Multimodal PSG foundation model integrating second to full-night scale

Anonymous Authors[1]

Abstract

Current sleep deep learning methods are typically limited by their focus on short, isolated segments of data, which overlooks the complex, multifaceted nature of full-night Polysomnography (PSG). To overcome these limitations, we present *PSG-M&m*, a foundation model designed to analyze both granular, second-by-second signal details and the broader, hour-long structural patterns of sleep. Our model employs a hierarchical dual-encoder architecture: a Macro-Encoder that evaluates long-term temporal trends throughout the night and a micro-Encoder that extracts localized characteristics from biosignals, optimized using a combination of masked autoencoding and multimodal contrastive learning. Macro-Encoder is refined through *Demographic-Guided Contrastive Learning*, which enhances its global representation by aligning sleep patterns with patient-demographics. Trained on a vast dataset (>20,000 PSG recordings, 158K hours), *PSG-M&m* significantly surpasses current foundation models. It offers improved generalizability for downstream clinical tasks, providing a more robust framework for comprehensive sleep analysis.

1. Introduction

Since humans spend about a third of their lives sleeping, it is a cornerstone of overall well-being. Recognizing this importance, researchers have invested heavily in understanding the intricate biology of sleep and combating the pervasive sleep disorders that impact millions globally (Benjafield et al., 2019). Currently, the clinical benchmark for evaluating sleep health is Polysomnography (PSG), a comprehensive diagnostic method that records a wide range of biological signals.

Because manual PSG analysis is extremely time-consuming, there has been a significant push toward using deep learning for automation. However, current models often struggle with limitations: they tend to be highly specialized and lack the flexibility needed to handle different types of tasks. Furthermore, their reliance on labeled data from experts restricts scalability and introduces errors stemming from inconsistent annotation (Perslev et al., 2021; Danker-Hopfe et al., 2004; 2009; Guillot et al., 2020). Most importantly, while clinical research highlights that sleep's "macro-architecture" is a critical indicator for long-term health and disease outcomes (Mander et al., 2017; Blackwell et al., 2011), existing models fail to adequately capture these long-term structural dynamics, hindering their utility for individualized clinical diagnosis and health forecasting.

To overcome these challenges, we introduce *PSG-M&m*, a comprehensive Sleep Foundation Model capable of understanding both the minute details (micro-structure) and the overall patterns (macro-structure) of sleep. Trained on a massive dataset of 20,964 PSG recordings—totaling 158,028 hours—*PSG-M&m* demonstrates exceptional versatility across various applications. Its architecture is divided into two main parts: **Micro-Encoder** uses a shared-private transformer design to interpret individual signal characteristics while simultaneously accounting for the relationships between different modalities. It is trained using a combination of masked autoencoding (MAE) (He et al., 2022) and contrastive learning (CL) (Oord et al., 2018) to ensure it learns resilient, high-resolution features. **Macro-Encoder** focuses on the broader context of a full night's sleep. By applying a Demographic-Guided Contrastive Learning technique, it aligns sleep sequence patterns with patient data (such as age, sex, and BMI). This allows the model to effectively map both healthy and pathological trajectories across the entire sleep cycle.

2. *PSG-M&m*

PSG-M&m's two encoders (Micro and Macro) are pretrained on large-scale PSG data with self-supervised learning. Independence from scored labels helps *PSG-M&m* to obtain generalizable embeddings and achieve scalability across diverse datasets, circumventing PSG scoring inconsistency.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

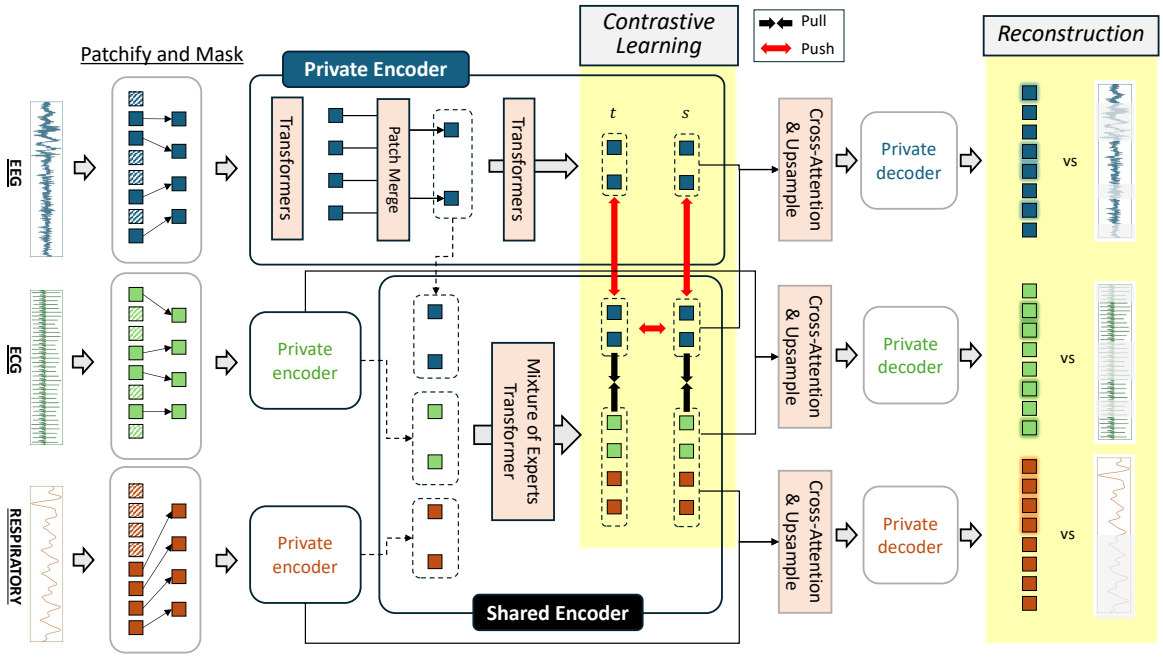


Figure 1. **Micro-Encoder design and pretraining method.** The Micro-Encoder adopts a private–shared encoder architecture. The model is trained with a hybrid objective of MAE and CL.

2.1. Micro-Encoder

Micro-Encoder architecture The Micro-Encoder serves to transform raw, multi-modal biosignals into a latent embedding space, allowing the system to analyze intricate micro-structural sleep patterns (see Figure 1). Initially, the system breaks raw signals into sequences of patches using convolutional embedding layers tailored to each modality. These patch sequences are then handled by a dual-pathway architecture: Modality-Private Encoders are dedicated to identifying patterns unique to specific signal types. Modality-Shared Encoder extracts physiological features that remain constant across modalities, while also identifying correlations between different signal types. Both pathways are built on a Transformer foundation (Vaswani et al., 2017), with the shared encoder utilizing a Mixture-of-Experts (MoE) design (Lepikhin et al., 2020) to boost both specialization and scalability.

The private encoders use a hierarchical structure to streamline temporal abstraction. In the initial layers, patches are processed individually. As the process moves deeper, every M consecutive patches are consolidated into a higher-level representation, which is passed both to the next private layer and to the shared encoder. Within the shared encoder, these combined representations from all private sources are merged to capture complex, inter-modal interactions.

Finally, the outputs from both the shared and private streams are integrated. The shared encoder's output is partitioned by modality and combined with its corresponding private encoder data through cross-attention. These fused embeddings are then up-sampled and sent to a decoder, which reconstructs the original biosignals from the masked segments.

Hybrid Self-supervised Learning The Micro-Encoder leverages a hybrid optimization strategy that combines MAE with CL to achieve superior performance. Within the MAE framework, the model masks a specific portion of patches across all modalities, requiring the encoders to derive information solely from the remaining visible data. These latent features are then transmitted to a decoder that attempts to reconstruct the missing signal segments. By successfully minimizing the error between the reconstructed output and the original raw data, the model gains a profound understanding of the core micro-structure and local textures inherent in the biosignals.

To further improve the latent space, we apply a contrastive learning objective to the embeddings produced by the shared encoder, ensuring that the resulting representations remain modality-agnostic and temporally consistent. We establish positive pairs by taking the average of shared embeddings derived from different modalities during the same time interval, a method inspired by SleepFM (Thapa et al., 2024; 2026). This approach compels the encoder to map various signals that correspond to an identical physiological state into close proximity within the latent space.

To refine these representations even further, we implement two distinct categories of negative pairs. Temporal negative pairs are formed by comparing shared embeddings from the same modality across different time slots, while representation negative pairs contrast a shared embedding with a private embedding originating from the same signal. Furthermore, we incorporate KoLeo regularization (Sablayrolles et al., 2018) to enhance the quality of the learned

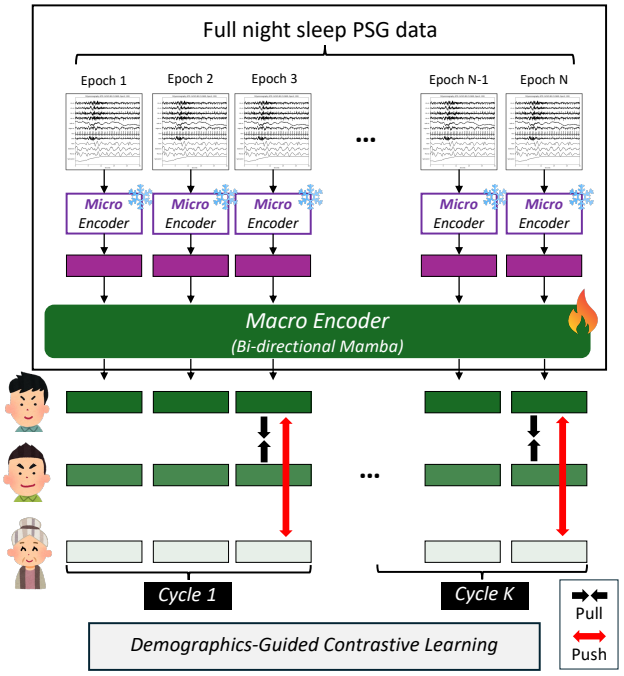


Figure 2. **Macro-Encoder design and pretraining.** Bi-directional Mamba is used for long-sequence modeling. DGCL aligns the sleep macro-structure between subjects with objective metadata.

representations. The final optimization objective is determined by calculating the total sum of the reconstruction, contrastive, and KoLeo losses.

2.2. Macro-Encoder

Macro-Encoder architecture As illustrated in Figure 2, the Macro-Encoder accepts the full sequence of epoch-level embeddings produced by the pretrained Micro-Encoder. To effectively manage these long-range dependencies, we utilize Mamba layers (Gu & Dao, 2024). We selected the Mamba architecture because it offers greater memory efficiency and superior scaling capabilities. Additionally, these layers are implemented bi-directionally to accommodate the significant variation in recording lengths between different subjects. By incorporating bi-directionality, the model is able to learn the sleep macro-structure by analyzing data from both the beginning and the end of the sleep session.

Demographic-guided Contrastive Learning (DGCL) To optimize the Macro-Encoder, we introduce a variant of contrastive learning that leverages age, sex and BMI as supervisory signals, which are known to influence sleep architecture. They provide objective ground truth, circumventing inter-scorer variability and evolving clinical guidelines associated with manual PSG labeling. Furthermore, the near-universal availability of demographic data ensures the scalability of DGCL.

To effectively capture sleep cycles, we partition the full-night sequence into 90-minute intervals. We then apply a soft-target contrastive objective that calculates the similarity between patient intervals based on their demographic profiles. Instead of traditional binary pairs, we adopt a weighted similarity approach (Yang et al., 2023). For a given interval c and subject pair (i, j), the loss for forward Mamba pass is formulated as follows:

$$\overrightarrow{\mathcal{L}}_{Ma,(i,j),c} = -w_{i,j} \log \frac{\exp(\langle \overrightarrow{Z}_i^c, \overrightarrow{Z}_j^c \rangle / \rho)}{\sum_{k=1}^{K} \exp(\langle \overrightarrow{Z}_i^c, \overrightarrow{Z}_k^c \rangle / \rho)} \quad (1)$$
$$w_{i,j} = \frac{\exp(-d_{i,j}/\upsilon)}{\sum_{k=1}^{K} \exp(-d_{i,k}/\upsilon)}$$

where $\overrightarrow{Z}_i^c$ and $\overrightarrow{Z}_j^c$ are the refined latent features of the subjects i and j at the end of c-th interval in the forward pass, ρ and υ are temperature scaling parameters and K is the total number of subjects in a batch. The *demographic distance* $d_{i,j}$ serves as the supervisory signal, calculated as following:

$$d_{i,j} = (|\text{age}_i - \text{age}_j| + |\text{BMI}_i - \text{BMI}_j|)/2 + \lambda_{sex} \quad (2)$$

Here, age and BMI are z-score normalized and λ_{sex} is a constant penalty applied only when subjects i and j are of different sex. The backward loss ($\overleftarrow{\mathcal{L}}_{Ma,(i,j),c}$) is defined symmetrically for the backward Mamba pass. The total Macro loss is aggregated over all intervals and subject pairs for forward and backward loss. This strategy effectively regularizes the latent space by pulling subjects with similar demographic profiles closer and pushing disparate subjects further apart.

2.3. Pretraining datasets and preprocessing

To pretrain *PSG-M&m*, we utilize the combination of open-sourced sleep datasets: SHHS1/2 (Zhang et al., 2018; Quan et al., 1997), KISS (Jeong et al., 2023), KVSS (Jang et al., 2025), Physionet 2018 (Goldberger et al., 2000; Ghassemi et al., 2018), MESA (Chen et al., 2015) and MrOS (Blackwell et al., 2011). The training split of SHHS1 and KISS are used for pretraining, with the partitioning scheme following the established protocols (Phan et al., 2021; Park et al., 2025). In total, we use 20,964 PSG recordings (158,028 hours) for pretraining. All raw biosignals are resampled at 100 Hz, bandpass and notch filtered, followed by z-score normalization prior to model input.

3. Experiments and Results

We evaluate *PSG-M&m* on three downstream tasks: sleep stage classification, sleep disordered breathing (SDB) segmentation and disease prediction. Two time series foundation models (MOMENT (Goswami et al., 2024) and UniTS (Gao et al., 2024)) and a sleep foundation model

Table 1. **Sleep stage classification results.** The results demonstrate the effectiveness of *PSG-M&m* compared to time series and sleep foundation models.

Dataset	Category	Models	Accuracy	Macro-F1	Kappa
SHHS1	Time series Foundation model	MOMENT-Base	79.4	65.6	**70.0**
		UniTS	64.2	59.2	53.3
	Sleep Foundation model	SleepFM-Disease	69.7	56.3	55.9
		PSG-M&m (Ours)	**81.9**	**74.1**	**70.0**
KISS	Time series Foundation model	MOMENT-Base	69.8	66.3	59.7
		UniTS	60.3	58.8	49.8
	Sleep Foundation model	SleepFM-Disease	57.6	58.4	46.1
		PSG-M&m (Ours)	**71.0**	**70.0**	**62.0**
CFS	Time series Foundation model	MOMENT-Base	71.4	52.4	56.7
		UniTS	64.1	58.1	53.3
	Sleep Foundation model	SleepFM-Disease	71.1	57.4	57.5
		PSG-M&m (Ours)	**80.7**	**71.2**	**72.9**
SOF	Time series Foundation model	MOMENT-Base	79.0	62.6	70.0
		UniTS	76.1	60.5	65.9
	Sleep Foundation model	SleepFM-Disease	60.1	47.0	42.3
		PSG-M&m (Ours)	**79.1**	**65.2**	**70.4**

(SleepFM-Disease (Thapa et al., 2026)) are chosen as baselines. Unless the baseline model's input format requires specific settings, we use the same input with identical preprocessing. Except disease prediction, the performance is measured on the test split of SHHS1 and KISS as well as two held-out datasets; CFS (Redline et al., 1995) and SOF (Spira et al., 2008). Disease prediction performance is only evaluated on SHHS1 where disease history is provided with PSG data.

Sleep stage classification Sleep staging refers to classifying each 30-second epoch into one of five sleep stages: Wake, REM, N1, N2 and N3. We conduct linear probing (Alain & Bengio, 2017) on target datasets. Weighted cross entropy loss is used to account for sleep stage imbalance. The results are provided in Table 1. Across all evaluation datasets, *PSG-M&m* outperforms all baselines, demonstrating its superior capability in capturing sleep-specific physiological features. Especially, *PSG-M&m* achieved 71.2% of Macro-F1 on CFS, and 65.2% of Macro-F1 on SOF, showing its generalizability to unseen datasets. Moreover, while KISS dataset is collected from different PSG systems (Nox and Embla) than other datasets (Compumedics), *PSG-M&m* shows robust sleep staging performance.

SDB segmentation SDB segmentation requires the high-resolution classification of each one-second interval into categories of normal or disordered breathing (hypopnea or apnea). We evaluate via a linear probing with weighted cross entropy loss. The results are summarized in Table 2. It demonstrates that *PSG-M&m*'s Macro-F1 is significantly higher than the established baselines. Although the raw accuracy on SHHS1 dataset appears slightly lower than some baselines, this metric is skewed by the severe class imbalance of SDB labels (10.6:1). In this context, the Macro-F1 score provides a more robust and equitable comparison. We attribute this superior performance to our Micro-Encoder's hybrid pretraining strategy.

Table 2. **SDB segmentation results.** The results demonstrate *PSG-M&m*'s superior capability to capture fine-grained details. Macro-F1 is emphasized as a more robust metric under severe class imbalance in SDB labels.

Dataset	Category	Models	Accuracy	Macro-F1
SHHS1	Time series Foundation model	MOMENT-Base	73.4	33.4
		UniTS	**88.2**	48.8
	Sleep Foundation model	SleepFM-Disease	77.5	39.4
		PSG-M&m (Ours)	77.3	**60.6**
KISS	Time series Foundation model	MOMENT-Base	76.0	58.5
		UniTS	79.8	63.6
	Sleep Foundation model	SleepFM-Disease	79.6	65.1
		PSG-M&m (Ours)	**81.8**	**75.4**
CFS	Time series Foundation model	MOMENT-Base	74.2	53.5
		UniTS	**85.5**	40.0
	Sleep Foundation model	SleepFM-Disease	82.1	46.6
		PSG-M&m (Ours)	79.9	**66.1**
SOF	Time series Foundation model	MOMENT-Base	70.3	18.3
		UniTS	**90.6**	35.7
	Sleep Foundation model	SleepFM-Disease	79.1	30.1
		PSG-M&m (Ours)	70.2	**52.5**

Table 3. **Disease prediction.** C-Index is reported on 6 selected diseases from SHHS1 dataset. *PSG-M&m* shows comparable performance to the SOTA PSG-based disease prediction model.

Models	Disease Outcomes					
	Angina	CVD death	CHF	CHD death	MI	Stroke
	C-Index					
SleepFM-Disease	0.632	**0.791**	0.764	**0.781**	0.636	**0.729**
PSG-M&m (Ours)	**0.778**	0.788	**0.793**	0.776	**0.662**	0.718

*CVD death: CardioVascular Disease death, CHF: Congestive Heart Failure, CHD death: Coronary Heart Disease death, MI: Myocardial Infarctions

Disease Prediction We utilize the SHHS dataset's disease histories to perform PSG-based disease prediction. We also employ linear probing for this task using the Cox Proportional Hazard loss. Following the evaluation protocol of SleepFM-Disease, we assess the predictive performance of *PSG-M&m* across six diseases. Performance is quantified via C-Index. The results are summarized in Table 3. *PSG-M&m* shows comparable performance to SleepFM-Disease, highlighting its potential for reliable clinical application in sleep-based healthcare.

4. Conclusion

In this study, we introduce *PSG-M&m*, a novel sleep foundation model pretrained on 20,964 PSG recordings. Our architecture utilizes a dual-encoder design: a Micro-Encoder with a private-shared transformer backbone optimized via hybrid of MAE and CL and a Mamba-based Macro-Encoder to model long-range temporal dependencies across full night. DGCL is employed to train the Macro-Encoder, which leverages objective metadata to encode sleep macro-structure . Extensive evaluations demonstrate that *PSG-M&m* outperforms existing foundation models across a wide spectrum of tasks.

References

Ahn, H. K., Na, Y., and Shin, H.-W. Refining sleep-disordered breathing annotations across multiple public sleep study datasets. *Journal of Sleep Research*, pp. e70264, 2025.

AIHub. Aihub. image of sleep quality assessment and sleep disorder diagnosis. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=210, 2020.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Benjafield, A. V., Ayas, N. T., Eastwood, P. R., Heinzer, R., Ip, M. S., Morrell, M. J., Nunez, C. M., Patel, S. R., Penzel, T., Pépin, J.-L., et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet respiratory medicine*, 7(8): 687–698, 2019.

Berry, R. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications. Version 2*, 2, 2012.

Berry, R. B., Brooks, R., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., Troester, M. T., and Vaughn, B. V. Aasm scoring manual updates for 2017 (version 2.4), 2017.

Blackwell, T., Yaffe, K., Ancoli-Israel, S., Redline, S., Ensrud, K. E., Stefanick, M. L., Laffan, A., Stone, K. L., and in Men Study Group, O. F. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *Journal of the American Geriatrics Society*, 59(12):2217–2225, 2011.

Chen, X., Wang, R., Zee, P., Lutsey, P. L., Javaheri, S., Alcántara, C., Jackson, C. L., Williams, M. A., and Redline, S. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38(6):877–888, 2015.

Choi, Y. R., Eo, G., Yoon, W., Lee, H., Jang, H., Kim, D. Y., Shin, H.-W., and Kim, H.-S. Poster: Home-based, on-device non-invasive obstructive sleep apnea monitoring with infrared video. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, pp. 708–709, 2024.

Danker-Hopfe, H., Kunz, D., Gruber, G., Klösch, G., Lorenzo, J. L., Himanen, S.-L., Kemp, B., Penzel, T., Röschke, J., Dorn, H., et al. Interrater reliability between scorers from eight european sleep laboratories in subjects with different sleep disorders. *Journal of Sleep Research*, 13(1):63–69, 2004.

Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., Heller, E., Loretz, E., Moser, D., Parapatics, S., et al. Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard. *Journal of Sleep Research*, 18(1):74–84, 2009.

Gao, S., Koker, T., Queen, O., Hartvigsen, T., Tsiligkaridis, T., and Zitnik, M. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37:140589–140631, 2024.

Ghassemi, M. M., Moody, B. E., Lehman, L.-W. H., Song, C., Li, Q., Sun, H., Mark, R. G., Westover, M. B., and Clifford, G. D. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference*, volume 45, pp. 1–4. IEEE, 2018.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, 2000.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, pp. 16115–16152, 2024.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Guillot, A., Sauvet, F., During, E. H., and Thorey, V. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28 (9):1955–1965, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jang, K., Choi, Y. R., Park, D., Shin, H.-W., and Kim, H.-S. *Poster: Home-based, On-Device, Non-contact Sleep Staging with Infrared Video*, pp. 1380–1382. Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400711299. URL https://doi.org/10.1145/3680207.3765693.

Jeong, J., Yoon, W., Lee, J.-G., Kim, D., Woo, Y., Kim, D.-K., and Shin, H.-W. Standardized image-based polysomnography database and deep learning algorithm for sleep-stage classification. *Sleep*, 46(12):zsad242, 2023.

Khalighi, S., Sousa, T., Santos, J. M., and Nunes, U. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Mander, B. A., Winer, J. R., and Walker, M. P. Sleep and human aging. *Neuron*, 94(1):19–36, 2017.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Park, K., Hong, J., Lee, W., Shin, H.-W., and Kim, H.-S. Distillsleep: real-time, on-device, interpretable sleep staging from single-channel electroencephalogram. *SLEEPJ*, 48(12):zsaf240, 2025.

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., and Igel, C. U-sleep: resilient high-frequency sleep staging. *NPJ Digital Medicine*, 4(1):72, 2021.

Phan, H., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., and De Vos, M. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2021.

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

Redline, S., Tishler, P. V., Tosteson, T. D., Williamson, J., Kump, K., Browner, I., Ferrette, V., and Krejci, P. The familial aggregation of obstructive sleep apnea. *American journal of respiratory and critical care medicine*, 151: 682–687, 1995.

Sablayrolles, A., Douze, M., Schmid, C., and Jégou, H. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018.

Spira, A. P., Blackwell, T., Stone, K. L., Redline, S., Cauley, J. A., Ancoli-Israel, S., and Yaffe, K. Sleep-disordered breathing and cognition in older women. *Journal of the American Geriatrics Society*, 56(1):45–50, 2008.

Thapa, R., He, B., Kjaer, M. R., Moore, H., Ganjoo, G., Mignot, E., and Zou, J. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals. *arXiv preprint arXiv:2405.17766*, 2024.

Thapa, R., Kjaer, M. R., He, B., Covert, I., Moore IV, H., Hanif, U., Ganjoo, G., Westover, M. B., Jennum, P., Brink-Kjaer, A., et al. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, pp. 1–11, 2026.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yang, Y., Liu, X., Wu, J., Borac, S., Katabi, D., Poh, M.-Z., and McDuff, D. Simper: Simple self-supervised learning of periodic targets. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=EKpMeEV0hOo.

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

A. Datasets

The brief description of the datasets used for this research is provided below. Table 4 provides the summary statistics extracted from each dataset used for pretraining. Table 5 provides the summary statistics from the data not used for pretraining, but used for downstream tasks. Table 6 lists the channels included in each modality from different datasets.

A.1. Physionet 2018 (PHY) Dataset

The PhysioNet 2018 dataset, originally curated for the 2018 PhysioNet/CinC Challenge (Goldberger et al., 2000; Ghassemi et al., 2018), was provided by the Computational Clinical Neurophysiology Laboratory and the Clinical Data Animation Laboratory at Massachusetts General Hospital. Although the challenge primarily focused on arousal detection, the dataset includes expert-labeled sleep stages for 994 subjects. An additional 991 recordings were reserved for testing purposes; however, their labels remain private. In this study, we utilized both the labeled training set and the unlabeled test set for pretraining *PSG-M&m*. All signals were sampled at 200 Hz, with sleep stages scored according to the American Academy of Sleep Medicine (AASM) guidelines.

A.2. Sleep Heart Health Study (SHHS) Dataset

The Sleep Heart Health Study (SHHS) (Zhang et al., 2018; Quan et al., 1997) is a multicenter cohort initiative organized by the National Heart, Lung, and Blood Institute. This study consists of data collected over two visits. SHHS1 was collected from the initial visit, conducted between 1995 and 1998, which involved 6,441 men and women aged 40 and older. SHHS2 was collected from the second visit, conducted between 2001 and 2003, which involved 3,295 participants. Polysomnography (PSG) was recorded in-home by trained technicians and included various physiological signals: EEG (C3-A2, C4-A1), dual-channel EOG, EMG, respiratory effort, airflow, oxygen saturation, ECG, and body position. In alignment with the experimental protocol of XSleepNet (Phan et al., 2021), we partitioned the dataset by reserving 30% for testing. From the remaining 70%, 100 subjects were set aside for validation, with the balance used for model training. The train split of SHHS1 is used for pretraining and finetuning for downstream tasks. The entire SHHS2 dataset is used for pretraining and not used for downstream tasks.

A.3. Korea Image-based Sleep Study (KISS) Dataset

The Korea Image-based Sleep Study (KISS) dataset (Jeong et al., 2023) is a standardized, image-based polysomnography (PSG) repository. Collected between 2013 and 2020 across four sleep centers, the dataset utilizes recordings from Embla and NOX-A1 PSG systems, totaling 10,253 records. Expert scoring was conducted in accordance with AASM version 2.6 guidelines (Berry, 2012; Berry et al., 2017). Each record captures 21 distinct biosignals including various EEG, EOG, and EMG channels alongside respiratory and movement data. The data is publicly accessible via AI Hub (AIHub, 2020). Following the experimental setup by Jeong *et al.* (Jeong et al., 2023), we selected 7,579 records and implemented an 80%/20% split for training and validation/test on a patient-wise basis. The train split of KISS is used for pretraining and finetuning for downstream tasks.

Table 4. The dataset statistics used for pretraining. Missing values result from study design or anonymized data.

Dataset	Records	Subjects	Age (years)	BMI	Sex % (Female/Male)	Stage count						Stage ratio (%)				
						W	N1	N2	N3	REM	Total	W	N1	N2	N3	REM
PHY-Train	993	993	55.2 ± 14.3	N/A	33 / 67	145,558	135,409	372,208	101,678	113,859	868,712	17	16	42	12	13
PHY-Test	989	989	54.8 ± 14.3	N/A	37 / 63	N/A						N/A				
SHHS1	3,667	3,667	63.1 ± 11.5	28.2 ± 5.2	52 / 48	739,301	136,407	1,519,573	472,529	516,768	3,384,578	22	4	45	14	15
SHHS2	2,554	2,554	67.6 ± 10.4	28.3 ± 5.0	54 / 46	683,291	106,964	1,103,742	303,068	395,437	2,592,502	26	4	43	12	15
KISS	6,064	6,064	44.8 ± 14.5	25.8 ± 4.3	20 / 80	981,362	665,497	1,556,469	601,227	660,186	4,464,741	22	15	35	13	15
KVSS	881	881	51.4 ± 14.2	26.8 ± 4.4	24 / 76	143,041	129,225	271,190	47,538	97,667	688,661	21	19	39	7	14
MrOS1	2,768	2,768	76.4 ± 5.5	27.1 ± 3.8	0 / 100	945,515	129,239	1,236,682	221,334	381,577	2,914,347	32	4	43	8	13
MrOS2	994	994	81.0 ± 4.4	26.9 ± 3.8	0 / 100	382,152	80,312	425,515	45,040	129,942	1,062,961	36	8	40	4	12
MESA	2,054	2,054	69.4 ± 9.1	28.7 ± 5.5	54 / 46	598,750	203,837	854,634	149,770	268,646	2,075,637	29	10	41	7	13

Table 5. The dataset statistics used for downstream evaluation.

Dataset	Records	Subjects	Age (years)	BMI	Sex % (Female/Male)	Stage count						Stage ratio (%)				
						W	N1	N2	N3	REM	Total	W	N1	N2	N3	REM
SHHS1-Val	99	99	62.0 ± 11.7	27.8 ± 4.2	54 / 46	20,178	3,544	40,377	11,831	13,740	89,670	23	4	45	13	15
SHHS1-Test	1,617	1,617	63.4 ± 11.5	28.0 ± 5.0	53 / 47	323,871	61,191	671,916	203,864	226,113	1,486,955	22	4	45	14	15
KISS-Val	748	748	44.7 ± 14.1	26.0 ± 4.3	22 / 78	121,319	87,032	187,347	71,362	78,128	545,188	22	16	34	13	14
KISS-Test	767	767	45.4 ± 14.3	26.0 ± 4.1	16 / 84	125,315	83,396	198,155	73,154	83,428	563,448	22	15	35	13	15
ISRUC-SG1	100	100	51.1 ± 15.9	N/A	44 / 56	20,979	11,513	28,287	17,480	11,928	90,187	23	13	31	19	13
ISRUC-SG2	16	8	46.9 ± 17.5	N/A	25 / 75	2,282	2,211	5,042	2,609	2,063	14,207	16	16	35	18	15
ISRUC-SG3	10	10	39.6 ± 9.6	N/A	10 / 90	1,817	1,248	2,678	2,035	1,111	8,889	20	14	30	23	12

Table 6. Channels and sampling rates included in different modalities across datasets. The manufacturer of PSG recording machine is also provided. When sampling rates are same across different channels in the modality, we only write once at the top row.

Modality	KISS / KVSS		SHHS1		SHHS2		PHY		MESA		MrOS1 / 2	
	Ch.	SR (Hz)	Ch.	SR (Hz)	Ch.	SR (Hz)	Ch.	SR (Hz)	Ch.	SR (Hz)	Ch.	SR (Hz)
EEG	C3-M2 C4-M1 O1-M2 O2-M1	200	C3-A2 C4-A1	125	C3-A2 C4-A1	125	C3-M2 C4-M1 O1-M2 O2-M1 F3-M2 F4-M1	200	C4-M1 Oz-Cz Fz-Cz	256	C3-M2 C4-M1 O1-M2 O2-M1	256
EOG	E1-M2 E2-M1	200	EOG (L) EOG (R)	125	EOG (L) EOG (R)	125	E1-M2	200	EOG (L) EOG (R)	256	EOG (L) EOG (R)	256
EMG	Chin EMG	200	EMG	125	EMG	125	Chin1-Chin2	200	Chin EMG	256	EMG (L)-EMG (R)	256
ECG	ECG	200	ECG	125	ECG	125	ECG	250	ECG	256	ECG1-ECG2	512
Respiratory	Flow Thermistor Thorax Abdomen	200	Flow	10	Flow Thorax Abdomen	10	Flow Thorax Abdomen	200	Flow Thorax Abdomen Press	32	Flow Thermistor Thorax Abdomen	64 16
Oxygen Sat.	Saturation	200	Oximetry	1	Oximetry	1	SaO2	200	SpO2	1	SpO2	1
Manufacturer	Nox, Embla		Compumedics		Compumedics		Unknown		Compumedics		Compumedics	

A.4. Korea Video Sleep Study (KVSS) Dataset

The Korea Video Sleep Study (KVSS) dataset is a retrospectively constructed, multi-center clinical cohort that provides synchronized infrared sleep video and polysomnography (PSG) with expert annotations (Choi et al., 2024). Data were collected under IRB-approved protocols from three hospitals (Chungnam National University Hospital, The Catholic University of Korea St. Vincent's Hospital, and Hallym University Hospital). Across sites, 936 PSG examinations with synchronized infrared video were identified, and 881 PSG video pairs were included after screening and quality control. Recordings were obtained during routine clinical studies for various sleep-related indications, including suspected OSA, insomnia, PLMD, and RBD, with infrared videos recorded in parallel with PSG to ensure temporal alignment. Despite minor site-specific differences in camera placement and illumination, all videos were standardized to MP4 at 640×480 resolution and 5 fps. PSG was stored in European Data Format (EDF), and sleep stages (Wake, N1, N2, N3, REM) and AASM-defined events were annotated by certified technologists and reviewed by sleep physicians. The dataset underwent de-identification and expert cross-checking to verify synchronization and address obvious scoring inconsistencies. In this work, KVSS is used only as a PSG dataset, and we leverage the recorded PSG signals and associated subject metadata without using expert annotations or the infrared videos.

A.5. Multi-Ethnic Study of Atherosclerosis (MESA) Dataset

The Multi-Ethnic Study of Atherosclerosis (MESA) (Chen et al., 2015) is a multicenter cohort of 6,814 adults aged 45–84 years from four racial/ethnic groups (White, Black, Hispanic, and Chinese-American). As part of Exam 5 (2010–2013), the MESA Sleep exam enrolled 2,237 participants who completed single-night unattended in-home polysomnography (PSG) and wrist actigraphy. PSG was set up during an in-home evening visit by trained staff, and sleep staging and respiratory events were scored at a centralized sleep reading center using standardized procedures. The National Sleep Research Resource (NSRR) release provides PSG recordings in European Data Format (EDF) and XML annotation files for sleep

staging and respiratory event scoring. Respiratory event annotations were harmonized via rule-based post-processing of the original labels to ensure consistent criteria across datasets, and details are described in (Ahn et al., 2025). The PSG includes EEG (Fz-Cz, Cz-Oz, C4-A1), EOG, EMG, ECG, nasal airflow, thoracic and abdominal respiratory effort, oxygen saturation, and body position. In this work, we utilize the C4-A1 EEG. Recordings missing any required channel were excluded. 2,054 participants were included in the final analytic cohort.

A.6. Osteoporotic Fractures in Men Study (MrOS) Dataset

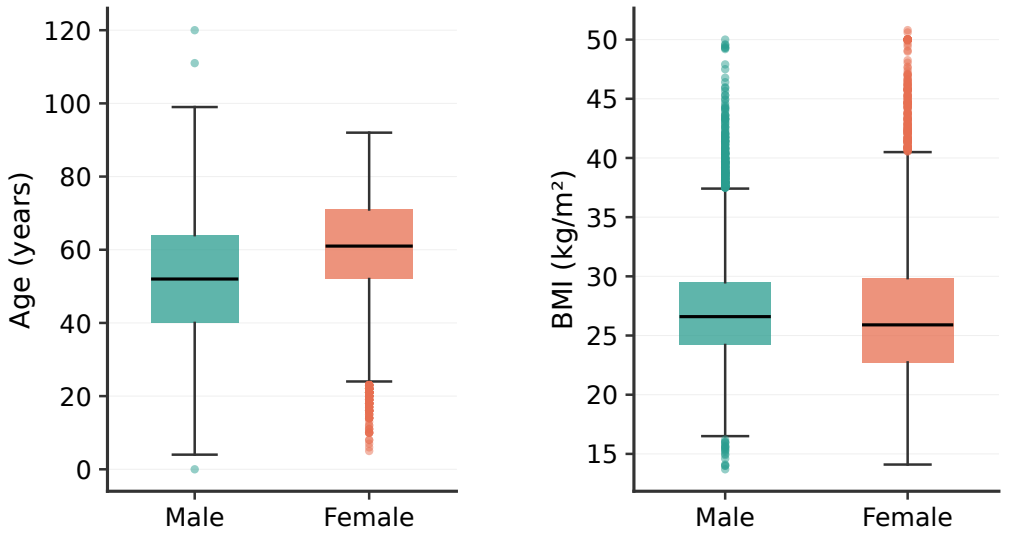
The Osteoporotic Fractures in Men Sleep Study (MrOS Sleep) (Blackwell et al., 2011) is a multicenter sleep cohort of older men (aged 65 years or older) that includes unattended in-home polysomnography (PSG), which was set up by trained technicians and annotated using standardized scoring procedures. The study was conducted between December 2003 and March 2005, during which 3,135 participants from the parent MrOS cohort of 5,994 men completed overnight unattended in-home PSG. The dataset available through the National Sleep Research Resource (NSRR) includes PSG recordings in European Data Format (EDF) and XML annotation files with multimodal biosignals, including EEG, EOG, EMG, ECG, nasal cannula (airflow), thoracic and abdominal respiratory effort, oxygen saturation, and body position. Respiratory event annotations were harmonized via rule-based post-processing of the original labels to ensure consistent criteria across datasets, and details are described in (Ahn et al., 2025). For the present study, we analyzed data from two sleep visits (Visit 1, 2003–2005; Visit 2, 2009–2012). We excluded recordings missing any of the required signals (EEG, EOG, EMG, ECG, airflow, thoracic/abdominal effort, or oxygen saturation). Of the 2,911 participants with successful PSG recordings, 2,678 (Visit 1) and 998 (Visit 2) were included in the final analytic cohort after quality control.

A.7. Institute of Systems and Robotics, University of Coimbra (ISRUC) Dataset

The Institute of Systems and Robotics, University of Coimbra (ISRUC) dataset (Khalighi et al., 2016) is a publicly available repository provided by the Sleep Medicine Center of the Hospital of Coimbra University (CHUC). The dataset is divided into three subgroups: SG1 and SG2, which feature patients with sleep disorders, and SG3, which consists of healthy control subjects. Each recording comprises 19 signals, including six EEG channels (F3-A2, C3-A2, O1-A2, F4-A1, C4-A1, O2-A1), dual-channel EOG and EMG, and various respiratory and cardiac sensors. Considering its relatively small size, this dataset is used to evaluate the adaptation efficiency of our model.

A.8. Demographics analysis

Figure 3 shows the age and BMI distribution stratified by sex based on our pretraining datasets.



(a) Age distribution by sex. *(b)* BMI distribution by sex.

Figure 3. Demographic distributions of the pre-training dataset. (a) Age distribution shows a median of approximately 52 years for males and 61 years for females. (b) BMI distribution shows similar medians across sexes (approximately 27 kg/m² for males and 26 kg/m² for females), with notable outliers in both groups.

B. Model architecture details

Table 7 summarizes the specification of the Micro-Encoder and the Macro-Encoder.

Table 7. Architecture Specification for the Micro- and Macro- Encoders. Norm: Normalization layer (RMSNorm), Exp: Expansion convolution, DW: Depthwise convolution, PW: Pointwise convolution

Encoder	Module	Item	Dim.	Kernel	Stride	Depth	Heads	Notes
Micro	Patch Emb.	Patch Emb. (Conv blocks)	[32,64,128]	[10,5,1]	[10,5,1]	–	–	Conv-Norm-GELU
	Private	Transformer blocks (Lower)	128	–	–	2	8	
		Patch Merge	384	10	5	–	–	Exp-DW-Norm-PW-GELU-PW
		Transformer blocks (Higher)	384	–	–	2	8	
	Shared	MoE Transformer blocks	384	–	–	4	8	4 Experts, 2 Activated
	Fusion	Transformer blocks	384	–	–	4	8	Cross-Attention
		Upsample	128	–	–	–	–	
	Decoder	Linear embedding	64	–	–	–	–	
		Transformer blocks	64	–	–	2	4	
		Projection (Linear)	50	–	–	–	–	Same as the patch size
Macro	Macro	Linear projection	512	–	–	–	–	
		Mamba blocks (Forward)	512	–	–	2	–	PreNorm-Mamba-PostNorm
		Mamba blocks (Backward)	512	–	–	2	–	PreNorm-Mamba-PostNorm

C. Training details

Pretraining is done in two-step process. The Micro-Encoder is trained in the first step and the Macro-Encoder is trained in the second step. Table 8 summarizes pretraining details for both Micro- and Macro-Encoder. When pretraining Macro-Encoder, we did not include PHY dataset because the dataset does not provide BMI information. Moreover, in other datasets, when the demographic attributes are not in proper format or unavailable, we exclude those records.

Table 9 summarizes the finetuning settings for downstream tasks used for Section 3.

When calculating the reconstruction loss for Micro-Encoder ($\mathcal{L}_{Mi}^{recon}$), the raw signal is smoothed using moving average of eleven adjacent points. This effectively eliminates noises and helps the model focus on more meaningful signals.

Table 8. Hyperparameter settings for pretraining for Micro and Macro-Encoders

Encoder	Item	Value	Notes
Micro-Encoder	Mask ratio	50%	
	Batch size	512	
	Input length	60 seconds	Equivalent to 120 epochs
	Optimizer	AdamW	
	β_1	0.9	
	β_2	0.99	
	Weight decay (λ)	0.05	
	Initial learning rate	5.00×10^{-4}	
	Learning rate schedule	Cosine annealing	
	Final learning rate	1.00×10^{-8}	
	Training epochs	3	
	Patch size	500 ms	Equivalent to 50 input points
	Temperature for contrastive loss (τ)	0.07	
	Sequence length for contrastive loss	30 seconds	
	Weight for contrastive loss (λ_{CL})	0.1	
	Weight for KoLeo loss (λ_{KoLeo})	0.01	
	Timespan for contrastive loss	30 seconds	Equivalent to 60 epochs
Macro-Encoder	Batch size	40	Subjects
	Maximum number of epochs per subject	1,080	Equivalent to 540 minutes
	Optimizer	AdamW	
	β_1	0.9	
	β_2	0.99	
	Weight decay (λ)	0.05	Not applied to SSM parameters
	Initial learning rate	1.00×10^{-4}	
	Learning rate schedule	Cosine annealing with warmup	
	Final learning rate	1.00×10^{-8}	
	Training epochs	4	
	Temperature for contrastive loss (ρ)	0.1	
	Temperature for weight calculation (υ)	0.5	
	Cycle length	90 minutes	180 epochs
	Demographic distance for sex difference (λ_{sex})	1	

Table 9. Training details used for downstream tasks.

Downstream tasks	Item	Value	Notes
Sleep staging	Loss	Weighted Cross Entropy	$w_k = \log_5(N/N_k)$ for class k
	Initial learning rate	1.00×10^{-2}	
	Batch size	4	Subjects
Apnea segmentation	Loss	Weighted Cross Entropy	$w_k = N/N_k$ for class k
	Initial learning rate	4.00×10^{-4}	
	Batch size	1024	Epochs
Disease prediction	Loss	Cox PH	
	Initial learning rate	1.00×10^{-2}	
	Batch size	4	Subjects
Age / AHI estimation	Loss	MAE	
	Initial learning rate	1.00×10^{-2}	
	Batch size	4	Subjects
Sex classification	Loss	Cross Entropy	
	Initial learning rate	1.00×10^{-2}	
	Batch size	4	Subjects
Common settings	Optimizer	AdamW	
	β_1	0.9	
	β_2	0.99	
	Weight decay (λ)	1×10^{-4}	
	Learning rate schedule	Cosine annealing	
	Final learning rate	1.00×10^{-8}	
	Training epochs	3	

D. More experimental results

D.1. Ablation study

Table 10 summarizes the sleep staging performance improvement of our principal components. The performance is evaluated on the SHHS1 test split. Experiments 1-3 is pretrained on SHHS1 train split and experiments 4-5 are done using the entire pretraining datasets. Table 11 presents the sleep staging accuracy measured on SHHS1 test split with varying demographic factors used for DGCL. DGCL is only done for the training split of SHHS1.

Table 10. **Ablation Study of *PSG-M&m* Components.** We evaluate the contribution of each module and training strategy to the final sleep staging accuracy on the SHHS1 dataset.

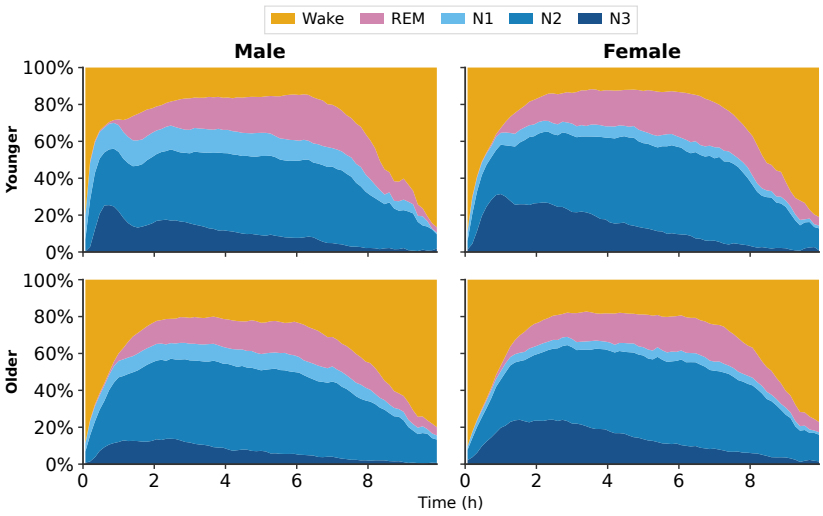
Id	Private encoder	Shared encoder	MAE	CL	Large scale pretraining	DGCL	Acc. (%)
1	✓		✓				74.8
2	✓	✓	✓				75.7
3	✓	✓	✓	✓			76.0
4	✓	✓	✓	✓	✓		79.8
5	✓	✓	✓	✓	✓	✓	**81.9**

Table 11. **Ablation Study of Demographic Factors in DGCL.** We investigate the impact of different demographic supervisory signals—Age, Sex, and BMI—on the final sleep staging accuracy. DGCL is only done on SHHS1 training split.

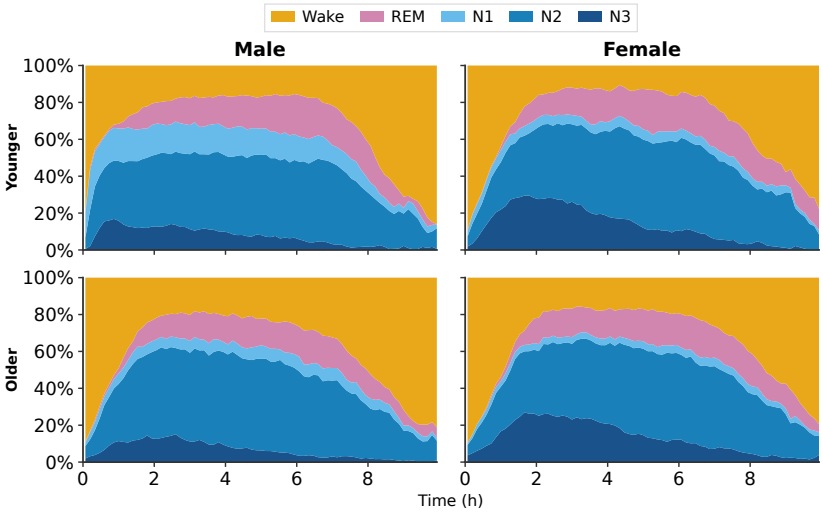
Id	Age	Sex	BMI	Acc. (%)
1				79.8
2	✓			80.6
3		✓		80.2
4			✓	80.1
5	✓	✓	✓	**80.7**

D.2. Sleep macro-structure analysis

To further investigate the demographic and clinical factors that influence sleep stage distributions, we present extended visualizations across various subgroups. Figure 4 shows sleep stage distributions by sex, age and BMI. Figure 5 shows sleep stage distributions by sex combined with age group, BMI category, and sleep apnea severity (AHI), respectively. Figure 6 presents the interactions between non-sex factors, revealing how multiple variables jointly influence sleep stage distributions throughout the night.

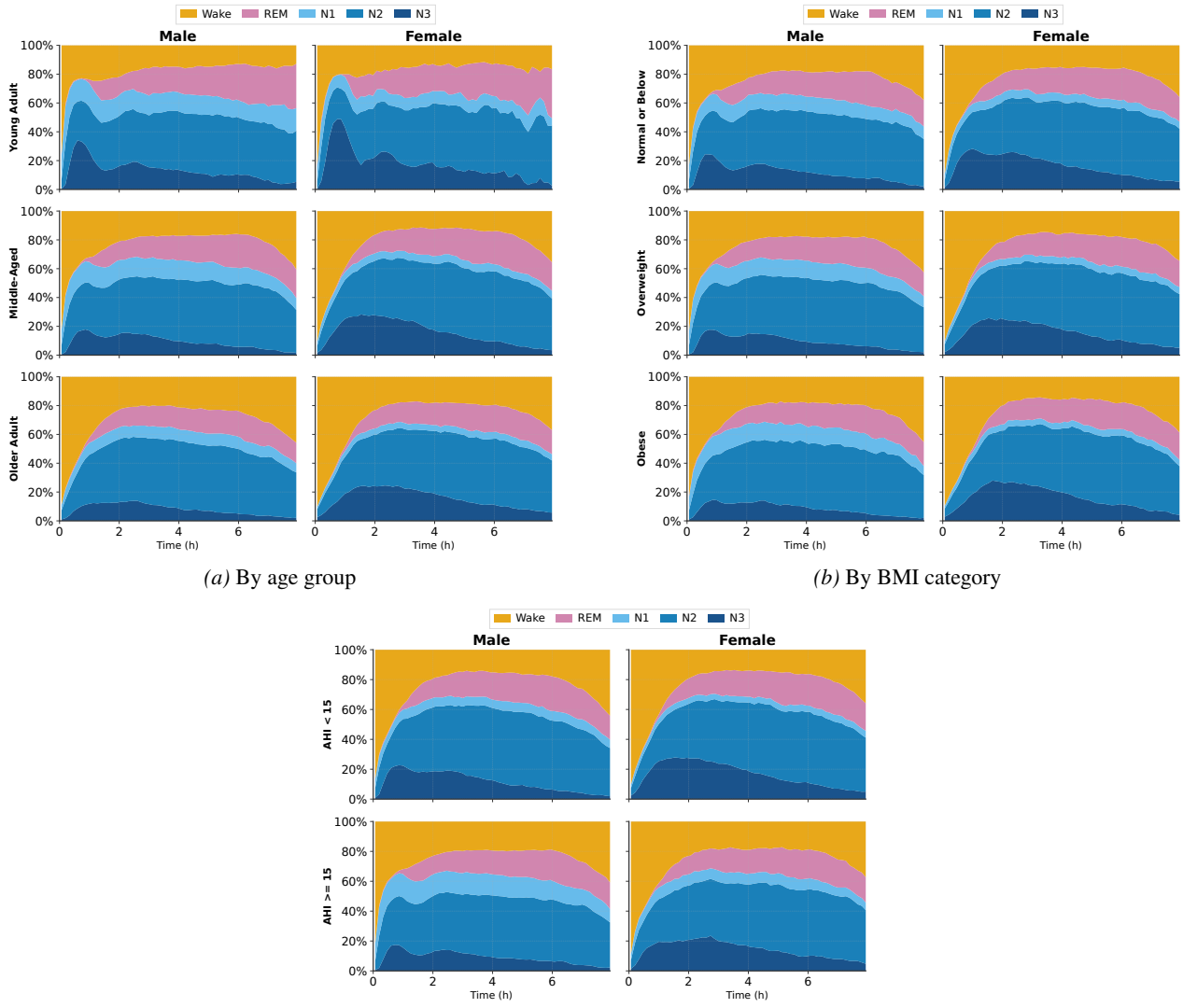


(a) Non-obese (BMI < 30 kg/m^2)



(b) Obese (BMI ≥ 30 kg/m^2)

Figure 4. **Sleep macro-structure variations across demographic groups**. Sleep stage distributions over full-night recordings by sex, age (Younger: < 60 yrs; Older: ≥ 60 yrs) and BMI. N3 proportion in the early sleep period or REM sleep proportion in later stage varies significantly across groups. These demographic dependent patterns motivate our Demographic-Guided Contrastive Learning objective.

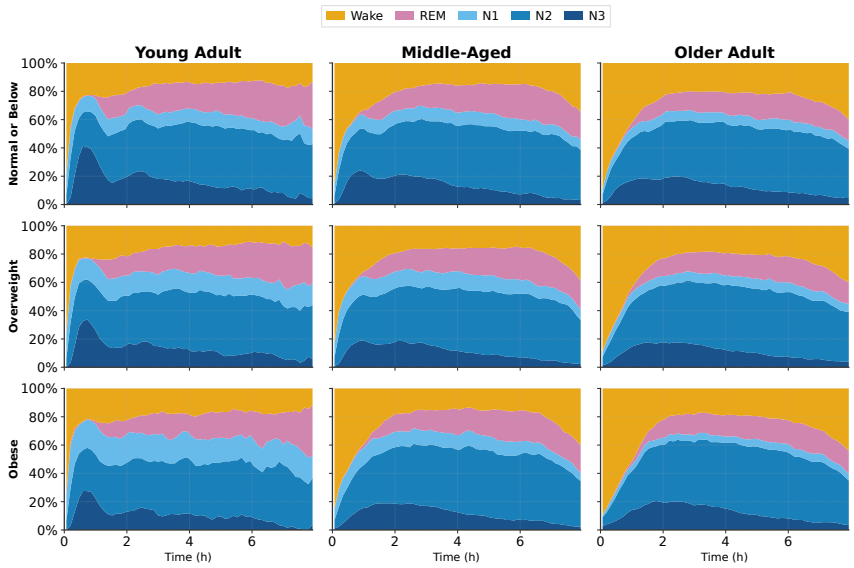


(a) By age group

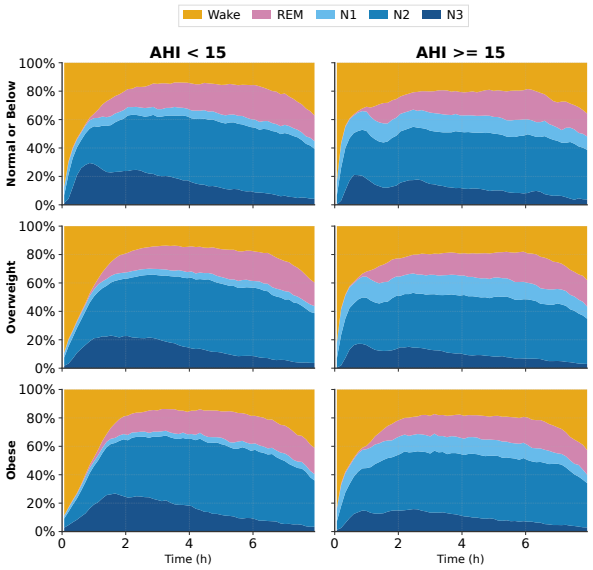
(b) By BMI category

(c) By sleep apnea severity

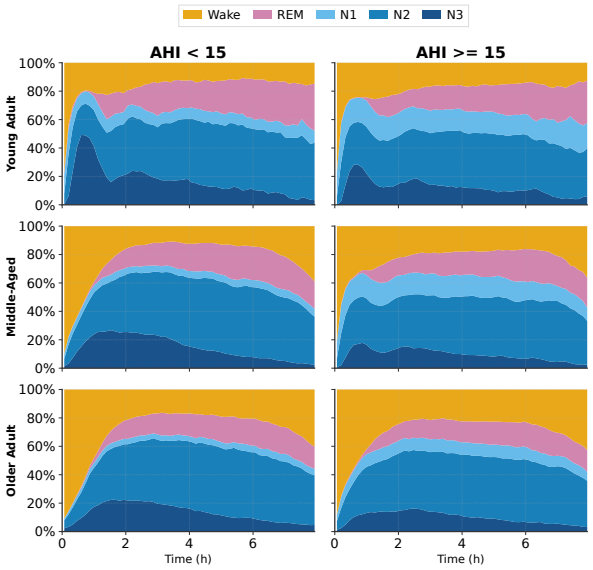
Figure 5. Sleep macrostructure variations across demographic and clinical groups. (a) Age groups: Young Adult (18–44 years), Middle-Aged (45–64 years), Older Adult ($\geq$ 65 years). (b) BMI categories: Normal or Below ($<$ 25 kg/m^2), Overweight (25–30 kg/m^2), Obese ($\geq$ 30 kg/m^2). (c) Sleep apnea severity based on Apnea-Hypopnea Index (AHI $\geq$ 15 events/h indicates moderate-to-severe). These demographic-dependent patterns motivate our Demographic-Guided Contrastive Learning objective.



(a) By BMI and age group



(b) By BMI and AHI

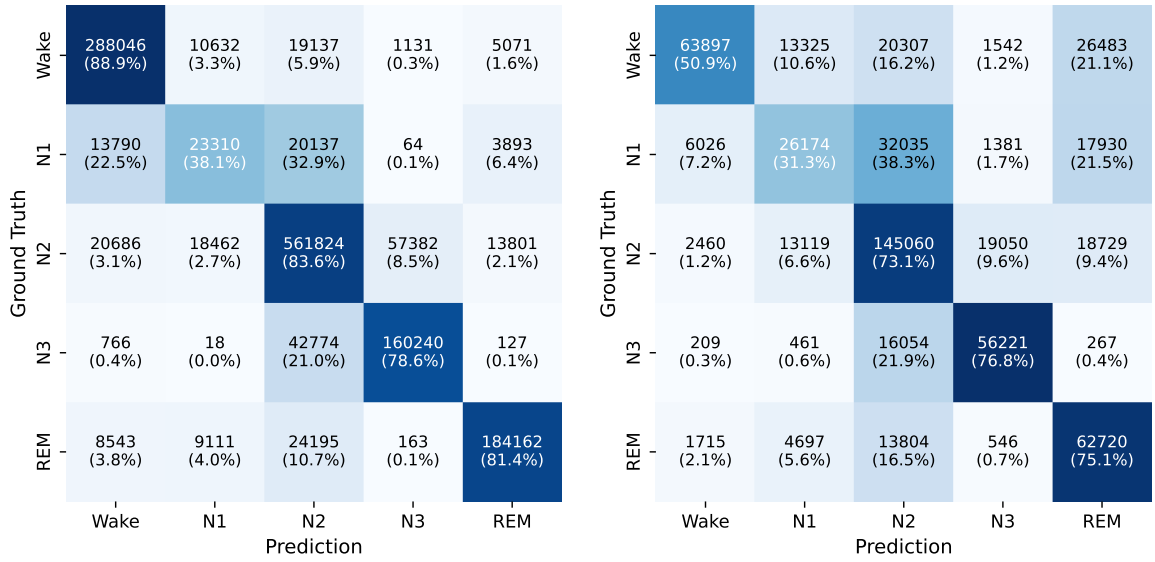


(c) By age group and AHI

Figure 6. Sleep macro-structure variations across combined demographic and clinical factors. (a) BMI categories (Normal or Below: $<$ 25 kg/m^2, Overweight: 25–30 kg/m^2, Obese: $\geq$ 30 kg/m^2) and age groups (Young Adult: 18–44 years, Middle-Aged: 45–64 years, Older Adult: $\geq$ 65 years). (b) BMI categories and sleep apnea severity (AHI $<$ 15 vs. AHI $\geq$ 15 events/h). (c) Age groups and sleep apnea severity. These combined factors jointly influence sleep stage distributions throughout the night.

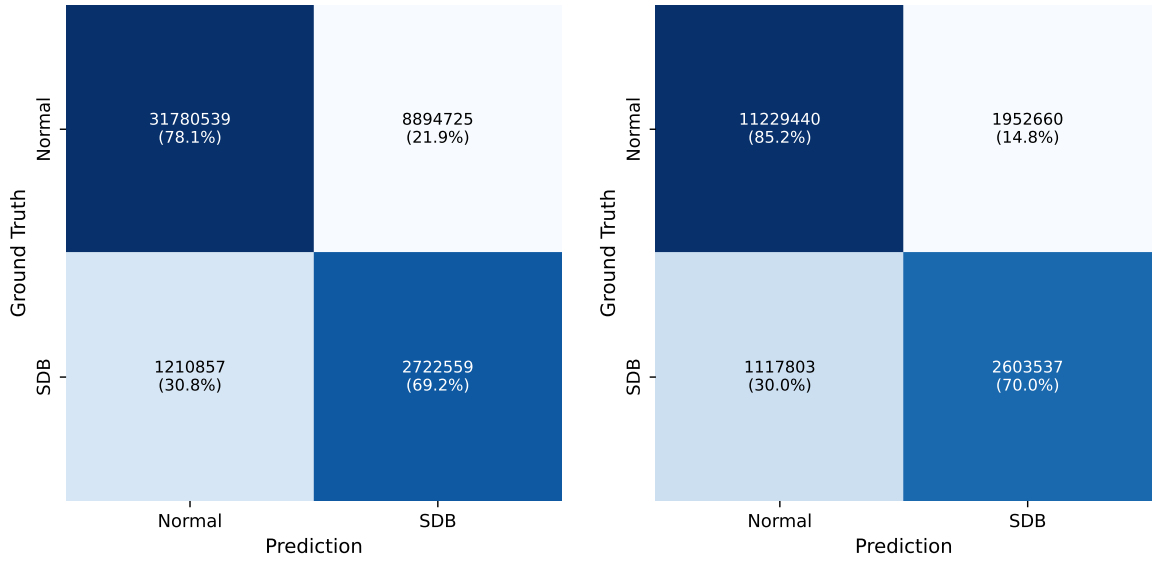
D.3. Confusion matrix

In Figure 7, we provide the confusion matrix for sleep staging and SDB segmentation evaluated on the test split of SHHS1 and KISS.



(a) SHHS1 - sleep staging

(b) KISS - sleep staging



(c) SHHS1 - SDB segmentation

(d) KISS - SDB segmentation

Figure 7. Confusion matrix for sleep staging and SDB segmentation on the test split of SHHS1 and KISS.

Table 12. **Ablation Study of *PSG-M&m* Components.** We evaluate the contribution of each module and training strategy to the final sleep staging accuracy on the SHHS1 dataset.

Id	Private encoder	Shared encoder	MAE	CL	Large scale pretraining	DGCL	Acc. (%)
1	✓		✓				74.8
2	✓	✓	✓				75.7
3	✓	✓	✓	✓			76.0
4	✓	✓		✓			67.7
5	✓	✓	✓	✓	✓		79.8
6	✓	✓	✓	✓	✓	✓	**81.9**