MULTIMODAL SELF-SUPERVISED LEARNING FOR WEARABLE SLEEP STAGING USING PHOTOPLETHYSMOGRAPHY AND ACCELEROMETER SIGNALS

Juhyeon Lee[1,2†], Simon A. Lee[1,3†], Cyrus Tanade[1], Viswam Nathan[1], Megha Thukral[1,4†]
Hao Zhou[1,5†], Keum San Chun[1], Sharanya Arcot Desai[1]

[1] Samsung Research America, [2] University of Massachusetts, Amherst,
[3] University of California, Los Angeles, [4] Georgia Institute of Technology,
[5] The Pennsylvania State University

ABSTRACT

Sleep staging is essential for quantifying sleep patterns and diagnosing sleep disorders, but the clinical gold standard, polysomnography (PSG), is obtrusive, manually scored, and impractical for large-scale or longitudinal monitoring. Wearable devices offer a minimally obtrusive alternative by using photoplethysmography (PPG) and accelerometry (ACC) for automated sleep staging, but existing supervised models depend on large-scale PSG-labeled datasets. To address this limitation, we propose a multimodal self-supervised learning (SSL) framework that jointly pretrains representations from PPG and ACC using 23,000 hours of unlabeled data. For downstream classification, we employ a causal Mamba-based sequence model that captures long-range temporal dependencies while preserving causal ordering, enabling potential real-time deployment. On the DREAMT dataset, our method improves AUROC from 0.59 to 0.72 and Cohen's κ from 0.06 to 0.30 over fully supervised baselines for four-class sleep staging, demonstrating improved generalization and scalability.

***Index Terms*—** Sleep Staging, Wearable Sensors, Self-supervised Learning, Multimodal Representation Learning

1. INTRODUCTION

Sleep is closely linked to numerous health outcomes, including cardiometabolic disease, mood disorders, cognitive decline, and accident risk [1]. An estimated 50–70 million people in the United States suffer from chronic sleep disorders such as insomnia and sleep apnea, underscoring the need for scalable and continuous sleep assessment [2].

Sleep staging is a central task in sleep assessment, providing quantitative measures of sleep pattern and supporting the diagnosis of sleep disorders [3]. The current clinical gold standard for sleep staging is overnight polysomnography (PSG), which records neural and physiological signals in a specialized laboratory setting and requires manual scoring by trained experts [3]. However, PSG is obtrusive, requiring numerous sensors and wires, disrupts habitual sleep, and is impractical for routine or longitudinal monitoring in home environments [4]. Manual scoring is also labor-intensive and not scalable, limiting the feasibility of continuous or population-scale sleep tracking [3].

Wearable devices, such as smartwatches, offer a minimally burdensome alternative for continuous and automated sleep staging in home settings and have already achieved widespread adoption [5, 6, 7]. Although they cannot directly measure brain activity, machine learning models infer sleep stages using cardiovascular dynamics from photoplethysmography (PPG) and movement patterns from accelerometry (ACC) [5]. Recent studies have shown that supervised models trained on wearable data can achieve moderate to substantial agreement with PSG [8, 9]. However, these models depend on large PSG-aligned datasets for training, which are costly and labor-intensive to collect. For example, WatchSleepNet achieved modest agreement with PSG labels on a challenging wrist-worn PPG dataset, but relied on transfer learning, pretraining on high-quality electrocardiogram (ECG) signals and PSG labels from over 7,000 participants [10]. Similarly, Silva *et al*. developed a supervised learning model on 1,522 PSG-labeled nights collected from commercial smartwatches and reported modest agreement with PSG-based staging [8].

To address the reliance on large PSG-annotated datasets, we propose a multimodal self-supervised learning (SSL) framework that jointly pretrains representations from PPG and ACC signals. Although several SSL approaches have been proposed for sleep staging, most focus on PSG-based datasets [11, 12] or learn unimodal representations using data augmentations or masked reconstruction objectives [13, 14]. Leveraging unlabeled data from commercial smartwatches, our framework learns robust and transferable representations for both modalities. For downstream classification, the pretrained representations are fine-tuned on a limited PSG-labeled dataset for four-class sleep staging using a causal Mamba-based sequence model. This model that captures long-range temporal dependencies across an entire night

†Work done while interning at Samsung Research America

while preserving causal ordering, making it well-suited for real-time applications such as closed-loop interventions (e.g., adaptive sounds, vibrations, lights, or temperature control based on detected sleep stage).

Our method outperforms supervised baselines, demonstrating that multimodal self-supervised learning improves generalization and robustness when labeled wearable datasets are relatively small. Combined with the rapid adoption of wearable devices, this approach shows potential for leveraging massive unlabeled datasets to learn more generalizable representations and enable scalable, robust, and continuous sleep staging in real-world settings.

2. METHODOLOGY

2.1. Overall Framework

Our model consists of two key components: (i) modality-specific 1D convolutional encoders that learn representations and serve as feature extractors for PPG and ACC signals, and (ii) a causal Mamba-based classifier for sleep staging. Fig. 1 illustrates the overall framework. The encoders are first pretrained using cross-modal contrastive learning on over 23,000 hours of unlabeled PPG and ACC data. The pretrained encoders are then combined with a causal Mamba classifier and fine-tuned with supervised learning on PSG-annotated wearable data.

2.2. Pretraining Dataset and Preprocessing

For pretraining, we used 23,000 hours of raw photoplethysmography (PPG) and three-axis accelerometry (ACC) data collected from 566 participants wearing Samsung Galaxy Watches. The unlabeled dataset includes both sleep and wake periods. Both signals were sampled at 25 Hz and band-pass filtered (0.5–12 Hz) to remove DC components and high-frequency noise. The filtered signals were segmented into non-overlapping 30-second windows, consistent with conventional PSG sleep staging, and z-score normalized on a per-segment basis.

2.3. Cross-Modal Contrastive Pretraining

To learn robust representations from PPG and ACC, we adopt a cross-modal contrastive learning approach that aligns the embeddings of temporally paired segments from the two modalities. This design is inspired by recent advances in multimodal representation learning in vision and health domains [15, 11].

Each 30-sec segment was independently processed by modality-specific encoders, implemented as ResNet1D-18 networks with 18 1D-convolutional blocks. Before encoding, inputs were projected using a linear layer to perform random input-projection masking, an augmentation method introduced in prior work [16]. At each training step, a random subset of time steps in the projected features was set to zero, encouraging the encoders to learn context-invariant and redundancy-tolerant representations. The encoders output d-dimensional embeddings $\mathbf{z}_i^{\mathrm{PPG}}, \mathbf{z}_i^{\mathrm{ACC}} \in \mathbb{R}^d$ for each segment i.

For a batch of N paired segments, the positive pair is defined as (z_i^{PPG}, z_i^{ACC}), and negatives are all other in-batch samples. Using cosine similarity $\mathrm{sim}(\cdot,\cdot)$ and a temperature parameter $\tau = 0.2$, the symmetric InfoNCE loss is:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[\log\frac{\exp(\mathrm{sim}(z_i^{PPG}, z_i^{ACC})/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(z_i^{PPG}, z_j^{ACC})/\tau)} + \log\frac{\exp(\mathrm{sim}(z_i^{ACC}, z_i^{PPG})/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(z_i^{ACC}, z_j^{PPG})/\tau)}\Big]. \tag{1}$$

This symmetric InfoNCE objective [17] encourages alignment between PPG and ACC representations corresponding to the same temporal window, while pushing apart embeddings from different temporal windows or subjects. The design leverages the shared underlying sleep state as the supervisory signal, enabling representation learning without PSG labels.

Pretraining was performed using the Adam optimizer [18] with a learning rate of 1×10^{-4}, batch size 128. Models were trained for 20 epochs, with 20 % of the pretraining data reserved for validation, and the model with the lowest validation loss was selected for downstream fine-tuning.

2.4. Downstream Classification: Night-Long Mamba

For supervised fine-tuning, the encoder outputs were concatenated, linearly projected, and layer-normalized to form fused per-segment representations. The full-night sequence for each subject, $(\tilde{z}_1, \ldots, \tilde{z}_T)$, was then passed through a one-layer causal Mamba state-space model. Mamba is a selective state-space architecture that efficiently captures long-range dependencies with linear-time complexity and constant memory usage [19]. Unlike bidirectional recurrent or transformer models, Mamba operates causally: the prediction at time t depends only on inputs up to t. This property ensures that the model is compatible with streaming and real-time deployment.

The Mamba classifier outputs per-segment logits $y_{1:T} \in \mathbb{R}^C$, where C is the number of sleep stages (Wake, Light, Deep, REM). We optimized the model using cross-entropy loss with the Adam optimizer [18], a learning rate of 5×10^{-5}. Training was run for up to 200 epochs with a patience of 20 epochs, and early stopping was applied based on the best validation Cohen's κ score, as this metric measures overall agreement and is less affected by class imbalance. In our experi-

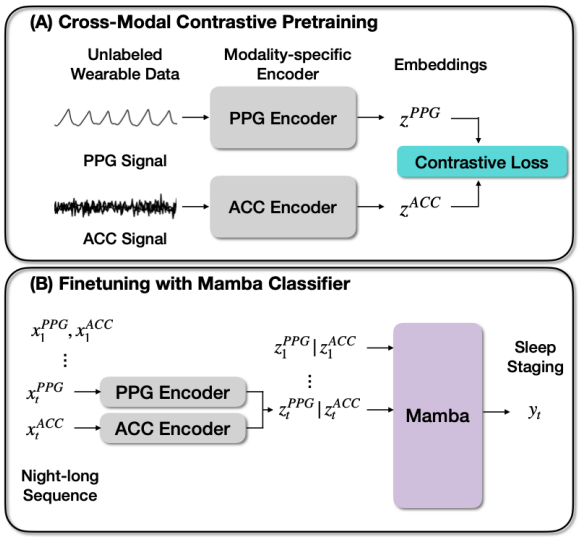

Fig. 1. Overall framework of the proposed approach. (A) cross-modal contrastive pretraining learns modality-specific representations from unlabeled PPG and ACC data. (B) the pretrained encoders are fine-tuned with a causal Mamba classifier for night-long sleep staging.

ments, Mamba was configured with an input dimension of 256 and a hidden state dimension of 64.

3. EXPERIMENTAL DESIGN

3.1. Evaluation Dataset and Preprocessing

For finetuning and performance evaluation, we used the DREAMT dataset [20], which is publicly available on PhysioNet [21]. DREAMT contains one-night recordings from 100 participants (55 female) aged 21–86 years (mean: 56.2 ± 16.6 years), of whom 92 were diagnosed with sleep disorders, making the dataset more challenging for sleep staging compared to healthy populations. Each subject wore an Empatica E4 wristband, which provides blood volume pulse (BVP) at 64 Hz, derived from raw PPG using a proprietary algorithm, and three-axis ACC signals at 32 Hz. Although our pretraining used raw PPG acquired from green-light exposure, DREAMT provides BVP signals derived from a combination of green- and red-light measurements. Because BVP reflects the same peripheral blood volume changes as PPG and preserves cardiovascular dynamics such as heart rate and heart-rate variability, our pretrained representations are expected to transfer effectively to BVP signals. Simultaneous PSG served as the clinical reference and was annotated into 30-second segments (wake, NREM1, NREM2, NREM3, REM). For four-class sleep staging, N1 and N2 are merged into a light-sleep stage, and N3 is treated as deep sleep [22].

Both BVP and ACC signals were band-pass filtered (0.5–12Hz) with a Butterworth filter, downsampled to 25Hz, and segmented into non-overlapping 30-second windows aligned with PSG annotations. This preprocessing yielded 48,695 Light, 2,704 Deep, 8,365 REM, and 18,709 Wake segments. Each segment was z-score normalized prior to model input.

3.2. Experimental Setup

We designed several experimental conditions to evaluate the effectiveness of our framework:

Fully Supervised Baseline vs. Cross-Modal Pretraining: We compared models trained fully supervised from scratch on the DREAMT dataset with models pretrained on large-scale unlabeled PPG–ACC data and then trained on DREAMT with supervised learning. For the supervised baseline, both the encoders and the Mamba classifier were randomly initialized and trained solely on the DREAMT dataset. For pretrained models, we evaluated two settings: (i) end-to-end fine-tuning of the entire network and (ii) freezing the pretrained encoders while training only the causal Mamba classifier.

Unimodal vs. Multimodal Fine-tuning: Although our SSL framework jointly learned PPG and ACC representations during pretraining, we conducted ablations by fine-tuning models using only the PPG encoder, only the ACC encoder, and the combined PPG+ACC encoders. This experiment isolates the contribution of each modality and quantifies the added value of multimodal fusion for downstream sleep staging.

Sequential Classifier Comparison: We compared the causal Mamba classifier with other causal sequential models, including causal temporal convolutional networks (TCN) and unidirectional LSTM, using identical pretrained encoder outputs to ensure a fair comparison. We configured the TCN with a kernel size of 3, dilation 2, and hidden dimension 256, and the LSTM with a hidden dimension of 64.

3.3. Evaluation Metrics

Model performance was evaluated using five-fold participant-independent cross-validation, ensuring no participant overlap across training, validation, and test splits. Within each fold, the training set was further divided into training and validation subsets (train:val:test = 60:20:20 participants) and used for early stopping. Performance metrics were computed at the segment level and averaged across folds. We report macro-averaged area under the receiver operating characteristic curve (AUROC), Cohen's κ, and weighted F1-score, which together reflect class-balanced ranking performance, agreement beyond chance, and class prevalance.

4. RESULTS

4.1. Effect of Cross-Modal Pretraining

Table 1 compares models trained from scratch with those initialized from cross-modal contrastive pretraining. Training from scratch on DREAMT achieved limited performance (AUROC = 0.59, κ = 0.06, Weighted-F1 = 0.52), confirming that purely supervised learning struggles to generalize from a relatively small wrist-worn dataset. Cross-modal pretraining improved downstream performance. Freezing the encoder and training only the Mamba classifier improved AUROC by 0.13, κ by 0.22, and Weighted-F1 by 0.10 relative to training from scratch. End-to-end fine-tuning achieved the highest performance (AUROC = 0.72, κ = 0.30, Weighted-F1 = 0.62), showing that cross-modal SSL pretraining yields robust, transferable representations that outperform purely supervised models.

Table 1. **Cross-Modal Pretraining vs. No pretraining.** Effect of cross-modal contrastive pretraining on four-class sleep staging performance. Best results are in **bold**.

Model	AUROC	κ	Weighted-F1
No pretraining	0.59	0.06	0.52
Trained only Mamba	**0.72**	0.28	**0.62**
Finetuned all layers	**0.72**	**0.30**	**0.62**

4.2. Effect of Multimodal Fine-tuning

Table 2 summarizes the performance of models fine-tuned using only PPG, only ACC, or both modalities. ACC-only model achieved the highest performance across all metrics (AUROC=0.74, κ=0.32, Weighted-F1=0.64), outperforming both the PPG-only and the PPG+ACC fusion models. We attribute the limited additional gain from PPG+ACC fusion to two factors: (1) the domain gap between the raw PPG used for pretraining and the BVP signal used for downstream fine-tuning, which may reduce the contribution of PPG representations, and (2) the fact that multimodal pretraining has already aligned PPG and ACC embeddings, therefore additional fusion during fine-tuning may provide no or only limited performance gains.

Table 2. **Unimodal vs. Multimodal Finetuning.** Effect of multimodal fine-tuning on sleep staging performance using the same pretrained encoders. Best results are in **bold**.

Model	AUROC	κ	Weighted-F1
PPG only	0.72	0.26	0.59
ACC only	**0.74**	**0.32**	**0.64**
PPG + ACC	0.72	0.30	0.62

4.3. Sequence Classifier Comparison

Table 3 presents results comparing different causal sequence models applied to identical pretrained encoder outputs. The Mamba classifier achieved the best overall performance (AUROC = 0.72, κ = 0.30, Weighted-F1 = 0.62), outperforming both the unidirectional LSTM and causal TCN baselines. These findings suggest that Mamba more effectively captures long-range temporal dependencies while preserving causality, making it well-suited for real-time sleep staging applications.

Table 3. **Sequential Classifier Comparison.** Comparison of causal sequential classifiers using identical pretrained encoders. Best results are in **bold**.

Model	AUROC	κ	Weighted-F1
TCN	0.68	0.25	0.60
LSTM	0.70	0.29	0.61
Mamba	**0.72**	**0.30**	**0.62**

5. DISCUSSION AND CONCLUSION

While our framework achieves fair agreement with PSG labels for four-class classification ($\kappa \geq 0.2$) [23], its performance is comparable to other wearable-based SSL models. Yuan *et al.* applied self-supervised pretraining to ACC data using contrastive learning between original and augmented signals, achieving $\kappa = 0.39$ on their internal validation set and $\kappa = 0.32$ on an external set for three-class classification (Wake/REM/NREM) [13]. When aggregating our results into the same three-class setting, we obtained $\kappa = 0.35$ for ACC-only and $\kappa = 0.37$ for PPG+ACC. Similarly, Logacjov *et al.* used a masked reconstruction objective on ACC signals and reported a binary wake–sleep AUROC of 0.83 on the same DREAMT dataset [14]. Our wake–sleep AUROC is 0.83 for ACC-only and 0.82 for PPG+ACC, confirming comparable performance. Notably, their pretraining ACC datasets comprised around 100,000 and 35,000 participants, respectively, whereas our results were achieved with fewer subjects, suggesting potential for further gains as larger and more diverse unlabeled datasets become available.

Although our model employs a causal architecture suitable for real-time sleep tracking and closed-loop interventions, we have not yet evaluated on-device performance, and it is not optimized for memory footprint or inference latency. Future work will explore model compression and quantization techniques to support on-device, real-time deployment.

In summary, our approach shows that multimodal SSL with PPG and ACC improves generalization for wearable sleep staging with limited labeled data. As wearable devices continue to generate vast amounts of continuous data, we envision that our SSL models will enable accurate, label-efficient, and scalable sleep staging in real-world settings.

6. REFERENCES

[1] Faith S Luyster and Patrick J Strollo Jr *et al.*, "Sleep: a health imperative," *Sleep*, vol. 35, no. 6, pp. 727–734, 2012.

[2] Bruce M Altevogt and Harvey R Colten *et al.*, "Sleep disorders and sleep deprivation: an unmet public health problem," 2006.

[3] Yun Ji Lee and Jae Yong Lee *et al.*, "Interrater reliability of sleep stage scoring: a meta-analysis," *Journal of Clinical Sleep Medicine*, vol. 18, no. 1, pp. 193–202, 2022.

[4] Jack D Edinger and Ana I Fins *et al.*, "Sleep in the laboratory and sleep at home: comparisons of older insomniacs and normal sleepers," *Sleep*, vol. 20, no. 12, pp. 1119–1126, 1997.

[5] Massimiliano De Zambotti and Nicola Cellini *et al.*, "Wearable sleep technology in clinical and research settings," *Medicine and science in sports and exercise*, vol. 51, no. 7, pp. 1538, 2019.

[6] Simon A Lee and Cyrus Tanade *et al.*, "Towards on-device foundation models for raw wearable signals," in *NeurIPS 2025 Workshop on Learning from Time Series for Health*, 2025.

[7] Simon A Lee and Cyrus Tanade *et al.*, "Himae: Hierarchical masked autoencoders discover resolution-specific structure in wearable time series," *arXiv preprint arXiv:2510.25785*, 2025.

[8] Fernanda B Silva and Luisa FS Uribe *et al.*, "Sleep staging algorithm based on smartwatch sensors for healthy and sleep apnea populations," *Sleep Medicine*, vol. 119, pp. 535–548, 2024.

[9] Kevin Kotzen and Peter H. Charlton *et al.*, "Sleepppg-net: A deep learning algorithm for robust sleep staging from continuous photoplethysmography," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 924–932, 2023.

[10] Will Ke Wang and Bill Chen *et al.*, "Watchsleepnet: A novel model and pretraining approach for advancing sleep staging with smartwatches," *Proceedings of Machine Learning Research*, vol. 287, pp. 1–20, 2025.

[11] Rahul Thapa and Bryan He *et al.*, "Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals," in *Forty-first International Conference on Machine Learning*.

[12] Chaoqi Yang and Cao Xiao *et al.*, "Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study," *JMIR AI*, vol. 2, no. 1, pp. e46769, 2023.

[13] Hang Yuan and Tatiana Plekhanova *et al.*, "Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality," *NPJ digital medicine*, vol. 7, no. 1, pp. 86, 2024.

[14] Aleksej Logacjov and Kerstin Bach *et al.*, "Long-term self-supervised learning for accelerometer-based sleep–wake recognition," *Engineering Applications of Artificial Intelligence*, vol. 141, pp. 109758, 2025.

[15] Alec Radford and Jong Wook Kim *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[16] Zhihan Yue and Yujing Wang *et al.*, "Ts2vec: Towards universal representation of time series," in *Proceedings of the AAAI conference on artificial intelligence*, 2022, vol. 36, pp. 8980–8987.

[17] Aaron van den Oord and Yazhe Li *et al.*, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[18] Diederik P Kingma and Diederik P Kingma *et al.*, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] Albert Gu and Tri Dao *et al.*, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[20] Ke Wang and Jiamu Yang *et al.*, "Dreamt: Dataset for real-time sleep stage estimation using multisensor wearable technology," *PhysioNet https://doi.org/10.13026/62AN-CB28*, 2024.

[21] Ary L Goldberger and Luis AN Amaral *et al.*, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[22] Vera Birrer and Mohamed Elgendi *et al.*, "Evaluating reliability in wearable devices for sleep staging," *NPJ Digital Medicine*, vol. 7, no. 1, pp. 74, 2024.

[23] J Richard Landis and Gary G Koch *et al.*, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.