PhysioME: A Robust Multimodal Self-Supervised Framework for Physiological Signals with Missing Modalities

Cheol-Hui Lee[1,2†], **Hwa-Yeon Lee**[1,2†], **Min-Kyung Jung**[1,2], **Dong-Joo Kim**[1,2,3*]

[1]Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea
[2]Interdisciplinary Program in Precision Public Health, Korea University, Seoul, South Korea
[3]Department of Neurology, Korea University College of Medicine, Seoul, South Korea
dlcjfgmlnasa28@korea.ac.kr, belita0152@korea.ac.kr, mkjung@korea.ac.kr, dongjookim@korea.ac.kr

Abstract

Missing or corrupted modalities are common in physiological signal-based medical applications owing to hardware constraints or motion artifacts. However, most existing methods assume the availability of all modalities, resulting in substantial performance degradation in the absence of any modality. To overcome this limitation, this study proposes PhysioME, a robust framework designed to ensure reliable performance under missing modality conditions. PhysioME adopts: (1) a multimodal self-supervised learning approach that combines contrastive learning with masked prediction; (2) a Dual-Path-NeuroNet backbone tailored to capture the temporal dynamics of each physiological signal modality; and (3) a restoration decoder that reconstructs missing modality tokens, enabling flexible processing of incomplete inputs. The experimental results show that PhysioME achieves high consistency and generalization performance across various missing modality scenarios. These findings highlight the potential of PhysioME as a reliable tool for supporting clinical decision-making in real-world settings with imperfect data availability.

Introduction

In real-world clinical environments, it is often challenging to continuously acquire all physiological signals because of sensor malfunctions, patient movements, or hardware limitations (Iranfar, Arza, and Atienza 2021). However, most existing methods are designed under the assumption that all modalities are available, which leads to a performance degradation when some modalities are missing (Reza, Prater-Bennette, and Asif 2024).

To overcome these limitations and enable continuous patient monitoring, specific model architectures that can robustly handle missing modality scenarios are required. There are two main strategies: "dedicated model" and "single-model" approaches (Lee and Kim 2023; Wu et al. 2024). The former trains independent networks for each possible combination of modalities. This strategy offers precise handling of specific cases, but suffers from excessive computational and resource demands, thereby limiting its applicability in practice (Bachmann et al. 2022). By contrast, the latter handles all scenarios within a unified architecture, thereby greatly improving its practicality in clinical environments. As a result of this advantage, recent studies have increasingly focused on the development of single-model approaches that are capable of handling diverse missing modality scenarios within a unified framework (Wu et al. 2024; Bachmann et al. 2022; Lee and Kim 2023; Deldari et al. 2024; Zhao et al. 2024).

However, most single-models focus on vision tasks (Bachmann et al. 2022; Lee and Kim 2023; Zhao et al. 2024), restricting the extensibility of using physiological signals or heterogeneous sensors to address structural challenges. First, vision models are designed to process spatial information within images and struggle to capture the temporal continuity of signals (Wang, Cao, and Philip 2020). Second, simply replacing missing signals with mask or class tokens often fails to capture the semantics of actual missing scenarios, thereby hampering the generalization performance (Bachmann et al. 2022; Zhao et al. 2024). Third, most missing modality models are built on supervised learning frameworks that require extensive labeled datasets, which are scarce in the medical domain (Esteva et al. 2019; Lee et al. 2024).

To overcome these limitations, we propose **PhysioME**, a single-model-based missing modality framework specifically designed for physiological signals. The aim of this study is to achieve robust performance in clinical settings where certain modalities may be missing. PhysioME has the following key characteristics:

- PhysioME is a multimodal self-supervised learning (SSL) framework tailored for heterogeneous physiological signals. It integrates contrastive learning and masked prediction to learn discriminative and generative representations jointly, thereby enabling robust and generalizable feature learning for physiological data.
- The SSL backbone of PhysioME, **Dual-Path (DP)-NeuroNet**, effectively encodes the temporal characteristics of physiological signals. It applies distinct temporal augmentations to two parallel paths that share the same NeuroNet (Lee et al. 2024) structure, thereby facilitating rich representation learning. The resulting modality-specific representations serve as inputs for PhysioME.
- PhysioME incorporates a dedicated **restoration decoder** module for each modality. These decoders are designed to reconstruct missing modality tokens by leveraging the observed modalities, allowing the model to robustly handle a wide range of missing modality scenarios in physiological

† Equal contribution
* Corresponding author

signals.

This proposed framework consistently maintains high performance under various missing modality conditions. This model demonstrates strong robustness and generalization in sleep stage classification and hypotension prediction using clinical datasets.

Methods: PhysioME

DP-NeuroNet

DP-NeuroNet comprises two identical instances of NeuroNet that share weights (Figure 2). NeuroNet (Lee et al. 2024) is an SSL framework for physiological signals that is designed to effectively learn generalizable representations from unlabeled data by combining masked prediction and contrastive learning.

NeuroNet consists of a multi-scale 1D ResNet-based frame network and a vision transformer (ViT)-based encoder-decoder architecture. A subset of the token sequence produced by the frame network is randomly sampled and reconstructed through the encoder-decoder, with the reconstruction optimized using the mean squared error (MSE) loss. In addition, contrastive learning is performed using the NT-Xent loss (Sohn 2016) applied to two independently sampled token sequences. (Appendix I for architectural details of NeuroNet).

Each augmented input, $\tilde{x}_i^{(1)}$ and $\tilde{x}_i^{(2)}$, is independently processed using NeuroNet. The representations $z^{(1)}$ and $z^{(2)}$ are then extracted using the NeuroNet encoder $f_{\text{neuronet\_enc}}$ (composed of the frame network and ViT encoder), followed by a multi-layer perceptron (MLP) layer. NT-Xent loss (Sohn 2016) is then applied to these representations to encourage similarities between different augmented views of the same input. The formula is as follows:

$$\mathcal{L}_{\text{NT-Xent}} = \frac{1}{2B} \sum_{b=1}^{B} \left[\ell\left(z_b^{(1)}, z_b^{(2)}\right) + \ell\left(z_b^{(2)}, z_b^{(1)}\right) \right] \tag{1}$$

$$\ell(z_a, z_b) = -\log \frac{\exp\left(\text{sim}(z_a, z_b)/\tau\right)}{\sum_{k=1}^{2B} \mathbf{1}_{[k \neq a]} \exp\left(\text{sim}(z_a, z_k)/\tau\right)} \tag{2}$$

where B denotes the batch size and $z_1^{(1)}$ refers to the first sample in the batch from $z^{(1)}$. The function $\text{sim}(\cdot)$ represents the cosine similarity.

PhysioME

PhysioME is a multimodal SSL framework that is specifically designed to handle missing modalities (Figure 1). Its core components are as follows:

(A) Training

Modality Encoder Each physiological signal is encoded using the pre-trained encoder $f_{\text{neuronet\_enc}}$ from DP-NeuroNet, which serves as the modality encoder. A total of M modality encoders are used, each responsible for extracting features specific to the m-th modality, where $m \in \{1, 2, \ldots, M\}$. Given an input signal $x^{(m)}$, the corresponding modality encoder produces a token sequence $e^{(m)} = \{e_i^{(m)}\}_{i=1}^{N} = f_{\text{modality\_enc}}^{(m)}(x^{(m)}) \equiv f_{\text{neuronet\_enc}}^{(m)}(x^{(m)})$, where N denotes the length of the token sequence. To adapt the modality encoders for PhysioME, low-rank adaptation (LoRA) (Hu et al. 2022) is applied to a subset of layers, whereas the remaining layers are frozen during training.

Multimodal Encoder Each M modality encoder produces a token sequence $e^{(m)}$, which is then fed into a ViT-based multimodal encoder to generate a unified sequence of tokens. This process consists of three main stages.

1. **Positional and Modality Encoding**: Each token sequence $e^{(m)}$ is passed through an MLP layer, after which positional encoding pe and modality encoding $mt^{(m)}$ are added. This results in a new token sequence: $z^{(m)} = \text{MLP}(e^{(m)}) + pe + mt^{(m)}$.
2. **Drop Modality and Random Sampling**: For each of the M token sequences $z^{(m)}$, either modality dropping or random sampling is randomly selected and applied.
   - **Drop modality** simulates a missing modality scenario by removing the entire token sequence $z^{(m)}$ and replacing it with a mask token μ.
   - **Random sampling** involves the selection of $\tilde{N}$ tokens from $z^{(m)}$, where $\tilde{N} < N$. The selected subset token sequence $\tilde{z}^{(m)}$ is constructed based on an index set $\mathcal{H} \subseteq \{1, \ldots, N\}$.
3. **Fusion Token**: The token sequence $\tilde{h}$ is input into the multimodal encoder $f_{\text{multimodal\_enc}}$ to generate the sampled fusion token sequence $\tilde{o}$.

Modality Decoder The ViT-based modality decoders are constructed independently for each modality. Each decoder takes only a subset of tokens $\tilde{o}^{(m)}$ corresponding to the set of modalities $\tilde{\mathcal{M}}$ that are not subjected to drop modality, extracted from the output $\tilde{o}$ of the multimodal encoder.

In this process, mask tokens μ are inserted into $\tilde{o}^{(m)}$ at the positions corresponding to the token indices not included in the random sampling subset $\mathcal{H}$. The resulting sequence is then passed through the modality decoder $f_{\text{modality\_dec}}^{(m)}$ and an MLP layer to generate a new token sequence: $d^{(m)} = \text{MLP}(f_{\text{modality\_dec}}^{(m)}(\mu,\ \tilde{o}^{(m)} \mid m \in \tilde{\mathcal{M}}))$. The output $d^{(m)}$ is used to reconstruct the original modality encoder output $e^{(m)}$, and this reconstruction is optimized using the MSE loss. After SSL is completed, the modality decoders are discarded.

Restoration Decoder The restoration decoder adopts a ViT-based architecture similar to the modality decoder but is designed with a deeper structure to enable more precise reconstruction. During training, a stop-gradient operation is applied to block backpropagation to the multimodal encoder and other networks, allowing the restoration decoder to be trained independently with the sole focus on reconstruction.

The restoration decoder $f_{\text{restoration\_dec}}^{(m)}$ takes only the modality-specific tokens $\tilde{o}^{(m)}$ corresponding to the modalities where the drop modality is applied, extracted from the output $\tilde{o}$ of the multimodal encoder. It then generates a new token sequence: $g^{(m)} = \text{MLP}(f_{\text{restoration\_dec}}^{(m)}(\tilde{o}^{(m)} \mid$

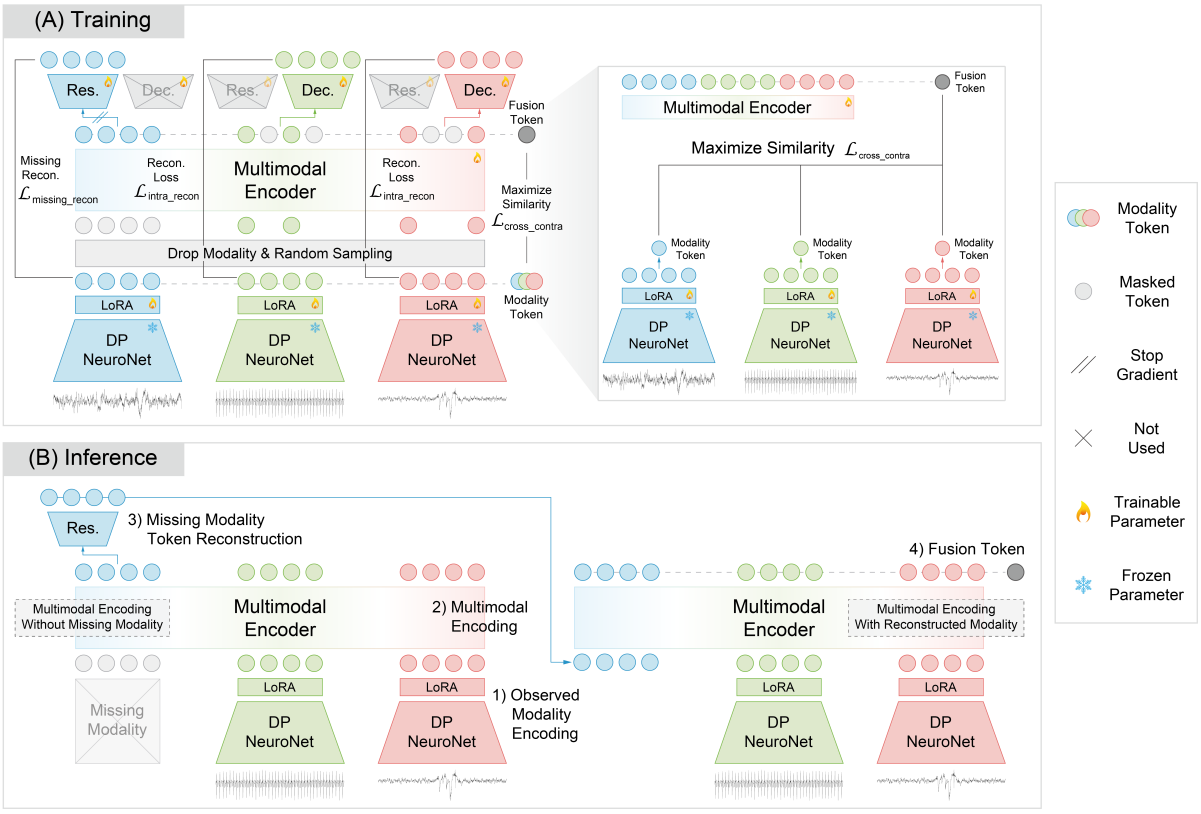

Figure 1: Overview of PhysioME architecture. (A) Training workflow of PhysioME: a multimodal SSL framework for missing modalities. (B) Inference procedure using the pretrained PhysioME.

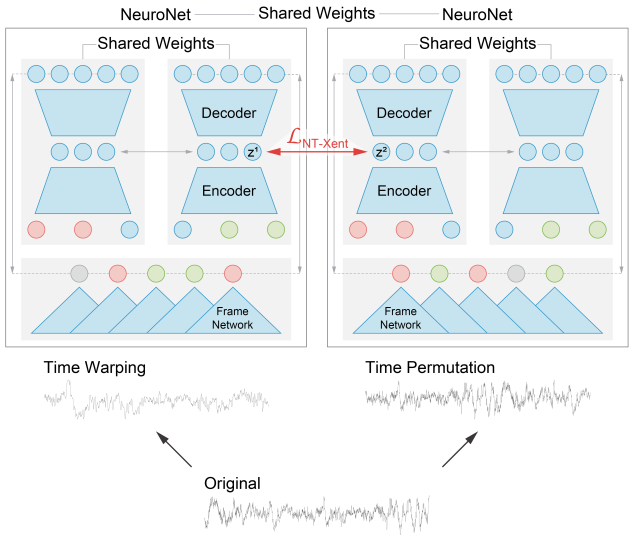

Figure 2: DP-NeuroNet structure.

$m \in \mathcal{M} \setminus \tilde{\mathcal{M}}$)), which is trained to reconstruct the original modality encoder output $e^{(m)}$ using the MSE loss. Once the training is complete, the restoration decoder is utilized during inference to estimate the token sequence corresponding to the missing modalities.

Objective Loss

1. **Reconstruction Loss**: The output $d^{(m)}$ from the modality decoder is defined only for modalities $m \in \tilde{\mathcal{M}}$ where drop modality is not applied. To reconstruct the corresponding modality encoder output $e^{(m)}$, the MSE loss is employed. Reconstruction is performed only on the token indices that were not selected during random sampling, denoted by the index set $\bar{\mathcal{H}}^{(m)}$. The loss is formulated as follows:

$$\mathcal{L}_{\text{intra\_recon}} = \frac{1}{|\tilde{\mathcal{M}}|} \sum_{m \in \tilde{\mathcal{M}}} \frac{1}{|\bar{\mathcal{H}}(m)|} \sum_{i \in \bar{\mathcal{H}}(m)} \left\| d_i^{(m)} - e_i^{(m)} \right\|_2^2 \tag{3}$$

2. **Missing Modality Reconstruction Loss**: The restoration decoder is trained only for the dropped modalities, i.e., $m \in \mathcal{M} \setminus \tilde{\mathcal{M}}$. The output $g^{(m)}$ is optimized to reconstruct the corresponding modality encoder output $e^{(m)}$. The loss is formulated as follows:

$$\mathcal{L}_{\text{missing\_recon}} = \frac{1}{|\mathcal{M} \setminus \tilde{\mathcal{M}}|} \sum_{m \in \mathcal{M} \setminus \tilde{\mathcal{M}}} \frac{1}{N} \sum_{i=1}^{N} \left\| g_i^{(m)} - e_i^{(m)} \right\|_2^2 \tag{4}$$

3. **Cross-Modality Contrastive Loss**: To align the representations between the modality and multimodal encoders, contrastive learning based on the NT-Xent loss is applied. This encourages the model to learn the relationship between the single-modal representation $e^{(m)}$, extracted from each modality encoder, and the fused representation o obtained from the multimodal encoder. Consequently, semantic consistency across modalities is preserved, and the quality of the fused representation is enhanced.

   Specifically, the representations $e^{(m)}$ and o are averaged along the sequence dimension, normalized, and passed through an MLP layer to obtain $em^{(m)}$ and om, respectively. The NT-Xent loss is computed based on the similarity between these two representations.

$$\mathcal{L}_{\text{NT-Xent}}^{(m)} = \frac{1}{2B} \sum_{b=1}^{B} \left[l\left(em_b^{(m)}, om_b\right) + l\left(om_b, em_b^{(m)}\right) \right] \tag{5}$$

$$l(z_a, z_b) = -\log \frac{\exp\left(\text{sim}(z_a, z_b)/\tau\right)}{\sum_{k=1}^{2B} \mathbf{1}_{[k \neq a]} \exp\left(\text{sim}(z_a, z_k)/\tau\right)} \tag{6}$$

   where B denotes the batch size, while $\text{sim}(\cdot)$ and τ represent the cosine similarity and temperature parameter, respectively. The final cross-modality contrastive loss is defined as the average loss across all modalities. The formula is as follows:

$$\mathcal{L}_{\text{cross\_contra}} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathcal{L}_{\text{NT-Xent}}^{(m)} \tag{7}$$

4. **Final Loss**: The model is optimized using a combined objective that integrates the three loss components as follows:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{intra\_recon}} + \beta \mathcal{L}_{\text{missing\_recon}} + \gamma \mathcal{L}_{\text{cross\_contra}} \tag{8}$$

   where α, β, and γ are hyperparameters that balance the contributions of the three losses.

(B) Inference To handle scenarios in which certain modalities are missing, PhysioME follows the inference procedure as follows:

Observed Modality Encoding For each observed modality $\mathcal{M}_{\text{obs}} \in \{1, \ldots, M\}$, the input $x^{(m)}$ is passed through the modality encoder $f_{\text{modality\_enc}}^{(m)}$ and an MLP layer, followed by the addition of positional encoding pe and a modality token $mt^{(m)}$.

$$e^{(m)} = f_{\text{modality\_enc}}^{(m)}(x^{(m)}), \quad \text{for } m \in \mathcal{M}_{\text{obs}} \tag{9}$$

$$z^{(m)} = \{z_i^{(m)}\}_{i=1}^{N} = \text{MLP}(e^{(m)}) + pe + mt^{(m)} \tag{10}$$

Multimodal Encoding The token sequences $z^{(m)}$ are merged with the masked tokens μ and fed into the multimodal encoder $f_{\text{multimodal\_enc}}$.

$$o = f_{\text{multimodal\_enc}}(\text{concat}(\mu,\ z^{(m)} \mid m \in \mathcal{M}_{\text{obs}})) \tag{11}$$

Missing Modality Token Reconstruction From the output o of $f_{\text{multimodal\_enc}}$, the token sequences $o^{(m)}$ corresponding to the unobserved modalities $\mathcal{M}_{\text{miss}} \subset \{1, \ldots, M\} \setminus \mathcal{M}_{\text{obs}}$ are extracted. These are then passed through the restoration decoder $f_{\text{restoration\_dec}}^{(m)}$ and an MLP layer to reconstruct the sequence tokens $g^{(m)}$ for the missing modalities.

$$g^{(m)} = \text{MLP}(f_{\text{restoration\_dec}}^{(m)}(o^{(m)})), \quad \text{for } m \in \mathcal{M}_{\text{miss}} \tag{12}$$

Generating a Fusion Token The token sequences reconstructed by the restoration decoder are merged with those extracted from the modality encoders, and then fed into the multimodal encoder. The resulting output token sequence ft includes a class token that represents the entire sequence and serves as a global representation for the final prediction in the downstream tasks.

$$ft = f_{\text{multimodal\_enc}}(\text{concat}(g^{(m)} \mid m \in \mathcal{M}_{\text{miss}}, e^{(m)} \mid m \in \mathcal{M}_{\text{obs}})) \tag{13}$$

Experiments

Dataset Description

Sleep-EDFX (Kemp et al. 2000) contains polysomnographic (PSG) recordings and aims to classify data into five sleep stages based on the American Academy of Sleep Medicine guidelines. This study utilized the SC subset and three channels: electroencephalography (EEG) Fpz-Cz, EEG Pz-Oz, and horizontal electrooculography (EOG). A 0-40 Hz band-pass filter was applied to all signals.

VitalDB (Lee et al. 2022) comprises intraoperative physiological recordings from 6,388 surgical cases, including 486,451 waveform segments and the corresponding perioperative patient information. This study aimed to predict intraoperative hypotension within the following 5 min using three types of data: arterial blood pressure (ABP), electrocardiography (ECG), and photoplethysmography (PPG) signals.

Baselines

Four multimodal SSL frameworks for missing modality were implemented as baselines. Except for CroSSL, the selected models were originally developed for vision tasks; thus, the signals were converted into short-time Fourier transform images and used as inputs for training. The baselines were as follows:

- **MultiMAE (Bachmann et al. 2022)** is an MAE-based approach that learns joint representations by reconstructing masked inputs across multiple modalities.
- **RobustSsF (Lee and Kim 2023)** employs four independent encoder-decoder branches to separately process each modality. To preserve inter-modality consistency, this model adopts an SSL scheme based on coupled regularization loss.
- **CroSSL (Deldari et al. 2024)** performs masking in the latent space and learns global representations from heterogeneous sensor modalities.

Modality			*Dedicated Model*				*Single Model*									
EEG Fpz-Cz	EEG Pz-Cz	EOG	ContraWR		SynthSleepNet		MultiMAE		RobustSsF		CroSSL		MaskMentor		PhysioME	
			ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
•	•	•	75.53	89.75	79.92	94.81	75.59	92.82	71.34	90.08	71.49	90.59	78.21	93.14	**79.64**	**94.86**
•			72.42 (-3.11)	88.57 (-1.18)	74.88 (-5.04)	90.02 (-4.79)	73.88 (-1.71)	91.50 (-1.32)	35.87 (-35.47)	60.93 (-29.15)	59.42 (-12.07)	82.23 (-8.36)	71.72 (-6.49)	90.35 (-2.79)	**76.31** (-3.33)	**93.65** (-1.21)
	•		71.73 (-3.80)	84.82 (-4.93)	71.76 (-8.16)	89.13 (-5.68)	65.92 (-9.67)	87.95 (-4.87)	34.34 (-37.00)	62.67 (-27.41)	58.30 (-13.19)	83.29 (-7.30)	**72.02** (-6.19)	88.15 (-4.99)	71.84 (-7.80)	**89.79** (-5.07)
		•	72.25 (-3.28)	88.49 (-1.26)	72.96 (-6.96)	91.99 (-2.82)	65.09 (-10.5)	88.53 (-4.29)	33.25 (-38.09)	60.28 (-29.80)	55.94 (-15.55)	80.51 (-10.08)	65.28 (-12.93)	90.21 (-2.93)	**72.33** (-7.31)	**91.61** (-3.25)
•	•		75.53 (-0.00)	89.76 (+0.01)	77.52 (-2.40)	93.57 (-1.24)	69.69 (-5.90)	89.22 (-3.60)	65.90 (-5.44)	86.40 (-3.68)	68.69 (-2.80)	89.22 (-1.37)	71.25 (-6.96)	88.19 (-4.95)	**77.08** (-2.56)	**93.55** (-1.31)
•		•	75.08 (-0.45)	90.50 (+0.75)	76.44 (-3.48)	93.46 (-1.35)	69.00 (-6.59)	89.91 (-2.91)	66.11 (-5.23)	87.73 (-2.35)	69.00 (-2.49)	89.91 (-0.68)	71.44 (-6.77)	91.01 (-2.13)	**78.34** (-1.30)	**94.47** (-0.39)
	•	•	75.12 (-0.41)	89.63 (-0.12)	75.30 (-4.62)	93.00 (-1.81)	70.27 (-5.32)	89.86 (-2.96)	58.39 (-12.95)	80.36 (-9.72)	70.27 (-1.22)	87.84 (-2.75)	72.72 (-5.49)	91.68 (-1.46)	**78.10** (-1.54)	**93.87** (-0.99)
Mean Absolute Value (MAV)			73.69 (1.84)	88.63 (1.38)	74.81 (5.11)	91.86 (2.95)	68.98 (6.62)	89.50 (3.33)	48.98 (22.36)	73.06 (17.02)	63.60 (7.89)	85.50 (5.09)	70.74 (7.47)	89.93 (3.21)	**75.67** (3.97)	**92.82** (2.04)

Table 1: Comparison with other methodologies for sleep stage classification using Sleep-EDFX.

Modality			*Dedicated Model*				*Single Model*									
ABP	ECG	PPG	ContraWR		SynthSleepNet		MultiMAE		RobustSsF		CroSSL		MaskMentor		PhysioME	
			ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
•	•	•	91.53	87.04	89.36	94.97	84.21	79.79	85.19	61.75	80.17	67.28	**92.52**	85.49	89.59	**94.71**
•			82.13 (-9.40)	90.35 (+3.31)	89.30 (-0.06)	94.81 (-0.16)	66.14 (-18.07)	73.49 (-6.30)	62.55 (-22.64)	53.88 (-7.87)	72.43 (-7.74)	51.50 (-15.78)	82.90 (-9.62)	60.01 (-25.48)	**88.90** (-0.69)	**94.38** (-0.33)
	•		85.61 (-5.92)	83.17 (-3.87)	86.97 (-2.39)	94.84 (-0.13)	71.70 (-12.51)	52.52 (-27.27)	75.56 (-9.63)	53.50 (-8.25)	71.86 (-8.31)	49.98 (-17.30)	**81.73** (-10.79)	59.85 (-25.64)	78.54 (-11.05)	**77.80** (-16.91)
		•	81.34 (-10.19)	74.52 (-12.52)	87.65 (-1.71)	94.58 (-0.39)	**87.62** (+3.41)	68.28 (-11.51)	77.58 (-7.61)	53.99 (-7.76)	71.49 (-8.68)	49.17 (-18.11)	82.70 (-9.82)	59.72 (-25.77)	81.14 (-8.45)	**79.92** (-14.79)
•	•		89.02 (-2.51)	90.16 (+3.12)	89.65 (+0.29)	94.04 (-0.93)	72.23 (-11.98)	60.47 (-19.32)	77.51 (-7.68)	55.07 (-6.68)	71.25 (-8.92)	51.11 (-16.17)	83.69 (-8.83)	63.28 (-22.21)	**89.40** (-0.19)	**94.63** (-0.08)
•		•	89.36 (-2.17)	92.96 (+5.92)	89.07 (-0.29)	94.98 (+0.01)	79.68 (-4.53)	72.63 (-7.16)	78.53 (-6.66)	55.86 (-5.89)	75.85 (-4.32)	58.88 (-8.40)	84.11 (-8.41)	63.86 (-21.63)	**89.38** (-0.21)	**94.36** (-0.35)
	•	•	88.34 (-3.19)	77.46 (-9.58)	86.81 (-2.55)	94.84 (-0.13)	83.10 (-1.11)	61.59 (-18.20)	79.01 (-6.18)	56.55 (-5.20)	74.10 (-6.07)	56.49 (-10.79)	88.39 (-4.13)	74.92 (-10.57)	**88.96** (-0.63)	**94.59** (-0.12)
Mean Absolute Value (MAV)			85.97 (5.56)	84.77 (6.39)	88.24 (1.22)	94.68 (0.29)	76.75 (8.60)	64.83 (14.96)	75.12 (10.07)	54.81 (6.94)	72.83 (7.34)	52.86 (14.43)	83.92 (8.60)	63.61 (21.88)	**86.05** (3.54)	**89.28** (5.43)

- The numbers in parentheses indicate changes in performance compared to the full-modality scenario, except for the last row.
- The MAV of each performance and the changes in performance are shown in the last row.
- The best results within single-models in each row are shown in bold.
- The best results in each row are underlined.
- The smallest change in performance within single-models in each row is shown in blue, compared with the full-modality scenario.

Table 2: Comparison with other methodologies for hypotension prediction using VitalDB.

- **MaskMentor (Zhao et al. 2024)** is a masked self-teaching framework in which a teacher model is trained using complete multimodal inputs, whereas a student model learns from partially missing inputs, using the predictions of the teacher as pseudo-labels.

Evaluation Scheme

To evaluate the performance of each method, 5-fold subject-group cross-validation was conducted. The datasets were divided into three subsets: pretrain, train, and test. The pretrain subset was utilized for SSL without labels and the train subset was used for linear evaluation with a limited amount of labeled data. A downstream classifier was attached to the pretrained network to perform two downstream tasks. Finally, the performance of the proposed model and baselines was investigated using the test subset.

Evaluation Metric

Model performance was evaluated using accuracy (ACC) and the area under the receiver operating characteristic (ROC)

curve (AUC). AUC is a reliable evaluation metric for tasks with class imbalance, such as sleep stage classification and hypotension prediction.

ACC indicates the proportion of correctly classified samples and provides an overall assessment of model performance across the entire dataset. ACC is defined as:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{14}$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

AUC quantifies the area under the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR). The ROC curve plots the TPR on the y-axis against the FPR on the x-axis. AUC is defined as follows:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d\mathrm{FPR} \tag{15}$$

where TPR and FPR are calculated as follows: $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ and $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$.

Results

Comparison of Other Methodologies

Two downstream tasks were conducted to evaluate the proposed model: (1) sleep stage classification using the Sleep-EDFX dataset (Table 1), and (2) hypotension prediction using the VitalDB dataset (Table 2). Its performance was compared with those of existing single-model approaches (i.e., MultiMAE (Bachmann et al. 2022), RobustSsF (Lee and Kim 2023), CroSSL (Deldari et al. 2024), and MaskMentor (Zhao et al. 2024)) and dedicated-model approaches (i.e., ContraWR (Yang et al. 2021) and SynthSleepNet (Lee et al. 2025)). In addition, various modality combination experiments were performed to assess the performance and robustness of the model under missing modality scenarios.

Comparison within Single-Model Approaches

(Performance under Missing Modalities) PhysioME consistently outperformed the existing methods across a wide range of missing modality scenarios. For both downstream tasks, it achieved the highest average ACC and AUC across all the missing modality combinations. In the sleep stage classification task, PhysioME demonstrated the best performance in nearly all scenarios, except for the single-modality case using EEG Pz-Cz (bold highlights in Table 1). In the hypotension prediction task, although some models showed slightly better ACC in specific cases (e.g., single-modality ECG and PPG), PhysioME achieved the highest AUC in all scenarios (bold highlights in Table 2).

(Robustness under Missing Modalities) PhysioME exhibited the smallest performance degradation compared to the full-modality setting in most missing modality scenarios, maintaining stable accuracy and AUC across conditions (blue highlights in Tables 1 and 2). When comparing the mean absolute difference in performance relative to the full-modality case, PhysioME showed the lowest values across all metrics: 3.97% (ACC) and 2.04 (AUC) for sleep stage classification, and 3.54% (ACC) and 5.43 (AUC) for hypotension prediction. These results indicate that PhysioME maintained stable performance even under missing modality conditions.

Comparison between Single-Model and Dedicated-Model Approaches

(Competitiveness Against Task-Specific Models) Despite being a single-model framework, PhysioME demonstrated performance comparable to or better than dedicated models across various modality combinations for both downstream tasks (underlines in Tables 1 and 2). This challenges the common assumption that dedicated models are inherently more advantageous, indicating that a single-model can achieve a robust and reliable performance.

Specifically, in the sleep stage classification task, PhysioME consistently outperformed ContraWR across all modality combinations and surpassed SynthSleepNet in most cases. In the hypotension prediction task, PhysioME achieved a performance on par with that of the dedicated models. Notably, in terms of the AUC, PhysioME outperformed ContraWR in all scenarios except for the single-modality ECG case. While SynthSleepNet exhibited a higher overall AUC, PhysioME achieved superior results for the ABP+PPG and ECG+PPG combinations.

Ablation Studies

Ablation studies were conducted on the Sleep-EDFX dataset to determine the optimal configuration for PhysioME. The performance was evaluated under three modality settings: (i) EEG Fpz-Cz, (ii) EEG Fpz-Cz + EOG, and (iii) EEG Fpz-Cz + EEG Pz-Cz + EOG, while being limited to 10 epochs for training efficiency.

Modality Backbone Network As listed in Table 3, the DP-NeuroNet backbone achieved the highest overall performance, confirming its superior ability to encode the distinctive characteristics of physiological signals. Specifically, CNN-based architectures (e.g., ResNet and EfficientNet) outperformed ViT-based architectures (e.g., ViT and Swin Transformer). Moreover, DP-NeuroNet surpassed the original NeuroNet across all modality combinations; in the single-modality setting, it improved ACC by 6.09% and AUC by 3.22.

Model Name	*Modality 1*		*Modality 2*		*Full*	
	ACC	AUC	ACC	AUC	ACC	AUC
ResNet	70.13	89.12	71.69	89.80	74.66	90.97
EfficientNet	68.66	86.21	72.01	88.54	74.38	89.53
ViT	42.68	66.08	47.16	66.99	48.18	65.28
Swin Transformer	51.22	73.76	55.95	75.79	57.19	75.65
NeuroNet	70.42	89.48	74.43	91.23	75.90	92.02
DP-NeuroNet	**76.51**	**92.70**	**77.77**	**93.85**	**78.83**	**94.46**

Table 3: Comparison of modality backbone networks.

Presence or Absence of a Gradient on Restoration Decoder This study investigated whether the gradients from the restoration decoder should be back-propagated to the upstream network components. As listed in Table 4, disabling

the gradient propagation (i.e., applying a stop-gradient operation) consistently outperformed the alternative across all modality combinations. These findings suggest that isolating the restoration decoder enables it to focus more effectively on reconstructing missing modality tokens, thereby yielding superior overall performance.

Gradient P/A	*Modality 1*		*Modality 2*		*Full*	
	ACC	AUC	ACC	AUC	ACC	AUC
with gradient	75.54	92.31	76.52	93.21	77.13	93.38
without gradient	**76.51**	**92.70**	**77.77**	**93.85**	**78.83**	**94.46**

* P/A = Presence or Absence.

Table 4: Performance based on the presence or absence of a gradient.

Restoration Strategy To determine the optimal strategy for handling missing modalities in PhysioME, three alternatives were evaluated: (i) masked tokens, (ii) memory tokens, and (iii) restoration decoder tokens (Table 5). In previous studies (Woo et al. 2023), memory tokens have been variously denoted as modality or learnable tokens. The results of this study show that the proposed restoration decoder approach delivered the highest performance across all metrics and modality configurations. Notably, in the single-modality scenario, the restoration decoder strategy outperformed the masked and memory token strategies by 4.94% and 2.29% in ACC, respectively.

Restoration Strategy	*Modality 1*		*Modality 2*		*Full*	
	ACC	AUC	ACC	AUC	ACC	AUC
Masked Token	71.57	85.47	73.02	87.16	74.50	87.63
Memory Token	74.22	91.63	75.79	91.80	77.08	93.43
Ours	**76.51**	**92.70**	**77.77**	**93.85**	**78.83**	**94.46**

* Ours: Restoration Decoder based Token

Table 5: Performance of the restoration strategy based on PhysioME.

Restoration Decoder Dimension and Depth Table 6 lists the effects of the decoder dimensions and depth on the model performance. Overall, increasing either parameter led to systematic improvements in all evaluation metrics. In particular, a configuration with 512 hidden dimensions and a depth of eight layers yielded the best results for most modality combinations. These findings suggest that a high-capacity restoration decoder can more effectively reconstruct information for missing modalities, thereby enhancing the overall system performance.

Discussion and Conclusion

This study proposed PhysioME, a physiological signal-specific framework designed to address the challenges of missing modalities frequently encountered in real-world clinical environments. Building upon SynthSleepNet (Lee et al. 2025), an SSL framework based on MultiMAE (Bachmann et al. 2022), PhysioME is specifically designed to capture

Dim	*Depth*	*Modality 1*		*Modality 2*		*Full*	
		ACC	AUC	ACC	AUC	ACC	AUC
128	4	75.29	92.40	76.86	93.54	77.37	93.88
	6	75.19	92.67	76.31	93.76	77.99	94.21
	8	75.48	92.29	76.80	93.89	77.95	94.27
256	4	75.36	92.38	76.62	93.67	77.87	94.03
	6	75.42	92.58	77.33	93.81	77.98	94.13
	8	75.48	92.02	77.48	93.68	77.82	94.05
512	4	75.59	92.43	77.16	93.62	78.04	94.19
	6	75.69	92.49	77.12	**93.88**	78.21	94.11
	8	**76.51**	**92.70**	**77.77**	93.85	**78.83**	**94.46**

Table 6: Performance of the restoration decoder based on dimensions and depth.

the temporal characteristics of physiological signals and heterogeneity across sensor types, enabling robust handling of various missing modality scenarios.

Leveraging this architectural design, PhysioME demonstrated strong predictive performance and generalization across diverse missing modality settings owing to several key structural components. First, the DP-NeuroNet serves as a core component in the success of PhysioME. By applying two different temporal augmentations to a dual-path network that shares weights across identical NeuroNet instances, DP-NeuroNet effectively captures the temporal dynamics of physiological signals and significantly enhances the quality of the modality-specific representations. These high-dimensional representations play a crucial role in improving both the prediction accuracy and generalization capability (Table 3). Second, the restoration decoder enables semantically informed reconstruction by modeling cross-modal interactions rather than relying on simple imputation methods such as mean substitution or input replication. In particular, the restoration decoder is optimized using the stop-gradient technique, which receives input from the multimodal encoder while preventing the gradients from flowing backward. This allows the decoder to focus solely on the restoration tasks. In addition, the decoder benefits from a deeper architecture that facilitates effective learning (Tables 5 and 6).

In clinical practice, it is often difficult to acquire all the physiological signals owing to various constraints. For example, EEG acquisition may be physically restricted during brain surgery (Bitar, Khan, and Rosenthal 2024) and ABP sensors may not be applicable to patients receiving anticoagulants or those with impaired clotting function (Puckett et al. 2003). Similarly, severe motion artifacts can compromise the reliability of ECG signals, and PPG measurements may be unreliable under conditions such as hypothermia or peripheral vasoconstriction (Littmann 2021). Despite these limitations, PhysioME consistently maintains predictive stability and robustness, demonstrating its potential to reliably support clinical decision-making in real-world healthcare environments.

However, this study has several limitations. First, the performance was evaluated under a pre-defined set of modality combinations. Generalization to entirely new combinations or domain transfer scenarios requires further investigation. Second, the architectures of DP-NeuroNet and the restora-

tion decoder are relatively complex in terms of the training efficiency. Specifically, a restoration decoder must be instantiated for each new modality to introduce scalability challenges. Future work will evaluate cross-hospital generalization using multi-center datasets collected from diverse institutions and patient populations, and further assess robustness not only to missing modalities but also to noise and artifacts commonly observed in clinical data.

Appendix

Background: NeuroNet

NeuroNet (Lee et al. 2024) is a self-supervised learning (SSL) framework that combines contrastive learning and masked prediction tasks to extract generalizable representations from unlabeled input data. In this study, NeuroNet was adopted as the core component to extract modality-specific features from each physiological signal.

The NeuroNet encoder is built around a frame network module. To effectively process time-series data, this module comprises a shared convolutional block and three parallel feature extractors with kernel sizes of 3, 5, and 7, each of which follows a convolutional residual-block architecture. The feature vectors are concatenated and passed through an MLP layer, producing frame-wise vectors.

While performing the masked prediction task based on the MAE strategy, a subset of the frame-wise vectors is randomly selected and input into a ViT-based encoder along with a class token. The encoder generates a latent representation from the observed frames, and the decoder reconstructs the masked frames using the latent representation and mask tokens. The reconstruction loss is calculated only over the masked frames using the MSE loss:

$$\mathcal{L}_{\text{inter\_recon\_loss}} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \left\| r^{(m)} - z^{(m)} \right\|_2^2 \quad (16)$$

where $\mathcal{M}$ denotes the set of masked frame indices. The vectors r and m correspond to the outputs of the decoder and the frame network, respectively.

In addition, NeuroNet employs contrastive learning based on the NT-Xent loss to ensure consistency across different views of the same input sequence. Two random samples are generated from a single input sample, and each sample is processed using the encoder and an MLP to produce the latent vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. The NT-Xent loss is defined as:

$$\mathcal{L}_{\text{NT-Xent\_loss}} = -\log \frac{\exp\left(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau\right)}, \quad (17)$$

where $\text{sim}(\cdot)$ denotes the cosine similarity and τ is the temperature parameter.

NeuroNet performs two independent masked prediction tasks using the same encoder and decoder. As a result, the total loss function combines two reconstruction losses and one contrastive loss:

$$\mathcal{L}_{\text{total}} = \frac{1}{2}\left(\mathcal{L}_{\text{inter\_recon\_loss}_1} + \mathcal{L}_{\text{inter\_recon\_loss}_2}\right) + \alpha \mathcal{L}_{\text{NT-Xent\_loss}}, \quad (18)$$

where α is a hyperparameter that balances the reconstruction and contrastive losses.

Hyperparameter Setting

The experimental environment was a machine equipped with an Intel Xeon Gold 6338 CPU 2.00GHz, 256GB of RAM, and two NVIDIA GPU Ada 6000. The codes, including data preprocessing and model development, were implemented based on Pytorch 2.7.1 and Python 3.11. The hyperparameter settings are listed in Appendix Table 7.

	Sleep Stage Classification	***Hypotension Prediction***
	DP-NeuroNet	
epoch	50	50
sampling rate	100	100
batch size	128	128
frame size (sec)	4	3
overlap step (sec)	1	3
encoder dim	512	512
encoder depth	8	8
encoder head	6	6
decoder dim	256	256
decoder depth	8	8
decoder head	4	4
mask ratio	0.8	0.8
balance scale	1.0	1.0
projection hidden	[1024, 512]	[1024, 512]
optimizer	AdamW	AdamW
optimizer momentum	(0.9, 0.999)	(0.9, 0.999)
learning rate	1e-05	1e-05
	PhysioME	
physiological signals	Fpz-Cz / Pz-Cz / EOG	ABP / ECG / PPG
epoch	50	50
batch size	512	512
encoder dim	512	512
encoder depth	6	4
encoder head	8	8
decoder dim	256	256
decoder depth	4	3
decoder head	8	8
recon decoder dim	256	256
recon decoder depth	8	8
recon decoder head	8	8
projection hidden	[1024, 512]	[1024, 512]
temperature scale	0.1	0.1
mask ratio	0.4	0.4
optimizer	AdamW	AdamW
optimizer momentum	(0.9, 0.999)	(0.9, 0.999)
learning rate	2e-4	2e-4
balance scale	α: 1.0 / β: 1.0 / γ: 1.0	α: 1.0 / β: 1.0 / γ: 1.0
lora (r)	4	4
lora (alpha)	16	16
lora (dropout)	0.05	0.05
	Downstream Task	
epoch	20	20
batch size	512	512
optimizer	AdamW	AdamW
optimizer momentum	(0.9, 0.999)	(0.9, 0.999)
learning rate	1e-05	1e-05

Table 7: Hyperparameter Settings

References

Bachmann, R.; Mizrahi, D.; Atanov, A.; and Zamir, A. 2022. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, 348–367. Springer.

Bitar, R.; Khan, U. M.; and Rosenthal, E. S. 2024. Utility and rationale for continuous EEG monitoring: a primer for the general intensivist. *Critical Care*, 28(1): 244.

Deldari, S.; Spathis, D.; Malekzadeh, M.; Kawsar, F.; Salim, F. D.; and Mathur, A. 2024. Crossl: Cross-modal self-supervised learning for time-series through latent masking. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 152–160.

Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; and Dean, J. 2019. A guide to deep learning in healthcare. *Nature medicine*, 25(1): 24–29.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Iranfar, A.; Arza, A.; and Atienza, D. 2021. Relearn: A robust machine learning framework in presence of missing data for multimodal stress detection from physiological signals. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 535–541. IEEE.

Kemp, B.; Zwinderman, A. H.; Tuk, B.; Kamphuisen, H. A.; and Oberye, J. J. 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9): 1185–1194.

Lee, C.-H.; Kim, H.; Han, H.-j.; Jung, M.-K.; Yoon, B. C.; and Kim, D.-J. 2024. Neuronet: A novel hybrid self-supervised learning framework for sleep stage classification using single-channel eeg. *arXiv preprint arXiv:2404.17585*.

Lee, C.-H.; Kim, H.; Yoon, B. C.; and Kim, D.-J. 2025. Toward Foundational Model for Sleep Analysis Using a Multimodal Hybrid Self-Supervised Learning Framework. *arXiv preprint arXiv:2502.17481*.

Lee, H.-C.; Park, Y.; Yoon, S. B.; Yang, S. M.; Park, D.; and Jung, C.-W. 2022. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1): 279.

Lee, J.; and Kim, D.-S. 2023. RobustSsF: Robust Missing Modality Brain Tumor Segmentation with Self-supervised Learning-Based Scenario-Specific Fusion. In *Workshop on Machine Learning for Multimodal Healthcare Data*, 43–53. Springer.

Littmann, L. 2021. Electrocardiographic artifact. *Journal of electrocardiology*, 64: 23–29.

Puckett, L. G.; Barrett, G.; Kouzoudis, D.; Grimes, C.; and Bachas, L. G. 2003. Monitoring blood coagulation with magnetoelastic sensors. *Biosensors and Bioelectronics*, 18(5-6): 675–681.

Reza, M. K.; Prater-Bennette, A.; and Asif, M. S. 2024. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Wang, S.; Cao, J.; and Philip, S. Y. 2020. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8): 3681–3700.

Woo, S.; Lee, S.; Park, Y.; Nugroho, M. A.; and Kim, C. 2023. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2776–2784.

Wu, R.; Wang, H.; Chen, H.-T.; and Carneiro, G. 2024. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*.

Yang, C.; Xiao, D.; Westover, M. B.; and Sun, J. 2021. Self-supervised EEG representation learning for automatic sleep staging. *arXiv preprint arXiv:2110.15278*.

Zhao, Z.; Li, J.; Wang, L.; Wang, Y.; and Lu, H. 2024. MaskMentor: Unlocking the Potential of Masked Self-Teaching for Missing Modality RGB-D Semantic Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1915–1923.