRO	.577±.011	.576±.011	.463±.019	.277±.002	.295±.005	.126±.015	.723±.003	.711±.005	.651±.006	3.643±.439	.386±.115
Ours (FT)	**.707**±.079	**.705**±.078	**.630**±.100	**.289**±.009	**.304**±.021	**.252**±.020	**.757**±.011	**.753**±.008	**.681**±.017	**2.633**±.109	**.668**±.030
Pretrained											
LSM	.580±.001	**.651**±.002	**.666**±.005	.222±.022	.231±.015	.129±.058	.728±.004	.710±.005	.645±.007	3.943±.079	.287±.026
LSM-2	.533±.008	.559±.010	.515±.018	.281±.007	**.316**±.005	**.339**±.007	.714±.002	.699±.002	.633±.005	3.711±.010	.411±.003
SimCLR	.510±.017	.514±.028	.354±.017	.230±.005	.238±.003	.113±.028	.621±.013	.620±.023	.504±.014	8.572±1.202	-3.796±2.688
MSN	.512±.003	.569±.003	.522±.006	.206±.003	.225±.006	.179±.025	.728±.003	.711±.003	.649±.003	3.970±.004	.297±.002
CMC	.563±.008	.595±.014	.534±.026	.258±.005	.266±.005	.213±.024	.686±.005	.678±.007	.618±.006	3.925±.295	.285±.076
MF-CLR	.506±.001	.470±.002	.297±.003	.227±.004	.221±.001	.069±.002	.406±.002	.339±.001	.297±.001	4.324±.038	.148±.018
FOCAL	.584±.004	.579±.011	.480±.025	.253±.003	.256±.003	.122±.008	.750±.004	.741±.004	**.688**±.005	3.769±.213	.370±.076
Ours (LP)	**.640**±.042	.639±.033	.590±.037	**.287**±.006	.290±.007	.248±.025	**.752**±.003	**.746**±.004	.673±.004	**2.606**±.067	**.693**±.018

backbone. All datasets are collected using the widely deployed Empatica E4 wristband, which records PPG at 64 Hz, EDA at 4 Hz, ACC at 32 Hz (3 axes), and TEMP at 4 Hz. We segment each stream into non-overlapping 10 s windows and use a 1 s patch size. To encode sensors with varying sampling rates and channel counts into a unified token space, each 10-second input window is first divided into $T = 10$ non-overlapping 1-second patches for every modality (i.e., 64-d for PPG, 3×32-d for 3-axis ACC, and 4-d for EDA and TEMP). A modality-specific 1D ResNet encoder then maps each raw patch into a token embedding $\tilde{\mathbf{h}} \in \mathbb{R}^d$ ($d = 256$), yielding an aligned sequence length and embedding dimension across all modalities.

Datasets and Tasks. We use publicly available datasets for pretraining and downstream evaluation, with no subject or dataset overlap between the two stages. Pretraining uses three daily-life datasets: BIG IDEAs [33], UE4W [34], and the non-cannabis-user subset of CAN-STRESS [35], totaling 3,565 hours from 56 subjects. Downstream evaluation includes WESAD [1] for 3-class emotion classification (15 subjects), DaLiA [36] for 8-class activity classification (15 subjects), AAUWSS [37] for 5-class sleep-stage classification (13 subjects), and WEEE [38] for VO_2 prediction (17 subjects). We adopted a subject-wise 5-fold cross-validation. Specifically, neither the baselines nor VCR have been exposed to the modality missingness in the downstream benchmark. Supervised training, fine-tuning, and linear probing are all conducted under the full-modal setting. During evaluation for a specific missing modality pattern, the designated modalities are removed from all test samples.

Baselines and Metrics. We compare VCR with 4 supervised baselines and 7 pretrained baselines, each configured with ∼50M parameters for capacity-matched comparison. The supervised baselines include ResNet [39], ViT [40], FlexMoE [22], and MAESTRO [12]. The pretrained baselines include SimCLR [28], MSN [20], CMC [19], MF-CLR [41], and FOCAL [26], which are based on contrastive learning, as well as LSM [18] and LSM-2 [11], which are masking-based methods specifically designed for wearable signals. Among them, FlexMoE and LSM-2 are designed to handle modality-missing scenarios. For evaluation, we employ Macro-F1, Balanced Accuracy (BAcc), and Cohen's Kappa (κ) for classification, and Mean Absolute Error (MAE) alongside R^2 for regression. We report results under two standard transfer settings. **FT** denotes supervised end-to-end fine-tuning of the pretrained model on the downstream task. **LP** denotes linear probing, where the pretrained encoder is frozen, and only a linear classifier/regressor is trained on top.

3.2 Experimental Results

Results on Full Modalities. Table 1 compares VCR with existing baselines when all modalities are available. VCR consistently outperforms both supervised and pretrained baselines. For classification, VCR improves macro-F1 by at least 0.097 on WESAD, 0.012 on AAUWSS, and 0.024 on DaLiA over the best supervised baselines, and by at least 0.056, 0.006, and 0.002 on the respective datasets over the best pretrained baselines. For regression, VCR reduces MAE by at least 0.406 and 1.105, while improving R^2 by 0.198 and 0.282 over the best supervised and pretrained baselines on WEEE, respectively. These results indicate that VCR learns transferable representations across diverse health monitoring tasks. Importantly, this full-modality performance indicates that robustness to missingness does not come at the cost of representation quality when all sensors are available; instead, preserving both modality-shared semantics and modality-specific residuals benefits downstream performance even when all sensors are available.

Results under Single Modality Missing. Table 2 presents the model performance when each modality is individually missing at test time across the four downstream tasks. Among the 32 primary metric comparisons (i.e., Macro-F1 for classification and MAE for regression $\times$ 4 missing-modality scenarios $\times$ 2 baseline types), VCR outperforms the baselines in 25 cases. Across all 88 total reported

Table 2: **Robustness under Single Modality Missing.** Comparison of performance when PPG, EDA, ACC, or TEMP is missing. Best results are marked in **bold**, second best are underlined.

	WESAD			AAUWSS			DaLiA			WEEE	
Method	↑F_1	↑BAcc	↑ κ	↑F_1	↑BAcc	↑ κ	↑F_1	↑BAcc	↑ κ	↓MAE	↑R^2
PPG Missing											
Supervised											
ResNet	.618±.064	.643±.053	.574±.050	.219±.021	.259±.017	.099±.035	**.735**±.007	.728±.007	.661±.012	3.405±.132	.415±.038
ViT	.590±.037	.627±.018	.578±.027	.247±.067	.271±.064	.153±.148	.686±.022	.685±.014	.629±.023	3.630±.103	.306±.055
FlexMoE	.562±.029	.621±.036	.625±.067	.192±.029	.229±.027	.052±.026	.717±.010	.705±.015	.641±.011	3.837±.342	.222±.135
MAESTRO	.582±.019	.576±.020	.442±.025	.240±.016	.273±.011	.122±.024	.699±.009	.700±.008	.628±.008	4.218±.465	.190±.135
Ours (FT)	**.727**±.030	**.729**±.045	**.631**±.047	**.250**±.006	**.277**±.006	**.161**±.008	.716±.048	**.741**±.027	**.663**±.028	**3.152**±.232	**.541**±.061
Pretrained											
LSM	.369±.032	.418±.026	.211±.061	.181±.020	.208±.014	.028±.046	.622±.017	.645±.013	.540±.014	4.062±.165	.272±.047
LSM-2	.510±.002	.541±.003	.473±.005	.226±.008	**.296**±.020	**.203**±.026	.688±.022	.669±.024	.599±.023	4.514±.026	.122±.009
SimCLR	.507±.019	.511±.031	.349±.018	.230±.005	.237±.003	.114±.030	.621±.012	.620±.022	.506±.010	3.685±.132	.338±.059
MSN	.495±.002	.548±.002	.484±.004	.189±.001	.204±.001	.076±.006	.704±.023	.674±.022	.630±.023	4.008±.006	.297±.002
CMC	.501±.053	.541±.044	.409±.098	.190±.007	.210±.006	.074±.032	.283±.066	.357±.065	.248±.064	5.035±.802	-.114±.257
MF-CLR	.489±.002	.513±.014	.391±.027	.195±.011	.245±.002	.170±.012	.291±.001	.313±.001	.270±.001	4.281±.124	.250±.039
FOCAL	.347±.131	.402±.084	.130±.150	.176±.036	.222±.017	.052±.037	.109±.025	.162±.019	.033±.017	26.241±1.714	-19.995±2.733
Ours (LP)	**.626**±.042	**.635**±.027	**.585**±.024	**.251**±.006	.281±.008	.102±.008	**.706**±.039	**.721**±.031	**.639**±.028	**3.191**±.056	**.497**±.029
EDA Missing											
Supervised											
ResNet	.512±.038	.537±.031	.393±.036	.208±.013	.229±.016	.057±.026	.486±.020	.491±.014	.376±.022	6.223±.653	-.726±.334
ViT	.554±.041	**.593**±.037	**.475**±.055	.217±.090	.233±.072	.060±.152	.502±.024	.500±.016	.412±.012	5.848±.901	-.530±.451
FlexMoE	.356±.055	.398±.049	.156±.108	**.253**±.029	.261±.027	.089±.038	.553±.025	.576±.027	.475±.038	3.171±.234	.458±.075
MAESTRO	.484±.012	.480±.011	.311±.017	.223±.006	.243±.002	.028±.007	.670±.011	.670±.015	.569±.015	4.725±.620	.001±.213
Ours (FT)	**.567**±.048	.588±.038	.375±.317	.249±.009	**.278**±.010	**.110**±.014	**.711**±.040	**.684**±.049	**.581**±.044	**2.985**±.132	**.578**±.024
Pretrained											
LSM	.506±.014	.556±.019	.484±.034	.187±.017	.210±.009	.047±.047	.456±.026	.491±.018	.395±.020	6.291±.088	-.457±.046
LSM-2	.253±.008	.337±.005	.021±.013	.217±.008	.250±.003	**.192**±.027	.514±.018	.543±.011	.440±.022	5.422±.027	-.116±.007
SimCLR	.445±.025	.451±.026	.240±.038	.175±.022	.204±.017	.049±.056	.461±.036	.463±.046	.331±.043	6.201±.316	-.489±.163
MSN	.515±.007	.573±.010	**.520**±.018	.197±.003	.212±.004	.119±.016	.481±.010	.512±.009	.404±.010	6.553±.027	-.633±.014
CMC	.465±.062	.517±.065	.397±.134	.213±.008	.236±.017	.152±.016	.352±.091	.426±.094	.269±.085	10.225±4.296	-3.229±3.313
MF-CLR	.345±.004	.336±.007	.005±.017	.177±.001	.199±.000	-.003±.001	.151±.004	.139±.000	.018±.001	5.446±.069	-.193±.038
FOCAL	.436±.022	.442±.029	.231±.042	**.254**±.005	.261±.013	.095±.005	.677±.028	.666±.027	.575±.027	5.550±1.671	-.336±.748
Ours (LP)	**.585**±.053	**.600**±.046	.494±.053	.241±.009	**.268**±.010	.073±.014	**.721**±.027	**.701**±.027	**.600**±.030	**2.925**±.049	**.615**±.020
ACC Missing											
Supervised											
ResNet	.460±.033	.500±.035	.250±.094	.176±.049	.186±.058	-.009±.154	.255±.033	.354±.029	.237±.051	4.410±.497	-.033±.202
ViT	.472±.043	.523±.057	.388±.118	.200±.037	.209±.044	.050±.067	.268±.041	.346±.040	.200±.036	4.428±1.097	.000±.415
FlexMoE	.505±.012	.556±.020	.503±.026	.216±.023	.233±.015	.114±.046	.356±.052	.427±.025	.352±.084	3.045±.309	.497±.093
MAESTRO	.504±.028	.504±.026	.341±.033	**.244**±.018	**.273**±.012	**.172**±.039	.379±.011	.378±.011	.327±.010	3.468±.339	.455±.113
Ours (FT)	**.618**±.075	**.642**±.218	**.513**±.078	.239±.006	.269±.009	.170±.023	**.416**±.091	**.452**±.076	**.414**±.082	**2.797**±.094	**.603**±.026
Pretrained											
LSM	.561±.005	**.636**±.004	**.626**±.009	.221±.017	.233±.015	.143±.057	.355±.045	.387±.032	.337±.039	3.759±.160	.354±.058
LSM-2	.510±.006	.556±.007	.502±.016	.213±.018	**.317**±.012	.238±.040	.380±.014	.445±.012	**.422**±.024	3.411±.003	.479±.001
SimCLR	.245±.098	.350±.099	.026±.157	.133±.000	.200±.000	.000±.000	.038±.028	.139±.029	.019±.039	19.524±13.451	-16.090±19.537
MSN	.284±.022	.357±.011	.094±.042	.230±.003	.258±.008	.265±.025	.290±.040	.332±.030	.291±.027	6.270±.031	-1.036±.024
CMC	.483±.032	.519±.032	.370±.060	.239±.007	.254±.004	.198±.013	.317±.029	.387±.027	.342±.046	5.729±.679	-.521±.355
MF-CLR	.502±.004	.477±.002	.315±.006	.225±.002	.218±.001	.067±.001	.377±.005	.294±.001	.250±.002	4.329±.058	.147±.020
FOCAL	.598±.069	.626±.036	.447±.099	.211±.001	.226±.004	.125±.007	.249±.032	.321±.034	.240±.056	4.516±.769	.111±.276
Ours (LP)	**.610**±.030	.608±.023	.514±.028	**.272**±.006	.270±.009	**.300**±.023	**.428**±.017	**.458**±.014	.405±.017	**3.358**±.041	**.545**±.017
TEMP Missing											
Supervised											
ResNet	.490±.057	.509±.056	.338±.065	.198±.016	.220±.021	.033±.029	.571±.035	.567±.025	.489±.040	3.792±.225	.338±.059
ViT	.553±.058	.580±.044	.486±.063	.235±.049	.256±.023	.130±.107	.504±.026	.510±.024	.420±.020	3.913±.242	.278±.107
FlexMoE	.531±.021	.592±.027	**.535**±.039	.208±.019	.256±.030	.099±.055	.516±.050	.518±.045	.450±.048	3.109±.278	.478±.086
MAESTRO	.563±.027	.572±.024	.471±.035	**.271**±.007	**.294**±.012	.105±.008	.621±.018	**.654**±.012	.537±.015	4.606±.592	.060±.183
Ours (FT)	**.566**±.158	**.608**±.024	.510±.063	.234±.034	.270±.023	**.132**±.059	**.648**±.014	.640±.017	**.553**±.013	**2.722**±.169	**.639**±.040
Pretrained											
LSM	.536±.015	**.604**±.020	**.573**±.034	.210±.017	.222±.012	.088±.040	.566±.007	.550±.011	.477±.011	3.869±.128	.297±.045
LSM-2	**.556**±.019	.570±.016	.514±.023	.227±.009	.230±.010	.127±.025	.458±.108	.475±.079	.424±.077	3.842±.012	.378±.003
SimCLR	.440±.023	.444±.023	.251±.039	.161±.013	.167±.014	-.048±.037	.401±.047	.409±.050	.327±.042	4.112±.135	.222±.059
MSN	.518±.013	.556±.008	.487±.010	.197±.001	.236±.007	.128±.014	.389±.021	.397±.016	.342±.013	4.028±.002	.343±.001
CMC	.515±.015	.544±.022	.414±.037	.237±.011	**.281**±.010	.148±.030	.591±.019	.597±.016	.514±.023	3.862±.619	.273±.175
MF-CLR	.492±.003	.470±.002	.297±.003	.226±.004	.221±.001	.069±.002	.291±.005	.339±.001	.297±.001	4.752±.035	.009±.020
FOCAL	.534±.027	.541±.033	.433±.056	.232±.003	.235±.005	.045±.009	**.660**±.007	**.653**±.005	**.558**±.014	3.527±.465	.423±.158
Ours (LP)	.540±.022	.537±.012	.438±.023	**.258**±.004	.264±.004	**.193**±.014	.625±.026	.618±.011	.540±.015	**2.681**±.108	**.675**±.027

metrics, VCR achieves the best score in 61 cases and the second-best in 19 cases. We highlight a representative challenging setting where a critical sensor is missing. On DaLiA, removing ACC substantially degrades activity recognition for competing methods, whereas VCR maintains F1 scores of 0.416 (FT) and 0.428 (LP). Moreover, when ACC is missing, VCR 's margin over supervised baselines increases from 0.024 to 0.037, while its margin over pretrained baselines rises from 0.002 to 0.048, indicating that VCR degrades more gracefully when a key sensor is unavailable.

Results under Multiple Modality Missing. To evaluate robustness under severe sensor absence, we test all cases where 2 or 3 modalities are simultaneously missing, as detailed in Appendix C. Across 3 classification tasks (10 missing patterns and 30 evaluation settings), VCR ranks Top-1 in 20/30 and Top-2 in 26/30 against pretrained baselines, and Top-1 in 19/30 and Top-2 in 27/30 against supervised baselines. For regression, VCR ranks Top-1 in 7/10 (MAE) and 7/10 settings (R^2) against pretrained baselines, and achieves Top-1 in 8/10 (MAE) and 7/10 (R^2) against supervised methods. These results show that VCR remains robust across all multiple missing-modality settings. Beyond whole-window modality absence, VCR-FT achieves 116 wins and 2 ties against supervised baselines across 120 intermittent-dropout settings; VCR-LP achieves 115 wins and 1 tie against LSM-2 (Appendix D).

Case Study of Hallucination Effects. VCR remains robust when key modalities are missing, primarily because the Orthogonal Tokenizer enforces strict information boundaries, which enables the Missing-aware MoE Backbone to learn how to accurately recover missing modality information, rather than relying on uncertain guesses that may lead to hallucinations. We support this intuition

through a comparison with LSM-2 , which directly reconstructs the entire raw data including both modality-shared and uninferable modality-specific components when crucial modality is missing, leading to semantic inconsistency. To visualize these effects, we use the DaLiA dataset. We freeze the encoders of VCR and LSM-2 and train an auxiliary decoder (without masking) to map global embeddings back to raw signals. For a full-modality sample $X = \{\mathbf{x}^{PPG}, \mathbf{x}^{EDA}, \mathbf{x}^{ACC}, \mathbf{x}^{TEMP}\}$, we obtain the reference embedding $\mathbf{emb}$ by feeding X into the frozen encoder. We then drop ACC, run the full pipeline to reconstruct $\hat{\mathbf{x}}^{ACC}$, form $X' = \{\mathbf{x}^{PPG}, \mathbf{x}^{EDA}, \hat{\mathbf{x}}^{ACC}, \mathbf{x}^{TEMP}\}$, and compute $\mathbf{emb}'$ from X' using the same encoder. If the reconstruction introduces unsupported modality-specific details, the resulting representation should deviate from the original one beyond what is justified by the observed context. We therefore visualize the t-SNE distributions of $\mathbf{emb}$ and $\mathbf{emb}'$ to quantify the induced representation shift.

Prior work [42, 43] characterizes hallucinations as plausible or coherent outputs that are incorrect or unfaithful to the conditioning input. Consistent with this view, although LSM-2 achieves a lower reconstruction MAE than VCR (0.426 vs. 0.509), it induces a larger embedding shift under missing ACC (cosine similarity: 0.732 vs. 0.813) shown in Figure 3(a). For LSM-2, this shift is significantly correlated with the drop in true-class probability (Pearson $r = 0.561$, $p < 0.001$) in Figure 3(b). Thus, waveform-close reconstructions can still introduce harmful task-relevant semantic errors, consistent with hallucination-like completions.

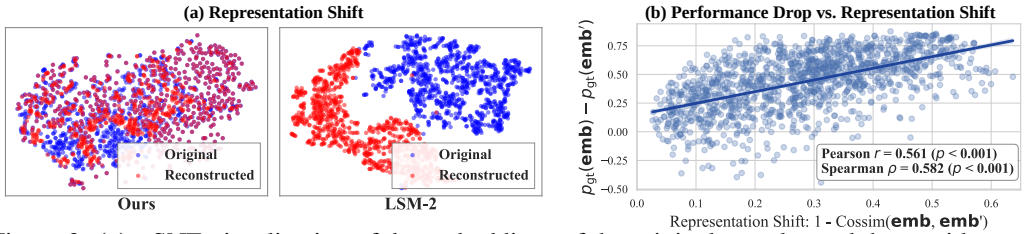

Figure 3: **(a)** t-SNE visualization of the embeddings of the original samples and those with reconstructed ACC signals. **(b)** Performance Drop vs. Representation Shift in LSM-2.

3.3 Analysis

Quantitative Assessment of Disentanglement. To quantify the independence between the modal-shared component $\mathbf{s}$ and the modal-specific component $\mathbf{p}$, we compute the Hilbert-Schmidt Independence Criterion (HSIC) using a Gaussian kernel. Experimental results show that the soft disentanglement baseline FOCAL [26], which relies on orthogonal loss, struggles to decouple low-frequency modalities such as EDA and TEMP. In contrast, VCR consistently achieves HSIC values between 0.003 and 0.005 across all modalities, with an average less than 40% of that of the baseline. Unlike typical methods that require explicit disentanglement losses between s and p, VCR achieves effective disentanglement purely through the geometric design of the tokenizer. Moreover, VCR demonstrates consistent robustness across different modalities, preventing high-frequency signals from dominating the disentanglement process, which is important for wearable data with large variations in sampling rate and spectral content.

Table 3: Independence evaluation.

Method	HSIC (Gaussian Kernel) ↓				
	PPG	EDA	ACC	TEMP	**Average**
FOCAL	0.0045	0.0140	0.0076	0.0174	0.0108
Ours	**0.0039**	**0.0046**	**0.0033**	**0.0048**	**0.0042**

Expert Specialization and Capacity Analysis. Figure 4(a) visualizes the different modality-to-expert activation patterns learned by the router. Taking EDA as an example, Experts 1–3 capture different dependencies between EDA and the other three modalities, while Expert 7 primarily handles cases where other modalities are missing. Figure 4(b) further analyzes the effect of the number of experts on DaLiA. Across different modality missingness scenarios, using 8 experts yields the best performance. When the number of experts is set to 1, the backbone degenerates into a dense architecture and performance drops, highlighting the necessity of the MoE structure for handling modality missingness effectively.

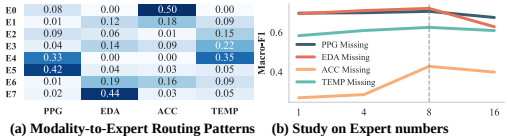

Figure 4: Analysis of the MoE Mechanism.

Ablation Studies. We conduct three groups of ablations to assess the contribution of each design choice. 1) replacing the structural disentanglement in the orthogonal tokenizer with a FOCAL-type

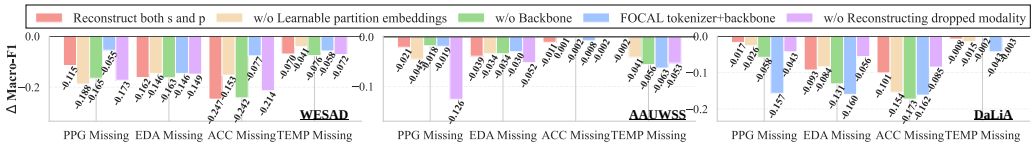

Figure 5: Macro-F1 drop of different ablations compared to VCR.

soft disentanglement based on orthogonal loss; 2) removing the missing-aware MoE backbone or replacing the learnable partition embedding with a fixed sinusoidal encoding; and 3) modifying the objective for structurally missing modalities from reconstructing only shared components to reconstructing both or reconstructing neither of the shared and modality-specific components. Figure 5 reports the **Macro-F1 drop** of each ablated variant relative to the full VCR pipeline. Each ablation causes a clear degradation across all three tasks and all four missing-modality settings (with only 1 exception among 60 metrics). Overall, the results suggest that soft disentanglement allows leakage between shared and modality-specific factors, weakening information separation; the missing-aware MoE backbone is necessary to adapt computation to heterogeneous modality availability, whereas fixed encodings are insufficient to capture modality-dependent semantic relations (e.g., the semantic affinity between EDA and TEMP may differ from that between EDA and ACC); and reconstructing modality-specific components for missing modalities incentivizes the model to generate unobserved content, increasing the risk of hallucination-like completion. Despite $\mathbf{s}$ being partially predictable from the observed modalities, explicitly reconstructing $\mathbf{s}$ improves robustness relative to skipping reconstruction, indicating a benefit from enforcing representation consistency.

Cross-device Evaluation. We further conduct a cross-device evaluation on WildPPG [44] using PPG signals collected from an unseen device during pretraining. Following the original study, we adopt heart rate prediction as the downstream task and report MAE, RMSE, and Pearson correlation coefficient ρ. As shown in Table 4, Frozen VCR with a linear probe achieves MAE/RMSE/ρ of 13.92/17.52/14.32, compared with 31.79/34.60/8.19 for LSM-2. As a contextual comparison, VCR outperforms 5/7 WildPPG-pretrained SSL methods reported in the original paper and approaches the best MAE result of 10.55.

Table 4: **Cross-device Performance Comparison.** With the better results highlighted in bold.

Model	Available modalities	MAE (bpm) ↓	RMSE (bpm) ↓	Pearson rho (%) ↑
TS2Vec (pretrained with WildPPG)†	PPG	10.55±0.37	16.52±0.39	26.31±0.25
Ours	PPG	**13.922±0.203**	**17.517±0.221**	**14.32±0.18**
LSM2	PPG	31.788±1.161	34.595±1.134	8.19±0.06

4 Conclusion

We presented VCR, a missingness-robust pretraining framework for learning valid representations for multimodal wearable signals. VCR utilizes a geometric orthogonal tokenizer to structurally decompose multimodal inputs into shared semantics and modality-specific residuals, which preserves information integrity while providing a rigorous inductive bias. Together with a missing-aware MoE backbone, VCR enables experts to adapt to diverse modality combinations without being forced to hallucinate unobservable specific details. Overall, VCR demonstrates significant advantages and strong generalization across multiple health monitoring tasks in both full-modality and modality-missing scenarios. VCR paves the way for general-purpose pretraining under missing modalities and highlights the importance of prioritizing representation validity, which is often overlooked when learning from incomplete observations.

5 Limitations

Our evaluation could be further expanded to larger cohorts and a broader range of devices. Comparisons between frozen-tokenizer backbone pretraining and joint tokenizer–backbone pretraining, along with controlled comparisons with separately pretrained unimodal models, remain future work.

Acknowledgments and Disclosure of Funding

We sincerely thank the anonymous reviewers for their insightful comments. This work was supported in part by the National Natural Science Foundation of China under Grant-62602541, Hong Kong RGC under ECS-26205226. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the funding agencies or others.

References

[1] 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*, pages 400–408, 2018.

[2] Rummana Bari, Md Mahbubur Rahman, Nazir Saleheen, Megan Battles Parsons, Eugene H Buder, and Santosh Kumar. Automated detection of stressful conversations using wearable physiological and inertial sensors. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4(4):1–23, 2020.

[3] Bing Zhai, Ignacio Perez-Pozuelo, Emma AD Clifton, Joao Palotti, and Yu Guan. Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–33, 2020.

[4] Abhay Koushik, Judith Amores, and Pattie Maes. Real-time sleep staging using deep learning on a smartphone for a wearable eeg. *arXiv preprint arXiv:1811.10111*, 2018.

[5] Taoran Sheng and Manfred Huber. Weakly supervised multi-task representation learning for human activity analysis using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–18, 2020.

[6] Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. Giobalfusion: A global attentional deep learning framework for multisensor information fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–27, 2020.

[7] 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 *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pages 15516–15541. PMLR, 2025.

[8] Oura Health Oy. Oura Ring 4. URL `https://support.ouraring.com/hc/en-us/articles/33045011508115-Oura-Ring-4`.

[9] Google. Fitbit Sense 2 Technical Specs. URL `https://store.google.com/us/product/fitbit_sense_2_specs`.

[10] Apple Inc. Apple Watch Series 9: Tech Specs. URL `https://support.apple.com/en-us/111833`.

[11] Maxwell A. Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A. Tailor, Ahmed Metwally, A. Ali Heydari, Yuwei Zhang, Jake Garrison, Samy Abdel-Ghaffar, Xuhai Xu, Ken Gu, Jacob Sunshine, Ming-Zher Poh, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Mark Malhotra, Shwetak Patel, Yuzhe Yang, James M. Rehg, Xin Liu, and Daniel McDuff. Lsm-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

[12] Payal Mohapatra, Yueyuan Sui, Akash Pandey, Stephen Xia, and Qi Zhu. Maestro: Adaptive sparse attention and robust learning for multimodal dynamic time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[13] Qi Shen, Junchang Xin, Bing T Dai, Shudi Zhang, and Zhiqiong Wang. Robust sleep staging over incomplete multimodal physiological signals via contrastive imagination. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:112025–112049, 2024.

[14] Siân Bladon, John Ainsworth, Roberto Cahuantzi, Matteo Cella, Richard J Drake, Emily Eisner, Richard Emsley, Sophie Faulkner, Kathryn Greenwood, Andrew Gumley, et al. Evaluating wearable devices for remote monitoring in psychosis: Pilot study nested within the connect cohort study. *JMIR Formative Research*, 10(1):e86049, 2026.

[15] 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.

[16] 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.

[17] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed A Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, et al. Sensorlm: Learning the language of wearable sensors. *Advances in neural information processing systems (NeurIPS)*, 2025.

[18] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam A. Tailor, Jacob Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[19] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision (ECCV)*, pages 776–794. Springer, 2020.

[20] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Computer Vision – ECCV 2022*, pages 456–473. Springer, 2022. doi: 10.1007/978-3-031-19821-2_26.

[21] 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.

[22] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-3135.

[23] Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: mixture-of-experts transformers for fleximodal fusion. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 67850–67900, 2024.

[24] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 343–351, 2016.

[25] Lun Huang, Qiang Qiu, and Guillermo Sapiro. Ssole: Rethinking orthogonal low-rank embedding for self-supervised learning. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

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

[27] 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)*, pages 12310–12320. PMLR, 2021.

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607, 2020.

[29] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

[30] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning (ICML)*, pages 1298–1312. PMLR, 2022.

[31] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 14668–14678, 2022.

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16000–16009, 2022.

[33] 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.

[34] 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*, pages 404–415. Springer, 2023.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[40] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[41] Jufang Duan, Wei Zheng, Yangzhou Du, Wenfa Wu, Haipeng Jiang, and Hongsheng Qi. Mf-clr: Multi-frequency contrastive learning representation for time series. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

[42] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882. IEEE, 2024.

[43] Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. Prevent the language model from being overconfident in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, pages 3456–3468, 2021.

[44] Manuel Meier, Berken Utku Demirel, and Christian Holz. Wildppg: a real-world ppg dataset of long continuous recordings. *Advances in Neural Information Processing Systems*, 37:2246–2266, 2024.

[45] Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[46] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning (ICML)*, pages 5606–5615. PMLR, 2021.

[47] 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.

[48] 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, pages 8980–8987, 2022.

[49] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[50] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, pages 16115–16152. PMLR, 2024.

[51] Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:29996–30025, 2023.

[52] 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)*, pages 2352–2359. International Joint Conferences on Artificial Intelligence Organization, 2021.

[53] Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[54] Zhenghong Lin, Yanchao Tan, Yunfei Zhan, Weiming Liu, Fan Wang, Chaochao Chen, Shiping Wang, and Carl Yang. Contrastive intra-and inter-modality generation for enhancing incomplete multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, pages 6234–6242, 2023.

[55] Shenghong Dai, Shiqi Jiang, Yifan Yang, Ting Cao, Mo Li, Suman Banerjee, and Lili Qiu. Babel: A scalable pre-trained model for multi-modal sensing via expandable modality alignment. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 240–253, 2025.

[56] Xiaomin Ouyang, Jason Wu, Tomoyoshi Kimura, Yihan Lin, Gunjan Verma, Tarek Abdelzaher, and Mani Srivastava. Mmbind: Unleashing the potential of distributed and heterogeneous data for multimodal learning in iot. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 491–503, 2025.

[57] Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[58] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:17117–17128, 2023.

[59] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:32971–32998, 2023.

[60] Andrea Hongn, Facundo Bosch, Lara Eleonora Prado, José Manuel Ferrández, and María Paula Bonomini. Wearable physiological signals under acute stress and exercise conditions. *Scientific Data*, 12(1):520, 2025.

A Related Works

Wearable / Physiological Foundation Models and Self-Supervised Pretraining. Self-supervised pretraining on large-scale wearable signals has enabled general-purpose representations across heterogeneous sensors [45, 18, 46]. Beyond wearable-specific pretraining, generic time-series SSL frameworks [47, 41, 48, 49, 50] have demonstrated strong label efficiency and transferability across diverse time-series domains, motivating the pursuit of general-purpose sensing representations. While such models demonstrate broad transfer [51, 52, 53] (e.g., imputation and downstream health/activity prediction), many formulations still train under fully observed inputs. While recent years have seen notable progress in incomplete multimodal learning, many existing works are tailored for specific tasks (e.g., recommendation systems [54] or sleep staging [13]), designed for particular modality combinations with a dominant junction modality [55], or focus primarily on handling modality missingness during training [56]. These approaches fall short of the goal of a modality-agnostic representation learning framework that can be fine-tuned for various health-related downstream tasks and effectively handle train-on-full, test-on-missing scenarios, especially in limited data settings. LSM-2 [11] with Adaptive and Inherited Masking (AIM) explicitly trains on naturally incomplete streams without pre-imputation, improving robustness to sensor dropouts. However, during training, for a missing modality that is entirely dropped, LSM-2 learns by reconstructing the raw data, which forces the model to generate unpredictable modality-specific details.

Supervised Methods for Missing Modalities. Recent routing/MoE architectures Flex-MoE [22] and FuseMoE [23] improve subset-robust inference by conditioning computation on observed modalities. MAESTRO [12] targets multimodal dynamic time series with arbitrary missing observations via symbolic tokenization, sparse cross-modal attention, and sparse MoE routing. These methods are difficult to transfer to new tasks, especially those with very limited labels [57, 58].

Shared-Specific Disentanglement, Orthogonality. Shared-specific factorization aims to separate cross-domain semantics from domain-specific variation [59], with early formulations such as DSN [24] combining task supervision, reconstruction, and soft separation constraints. In multimodal SSL for sensing, FOCAL [26] factorizes each modality into shared and private embeddings and encourages orthogonality while optimizing contrastive objectives. These methods rely on soft constraints, which fail to strictly disentangle the modalities and thus cannot prevent the model from generating inherently unpredictable modality-specific information.

B Downstream Dataset Details

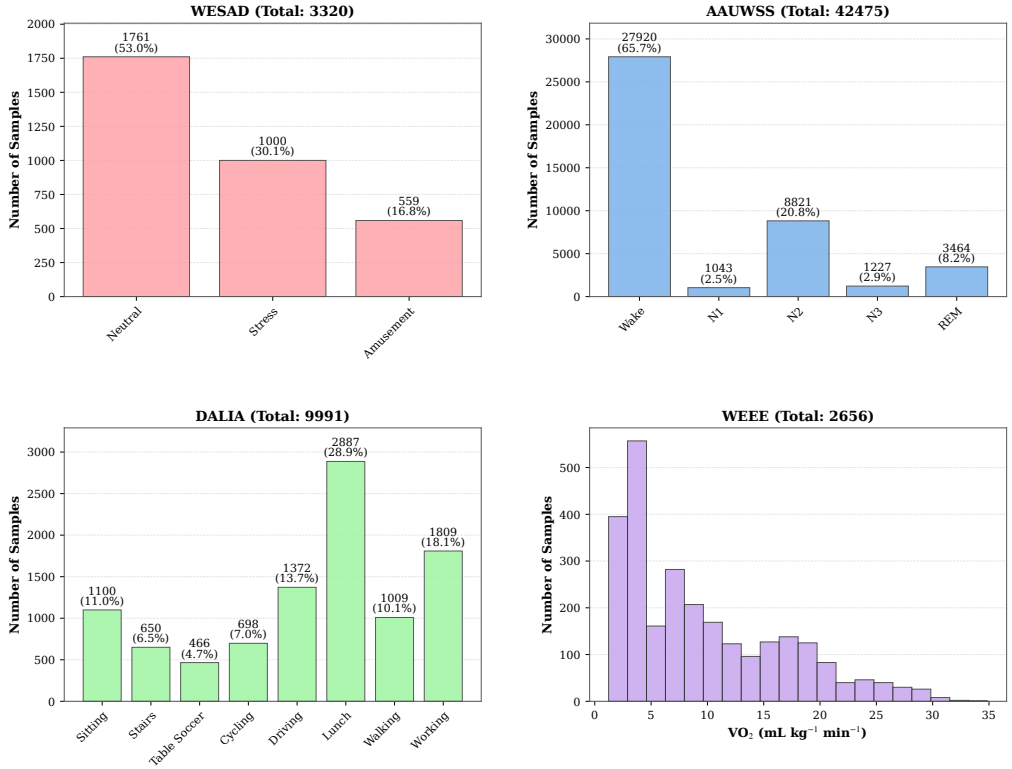

Figure 6: Distributions of Four Downstream Tasks (WESAD, AAUWSS, DaLiA, and WEEE).

C Evaluation with Multiple Simultaneous Missing Modalities

Table 5: **Multiple Modality Missing Results (Macro F1).** Best results in each category (Supervised/Pretrained) are marked in **bold**, second best are underlined. W=WESAD, A=AAUWSS, D=DALIA.

Missing Modality	Task	ResNet	ViT	FlexMoE	MAESTRO	**Ours (FT)**	LSM	LSM-2	SimCLR	MSN	CMC	MFCLR	FOCAL	**Ours (LP)**
PPG EDA	W	0.469	**0.512**	0.403	0.455	0.371	**0.451**	0.411	0.393	0.231	0.318	0.263	0.268	0.376
	A	0.180	0.200	**0.215**	0.196	0.208	0.182	0.170	**0.216**	0.212	0.193	0.173	0.157	0.212
	D	0.504	0.482	0.547	0.604	**0.720**	0.281	0.522	0.451	0.136	0.148	0.159	0.054	**0.660**
PPG ACC	W	0.597	0.529	0.530	**0.655**	0.560	0.420	0.530	0.209	0.514	0.420	0.444	0.352	**0.588**
	A	0.179	0.189	0.211	**0.282**	0.216	0.176	0.115	0.169	0.228	0.186	0.183	0.187	**0.284**
	D	0.216	0.171	0.354	0.255	**0.405**	0.168	0.401	0.032	**0.408**	0.087	0.302	0.054	0.399
PPG TEMP	W	0.466	0.566	0.564	0.592	**0.608**	0.522	**0.639**	0.493	0.231	0.497	0.479	0.293	0.582
	A	0.198	0.228	0.199	0.201	**0.229**	0.206	0.217	0.227	0.190	0.240	0.196	0.158	**0.303**
	D	0.558	0.509	0.483	0.577	**0.646**	0.388	0.526	0.400	0.399	0.189	0.228	0.064	**0.616**
EDA ACC	W	**0.399**	0.279	0.324	0.332	0.333	0.335	**0.418**	0.211	0.231	0.340	0.326	0.266	0.354
	A	0.158	0.190	0.252	0.204	**0.268**	0.182	0.190	0.169	0.228	**0.245**	0.175	0.181	0.204
	D	0.156	0.128	0.173	0.331	**0.352**	0.084	0.174	0.044	0.148	0.123	0.142	0.193	**0.306**
EDA TEMP	W	0.435	0.411	0.425	**0.481**	0.464	0.413	0.264	0.383	0.231	0.255	0.325	0.432	**0.460**
	A	0.161	0.201	0.201	0.236	**0.247**	0.225	0.173	0.224	0.173	0.150	0.178	0.226	**0.230**
	D	0.303	0.259	0.274	0.499	**0.561**	0.223	0.252	0.298	0.095	0.263	0.090	0.463	**0.493**
ACC TEMP	W	0.503	0.505	0.514	0.472	**0.563**	0.441	0.549	0.242	0.231	0.487	0.490	**0.561**	**0.561**
	A	0.134	0.165	0.193	0.240	**0.277**	0.235	0.230	0.165	0.183	0.258	0.225	0.194	**0.271**
	D	0.102	0.187	0.181	0.310	**0.437**	0.237	0.241	0.024	0.232	0.222	0.277	0.199	**0.387**
PPG EDA ACC	W	**0.408**	0.270	0.312	0.353	0.279	0.225	**0.448**	0.211	0.231	0.299	0.231	0.232	0.303
	A	0.172	0.185	0.220	0.221	**0.231**	0.173	0.185	0.168	0.183	0.180	0.173	0.177	**0.215**
	D	0.130	0.062	0.155	0.200	**0.213**	0.054	0.158	0.026	0.098	0.053	0.181	0.054	**0.183**
PPG EDA TEMP	W	0.439	0.414	0.418	**0.480**	0.471	0.399	0.421	0.383	0.231	0.250	0.258	0.239	**0.423**
	A	0.159	**0.193**	0.153	0.167	0.190	0.160	0.222	**0.226**	0.192	0.171	0.173	0.149	0.180
	D	0.307	0.282	0.268	0.426	**0.531**	0.199	0.258	0.300	0.201	0.103	0.065	0.054	**0.458**
PPG ACC TEMP	W	0.491	0.501	0.520	**0.575**	0.567	0.292	**0.646**	0.242	0.231	0.382	0.415	0.314	0.590
	A	0.150	0.140	0.178	**0.244**	0.231	0.173	0.151	0.165	0.270	0.203	0.199	0.180	**0.304**
	D	0.072	0.096	0.165	0.191	**0.262**	0.167	0.261	0.024	0.165	0.039	0.228	0.054	**0.268**
EDA ACC TEMP	W	0.151	0.196	0.257	0.357	**0.391**	0.304	0.260	0.173	0.231	0.216	0.308	0.253	**0.364**
	A	0.119	0.180	0.167	0.233	**0.242**	0.182	0.157	0.173	0.179	**0.230**	0.176	0.178	0.214
	D	0.037	0.066	0.090	0.238	**0.311**	0.082	0.066	0.024	0.039	0.083	0.092	0.133	**0.230**

Table 6: **WEEE Modality Missing Results (MAE ↓).** Best results in each category (Supervised/Pretrained) are marked in **bold**, second best are underlined. To further validate generalizability beyond classification, we introduced a new regression benchmark, WEEE (predicting VO_2), evaluated under all 10 possible missing patterns ($C_4^2 + C_4^3$). VCR achieves strong performance.

Missing Modality	ResNet	ViT	FlexMoE	MAESTRO	**Ours (FT)**	LSM	LSM-2	SimCLR	MSN	CMC	MFCLR	FOCAL	**Ours (LP)**
PPG EDA	5.782	6.451	4.038	5.981	**3.518**	6.654	7.838	6.196	6.728	11.839	4.938	24.576	**3.756**
PPG ACC	4.760	5.102	4.368	4.749	**4.028**	**3.907**	6.246	19.697	5.036	6.115	4.249	26.490	4.448
PPG TEMP	3.627	3.878	3.781	5.606	**3.278**	3.810	6.031	4.105	4.392	5.159	4.905	26.953	**3.382**
EDA ACC	5.125	3.955	3.106	3.945	**2.985**	5.263	4.658	20.444	17.254	7.631	5.524	6.613	**3.621**
EDA TEMP	4.494	6.112	3.529	6.687	**3.072**	5.945	5.090	7.243	5.598	9.957	6.153	4.926	**2.973**
ACC TEMP	4.260	4.391	2.958	3.848	**2.899**	3.741	3.678	21.312	4.018	6.205	4.782	3.754	**3.421**
PPG EDA ACC	5.988	5.046	**4.200**	6.538	4.545	5.352	6.917	20.536	11.968	9.373	**4.761**	24.825	4.764
PPG EDA TEMP	4.218	6.509	4.671	10.495	**4.081**	6.892	9.449	7.225	6.111	11.641	5.532	25.288	**4.178**
PPG ACC TEMP	4.495	5.232	**4.205**	4.673	5.058	**3.815**	14.385	21.308	4.322	6.619	4.861	27.203	4.642
EDA ACC TEMP	4.938	5.559	3.623	5.606	**2.982**	5.095	4.929	22.162	6.890	7.477	6.245	5.587	**3.596**

Table 7: **WEEE Modality Missing Results (R^2 ↑).** Best results in each category (Supervised/Pretrained) are marked in **bold**, second best are underlined. To further validate generalizability beyond classification, we introduced a new regression benchmark, WEEE (predicting VO_2), evaluated under all 10 possible missing patterns ($C_4^2 + C_4^3$). VCR achieves strong performance.

Missing Modality	ResNet	ViT	FlexMoE	MAESTRO	**Ours (FT)**	LSM	LSM-2	SimCLR	MSN	CMC	MFCLR	FOCAL	**Ours (LP)**
PPG EDA	-0.538	-0.892	0.098	-0.624	**0.423**	-0.581	-1.139	-0.487	-0.734	-4.939	0.042	-18.182	**0.240**
PPG ACC	-0.214	-0.353	0.119	-0.046	**0.222**	**0.348**	-0.578	-16.357	-0.282	-0.690	0.278	-20.681	0.144
PPG TEMP	0.377	0.289	0.300	-0.395	**0.487**	0.354	-0.512	0.223	0.237	-0.148	0.036	-21.087	**0.432**
EDA ACC	-0.250	0.245	0.492	0.294	**0.597**	-0.057	0.018	-17.426	-8.536	-1.581	-0.221	-0.925	**0.443**
EDA TEMP	0.134	-0.548	0.356	-0.943	**0.547**	-0.364	-0.071	-1.007	-0.155	-2.813	-0.501	-0.047	**0.597**
ACC TEMP	0.073	0.059	0.539	0.312	**0.576**	0.378	0.403	-18.945	0.303	-0.694	-0.000	0.380	**0.529**
PPG EDA ACC	-0.842	-0.316	**0.210**	-0.861	0.161	-0.091	-0.889	-17.565	-3.929	-3.061	**0.127**	-19.198	-0.021
PPG EDA TEMP	**0.227**	-0.730	-0.059	-3.865	0.184	-0.678	-2.200	-0.995	-0.373	-4.546	-0.182	-19.038	**0.028**
PPG ACC TEMP	-0.050	-0.326	**0.250**	-0.074	-0.115	**0.344**	-7.101	-18.911	0.239	-0.886	0.080	-21.557	0.078
EDA ACC TEMP	-0.124	-0.269	0.310	-0.508	**0.612**	-0.012	-0.196	-20.357	-0.834	-1.327	-0.539	-0.421	**0.428**

D Intermittent Missingness Evaluation

Beyond entire-modality missingness, another common scenario is intermittent missingness, where certain modalities become unavailable for short periods of time. To evaluate the robustness of VCR under this scenario, we conduct a contiguous within-window dropout evaluation. Specifically, for each nonempty modality subset, we mask a contiguous 2-second (20%) or 5-second (50%) segment within the input window, without exposing the models to such dropout patterns during training or finetuning. We compare VCR against three missingness-aware baselines: two supervised methods, FlexMoE and MAESTRO, and an SSL-based method, LSM-2. Across 120 evaluation settings, VCR-FT achieves the best performance in 116 settings and ties for the best in 2, while VCR-LP outperforms LSM-2 in 115 settings and ties in 1. These results demonstrate the robustness of VCR to short, temporally localized modality missingness, extending its effectiveness beyond complete whole-window modality absence.

Table 8: **Intermittent Drop Results for Supervised Methods**. FT is compared with the supervised baselines FlexMoE and MAESTRO under each missing-modality combination and drop ratio. Best results are marked in **bold**.

		WESAD $\uparrow F_1$			AAUWSS $\uparrow F_1$			DALIA $\uparrow F_1$			WEEE $\downarrow$MAE		
Missing modalities	Drop	FlexMoE	MAESTRO	Ours FT	FlexMoE	MAESTRO	Ours FT	FlexMoE	MAESTRO	Ours FT	FlexMoE	MAESTRO	Ours FT
PPG	20%	.582$_{\pm.030}$	.578$_{\pm.012}$	**.622**$_{\pm.056}$	.231$_{\pm.023}$	.270$_{\pm.011}$	**.289**$_{\pm.018}$	.714$_{\pm.008}$	.730$_{\pm.003}$	**.763**$_{\pm.006}$	3.147$_{\pm.207}$	3.680$_{\pm.434}$	**2.690**$_{\pm.107}$
	50%	.586$_{\pm.029}$	.574$_{\pm.018}$	**.627**$_{\pm.064}$	.219$_{\pm.019}$	.261$_{\pm.015}$	**.277**$_{\pm.020}$	.714$_{\pm.007}$	.731$_{\pm.002}$	**.768**$_{\pm.007}$	3.394$_{\pm.260}$	3.776$_{\pm.417}$	**2.800**$_{\pm.137}$
EDA	20%	.570$_{\pm.028}$	.559$_{\pm.016}$	**.601**$_{\pm.054}$	.230$_{\pm.019}$	.276$_{\pm.006}$	**.291**$_{\pm.016}$	.721$_{\pm.009}$	.723$_{\pm.005}$	**.765**$_{\pm.008}$	3.031$_{\pm.202}$	3.785$_{\pm.476}$	**2.613**$_{\pm.096}$
	50%	.548$_{\pm.023}$	.527$_{\pm.011}$	**.563**$_{\pm.049}$	.226$_{\pm.013}$	.262$_{\pm.014}$	**.284**$_{\pm.016}$	.726$_{\pm.009}$	.713$_{\pm.008}$	**.767**$_{\pm.010}$	2.990$_{\pm.176}$	4.078$_{\pm.533}$	**2.643**$_{\pm.100}$
ACC	20%	.568$_{\pm.028}$	.581$_{\pm.005}$	**.621**$_{\pm.053}$	.238$_{\pm.025}$	.270$_{\pm.005}$	**.293**$_{\pm.017}$	.705$_{\pm.007}$	.703$_{\pm.006}$	**.750**$_{\pm.008}$	2.905$_{\pm.156}$	3.423$_{\pm.399}$	**2.647**$_{\pm.097}$
	50%	.555$_{\pm.026}$	.589$_{\pm.016}$	**.615**$_{\pm.029}$	.239$_{\pm.020}$	.252$_{\pm.011}$	**.288**$_{\pm.020}$	.649$_{\pm.039}$	.638$_{\pm.009}$	**.724**$_{\pm.010}$	2.820$_{\pm.172}$	3.194$_{\pm.358}$	**2.662**$_{\pm.097}$
TEMP	20%	.591$_{\pm.036}$	.575$_{\pm.012}$	**.619**$_{\pm.064}$	.241$_{\pm.023}$	.274$_{\pm.002}$	**.294**$_{\pm.015}$	.701$_{\pm.005}$	.713$_{\pm.004}$	**.748**$_{\pm.010}$	3.029$_{\pm.203}$	3.756$_{\pm.459}$	**2.596**$_{\pm.083}$
	50%	.604$_{\pm.031}$	.571$_{\pm.021}$	**.621**$_{\pm.070}$	.233$_{\pm.016}$	.274$_{\pm.005}$	**.297**$_{\pm.017}$	.680$_{\pm.007}$	.689$_{\pm.010}$	**.721**$_{\pm.010}$	3.035$_{\pm.213}$	4.012$_{\pm.496}$	**2.556**$_{\pm.064}$
PPG+EDA	20%	.573$_{\pm.027}$	.566$_{\pm.016}$	**.604**$_{\pm.056}$	.227$_{\pm.027}$	.273$_{\pm.009}$	**.288**$_{\pm.018}$	.724$_{\pm.008}$	.730$_{\pm.006}$	**.770**$_{\pm.006}$	3.111$_{\pm.209}$	3.832$_{\pm.473}$	**2.667**$_{\pm.108}$
	50%	.551$_{\pm.019}$	.528$_{\pm.017}$	**.565**$_{\pm.056}$	.221$_{\pm.015}$	.260$_{\pm.019}$	**.264**$_{\pm.019}$	.730$_{\pm.011}$	.720$_{\pm.011}$	**.782**$_{\pm.006}$	3.324$_{\pm.232}$	4.270$_{\pm.537}$	**2.805**$_{\pm.149}$
PPG+ACC	20%	.570$_{\pm.027}$	.594$_{\pm.011}$	**.631**$_{\pm.052}$	.236$_{\pm.020}$	.275$_{\pm.005}$	**.291**$_{\pm.017}$	.704$_{\pm.005}$	.714$_{\pm.007}$	**.756**$_{\pm.010}$	3.000$_{\pm.181}$	3.414$_{\pm.379}$	**2.695**$_{\pm.113}$
	50%	.559$_{\pm.020}$	**.635**$_{\pm.010}$	.628$_{\pm.042}$	.221$_{\pm.012}$	**.282**$_{\pm.014}$	.281$_{\pm.020}$	.638$_{\pm.067}$	.665$_{\pm.005}$	**.730**$_{\pm.011}$	3.159$_{\pm.243}$	3.278$_{\pm.335}$	**2.898**$_{\pm.174}$
PPG+TEMP	20%	.594$_{\pm.032}$	.582$_{\pm.013}$	**.626**$_{\pm.068}$	.232$_{\pm.023}$	.269$_{\pm.010}$	**.291**$_{\pm.017}$	.705$_{\pm.006}$	.717$_{\pm.007}$	**.754**$_{\pm.008}$	3.134$_{\pm.212}$	3.795$_{\pm.456}$	**2.658**$_{\pm.093}$
	50%	.611$_{\pm.033}$	.591$_{\pm.022}$	**.633**$_{\pm.088}$	.204$_{\pm.019}$	.254$_{\pm.011}$	**.279**$_{\pm.018}$	.681$_{\pm.007}$	.700$_{\pm.006}$	**.734**$_{\pm.013}$	3.332$_{\pm.279}$	4.167$_{\pm.476}$	**2.727**$_{\pm.094}$
EDA+ACC	20%	.561$_{\pm.025}$	.566$_{\pm.016}$	**.607**$_{\pm.047}$	.236$_{\pm.020}$	.268$_{\pm.004}$	**.291**$_{\pm.016}$	.714$_{\pm.006}$	.710$_{\pm.004}$	**.757**$_{\pm.010}$	2.875$_{\pm.152}$	3.542$_{\pm.441}$	**2.611**$_{\pm.092}$
	50%	.531$_{\pm.027}$	.541$_{\pm.022}$	**.558**$_{\pm.035}$	.244$_{\pm.018}$	.242$_{\pm.011}$	**.284**$_{\pm.016}$	.672$_{\pm.046}$	.655$_{\pm.009}$	**.736**$_{\pm.012}$	2.699$_{\pm.143}$	3.421$_{\pm.470}$	**2.613**$_{\pm.090}$
EDA+TEMP	20%	.585$_{\pm.037}$	.560$_{\pm.014}$	**.608**$_{\pm.059}$	.238$_{\pm.022}$	.276$_{\pm.004}$	**.292**$_{\pm.016}$	.710$_{\pm.010}$	.714$_{\pm.006}$	**.753**$_{\pm.008}$	3.011$_{\pm.201}$	3.924$_{\pm.496}$	**2.584**$_{\pm.085}$
	50%	.577$_{\pm.040}$	.529$_{\pm.019}$	**.579**$_{\pm.049}$	.236$_{\pm.019}$	.273$_{\pm.004}$	**.285**$_{\pm.018}$	.683$_{\pm.017}$	.678$_{\pm.006}$	**.726**$_{\pm.012}$	3.012$_{\pm.210}$	4.603$_{\pm.591}$	**2.569**$_{\pm.084}$
ACC+TEMP	20%	.583$_{\pm.034}$	.578$_{\pm.017}$	**.624**$_{\pm.057}$	.245$_{\pm.021}$	.267$_{\pm.005}$	**.295**$_{\pm.015}$	.700$_{\pm.004}$	.705$_{\pm.007}$	**.740**$_{\pm.009}$	2.899$_{\pm.166}$	3.513$_{\pm.410}$	**2.608**$_{\pm.085}$
	50%	.589$_{\pm.037}$	.564$_{\pm.024}$	**.626**$_{\pm.052}$	.223$_{\pm.016}$	.250$_{\pm.006}$	**.291**$_{\pm.020}$	.620$_{\pm.036}$	.629$_{\pm.010}$	**.695**$_{\pm.010}$	2.782$_{\pm.163}$	3.380$_{\pm.362}$	**2.601**$_{\pm.065}$
PPG+EDA+ACC	20%	.559$_{\pm.024}$	.573$_{\pm.016}$	**.613**$_{\pm.046}$	.231$_{\pm.018}$	.276$_{\pm.006}$	**.291**$_{\pm.019}$	.715$_{\pm.004}$	.719$_{\pm.006}$	**.766**$_{\pm.007}$	2.964$_{\pm.170}$	3.560$_{\pm.416}$	**2.672**$_{\pm.110}$
	50%	.526$_{\pm.016}$	.568$_{\pm.021}$	**.579**$_{\pm.042}$	.240$_{\pm.013}$	.265$_{\pm.020}$	**.274**$_{\pm.019}$	.656$_{\pm.067}$	.688$_{\pm.004}$	**.750**$_{\pm.011}$	2.888$_{\pm.189}$	3.479$_{\pm.441}$	**2.781**$_{\pm.155}$
PPG+EDA+TEMP	20%	.588$_{\pm.032}$	.559$_{\pm.018}$	**.610**$_{\pm.058}$	.231$_{\pm.025}$	.274$_{\pm.009}$	**.286**$_{\pm.015}$	.710$_{\pm.007}$	.723$_{\pm.006}$	**.761**$_{\pm.010}$	3.086$_{\pm.221}$	3.966$_{\pm.492}$	**2.624**$_{\pm.097}$
	50%	**.590**$_{\pm.033}$	.540$_{\pm.026}$	**.590**$_{\pm.056}$	.217$_{\pm.016}$	**.263**$_{\pm.009}$	**.263**$_{\pm.020}$	.681$_{\pm.017}$	.692$_{\pm.013}$	**.745**$_{\pm.014}$	3.269$_{\pm.250}$	4.876$_{\pm.586}$	**2.667**$_{\pm.121}$
PPG+ACC+TEMP	20%	.588$_{\pm.030}$	.591$_{\pm.003}$	**.629**$_{\pm.057}$	.236$_{\pm.018}$	.273$_{\pm.005}$	**.293**$_{\pm.016}$	.698$_{\pm.002}$	.710$_{\pm.004}$	**.747**$_{\pm.010}$	2.952$_{\pm.182}$	3.528$_{\pm.403}$	**2.667**$_{\pm.099}$
	50%	.601$_{\pm.036}$	.622$_{\pm.019}$	**.638**$_{\pm.059}$	.206$_{\pm.010}$	.270$_{\pm.010}$	**.287**$_{\pm.016}$	.609$_{\pm.059}$	.664$_{\pm.012}$	**.710**$_{\pm.013}$	3.059$_{\pm.214}$	3.351$_{\pm.292}$	**2.838**$_{\pm.114}$
EDA+ACC+TEMP	20%	.577$_{\pm.030}$	.564$_{\pm.009}$	**.610**$_{\pm.051}$	.242$_{\pm.018}$	.268$_{\pm.006}$	**.291**$_{\pm.014}$	.707$_{\pm.009}$	.707$_{\pm.002}$	**.747**$_{\pm.008}$	2.895$_{\pm.167}$	3.657$_{\pm.446}$	**2.571**$_{\pm.084}$
	50%	.558$_{\pm.032}$	.528$_{\pm.017}$	**.585**$_{\pm.040}$	.244$_{\pm.015}$	.242$_{\pm.008}$	**.287**$_{\pm.015}$	.681$_{\pm.022}$	.658$_{\pm.010}$	**.713**$_{\pm.010}$	2.684$_{\pm.162}$	3.742$_{\pm.494}$	**2.487**$_{\pm.059}$
PPG+EDA+ACC+TEMP	20%	.581$_{\pm.032}$	.575$_{\pm.013}$	**.619**$_{\pm.054}$	.236$_{\pm.015}$	.276$_{\pm.003}$	**.291**$_{\pm.015}$	.711$_{\pm.006}$	.721$_{\pm.003}$	**.758**$_{\pm.007}$	2.937$_{\pm.169}$	3.657$_{\pm.440}$	**2.624**$_{\pm.095}$
	50%	.574$_{\pm.021}$	.589$_{\pm.009}$	**.612**$_{\pm.040}$	.239$_{\pm.008}$	.280$_{\pm.007}$	**.290**$_{\pm.015}$	.669$_{\pm.044}$	.713$_{\pm.009}$	**.745**$_{\pm.011}$	2.891$_{\pm.181}$	3.765$_{\pm.458}$	**2.622**$_{\pm.099}$

Table 9: **Intermittent Drop Results for SSL Methods**. LP is compared with the SSL baseline LSM2 under each missing-modality combination and drop ratio. Best results are marked in **bold**.

Missing modalities	Drop	WESAD ↑F_1		AAUWSS ↑F_1		DALIA ↑F_1		WEEE ↓MAE	
		LSM2	Ours LP	LSM2	Ours LP	LSM2	Ours LP	LSM2	Ours LP
PPG	20%	.597±.004	**.626**±.051	.270±.002	**.278**±.011	.716±.004	**.755**±.003	3.646±.077	**2.638**±.062
	50%	.592±.001	**.628**±.050	**.261**±.004	**.261**±.007	.709±.003	**.759**±.006	3.747±.222	**2.703**±.099
EDA	20%	.572±.011	**.610**±.046	.248±.002	**.303**±.009	.719±.003	**.761**±.004	3.653±.067	**2.585**±.056
	50%	.530±.015	**.578**±.040	.229±.004	**.298**±.022	.718±.005	**.770**±.007	3.824±.282	**2.587**±.065
ACC	20%	.602±.010	**.629**±.050	.268±.001	**.294**±.004	.715±.001	**.739**±.008	3.560±.099	**2.640**±.051
	50%	.598±.012	**.620**±.060	.268±.003	**.289**±.002	**.713**±.003	.701±.017	3.571±.129	**2.692**±.047
TEMP	20%	.607±.008	**.626**±.050	.278±.002	**.295**±.005	.717±.005	**.751**±.005	3.648±.167	**2.577**±.064
	50%	.623±.021	**.624**±.052	.276±.004	**.293**±.006	.714±.008	**.736**±.006	3.791±.331	**2.545**±.083
PPG+EDA	20%	.571±.012	**.616**±.050	.249±.003	**.286**±.011	.718±.002	**.765**±.004	3.705±.130	**2.609**±.055
	50%	.532±.007	**.580**±.045	.227±.002	**.261**±.008	.718±.005	**.773**±.006	4.051±.610	**2.657**±.068
PPG+ACC	20%	.601±.006	**.629**±.051	.266±.001	**.279**±.011	.711±.002	**.741**±.006	3.580±.058	**2.662**±.060
	50%	.600±.009	**.634**±.048	**.266**±.005	.251±.002	.698±.006	**.703**±.017	3.707±.146	**2.859**±.129
PPG+TEMP	20%	.611±.011	**.625**±.050	.273±.002	**.277**±.010	.716±.004	**.753**±.005	3.673±.154	**2.613**±.066
	50%	.623±.019	**.632**±.054	.255±.002	**.260**±.008	.709±.008	**.742**±.007	3.943±.427	**2.641**±.098
EDA+ACC	20%	.582±.007	**.612**±.044	.249±.002	**.307**±.009	.718±.003	**.748**±.008	3.566±.117	**2.614**±.051
	50%	.536±.005	**.582**±.041	.242±.002	**.292**±.013	.712±.004	**.718**±.013	3.595±.274	**2.655**±.059
EDA+TEMP	20%	.588±.009	**.613**±.050	.266±.005	**.304**±.010	.720±.003	**.759**±.006	3.692±.142	**2.562**±.068
	50%	.540±.030	**.590**±.043	.238±.002	**.305**±.013	.715±.004	**.753**±.007	3.969±.303	**2.547**±.088
ACC+TEMP	20%	.609±.011	**.625**±.051	.275±.003	**.297**±.004	.714±.004	**.737**±.003	3.569±.136	**2.609**±.062
	50%	.614±.012	**.629**±.051	.277±.006	**.283**±.005	**.704**±.007	.688±.011	3.610±.138	**2.621**±.077
PPG+EDA+ACC	20%	.582±.011	**.623**±.049	.254±.003	**.292**±.008	.714±.002	**.751**±.006	3.606±.098	**2.633**±.051
	50%	.534±.009	**.597**±.034	.237±.002	**.274**±.007	.704±.001	**.725**±.018	3.720±.332	**2.720**±.070
PPG+EDA+TEMP	20%	.593±.009	**.612**±.050	.264±.002	**.286**±.007	.718±.002	**.760**±.005	3.741±.151	**2.569**±.064
	50%	.569±.013	**.597**±.050	.232±.005	**.263**±.010	.716±.004	**.761**±.003	4.123±.512	**2.528**±.086
PPG+ACC+TEMP	20%	.609±.011	**.632**±.052	.274±.005	**.278**±.012	.717±.002	**.740**±.004	3.575±.103	**2.640**±.068
	50%	.621±.014	**.643**±.048	**.268**±.003	.245±.005	.699±.011	**.703**±.007	3.733±.154	**2.792**±.117
EDA+ACC+TEMP	20%	.593±.006	**.615**±.048	.266±.004	**.305**±.008	.718±.003	**.745**±.004	3.602±.126	**2.579**±.063
	50%	.562±.008	**.602**±.046	.252±.002	**.302**±.009	.715±.004	**.723**±.007	3.632±.131	**2.563**±.108
PPG+EDA+ACC+TEMP	20%	.597±.008	**.618**±.048	.265±.001	**.292**±.007	.716±.001	**.752**±.005	3.628±.064	**2.602**±.054
	50%	.592±.008	**.620**±.041	.265±.003	**.282**±.007	.716±.004	**.747**±.002	3.752±.137	**2.606**±.060

E Ablation on Shared/Specific Features.

To further validate the necessity of preserving both modality-shared and modality-specific information, we conduct an ablation study on the feature types generated by our orthogonal tokenizer. As shown in Table 10, we evaluate the representation quality directly under the full-modality setting using different subsets of the tokens: only the shared components ($\mathbf{s}$_only), only the specific residuals ($\mathbf{p}$_only), and the combination of both (both ($\mathbf{s}$+$\mathbf{p}$)).

The results demonstrate that utilizing both components consistently yields the best Macro-F1 performance across three downstream health tasks (WESAD, AAUWSS, and DALIA). Dropping either the modal-shared or the modality-specific tokens leads to notable performance degradation. This confirms that these two part of features are both crucial for accurate human state estimation.

Table 10: **Ablation Study on Feature Types (Macro F1 ↑).** Macro-F1 results evaluated directly under the full modality setting, using shared ($\mathbf{s}$) and specific ($\mathbf{p}$) tokens generated by the tokenizer.

Ablation Mode	WESAD	AAUWSS	DALIA
both ($\mathbf{s}$+$\mathbf{p}$)	**0.6593±0.0152**	**0.3178±0.0094**	**0.7425±0.0105**
$\mathbf{s}$_only	0.5842±0.0183	0.2641±0.0112	0.7176±0.0132
$\mathbf{p}$_only	0.5400±0.0215	0.2968±0.0145	0.6988±0.0167

F Ablation on $\lambda_{recon}/\lambda_{align}$:

Table 11: Analysis of the weight ratio $\lambda_{recon}/\lambda_{align}$ in the Tokenizer loss.

Ratio	WESAD (F_1)	AAUWSS (F_1)	DaLiA (F_1)
0.1	0.640 ± 0.042	0.287 ± 0.006	0.752 ± 0.003
1.0	0.583 ± 0.035	0.275 ± 0.009	0.662 ± 0.007
10.0	0.623 ± 0.048	0.285 ± 0.007	0.685 ± 0.005

Given the disparate scales and optimization difficulties of the reconstruction and contrastive losses, we evaluate various $\lambda_{recon}/\lambda_{align}$ ratios, as reported in Table 11. We observe that optimal performance is achieved at a ratio of 0.1. An excessively large λ_{recon} incentivizes the model to offload the majority of information into the modality-specific residual $\mathbf{p}$ to minimize reconstruction error. Consequently, the shared representation $\mathbf{s}$ degenerates into modality-consistent noise rather than capturing meaningful semantics.

G Ablation on MoE Block Placement

Table 12: Impact of MoE layer index in the backbone.

Layer	WESAD (F_1)	AAUWSS (F_1)	DaLiA (F_1)
Layer 0	0.602 ± 0.033	0.236 ± 0.005	0.711 ± 0.008
Layer 2	0.640 ± 0.042	0.287 ± 0.006	0.752 ± 0.003
Layer 4	0.623 ± 0.046	0.249 ± 0.008	0.730 ± 0.006

We evaluate the impact of MoE layer placement within the missing-aware backbone and find that Layer 2 (0-indexed) yields optimal performance, as shown in Table 12. We attribute this to a trade-off between contextual awareness and specialization: if the MoE layer is placed too early, tokens lack sufficient self-attention depth to perceive global modality availability, rendering routing decisions susceptible to initialization noise. Conversely, if placed too late, cross-modal representations become excessively fused, diminishing the experts' ability to specialize in handling distinct modality combinations.

H Ablation on Shared Subspace Rank

In our framework, the rank r governs the allocation of information capacity between the shared and specific subspaces. To regulate this, we employ a gating mechanism that determines the activation of each orthogonal basis vector within the projection operator $\mathcal{P}$, thereby adaptively learning its effective rank. We evaluate various initial rank-to-dimension ratios (r/d) and observe that an initialization of 0.9 yields optimal performance, as shown in Table 13. Furthermore, visualizing the rank evolution for initializations of 0.75 and 0.9 in Figure 7 reveals a consistent trend: the effective rank of $\mathcal{P}$ across all modalities converges to a value exceeding 0.8 during training.

Table 13: Ablation study on the initial ratio of shared subspace rank r/d.

Init Ratio	WESAD (F_1)	AAUWSS (F_1)	DaLiA (F_1)
0.5	0.594 ± 0.038	0.251 ± 0.008	0.668 ± 0.006
0.75	0.644 ± 0.045	0.282 ± 0.005	0.671 ± 0.005
0.9	0.640 ± 0.042	0.287 ± 0.006	0.752 ± 0.003

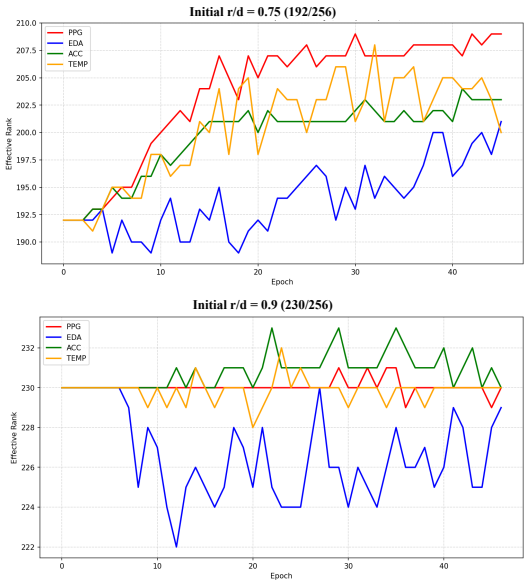

Figure 7: Rank evolution during training.

I Hyperparameter Settings

For the loss weights tuning, we fix λ_{align} = 1 and λ_{rand} = 1, so the main coefficients are: λ_{recon}, λ_{white}, and λ_{struct}. For Tokenizer, λ_{recon} / λ_{align} controls the trade-off between reconstructability and cross-modal shared semantics. If λ_{recon} is too large, the model can offload part of the shared information into the specific branch **p**, because reconstructing from **p** is easier to optimize than aligning the shared branch **s** across modalities. This weakens the semantic content of **s**. This behavior is consistent with our ablation in Appendix Table 11, where: λ_{recon} / λ_{align} = 0.1 performs best across all three downstream tasks. We keep λ_{white} small (0.05) because it serves as a regularizer for approximate whitening/independence rather than a primary task objective, and thus should not dominate semantic learning. For Backbone, $\mathcal{L}_{rand}$ and $\mathcal{L}_{struct}$ are both latent-space reconstruction losses with matched form, so their scales are much more comparable than the tokenizer objectives. They play complementary roles: $\mathcal{L}_{rand}$ learns general contextual semantics $\mathcal{L}_{struct}$ explicitly trains robustness to modality-level missingness Since both are essential in our setting, we use the balanced default: λ_{rand} : λ_{struct} = 1:1.

Table 14: Hyperparameters for Orthogonal Tokenizer pre-training.

Hyperparameters		Values
Input Configuration	Sampling Rates (Hz)	{PPG:64, ACC:32, EDA:4, TEMP:4}
	Input Channels	{1, 3, 1, 1}
	Window Duration	10s
	Patch Size	1s
Encoder & Projector	Encoder Architecture	ResNet-1D
	Embedding Dimension (d_{model})	256
	Initial Rank Ratio (r/d)	0.9
Optimization	Batch Size	256
	Optimizer	AdamW
	Learning Rate	1×10^{-4}
	Weight Decay	1×10^{-5}
	Warmup Epochs	10
	Reconstruction Weight (λ_{recon})	0.1
	Whitening Weight (λ_{white})	0.05

Table 15: Hyperparameters for Missing-Aware MoE Backbone pre-training.

Hyperparameters		Values
Transformer Architecture	Encoder Layers	6
	Decoder Layers	2
	Hidden Size (d_{model})	512
	Attention Heads	8
	Head Dimension	64
	Dropout	0.1
Mixture of Experts (MoE)	MoE Layer Index	2 (0-indexed)
	Number of Experts	8
	Top-k Routing	2
Optimization & Masking	Masking Ratio	0.75
	Batch Size	1024
	Optimizer	AdamW
	Learning Rate	1×10^{-4}
	Weight Decay	0.05
	Adam β	(0.9, 0.999)
	Warmup Epochs	10

J Evaluation on Larger Scale Dataset

To further address dataset scale, we include the Wearable Device Dataset [60]. It contains synchronized PPG, EDA, ACC, and TEMP signals collected with the E4 from 36 participants and supports downstream 2-class stress detection. Table 16 shows the comparison with LSM2. VCR outperforms LSM2 in 12 of the 15 scenarios.

Table 16: Macro-F1 (↑) comparison under different missing-modality scenarios. Better results are marked in **bold**.

Scenario	OursLP	LSM2
FULL	$\mathbf{0.712 \pm 0.010}$	0.652 ± 0.008
ACC Missing	$\mathbf{0.720 \pm 0.012}$	0.668 ± 0.005
EDA Missing	$\mathbf{0.626 \pm 0.024}$	0.528 ± 0.016
PPG Missing	$\mathbf{0.644 \pm 0.032}$	0.524 ± 0.004
TEMP Missing	$\mathbf{0.651 \pm 0.006}$	0.622 ± 0.008
ACC TEMP Missing	0.643 ± 0.008	$\mathbf{0.649 \pm 0.009}$
EDA ACC Missing	$\mathbf{0.688 \pm 0.019}$	0.599 ± 0.020
EDA TEMP Missing	0.588 ± 0.044	$\mathbf{0.610 \pm 0.024}$
PPG ACC Missing	$\mathbf{0.597 \pm 0.036}$	0.574 ± 0.010
PPG EDA Missing	$\mathbf{0.594 \pm 0.023}$	0.431 ± 0.012
PPG TEMP Missing	$\mathbf{0.553 \pm 0.050}$	0.507 ± 0.003
EDA ACC TEMP Missing	$\mathbf{0.626 \pm 0.031}$	0.439 ± 0.005
PPG ACC TEMP Missing	0.482 ± 0.054	$\mathbf{0.576 \pm 0.014}$
PPG EDA ACC Missing	$\mathbf{0.555 \pm 0.028}$	0.400 ± 0.004
PPG EDA TEMP Missing	$\mathbf{0.465 \pm 0.024}$	0.438 ± 0.008

K System Cost and Training Schedule

Table 17 reports training costs. The training times for VCR and the SSL baselines correspond to pretraining, while those of supervised baselines are reported separately for each dataset. VCR pretrains the orthogonal tokenizer for 80 epochs and the missing-aware backbone for another 80 epochs, followed by 100 epochs of downstream fine-tuning. The supervised baselines are trained from scratch for 100 epochs.

Table 17: System cost comparison. All times are in GPU-hours; the last four columns report supervised downstream training time.

Method	GFLOPs	Peak memory (GiB)	Pretraining time	WESAD	AAUWSS	DaLiA	WEEE
VCR	5.64	2.09	11.67	–	–	–	–
LSM	3.07	1.55	11.06	–	–	–	–
LSM-2	2.74	2.72	50.34	–	–	–	–
SimCLR	4.08	6.50	16.73	–	–	–	–
CMC	1.01	1.38	8.12	–	–	–	–
MSN	2.04	1.29	4.98	–	–	–	–
MF-CLR	0.98	2.55	11.77	–	–	–	–
FOCAL	1.61	6.73	30.27	–	–	–	–
ResNet	15.11	5.65	–	0.4165	4.2028	1.0942	0.3707
ViT	1.12	1.48	–	0.2518	2.4399	0.5684	0.2546
FlexMoE	0.86	1.68	–	1.0092	10.7302	2.7415	0.8764
MAESTRO	5.29	4.82	–	0.6543	5.9700	1.4523	0.5482

L Baseline Hyperparameters and Reproduction Details

Table 18: Final training and downstream hyperparameters for all methods.

Method	Training learning rate	Weight decay	Dropout	Learning-rate warm-up	Downstream learning rate
VCR tokenizer	10^{-4}	10^{-5}	0.1	10 epochs	10^{-3}
VCR backbone	10^{-4}	5×10^{-2}	0.1	10 epochs	10^{-3}
LSM	5×10^{-3}	10^{-4}	0.1	5 epochs	10^{-3}
LSM-2	5×10^{-3}	10^{-4}	0.1	5 epochs	10^{-3}
SimCLR	10^{-4}	10^{-4}	0.1	10 epochs	10^{-3}
CMC	10^{-3}	10^{-4}	0.1	10 epochs	10^{-3}
MSN	10^{-3}	10^{-4}	0.1	10 epochs	10^{-3}
MF-CLR	10^{-3}	5×10^{-2}	0.1	10 epochs	10^{-3}
FOCAL	10^{-4}	10^{-5}	0.1	10 epochs	10^{-3}
ResNet	10^{-3}	10^{-4}	0.1	10 epochs	N/A
ViT	10^{-3}	5×10^{-3}	0.1	10 epochs	N/A
FlexMoE	10^{-4}	10^{-4}	0.1	5 epochs	N/A
MAESTRO	10^{-3}	10^{-4}	0.1	10 epochs	N/A

Training and hyperparameter tuning. All baselines were trained from scratch using the same data, subject-independent splits, input windows, and evaluation protocol as VCR. We tuned learning rate, weight decay, dropout, warm-up ratio, and downstream learning rate using method-specific search spaces centered on the values recommended by the original papers. All methods received the same number of search trials, the same subject-wise validation protocol, and comparable computational budgets, without access to test subjects. Table 18 reports the final settings.

LSM and LSM-2 reproduction. Since official implementations were unavailable, we reimplemented LSM and LSM-2 following the architectures and training procedures reported in the original papers. For capacity matching, we preserved the method-defining topology and objectives while scaling model width to approximately 50M parameters. LSM retains a 12-layer encoder and an 8-layer decoder, while LSM-2 retains a 12-layer encoder and a 4-layer decoder. The parameter counts of LSM, LSM-2, and VCR are 46.96M, 52.20M, and 50.68M, respectively.

Pretraining data and masking. Neither the original LSM-2 pretraining data nor its pretrained parameters were publicly available, so we trained our LSM-2 implementation on exactly the same complete pretraining data as VCR. Since the training data of LSM and LSM-2 were not public, we did not apply aggregation statistics. Masking was applied to complete patch tokens rather than individual raw samples, following the original papers. For LSM-2, each sample uses, with equal probability, 80% random masking, 50% temporal-slice masking, or 50% signal-slice masking.

M Compute Resources

All experiments, including pretraining and downstream evaluations, are conducted on a NVIDIA RTX 5090 GPU (32GB VRAM), an Intel(R) Xeon(R) Platinum 8470Q CPU (25 vCPUs), and 90GB of system memory. The complete pretraining pipeline of our VCR framework requires approximately 11.67 GPU-hours (including both the orthogonal tokenizer and the missing-aware MoE backbone). Taking into account the evaluation of baselines, cross-validation on multiple downstream tasks, and failed experiments, the total computational budget of this research project is estimated to be approximately 2400 GPU hours.

N Broader Impacts

Positive Societal Impacts: The primary positive impact of VCR lies in advancing continuous, non-invasive health monitoring. By significantly improving model robustness against real-world

modality missingness (e.g., sensor failure, user non-compliance, or hardware limitations), VCR makes intelligent wearable health applications more reliable and accessible. This can potentially lower the barrier to entry for digital health technologies, allowing users with cheaper or partial sensor configurations to still benefit from high-quality physiological tracking. Furthermore, by relying on self-supervised learning, VCR reduces the reliance on large-scale expert annotations, which are notoriously expensive and time-consuming to obtain in the medical and healthcare domains.

Potential Negative Impacts and Risks: Wearable signals can infer highly sensitive personal states. If misused, such robust tracking systems could be deployed for continuous surveillance by employers or insurance companies without the user's explicit consent, potentially leading to privacy violations.

O License

The license information for multiple datasets: BIG IDEAs (ODC-By 1.0), UE4W (CC-BY 4.0), CAN-STRESS (CC-BY 4.0), AAUWSS (CC-BY 4.0), WEEE (CC-BY 4.0), DaLiA (CC-BY 4.0) and WESAD (CC-BY 4.0).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. We explicitly state our main contributions and claims at the end of the Introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations section acknowledges limitations in population scale and device coverage, the missing freeze-versus-joint-pretraining comparison, and future unimodal pretraining comparisons.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper is primarily empirical and methodological. The mathematical derivations provided in Section 2.3 serve to provide intuition and justify the architectural design (i.e., structural disentanglement).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experimental setup, datasets, metrics, training details, and hyperparameters are fully described in Section 3.1. and Appendix H.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We commit to fully open-sourcing the codebase, pre-trained weights, and reproduction instructions upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We detailed the data split, evaluation protocol, how to select hyperparameters, and the final hyperparameters/optimizers we used in Section 3.1 and Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We reported the mean $\pm$ standard deviation in our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We detail the hardware configuration, the exact execution time and an aggregated estimate of the total compute resources used for the entire project in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. We utilize publicly available, anonymized datasets for wearable sensor data pretraining, and the work poses no risks of harm, discrimination, or exploitation.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential positive and negative societal impacts of our work in Appendix J. We acknowledge the benefits of accessible health monitoring as well as the potential risks regarding data privacy.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work introduces a representation learning framework for wearable physiological signals. It does not release high-risk models (e.g., generative language models, deepfake generators) or scraped internet datasets that require specific access safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the datasets used in this paper, and detailed their license in Appendix K.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce or release any new datasets or similar standalone assets. The source code and pre-trained models will be released upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or new research with human subjects. All experiments are conducted on existing, publicly available, and de-identified benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Since the study relies entirely on pre-existing, publicly available datasets that have already been de-identified by their original authors, no new IRB approval was required for this secondary analysis.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methods in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.