SLEEP2VEC: UNIFIED CROSS-MODAL ALIGNMENT FOR HETEROGENEOUS NOCTURNAL BIOSIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tasks ranging from sleep staging to clinical diagnosis traditionally rely on standard polysomnography (PSG) devices, bedside monitors and wearable devices, which capture diverse nocturnal biosignals (e.g., EEG, EOG, ECG, SpO_2). However, heterogeneity across devices and frequent sensor dropout pose significant challenges for unified modelling of these multimodal signals. We present `sleep2vec`, a foundation model for diverse and incomplete nocturnal biosignals that learns a shared representation via cross-modal alignment. `sleep2vec` is contrastively pre-trained on 42,249 overnight recordings spanning nine modalities using a *Demography, Age, Site & History-aware InfoNCE* objective that incorporates physiological and acquisition metadata (*e.g.*, age, gender, recording site) to dynamically weight negatives and mitigate cohort-specific shortcuts. On downstream sleep staging and clinical outcome assessment, `sleep2vec` consistently outperforms strong baselines and remains robust to any subset of available modalities and sensor dropout. We further characterize, to our knowledge for the first time, scaling laws for nocturnal biosignals with respect to modality diversity and model capacity. Together, these results show that unified cross-modal alignment, coupled with principled scaling, enables label-efficient, general-purpose modelling of real-world nocturnal biosignals.

1 INTRODUCTION

Sleep is a central determinant of human health, it shapes cognition, metabolism, cardiovascular function, and mental well-being, and its disruption both signals and drives disease (Irwin, 2015; Mukherjee et al., 2015; Leng et al., 2019; Lim et al., 2023). Sleep is clinically assessed with polysomnography (PSG) (Bloch, 1997; Boulos et al., 2019), which is a gold standard multi-sensor recording that jointly measures neural and ocular electrophysiology, muscle tone, cardiorespiratory dynamics, and oxygen saturation. Outside the clinical facilities, a growing range of bedside monitors and wearable devices captures subsets of these PSG modalities, creating a fragmented landscape across devices and care settings (Paalasmaa et al., 2012; Sadek et al., 2020; Birrer et al., 2024; Yu et al., 2025; Pillai et al., 2025). This reality motivates the question:

" Can cross-modal alignment of nocturnal biosignals enable a unified physiological representation that generalizes robustly across heterogeneous sensor sets in sleep medicine? "

Physiological signal pre-training offers a promising paradigm by learning generalized representations from diverse biosignals with minimal supervision (Thapa et al., 2024; 2025; Pillai et al., 2025; Fox et al., 2025). Yet real-world data bring hard constraints, sensor montages vary across centers and devices, sampling rates differ, entire channels are often missing, and large-scale expert annotation remains costly, making such a foundation both necessary and challenging.

We posit that concurrent nocturnal signals represent multiple perspectives of the same latent physiological state (Rechtschaffen & Kales, 1968; Berry et al., 2012; 2017). A proper alignment of these heterogeneous views into a unified representation space enables downstream tasks to flexibly operate on arbitrary modalities without retraining specialized pipelines. Such a space must yield modality-agnostic representations robust enough to ensure reliable inference even when modality missing occurs. This leads to a scaling hypothesis, suggesting that increasing modality diversity and model capacity can enrich semantic coverage and regularize modality-specific nuances. Although scaling laws have been extensively studied in language and vision, their implications remain largely

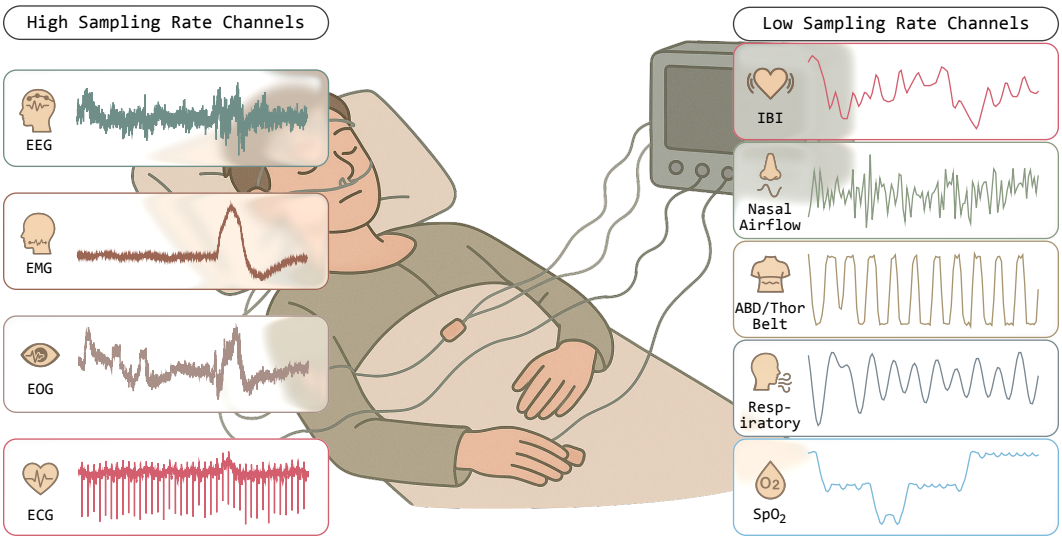

Figure 1: Polysomnography (PSG) captures diverse physiological signals, illustrated here as 30-second segments for each modality. High sampling rate electrophysiological channels include EEG, EMG, EOG, and ECG, while lower sampling rate cardiopulmonary and oximetry channels encompass Nasal Airflow, Abdominal/Thoracic Belt (ABD/Thor Belt), and SpO_2. Inter-Beat Interval (IBI) and Respiratory effort (RESP) signals, although not directly recorded by PSG, are derived from ECG and ABD/Thor Belt signals, respectively, and can also be measured via wearable devices. Together, these concurrent nocturnal signals provide complementary perspectives on a shared latent physiological state, highlighting the multimodal complexity inherent to sleep monitoring.

unexplored in physiological signal contexts. We therefore propose and evaluate a framework demonstrating predictable benefits of scaling PSG foundation models along both modality and parameter axes, especially for cross-center generalization where variations in sensor montage, demographics, and acquisition protocols are prevalent.

Prior work only partially addresses these needs. Existing models are typically trained for a specific downstream task (Wang et al., 2024; Shen et al., 2024; Carter et al., 2024; Pan et al., 2024; Shen et al., 2024; Lee et al., 2025; Ma et al.; Fox et al., 2025), lacking the generality required of a foundation model capable of supporting multiple tasks. Contrastive pre-training has shown promise on limited sets of physiological, typically one to three channels (*e.g.*, EEG and ECG) (Wang et al., 2024; Mathew et al., 2024; Thapa et al., 2024; Zhou et al., 2025; Thapa et al., 2025), but has not scaled to the full palette of PSG sensors. When more modalities are involved, objectives often prioritize reconstruction (Narayanswamy et al., 2024; Luo et al., 2024; Mathew et al., 2024; Nie et al., 2025) rather than explicit cross-modal alignment. Reconstruction encourages fidelity to modality-specific details but does not enforce that heterogeneous inputs map to a shared semantic manifold. As a result, inference typically assumes access to the same modality set used in training, degrading under realistic sensor missing scenarios. Moreover, systematic analyses of how performance scales with modalities and parameters are scarce.

We address these gaps with `sleep2vec`, a PSG foundation model that aligns heterogeneous nocturnal signals into a unified embedding space. Our framework jointly leverages nine modalities, waveform channels including EEG, EOG, EMG, ECG, Nasal airflow, Abdominal/Thoracic Belt (ABD/Thor Belt) and SpO_2; and interval-derived features including Inter-beat Interval (IBI) and Respiratory effect, from 42,249 nights of physiological recordings. A context-aware InfoNCE objective, explicitly modelling physiological similarity (age, gender, recording center) to dynamically weight samples, effectively distinguishes hard from easy negatives, mitigating overfitting to dataset-specific nuances.

Our work makes the following contributions:

(i) Unified multimodal PSG pre-training: We propose, to our knowledge, the largest scale multimodal contrastive pre-training framework for PSG foundation models, jointly aligning waveform and interval-based modalities, uncovering comprehensive inter-modal physiological correlations.

(ii) Scaling law investigation: We systematically explore scaling PSG foundation models along modality diversity and parameter dimensions, demonstrating predictable improvements in cross-cohort generalization with minimal task-specific labels.

(iii) Cross-modal training objective: We propose *Demography, Age, Site & History-aware InfoNCE* (DASH-InfoNCE), a context-aware contrastive objective that conditions negative-sample weighting on demographic, age, acquisition-site, and recording-history metadata. This metadata-guided weighting suppresses cohort-specific shortcuts and improves robustness and cross-site generalization across heterogeneous PSG sensor montages.

(iv) Comprehensive downstream evaluation: We extensively evaluate sleep2vec on both SHHS and WSC datasets, spanning tasks such as sleep staging, demographic prediction, and diagnostic outcomes, representing the broadest evaluation of a PSG foundation model to date.

2 Related Work

Multimodal alignment for flexible inference. Contrastive alignment maps heterogeneous inputs into a shared embedding space, enabling zero-shot transfer, retrieval, and robustness to input permutations. In vision–language, CLIP popularized large-scale image–text alignment (Radford et al., 2021), and many-to-one binding across six modalities in ImageBind (Girdhar et al., 2023). *In particular, PSG contains tens of synchronized channels (EEG, EOG, EMG, ECG, Nasal airflow, Respiratory effort, SpO_2, etc.), yet prior multimodal alignment works seldom extend beyond EEG-only or a few paired channels, and rarely handle montage shifts at this scale.*

Self-supervised learning for sleep and PSG data. Self-supervised learning (SSL) for sleep data has evolved from early approaches targeting task-specific objectives such as sleep staging using fixed PSG montages (Supratak et al., 2017; Perslev et al., 2021), toward broader pre-training frameworks. Recent works emphasize constructing foundational models but typically remain limited by: (i) pre-training strategies narrowly tailored to single downstream tasks or restricted label sets (Fox et al., 2025), (ii) alignment restricted to selected subsets of PSG channels, thus failing to address comprehensive multimodal integration (Fang et al., 2024; Narayanswamy et al., 2024; Luo et al., 2024; Thapa et al., 2024; 2025), or (iii) employing cross-modal generative methods that prioritize modality-specific signal fidelity rather than explicitly aligning heterogeneous modalities (Chen et al., 2024; Nie et al., 2025). *Consequently, cross-modal alignment covering the full PSG spectrum remains largely unexplored, with most pre-training focusing on fixed, small montages.*

Scaling and generalization. In language and vision, performance follows predictable trends as model and data scale. Despite rapid progress, systematic studies of scaling laws for physiological time series and PSG remain sparse. Existing PSG SSL studies seldom probe modality-diversity scaling or parameter scaling. *To our knowledge, systematic modality-diversity scaling has not been charted in sleep; existing studies also under-report parameter/data scaling for PSG SSL, leaving open how capability grows with both model size and channel count.*

3 Method

3.1 Dataset and Preprocessing

We leveraged publicly available PSG datasets for pre-training, including the Human Sleep Project (HSP) (Sun et al., 2023) and four cohorts obtained from the National Sleep Research Resource (NSRR) (Zhang et al., 2018): the Sleep Heart Health Study (SHHS) (Quan et al., 1997), Osteoporotic Fractures in Men Study (MrOS) (Blackwell et al., 2011), Multi-Ethnic Study of Atherosclerosis (MESA) (Chen et al., 2015), and Wisconsin Sleep Cohort (WSC) (Young et al., 2009), collectively encompassing multi-center, multi-device acquisitions from diverse demographic populations (age range: 1-109; recording span: 1995-present). Table 4 presents an overview of the five datasets involved. The five datasets were harmonized into a unified corpus comprising 42,249 overnight

recordings from 30,852 subjects, processed through a standardized pipeline to minimize cohort-specific biases and ensure symmetrical handling during batching and evaluation.

A pool of nine PSG channels across cohorts was established as shown in Figure 1: comprising two groups of signals differentiated by sampling rates: higher sampling rate electrophysiological signals, including EEG, EOG, EMG, ECG uniformly resampled to 128 Hz, and lower sampling rate physiological signals, including Nasal airflow, ABD/Thor belt, SpO_2, Inter-Beat Interval (IBI) and respiratory signals uniformly resampled to 4 Hz. Only minimal preprocessing is applied to the higher sampling rate signals to preserve raw signal characteristics crucial for downstream physiological interpretation, involving temporal resampling to the target frequency and z-score normalization with cohort-invariant statistics. The IBI channel is derived from ECG R-peak detection, with raw inter-beat intervals cleaned for outliers and artifacts and then linearly interpolated to a continuous 4 Hz sequence. The Respiratory effect reflects breathing cycles extracted from either Nasal Airflow or Abdominal Belt, standardized by band-limiting, and resampling to 4 Hz. Both IBI and Respiratory effect can be captured not only by PSG but also using simpler, low-burden hardware such as ballistocardiography mats or other contactless sensors (Chen et al., 2025).

Participant information is retained when available, including age, gender, and recording site, to facilitate cohort-aware analysis and difficulty estimation during pre-training. Participant-level data partitions are established to prevent data leakage across splits. A dedicated pre-training split ($N_{pre\text{-}train}$=23,934 participants) is exclusively reserved for foundation model learning, while downstream splits follow an 8:1:1 ratio ($N_{train}/N_{val}/N_{test}$=8,792/1,102/1,116), ensuring no participant overlap and identical modality coverage. For downstream evaluation, sleep staging labels on SHHS and WSC as well as clinical diagnosis labels on SHHS were aligned with the same participant-level splits as in pre-training, ensuring consistency and preventing any data leakage.

3.2 Model Architecture

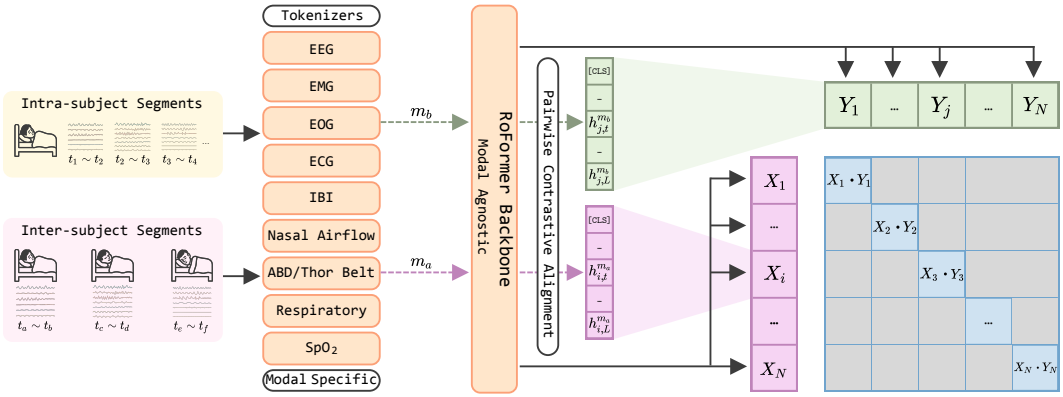

Figure 2: An illustration of the multimodal pre-training framework. Each overnight PSG recording is partitioned into intra-subject segments (different temporal slices from the same individual) and inter-subject segments (slices from different individuals), which are independently tokenized via modality-specific MLP tokenizers. A learnable `[CLS]` token is prepended to each masked sequence before processing through a modality-agnostic RoFormer backbone. Hidden states from the backbone at each timestep are projected into a shared alignment space, enabling timestep-wise pairwise contrastive alignment across modalities.

A minimalistic tokenizer based on a multi-layer perceptron (MLP) was implemented, comprising two feed-forward layers and a residual connection. The tokenizer maps input 30-second fragments into embeddings of dimension D through an initial linear transformation that projects inputs into an intermediate hidden representation of dimension $2D$, activated by the SiLU nonlinearity (Elfwing et al., 2018) and regularized using dropout with a 0.1 probability. The choice of 30-second segments aligns with the standard epoch duration recommended by the American Academy of Sleep Medicine (AASM) guidelines (Berry et al., 2012; 2017) for polysomnographic analysis. Subsequently, this hidden representation is linearly transformed into the final embedding space (D). In parallel, a residual linear transformation directly maps the inputs into the output embedding dimension, enhancing

gradient flow and training stability. A LayerNorm is further applied to the final embedding for normalization. Cross-modal sampling rate differences are resolved by encoding 30-second tokens into tokens of equal embedding dimension using modality-specific tokenizers, each operating directly on the original sampling rates, resulting in temporally aligned embeddings across modalities to be fed into the modality-agnostic backbone.

As shown in Figure 2, a simple yet effective sampling strategy is employed to maintain stable optimization as the number of modalities grows during pre-training: exactly two modalities (m_a, m_b) are randomly selected for each mini-batch, with one instance drawn from each modality, as shown in Figure 2. Independent time-step masking is then applied to these paired instances to enhance robustness and mitigate shortcut learning. Each 30-second token is of a 15% probability of being replaced by a learnable, modality-specific mask token, after which alignment is conducted exclusively between these masked segments. A dedicated learnable `[CLS]` token is prepended to the sequence, the resulting input is then processed through a modality-agnostic RoFormer backbone (Su et al., 2024). It is important to emphasize that the RoFormer backbone in sleep2vec should be viewed as one concrete instantiation of a generic modality-agnostic sequence encoder rather than a core contribution in isolation. The aim is not to advocate RoFormer as the uniquely optimal architecture for PSG, but to show that pairing a flexible backbone capable of ingesting arbitrary channel subsets with a metadata-aware contrastive alignment objective provides a simple and effective recipe for handling heterogeneous PSG montages. In principle, other Transformer style or state-space sequence encoders could be substituted without changing the overall framework, and we expect the benefits of unified cross-cohort pre-training and metadata-aware alignment to transfer across such choices. The backbone outputs hidden states for each timestep, as well as a global nocturnal representation at the `[CLS]` position. These hidden states are projected into a shared 128-dimensional alignment space via a shared three-layer MLP projection head, enabling the application of a cross-modal contrastive loss at each timestep.

During fine-tuning, both masking and the contrastive learning projection head are removed, and modal configurations remain fixed per downstream task. Task-specific heads directly operate on backbone features. Sequence-level tasks (*e.g.*, sleep staging) use per-time-step hidden states, while aggregate tasks (*e.g.*, gender, age, or clinical diagnosis) rely on the global nocturnal representation from the `[CLS]` position. When multiple modalities are available at inference, their representations are aggregated using simple fusion strategies such as averaging, concatenation, or a small gating module. Specific fusion methods employed per task are detailed in the experimental results section.

3.3 Cross-Modal Alignment Objective: DASH-InfoNCE

During pre-training, each mini-batch contains B paired segments, each of length L timesteps. For segment index $i \in \{1, \dots, B\}$, time index $t \in \{1, \dots, L\}$, and modality $m \in \{m_a, m_b\}$, denote $\mathbf{v}_{i,t}^{(m)} \in \mathbb{R}^d$ as the corresponding d-dimensional embedding. Given $\widehat{\mathbf{v}}_{i,t}^{(m)}$ as its ℓ_2 normalized product given by $\widehat{\mathbf{v}}_{i,t}^{(m)} = \mathbf{v}_{i,t}^{(m)} / \|\mathbf{v}_{i,t}^{(m)}\|_2$, the cosine similarity is

$$s_{i,j,t} \;=\; \left\langle \widehat{\mathbf{v}}_{i,t}^{(m_a)},\, \widehat{\mathbf{v}}_{j,t}^{(m_b)} \right\rangle \in [-1, 1], \qquad i, j \in \{1, \dots, B\},\; t \in \{1, \dots, L\}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product. The index mapping $\pi : 1, \dots, B \to 1, \dots, B$ specifies the number of paired segments in modality m_b for an anchor in modality m_a, where batches are typically aligned such that $\pi(i) = i$. Demographic and acquisition metadata for segment i are denoted by (a_i, g_i, c_i, u_i), where $a_i \in \mathbb{R}_+$, $g_i \in \mathcal{G}$, $c_i \in \mathcal{C}$, and u_i represent age, gender, acquisition site, and the subject-night identifier, respectively. These variables are used solely for weighting and modulation below and never as labels in the learning objective.

3.3.1 Base Formulation: Temporal InfoNCE

With temperature $\tau > 0$, the baseline timestep InfoNCE loss aligning m_a to m_b is

$$\mathcal{L}_{\text{base}}^{(t)} \;=\; \frac{1}{B} \sum_{i=1}^{B} \left[-\log \frac{\exp\big(s_{i,\pi(i),t}/\tau\big)}{\sum_{j=1}^{B} \exp\big(s_{i,j,t}/\tau\big)} \right]. \tag{2}$$

This objective encourages the similarity between the paired cross-modal embeddings $(i, \pi(i), t)$ to exceed the similarities to all in-batch, same-time candidates (i, j, t) with $j \neq \pi(i)$. The temperature coefficient τ controls the concentration of the induced softmax distribution.

3.3.2 PROPOSED DASH-INFONCE LOSS

A novel DASH-InfoNCE loss that reshapes the negative set by (i) metadata-driven sample weighting and (ii) margin-based modulation of *pseudo-negatives* (*i.e.*, negatives from the same subject-night) is introduced in this section. For anchor (i, t), define

$$\ell_{\mathrm{DASH}}(i,t) \;=\; -\log \frac{\exp\big(s_{i,\pi(i),t}/\tau\big)}{\sum_{j=1}^{B} \omega_{i,j}\; \exp\big(\big[s_{i,j,t} - \gamma\,\psi(d_{i,j}, p_{i,j,t})\big]/\tau\big)}, \tag{3}$$

where $\omega_{i,j} \geq 0$ are segment-specific weights satisfying $\sum_{j=1}^{B} \omega_{i,j} = 1$, $\gamma \geq 0$ is a modulation strength, and $\psi(\cdot,\cdot) \geq 0$ reduces the effective logit of designated pseudo-negatives before the softmax. The binary indicator $d_{i,j} \in 0, 1$ selects which pairs are margin-modulated, with the convention $d_{i,\pi(i)} = 0$ ensuring that positives are not penalized. The optional factor $p_{i,j,t} \in [0, 1]$ encodes time-specific signals, and in our instantiation below, we set ψ to a fixed margin with $p_{i,j,t}$ absorbed into that choice. Relative to Eq. (2), the numerator is unchanged while the denominator concentrates probability mass on demographically similar, presumably harder negatives via $\omega_{i,j}$ and diminishes the competitive strength of same-subject-night negatives through the subtractive margin $\gamma\,\psi$.

3.3.3 SAMPLE WEIGHTING MECHANISM

Let $\kappa : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ be a non-negative, symmetric kernel that decreases with the age difference $|a_i - a_j|$. We further define similarity factors for gender and acquisition site as $s_{i,j}^{(g)} \in \{\gamma_{\mathrm{same}},\ \gamma_{\mathrm{diff}}\}$ and $s_{i,j}^{(c)} \in \{\delta_{\mathrm{same}},\ \delta_{\mathrm{diff}}\}$, with $\gamma_{\mathrm{same}} > \gamma_{\mathrm{diff}} \geq 0$ and $\delta_{\mathrm{same}} > \delta_{\mathrm{diff}} \geq 0$, where the value is chosen according to whether $g_i = g_j$ and $c_i = c_j$, respectively.

Given the *pseudo-negative* indicator $h_{i,j} \;=\; \mathbb{I}[u_i = u_j \;\wedge\; j \neq \pi(i)] \;\in\; \{0, 1\}$, the unnormalized weights are defined as $\alpha_{i,j} \;=\; \kappa(a_i, a_j)\, s_{i,j}^{(g)}\, s_{i,j}^{(c)} \;\;+\;\; \varepsilon\, h_{i,j}$, where $\varepsilon \;=\; 10^{-6}$. The normalized weights are computed by

$$\omega_{i,j} = \frac{\alpha_{i,j}}{\sum_{k=1}^{B} \alpha_{i,k}}, \qquad \sum_{j=1}^{B} \omega_{i,j} = 1. \tag{4}$$

This weighting scheme assigns higher values to negatives closely matched by age, gender, and acquisition site. Constant ε ensures negatives from the same subject-night retain non-zero weights, stabilizing the denominator in Eq. (3) when closely matched demographic negatives are rare.

3.3.4 PSEUDO-NEGATIVE MODULATION

We modulate only negatives drawn from the same subject-night. Let $d_{i,j} = h_{i,j}$ and take a margin-only instantiation of ψ:

$$\psi(d_{i,j}, p_{i,j,t}) \;=\; \begin{cases} m, & d_{i,j} = 1, \\ 0, & d_{i,j} = 0, \end{cases} \qquad m > 0. \tag{5}$$

By combining Eq. (5) and Eq. (3), the fixed margin γm is subtracted from the logits of same-subject-night negatives prior to the softmax. This reduces the tendency to over-penalize semantically close negatives originating from the same subject-night while preserving their presence in the denominator through Eq. (4).

3.3.5 FINAL OBJECTIVE

The DASH-InfoNCE objective averages the per-anchor loss Eq. (3) over instances and time:

$$\mathcal{L}_{\mathrm{DASH}}^{(t)} \;=\; \frac{1}{B} \sum_{i=1}^{B} \ell_{\mathrm{DASH}}(i,t), \qquad \mathcal{L}_{\mathrm{DASH}} \;=\; \frac{1}{L} \sum_{t=1}^{L} \mathcal{L}_{\mathrm{DASH}}^{(t)}. \tag{6}$$

Averaging across t enforces alignment at each timestep. Note that every component of Eq. (3)–Eq. (6) depends only on demographic and acquisition metadata (a_i, g_i, c_i) and identifiers u_i via Eq. (4), no downstream task labels are used during pre-training.

3.4 Feature Fusion

In multimodal physiological tasks, the way modality-specific features are fused/aggregated has a direct impact on performance. A naïve **Concat** strategy (*i.e.* concatenating embeddings before classification) produces high-dimensional, sparse representations that inflate computation and sample complexity, exacerbating overfitting. Conversely, **Mean** aggregation (*i.e.* element-wise averaging) assumes equal informativeness and reliability across channels; in practice, physiological streams differ in SNR and complementary content, so uniform averaging washes out modality-specific cues and is brittle under missing sensors.

To address these limitations, we adopt the **Gating Mechanism**, which introduces learnable scalar weights assigned to each modality. This approach adaptively emphasizes modalities based on their informativeness, dynamically adjusting the contribution of each PSG channel. Consequently, it yields a more expressive, compact, and task-oriented aggregated representation, enabling efficient downstream learning.

4 Experiments

4.1 Pre-training Diagnostics: Alignment & Retrieval

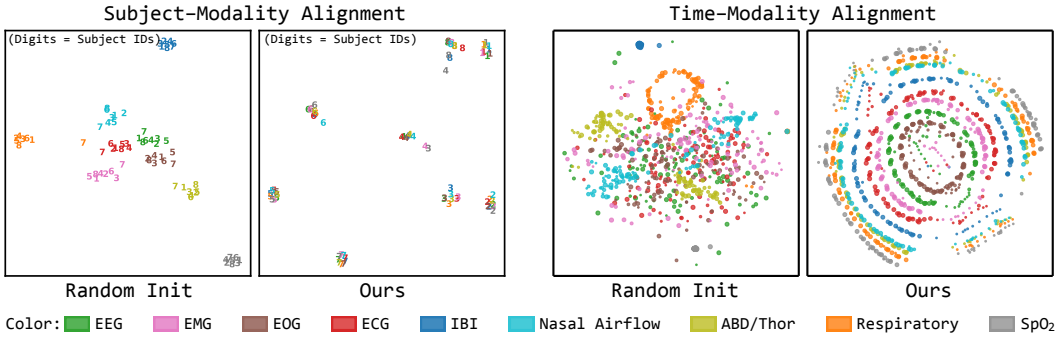

Figure 3: t-SNE visualization of encoder embeddings comparing random initialization and post-pre-training results. **Left Panel (Subject–Modality Alignment):** Visualization of `[CLS]` token embeddings shows that pre-training effectively clusters embeddings from different modalities into distinct, subject-specific groups, indicating aligned subject-level physiological states. **Right Panel (Time–Modality Alignment):** Visualization of timestep-level embeddings, dot sizes indicate temporal ordering (larger $\rightarrow$ later). Pre-trained embeddings form structured trajectories, contrasting with the scattered distribution observed prior to training.

To assess the effectiveness of multimodal alignment, Figure 3 (left panel) visualizes the `[CLS]` token embeddings using t-SNE both prior to and following pre-training. Initially, embeddings cluster by modality, reflecting intrinsic modality-specific biases and heterogeneous signal characteristics. After pre-training, embeddings from distinct modalities corresponding to the same subject are coherently grouped, indicating improved alignment and preservation of subject-specific structures.

Further analysis of timestep-level embeddings from a random subject (Figure 3, right panel) reveals structured trajectories emerging post-training, indicating an effective modality alignment at a finer temporal resolution. The coordinated variation in dot sizes across concentric rings emphasizes temporal consistency within the representations. Such consistency is beneficial for downstream sequential tasks like sleep staging, underscoring the practical advantages of the temporally aligned `sleep2vec` embeddings.

Table 1: Performance of five-class sleep staging (W/N1/N2/N3/REM) across PSG channel sets and models on **SHHS**. Reported metrics regarding overall performance including Accuracy (Acc., %), Cohen Kappa (κ), Macro-F1 (MF1, %), Sensitivity (Sens., %) and Specificity (Spec., %). Class-wise F1 (%) is also listed. Baselines reproduced by us for fair comparison are marked with †. Note that these foundation model (FM) baselines were individually pre-trained for each PSG channel subset, whereas `sleep2vec` was pre-trained only once across all modalities. "FULL CHANNELS" refers to the fixed channel configuration that each model is designed for and individually pre-trained on. Underlined numbers indicate the best overall performance within each channel set; **bold numbers** denote the best performance among FMs; **bold-underlined numbers** indicate cases where the FM surpasses specialized models.

PSG Channel Set		Overall Performance (↑)					Class-wise F1 (↑)				
Inference Subset	Model	Acc.	κ	MF1	Sens.	Spec.	W	N1	N2	N3	REM
EEG	***Specialized (non-FM) Model***										
	DeepSleepNet (Supratak et al., 2017)	81.0	0.73	–	–	73.9	85.4	40.5	82.5	79.3	81.9
	SleepEEGNet (Mousavi et al., 2019)	73.9	0.65	–	–	68.4	81.3	34.4	73.4	75.9	77.0
	AttnSleep (Eldele et al., 2021)	84.2	0.78	–	–	75.3	86.7	33.2	87.1	87.1	82.1
	XSleepNet1 (Phan et al., 2021)	87.6	0.83	80.7	79.7	96.5	91.6	51.4	88.5	85.0	88.4
	XSleepNet2 (Phan et al., 2021)	87.5	0.83	81.0	80.4	96.5	92.0	49.9	88.3	85.0	88.2
	L-SeqSleepNet (Phan et al., 2023)	87.6	0.83	80.3	79.4	96.5	92.4	48.6	88.2	83.9	88.5
	SleepTransformer (Phan et al., 2022)	87.7	0.83	80.1	78.7	96.5	92.2	46.1	88.3	85.2	88.6
	Foundation Model										
	SleepFM (Thapa et al., 2024; 2025) †	86.3	0.81	76.3	75.3	**96.1**	**93.2**	36.6	86.3	77.3	88.1
	sleep2vec	**87.4**	**0.82**	**77.3**	**76.2**	**96.1**	92.4	**40.1**	**86.5**	**77.7**	**88.7**
IBI & RESP	***Specialized (non-FM) Model***										
	Sun et al. (2019) †	71.3	0.59	57.3	56.9	91.7	85.2	4.8	70.1	49.9	76.4
	Goldammer et al. (2022) †	77.2	0.68	63.6	62.7	93.4	88.2	15.5	76.4	55.5	82.4
	Foundation Model										
	SleepFM (Thapa et al., 2024; 2025) †	79.7	0.71	65.7	65.4	94.2	90.4	12.9	78.4	**61.7**	84.8
	SleepFounder (Nie et al., 2025) †	80.9	0.73	**68.3**	**67.0**	94.5	**91.3**	**22.3**	80.0	61.1	**85.9**
	sleep2vec	**83.0**	**0.75**	65.9	65.8	**95.1**	86.6	5.3	**80.3**	60.9	84.9
ECG & ABD	***Foundation Model***										
	SleepFM (Thapa et al., 2024; 2025) †	77.9	0.68	62.7	62.7	93.6	88.4	6.6	76.9	60.9	80.4
	sleep2vec	**82.7**	**0.75**	**65.6**	**65.2**	**95.0**	**92.6**	**6.2**	**80.6**	**62.4**	**86.1**
EEG & EOG & EMG	***Specialized (non-FM) Model***										
	SeqSleepNet (Phan et al., 2019)	87.2	0.82	80.2	78.7	96.3	91.8	49.1	88.2	83.5	88.2
	XSleepNet1 (Phan et al., 2021)	89.1	0.85	82.3	81.2	96.9	–	–	–	–	–
	XSleepNet2 (Phan et al., 2021)	89.1	0.85	82.2	81.4	96.9	–	–	–	–	–
	Olesen et al. (2021)	85.8	0.79	–	–	–	–	–	–	–	–
	Foundation Model										
	SleepFM (Thapa et al., 2024; 2025) †	87.0	0.82	78.0	77.8	96.4	93.6	**40.7**	86.8	77.8	**90.9**
	sleep2vec	**88.3**	**0.83**	**78.7**	**77.9**	**96.8**	**94.5**	40.6	**87.8**	**79.8**	89.0
FULL CHANNELS	***Foundation Model***										
	SleepFM (Thapa et al., 2024; 2025) †	86.7	0.81	77.3	76.9	96.3	93.4	39.2	86.7	77.1	90.3
	PFTSleep (Fox et al., 2025)	87.7	0.83	**80.8**	**82.3**	96.7	93.3	**48.6**	87.8	**82.7**	**91.5**
	sleep2vec (InfoNCE)	88.4	**0.84**	78.6	77.9	96.8	94.7	39.8	87.9	80.0	90.8
	sleep2vec	**88.6**	**0.84**	79.5	78.4	**96.8**	**94.8**	44.1	**88.2**	79.2	91.2

4.2 DOWNSTREAM FINE-TUNING RESULTS

4.2.1 SLEEP STAGING

We first assess the quality of the learned representations on sleep staging. Experiments are performed on the SHHS and WSC datasets, and the results are presented in Tables 1 and 15, respectively. Several trends can be observed from Table 1:

(i) There remains a very limited number of comprehensive works on PSG data, as the majority of existing methods focus narrowly on single-channel EEG or small subsets of physiological signals. Specialized methods typically achieve top performance across available channel sets, setting a challenging baseline for foundation models.

(ii) Foundation models generally exhibit lower performance compared to specialized sleep staging approaches optimized specifically for sleep data. This is evident in EEG-only scenarios, where specialized models consistently hold slight edges in overall metrics compared to baseline FM SleepFM (Acc. 86.6%) and sleep2vec (Acc. 87.4%). However, the gap is marginal, with sleep2vec nearly matching specialized models in certain metrics (κ of 0.82 *vs.* 0.83).

(iii) sleep2vec consistently outperforms baseline foundation models across all PSG channel subsets. Notable improvements appear in configurations such as "IBI & RESP", where sleep2vec exceeds baseline FMs (Acc.: 83.0% *vs.* SleepFM 79.4% and SleepFounder 80.9%). sleep2vec can also achieve performance comparable to, and in certain cases surpassing, specialized models.

(iv) Increasing modality diversity appears beneficial, with sleep2vec consistently demonstrating performance gains when additional physiological signals are included. This trend highlights the scalability and utility of incorporating diverse modalities into foundational model frameworks, further underscoring the capability of sleep2vec to effectively leverage multimodal physiological signals.

To further assess cross-cohort generalization, we evaluate models fine-tuned on SHHS directly on the APPLES cohort, which is unseen during both pre-training and fine-tuning. As shown in Table 2, sleep2vec preserves strong robustness under distribution shift and consistently outperforms baseline methods.

Table 2: Cross-cohort evaluation of five-class sleep staging (W/N1/N2/N3/REM) across PSG channel sets and models on **unseen APPLES**. Models are fine-tuned on SHHS without seeing any data from APPLES during both pre-training and fine-tuning. "FULL CHANNELS" refers to the fixed channel configuration that each model is designed for and individually pre-trained on. Underlined numbers indicate the best overall performance within each channel set; **bold numbers** denote the best performance among FMs; **bold-underlined numbers** indicate cases where the FM surpasses specialized models.

PSG Channel Set		Overall Performance (↑)					Class-wise F1 (↑)				
Inference Subset	Model	Acc.	κ	MF1	Sens.	Spec.	W	N1	N2	N3	REM
IBI & RESP	***Specialized (non-FM) Model***										
	(Sun et al., 2019) †	63.6	0.46	48.8	56.4	89.0	78.6	1.7	68.5	20.2	75.2
	(Goldammer et al., 2022) †	67.7	0.53	53.1	61.2	90.4	79.9	7.5	72.6	25.8	79.5
	Foundation Model										
	SleepFM (Thapa et al., 2024; 2025) †	69.1	0.55	54.3	65.2	91.0	82.2	4.6	73.5	28.1	82.8
	SleepFounder (Nie et al., 2025) †	68.8	0.55	55.6	**67.3**	91.0	85.5	8.8	71.9	27.2	**84.6**
	sleep2vec (InfoNCE)	71.5	0.59	56.5	66.7	91.7	**86.8**	7.9	74.1	29.5	84.1
	sleep2vec	**73.2**	**0.61**	**57.8**	66.2	**92.1**	86.5	**10.2**	**76.5**	**31.5**	84.2
FULL CHANNELS	***Foundation Model***										
	SleepFM (Thapa et al., 2024; 2025) †	71.4	0.59	60.0	58.9	91.6	85.5	24.5	75.8	32.0	81.5
	sleep2vec (InfoNCE)	76.8	0.67	63.5	**73.3**	93.4	**90.5**	24.0	79.5	35.5	88.1
	sleep2vec	**78.4**	**0.69**	**65.2**	72.0	**93.8**	89.6	**27.3**	**81.8**	**39.0**	**88.2**

4.2.2 LEAVE-ONE-OUT ANALYSIS

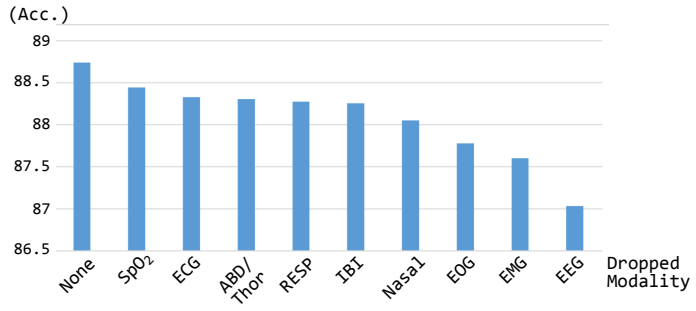

Figure 4: Leave-one-out analysis on the SHHS sleep staging task. Each bar represents model accuracy when one of the nine modalities is excluded during both pre-training and fine-tuning. The observed drop in accuracy relative to the full channels baseline (labeled "None") reflects the contribution and relative importance of each individual modality to the overall model performance.

To further examine the role of individual modalities, we performed a leave-one-out (LOO) study, where one modality was excluded during both pre-training and fine-tuning. Using the same setup as in Section 4.2.1, we evaluated the model on the SHHS dataset. As shown in Figure 4, excluding modalities such as EEG or IBI leads to a substantial accuracy drop, while others (e.g., SpO_2, EOG) have relatively minor effects.

4.2.3 CLINICAL DISEASE PREDICTION AND MODALITY SCALING

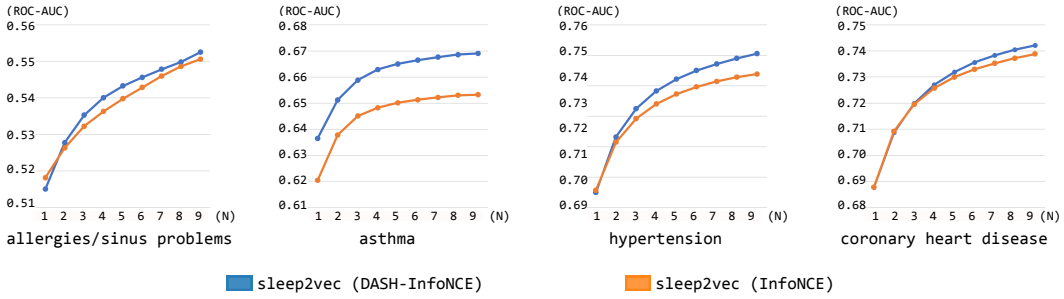

Figure 5: ROC-AUC scores for disease prediction tasks using varying numbers of modalities (N) on the SHHS dataset. Results are averaged across all possible modality combinations of size N

For clinical evaluation, four prevalent and clinically significant conditions are selected from the SHHS dataset, including *allergies/sinus problems*, *asthma*, *hypertension* and *coronary heart disease*, as shown in Figure 5. These conditions span two major physiological systems directly monitored by PSG, the respiratory system and the cardiovascular system. By including these diverse clinical outcomes, we explicitly test whether cross-modal embeddings generalize robustly across organ systems and sensor subsets.

Specifically, for a given number of modalities N, we enumerate all possible modality combinations, build corresponding ensemble models, and report the average ROC-AUC. Results presented in Figure 5 demonstrate: **(i)** clear modality-scaling effects, as performance consistently improves as more modalities are incorporated, suggesting a robust scaling law across clinical prediction tasks; **(ii)** the proposed DASH-InfoNCE loss consistently outperforms the standard InfoNCE baseline, indicating its effectiveness in harnessing richer inter-modal physiological correlations. This performance advantage of DASH-InfoNCE becomes increasingly pronounced with additional modalities, underscoring its efficacy in large-scale multimodal pre-training scenarios.

5 CONCLUSION

In this work, we introduced `sleep2vec`, a foundation model aligning multimodal polysomnography (PSG) signals into a unified embedding space for robust physiological representation learning. Leveraging over 42,000 overnight recordings and our novel DASH-InfoNCE loss, which accounts for demographic, age, site, and history variations, we demonstrated significant performance improvements on sleep staging and clinical prediction tasks. Experiments confirmed `sleep2vec`'s robustness to incomplete sensor data and revealed clear scaling laws with increased modality diversity and larger model sizes. Our results establish `sleep2vec` as a scalable and versatile tool, enabling generalized physiological monitoring and clinical decision support in sleep medicine.

6 ETHICS STATEMENT

The datasets employed consist of anonymized PSG recordings from publicly available sources. Ethical approval and informed consent for the original data collection were secured by the institutions responsible for the individual studies. All subject identifiers were removed prior to dataset acquisition, ensuring complete anonymization and protecting participants' privacy.

7 REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our research, we provide the following details. A comprehensive description of our data processing pipeline is provided in Section 3.1. Details of the datasets involved and the training configurations for the proposed model are presented in Appendices A.2 and A.3, respectively. Furthermore, the specific configuration used for fine-tuning is provided in Appendix A.7.

REFERENCES

Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, Carole Marcus, Bradley V Vaughn, et al. The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176 (2012):7, 2012.

Richard B Berry, Rita Brooks, Charlene Gamaldo, Susan M Harding, Robin M Lloyd, Stuart F Quan, Matthew T Troester, and Bradley V Vaughn. AASM scoring manual updates for 2017 (version 2.4), 2017.

Vera Birrer, Mohamed Elgendi, Olivier Lambercy, and Carlo Menon. Evaluating reliability in wearable devices for sleep staging. *NPJ Digital Medicine*, 7(1):74, 2024.

Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E Ensrud, Marcia L Stefanick, Alison Laffan, Katie L Stone, and Osteoporotic Fractures in Men Study Group. 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.

Konrad E Bloch. Polysomnography: a systematic review. *Technology and Health Care*, 5(4):285–305, 1997.

Mark I Boulos, Trevor Jairam, Tetyana Kendzerska, James Im, Anastasia Mekhael, and Brian J Murray. Normal polysomnography parameters in healthy adults: a systematic review and meta-analysis. *The Lancet Respiratory Medicine*, 7(6):533–543, 2019.

Jonathan F Carter, João Jorge, Oliver Gibson, and Lionel Tarassenko. Sleepvst: sleep staging from near-infrared video signals using pre-trained transformers. In *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 12479–12489, 2024.

Shigeng Chen, Xuesong Chen, Weijun Huang, Fei Lei, Chuxuan Shan, Zengrui Jin, Yunhan Shi, Yichen Wang, Rui Zhao, Xing Xu, et al. BCGNet: An AI Model Trained on 600K Hours of Sleep Data for BCG-based Contactless Monitoring. *medRxiv*, pp. 2025–08, 2025.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, 2015.

Zhaoliang Chen, Cheng Ding, Nirbhay Modhe, Jiaying Lu, Carl Yang, and Xiao Hu. Adapting a Generative Pretrained Transformer Achieves SOTA Performance in Assessing Diverse Physiological Functions Using Only Photoplethysmography Signals: A GPT-PPG Approach. In *Proc. AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An Attention-Based Deep Learning Approach for Sleep Stage Classification with Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29: 809–818, 2021.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *Neural Networks*, 107:3–11, 2018.

Ching Fang, Christopher Michael Sandino, Behrooz Mahasseni, Juri Minxha, Hadi Pouransari, Erdrin Azemi, Ali Moin, and Ellen L. Zippi. Promoting cross-modal representations to improve multimodal foundation models for physiological signals. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*, 2024.

Benjamin Fox, Joy Jiang, Sajila Wickramaratne, Patricia Kovatch, Mayte Suarez-Farinas, Neomi A Shah, Ankit Parekh, and Girish N Nadkarni. A foundational transformer leveraging full night, multichannel sleep study data accurately classifies sleep stages. *Sleep*, 48(8):zsaf061, 03 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 15180–15190, 2023.

Miriam Goldammer, Sebastian Zaunseder, Moritz D. Brandt, Hagen Malberg, and Felix Gräßer. Investigation of automated sleep staging from cardiorespiratory signals regarding clinical applicability and robustness. *Biomedical Signal Processing and Control*, 71:103047, 2022.

Michael R Irwin. Why sleep is important for health: a psychoneuroimmunology perspective. *Annual Review of Psychology*, 66(1):143–172, 2015.

Samuel T Kuna, Ruth Benca, Clete A Kushida, James Walsh, Magdy Younes, Bethany Staley, Alexandra Hanlon, Allan I Pack, Grace W Pien, and Atul Malhotra. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep*, 36(4):583–589, 2013.

Hyojin Lee, You Rim Choi, Hyun Kyung Lee, Jaemin Jeong, Joopyo Hong, Hyun-Woo Shin, and Hyung-Sin Kim. Explainable vision transformer for automatic visual sleep staging on multimodal psg signals. *npj Digital Medicine*, 8(1):55, 2025.

Yue Leng, Erik S Musiek, Kun Hu, Francesco P Cappuccio, and Kristine Yaffe. Association between circadian rhythms and neurodegenerative diseases. *The Lancet Neurology*, 18(3):307–318, 2019.

Junxin Li, Michael V Vitiello, and Nalaka S Gooneratne. Sleep in normal aging. *Sleep medicine clinics*, 17(2):161–171, 2022.

Diane C Lim, Arezu Najafi, Lamia Afifi, Claudio LA Bassetti, Daniel J Buysse, Fang Han, Birgit Högl, Yohannes Adama Melaku, Charles M Morin, Allan I Pack, et al. The need to promote sleep health in public health agendas across the globe. *The Lancet Public Health*, 8(10):e820–e826, 2023.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals. *arXiv preprint arXiv:2412.09758*, 2024.

Shuo Ma, Yingwei Zhang, Yiqiang Chen, Hualei Wang, Yuan Jin, Wei Zhang, and Ziyu Jia. SleepSMC: Ubiquitous Sleep Staging via Supervised Multimodal Coordination. In *Proc. ICLR*.

George Mathew, Daniel Barbosa, John Prince, and Subramaniam Venkatraman. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. *npj Cardiovascular Health*, 1(1):25, 2024.

Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. SleepEEGNet: Automated Sleep Stage Scoring with Sequence to Sequence Deep Learning Approach. *PloS one*, 14(5):e0216456, 2019.

Sutapa Mukherjee, Sanjay R Patel, Stefanos N Kales, Najib T Ayas, Kingman P Strohl, David Gozal, and Atul Malhotra. An official american thoracic society statement: the importance of healthy sleep. recommendations and future priorities. *American Journal of Respiratory and Critical Care Medicine*, 191(12):1450–1458, 2015.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024.

Guangkun Nie, Xuesong Chen, Yichen Wang, Jingxu Chen, Yunhan Shi, Jianwen Zhong, Weijun Huang, Zengrui Jin, Fei Lei, Leilei Wang, et al. A Low-Burden Sleep Foundation Model Built on Respiratory and Heartbeat Signals from 780,000+ Hours of Multi-Ethnic Sleep Recordings. *medRxiv*, pp. 2025–09, 2025.

Maurice M Ohayon, Mary A Carskadon, Christian Guilleminault, and Michael V Vitiello. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep*, 27(7):1255–1273, 2004.

Alexander Neergaard Olesen, Poul Jørgen Jennum, Emmanuel Mignot, and Helge Bjarup Dissing Sorensen. Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. *Sleep*, 44(1):zsaa161, 2021.

Joonas Paalasmaa, Mikko Waris, Hannu Toivonen, Lasse Leppäkorpi, and Markku Partinen. Unobtrusive Online Monitoring of sleep at Home. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3784–3788, 2012.

Jiahui Pan, Yangzuyi Yu, Man Li, Wanxin Wei, Shuyu Chen, Heyi Zheng, Yanbin He, and Yuanqing Li. A multimodal consistency-based self-supervised contrastive learning framework for automated sleep staging in patients with disorders of consciousness. *IEEE Journal of Biomedical and Health Informatics*, 29(2):1320–1332, 2024.

Paul E Peppard, Terry Young, Jodi H Barnet, Mari Palta, Erika W Hagen, and Khin Mae Hla. Increased prevalence of sleep-disordered breathing in adults. *American journal of epidemiology*, 177(9):1006–1014, 2013.

Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-Sleep: Resilient High-Frequency Sleep Staging. *NPJ digital medicine*, 4(1):72, 2021.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2021.

Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. SleepTransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, 2022.

Huy Phan, Kristian P Lorenzen, Elisabeth Heremans, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, Mathias Baumert, Kaare B Mikkelsen, and Maarten De Vos. L-SeqSleepNet: Whole-Cycle Long Sequence Modeling for Automatic Sleep Staging. *IEEE Journal of Biomedical and Health Informatics*, 27(10):4748–4757, 2023.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open Foundation Models for Optical Physiological Signals. In *Proc. ICLR*, 2025.

Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

Stuart F. Quan, Cynthia S. Chan, William C. Dement, Alan Gevins, James L. Goodwin, Daniel J. Gottlieb, Sylvan Green, Christian Guilleminault, Max Hirshkowitz, Pamela R. Hyde, Gary G. Kay, Eileen B. Leary, Deborah A. Nichols, Paula K. Schweitzer, Richard D. Simon, James K. Walsh, and Clete A. Kushida. The association between obstructive sleep apnea and neurocognitive performance—the apnea positive pressure long-term efficacy study (apples). *Sleep*, 34(3):303–314, 03 2011. ISSN 0161-8105. doi: 10.1093/sleep/34.3.303. URL https://doi.org/10.1093/sleep/34.3.303.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. ICML*, pp. 8748–8763, 2021.

Allan Rechtschaffen and Anthony Kales (eds.). *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Brain Information Service/Brain Research Institute, University of California, Los Angeles, Los Angeles, 1968. NIH Publication No. 204.

Richard S Rosenberg and Steven Van Hout. The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring. *Journal of clinical sleep medicine*, 9(1):81–87, 2013.

Ibrahim Sadek, Terry Tan Soon Heng, Edwin Seet, and Bessam Abdulrazak. A new approach for detecting sleep apnea using a contactless bed sensor: Comparison study. *Journal of Medical Internet Research*, 22(9):e18297, 2020.

Qi Shen, Junchang Xin, Bing Dai, Shudi Zhang, and Zhiqiong Wang. Robust sleep staging over incomplete multimodal physiological signals via contrastive imagination. In *Proc. NeurIPS*, volume 37, pp. 112025–112049, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Haoqi Sun, Wolfgang Ganglberger, Ezhil Panneerselvam, Michael J Leone, Syed A Quadri, Balaji Goparaju, Ryan A Tesh, Oluwaseun Akeju, Robert J Thomas, and M Brandon Westover. Sleep staging from electrocardiography and respiration with deep learning. *Sleep*, 43(7):zsz306, 12 2019.

Haoqi Sun, Wolfgang Ganglberger, Samaneh Nasiri, Aditya Gupta, Manohar Ghanta, Valdery Moura Junior, Sydney Cash, Katie Stone, Zhiyong Zhang, Gauri Ganjoo, Thijs E. Nassi, Ruoqi Wei, Erik-Jan Meulenbrugge, Rhoda Au, Gari Clifford, Lynn Marie Trotti, Dennis Hwang, Emmanuel Mignot, Umakanth Katwa, and M. Brandon Westover. The human sleep project (version 2.0), 2023. URL `https://doi.org/10.60508/qjbv-hg78`.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, Nov 2017.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Y Zou. SleepFM: Multi-Modal Representation Learning for Sleep Across ECG, EEG and Respiratory Signals. In *Proc. AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Rahul Thapa, Magnus Ruud Kjær, Bryan He, Ian Covert, Hyatt Moore, Umaer Hanif, Gauri Ganjoo, Brandon M Westover, Poul Jennum, Andreas Brink-Kjær, et al. A Multimodal Sleep Foundation Model Developed with 500K Hours of Sleep Recordings for Disease Predictions. *medRxiv*, pp. 2025–02, 2025.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *Proc. AAAI*, pp. 265–273, 2024.

Terry Young, Mari Palta, Jerome Dempsey, Paul E Peppard, F Javier Nieto, and K Mae Hla. Burden of sleep apnea: rationale, design, and major findings of the Wisconsin Sleep Cohort study. *WMJ: official publication of the State Medical Society of Wisconsin*, 108(5):246, 2009.

Baoxian Yu, Yaosheng Chen, Dongli Cai, and Han Zhang. A Proof-of-Concept Study on Noncontact BCG-Based Cardiac Monitoring for In-Patients With Sleep Apnea Syndrome Using Piezoelectric Ceramics. *IEEE Transactions on Instrumentation and Measurement*, 74, 2025.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: Towards a Sleep Data Commons. *Journal of the American Medical Informatics Association*, 25(10): 1351–1358, 2018.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Haiteng Jiang, Shijian Li, Benyan Luo, Tao Li, and Gang Pan. Personalized sleep staging leveraging source-free unsupervised domain adaptation. In *Proc. AAAI*, pp. 14529–14537, 2025.

A Appendix

A.1 Physiological Signals in Polysomnography (PSG)

Polysomnography (PSG) is a comprehensive overnight test performed using a polysomnograph that records multiple physiological signals during sleep, including brain activity, eye movements, muscle tone, heart rhythm, breathing patterns, and oxygen levels. An illustration of a subject wearing PSG device for nocturnal sleep recording is presented in Figure 1. It is used in clinical and research settings to diagnose and study sleep disorders, including but not limited to sleep apnea, narcolepsy, and insomnia, by providing an objective assessment of sleep stages and abnormalities.

Table 3: Polysomnography (PSG) channels and derived interval-based features used in this study. High sampling rate electrophysiological channels include EEG, EMG, EOG, and ECG; lower sampling rate cardiopulmonary and oximetry channels encompass Nasal airflow, Abdominal/Thoracic belt (ABD/Thor belt), and SpO_2. Respiratory effort (RESP) and Inter-Beat Interval (IBI) are interval-derived features, obtained from ABD/Thor belt and ECG channels, respectively, and can also be measured via wearable devices. *Sampling rate ranges summarize AASM Berry et al. (2012; 2017) minimum recommended digital sampling rates, actual device settings may vary.*

Physiological Signals	Typical Placement	Sampling (Hz)	Common Usage
PSG Channel			
EEG	Scalp	200–500	Detecting sleep stages, brain activity patterns, and brief arousals
EOG	Around the eyes	200–500	Identifying eye movements, especially for REM sleep detection and stage transitions
Chin EMG	Under the chin	200–500	Measuring muscle tone, useful for distinguishing REM and detecting disorders such as bruxism
ECG	Chest leads	200–500	Heart activity and variability (HR/HRV), used to study arousals and cardiorespiratory patterns
Nasal airflow	Under the nose	25–100	Detecting apnoeas/hypopnoeas and breathing irregularities
ABD/Thor belt	Around abdomen & chest	25–100	Tracking breathing effort, helping to classify types of sleep-disordered breathing
SpO_2	Finger probe	10–25	Monitoring blood oxygen drops during breathing events, used to measure severity
Interval-derived Feature			
Respiratory effort (RESP)	–	4–10	Breath-to-breath timing, used for variability analysis and detecting abnormal breathing cycles
Inter-Beat Interval (IBI)	–	4–10	Beat-to-beat timing, used to compute HRV and study autonomic regulation during sleep

Table 4: Overview of the PSG datasets used in this study.

Dataset	Age Span[1]	Duration	Recording Span	# Recordings	# Subjects	Total Hours
Sleep Heart Health Study (SHHS) (Quan et al., 1997)	39-90	8.9 ± 1.1	1995-2003	8,440	5,795	75,431
Wisconsin Sleep Cohort (WSC) (Young et al., 2009)	37-85	8.0 ± 0.8	2000-2015	2,570	1,123	20,520
Osteoporotic Fractures in Men Study (MrOS) (Blackwell et al., 2011)	67-90	11.5 ± 2.3	2000-2005	3,930	2,905	45,110
Multi-Ethnic Study of Atherosclerosis (MESA) (Chen et al., 2015)	54-94	10.6 ± 1.6	2010-2012	2,056	2,056	21,745
Human Sleep Project (HSP) (Sun et al., 2023)	1-109	7.6 ± 1.1	2007-present	25,253	18,973	190,732
Apnea Positive Pressure Long-term Efficacy Study (APPLES) (Quan et al., 2011)	18-83	8.2 ± 1.2	2003-2004	1,096	1,096	8,955

A.2 Overview of Datasets

Table 4 compiles six large publicly available PSG cohorts spanning children to older-adult populations (ages 1–109) and nearly three decades of acquisition (1995–present). As indicated by the Duration column in Table 4, all recordings correspond to full-night PSG studies. The corpus of training data comprises 42,249 overnight recordings from 30,852 subjects across these five cohorts (SHHS, WSC, MrOS, MESA, and HSP). In addition, the APPLES cohort, consisting of 1,096 recordings, is used as an external validation cohort. HSP contributes the broadest age range and largest share of data, while SHHS, WSC, MrOS, MESA and APPLES provide well-characterized adult cohorts. The diversity in demographics and collection periods enables robust pre-training and evaluation under heterogeneous sensors and montages.

A.3 Pre-training Configurations

During pre-training, we ensured consistency by using a fixed batch size of 320 across all models. Including the prepended `[CLS]` token, the maximum sequence length was capped at $L = 121$. For contrastive learning, the temperature parameter was set to $\tau = 0.2$. Optimization was performed

[1] In SHHS and MrOS, ages greater than 90 are top-coded and recorded as 90.

using AdamW (learning rate 5×10^{-5}, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and weight decay of 0.01 for non-normalization weights), with a linear warmup over 3% of steps followed by cosine decay.

Table 5: Configurations of the `sleep2vec` model across different sizes.

Configuration	Small	Medium	Large
Number of parameters	63.5M	133.7M	238.2M
Hidden dimension	512	768	1024
Number of layers	8	12	16
Attention heads	16	16	16
Segment duration (pre-training)	1 hour	1 hour	1 hour
Segment duration (fine-tuning)	Whole night	Whole night	Whole night

The architectural hyper-parameters are adjusted to yield `sleep2vec` variants with varying numbers of parameters, as detailed in Table 5. Unless otherwise stated, all experiments were conducted using the $\texttt{sleep2vec}_{\text{medium}}$ variant.

For the sample-weighting mechanism introduced in Section 3.3.3, we employed a Laplace kernel

$$\kappa(a_i, a_j) = \exp\Big(-\tfrac{|a_i - a_j|}{\sigma_{\text{age}}}\Big),$$

with bandwidth $\sigma_{\text{age}} = 20.0$. This choice reflects clinical observations that sleep physiology and sleep-disordered breathing vary gradually with age rather than abruptly. A 20-year scale captures meaningful across-lifespan differences without over-penalizing small age gaps (Ohayon et al., 2004; Li et al., 2022).

Gender coefficients are set to $\gamma_{\text{same}} = 1.0$ and $\gamma_{\text{diff}} = 0.8$, acknowledging sex differences in sleep architecture and in the prevalence/severity of sleep-disordered breathing that are present but not dominant at the individual-record level (Peppard et al., 2013). The site coefficients were set to $\delta_{\text{same}} = 1.3$ and $\delta_{\text{diff}} = 0.8$ to account for systematic inter-site variation (device, montage, scoring protocol) that is frequently larger than gender effects in multi-center cohorts (Rosenberg & Van Hout, 2013; Kuna et al., 2013). We used $\varepsilon = 10^{-6}$ for numerical stability.

Finally, a fixed margin term $\gamma m = 0.1$ (Eq. 4 and Eq. 5 in Section 3) was applied when modulating pseudo-negatives from the same subject-night, reflecting the high correlation of repeated segments within a recording and discouraging them from being treated as fully independent negatives.

Each pre-training run used two high-memory GPUs, the largest configuration trained for up to 48 hours.

We intentionally avoid any cross-modal reconstruction objective during pre-training. `sleep2vec` is trained solely with the InfoNCE and DASH-InfoNCE losses described above, which encourage alignment between heterogeneous PSG montages and associated metadata without requiring an explicit generative decoder. In preliminary experiments, a variant that replaced the contrastive objective with a generic cross-modal reconstruction module was implemented. Under matched data and compute budgets, this reconstruction-based variant was substantially harder to optimize and frequently failed to converge to competitive solutions. These observations, combined with the additional computational overhead of large reconstruction decoders, motivated our design choice to focus on contrastive alignment as a more stable and scalable route to robust missing-modality generalization.

A.4 Ablation Study

A.4.1 Ablation of Feature Fusion Strategies

To further assess the influence of different feature-fusion strategies, we perform an ablation study comparing the three representative designs incorporated in our framework: Concatenation, Mean and the adopted Gating mechanism.

In practice, Concatenation rapidly becomes computationally prohibitive as the number of modalities increases, since it expands the hidden representation dimensionality and consequently inflates

Table 6: Ablation study of different feature-fusion strategies and their impact on five-class sleep-staging performance (W/N1/N2/N3/REM). The evaluated model is the medium-sized sleep2vec variant, fine-tuned on SHHS and evaluated on **unseen APPLES**. **Bold numbers** denote the best performance among FMs.

	Overall Performance (↑)					Class-wise F1 (↑)				
Feature Fusion	Acc.	κ	MF1	Sens.	Spec.	Wake	N1	N2	N3	REM
Concatenation	76.9	0.66	63.1	71.8	93.3	89.9	21.0	79.9	37.3	87.4
Mean	78.1	0.68	64.7	71.9	93.7	89.3	26.2	81.5	38.4	88.1
Gating	**78.4**	**0.69**	**65.2**	**72.0**	**93.8**	**89.6**	**27.3**	**81.8**	**39.0**	**88.2**

both the parameter count and VRAM usage of subsequent layers. Additionally, it does not provide measurable performance benefits over the lightweight alternatives and is therefore not used as our default fusion approach.

Our analysis thus focuses on Mean and Gating, two scalable and computationally efficient paradigms. Across representative downstream sleep staging tasks, both strategies achieve competitive performance. Nonetheless, the Gating mechanism consistently yields small but robust improvements over Mean, and further offers enhanced interpretability through modality-specific gating coefficients that quantify the contribution of each input signal.

The results of this ablation study are reported in Table 6. Collectively, these findings justify our choice of Gating as the default fusion strategy, as it provides a balanced combination of scalability, empirical performance and interpretability.

A.4.2 Ablation of Metadata Components in DASH-InfoNCE

Table 7: Ablation study of metadata-aware contrastive objectives. Four contrastive formulations are compared during pre-training: (i) vanilla InfoNCE, (ii) single-metadata–aware variants that incorporate one metadata factor at a time (Age-aware, Gender-aware, Site-aware InfoNCE), and (iii) the proposed DASH-InfoNCE. All medium-sized sleep2vec models are pre-trained on the full multimodal corpus and subsequently fine-tuned and evaluated on **SHHS** for five-class sleep staging. "Retrieval Acc." corresponds to recall@1 in a cross-modal retrieval task, given a query embedding from one modality, the model must retrieve the correctly paired PSG segment from a pool of candidates drawn from other modalities, and Retrieval Acc. is the fraction of queries for which the true pair is ranked first. **Bold numbers** denote the best performance among FMs.

	Overall Performance (↑)					Class-wise F1 (↑)					Retrieval Acc. (↑)
Method	Acc.	κ	MF1	Sens.	Spec.	Wake	N1	N2	N3	REM	
Vanilla InfoNCE	88.4	**0.84**	78.6	77.9	**96.8**	94.7	39.8	87.9	**80.0**	90.8	0.351
Age-aware	88.3	0.83	78.7	77.2	96.7	94.6	41.6	88.0	78.1	91.0	0.355
Gender-aware	88.5	**0.84**	79.2	78.1	**96.8**	94.7	43.0	88.1	79.3	91.1	0.356
Site-aware	88.1	0.83	78.0	75.9	96.6	94.6	39.9	87.8	76.8	90.9	0.363
DASH-InfoNCE	**88.6**	**0.84**	**79.5**	**78.4**	**96.8**	**94.8**	**44.1**	**88.2**	79.2	**91.2**	**0.368**

To isolate the contribution of each metadata component within DASH-InfoNCE, an additional ablation study was conducted to examine the model's generalization behavior under distribution shift. Specifically, the impact of incorporating individual metadata, age, gender, and site, on performance in the unseen APPLES cohort.

For this analysis, models are pre-trained with only one metadata enabled at a time, followed by fine-tuning on SHHS and direct evaluation on APPLES without any further adaptation. Downstream sleep staging performance as well as cross-modal retrieval accuracy (Recall@1) are both reported, quantifying the quality of modality alignment in the shared embedding space.

Results presented in Table 7 and Table 8 suggest that activating any single metadata consistently improves performance over the vanilla InfoNCE baseline on the unseen cohort, either through higher

Table 8: Ablation study of metadata-aware contrastive objectives. Four contrastive formulations are compared during pre-training: (i) vanilla InfoNCE, (ii) single-metadata–aware variants that incorporate one metadata factor at a time (Age-aware, Gender-aware, Site-aware InfoNCE), and (iii) the proposed DASH-InfoNCE. All medium-sized sleep2vec models are pre-trained on the full multimodal corpus and subsequently fine-tuned on SHHS and evaluated on **unseen APPLES** for five-class sleep staging. "Retrieval Acc." corresponds to recall@1 in a cross-modal retrieval task, given a query embedding from one modality, the model must retrieve the correctly paired PSG segment from a pool of candidates drawn from other modalities, and Retrieval Acc. is the fraction of queries for which the true pair is ranked first. **Bold numbers** denote the best performance among FMs.

	Overall Performance (↑)					Class-wise F1 (↑)					Retrieval Acc. (↑)
Method	Acc.	κ	MF1	Sens.	Spec.	Wake	N1	N2	N3	REM	
Vanilla InfoNCE	76.8	0.67	63.5	**73.3**	93.4	**90.5**	24.0	79.5	35.5	88.1	0.351
Age-aware	78.3	0.68	64.9	72.0	93.7	89.5	26.6	81.7	39.0	88.0	0.355
Gender-aware	78.3	**0.69**	**65.5**	72.4	93.7	89.7	**29.1**	81.7	**39.2**	87.8	0.356
Site-aware	78.1	0.68	64.8	72.1	93.6	90.3	26.2	81.2	38.4	87.8	0.363
DASH-InfoNCE	**78.4**	**0.69**	65.2	72.0	**93.8**	89.6	27.3	**81.8**	39.0	**88.2**	**0.368**

Macro-F1 (MF1) or improved retrieval alignment. Combining all three metadata factors in the full DASH-InfoNCE formulation provides the strongest overall performance.

These findings demonstrate that age, gender and site information contribute complementary signals that enhance robustness under distribution shift. Together, they strengthen cross-modal alignment and cross-cohort generalization, highlighting DASH-InfoNCE as an effective strategy for improving model stability and transferability in unseen clinical cohorts.

A.4.3 Ablation of Masking Strategies

Table 9: Ablation study on the effect of masking ratios during pre-training. All medium-sized sleep2vec models are fine-tuned and evaluated on **SHHS** for five-class sleep staging. **Bold numbers** denote the best performance among FMs.

	Overall Performance (↑)					Class-wise F1 (↑)					Retrieval Acc. (↑)
Mask Ratio	Acc.	κ	MF1	Sens.	Spec.	Wake	N1	N2	N3	REM	
0%	88.5	**0.84**	78.8	77.2	96.7	94.7	41.9	**88.2**	78.2	91.1	**0.403**
15%	**88.6**	**0.84**	**79.5**	**78.4**	**96.8**	**94.8**	**44.1**	**88.2**	**79.2**	**91.2**	0.368
30%	88.4	0.83	78.8	77.2	96.7	94.7	41.6	88.1	78.6	91.1	0.281

Table 10: Ablation study on the effect of masking ratios during pre-training. All medium-sized sleep2vec models are fine-tuned on **SHHS** and evaluated on **unseen APPLES** for five-class sleep staging. **Bold numbers** denote the best performance among FMs.

	Overall Performance (↑)					Class-wise F1 (↑)					Retrieval Acc. (↑)
Mask Ratio	Acc.	κ	MF1	Sens.	Spec.	Wake	N1	N2	N3	REM	
0%	77.9	0.68	64.0	71.4	93.6	89.4	23.2	81.5	38.6	87.4	**0.403**
15%	**78.4**	**0.69**	**65.2**	72.0	**93.8**	89.6	**27.3**	**81.8**	**39.0**	**88.2**	0.368
30%	77.5	0.67	63.4	**72.2**	93.5	**89.8**	20.0	80.8	38.7	87.9	0.281

The role of masking strength during contrastive pre-training is also investigated to assess how different corruption levels affect the robustness of the learned representations. Specifically, downstream sleep staging performance of three masking ratios (0%, 15% and 30%) is reported in Table 9 (in-domain SHHS) and Table 10 (cross-cohort APPLES).

A moderate masking ratio of 15% yields the most favorable balance between representational robustness and downstream accuracy. Compared to the no-masking condition, moderate masking leads to consistent improvements in Macro-F1 (MF1) and class-wise metrics across both evaluation settings. In contrast, increasing the masking ratio to 30% provides no additional generalization benefits,

suggesting that overly aggressive masking may overly corrupt physiologically meaningful temporal structure. These results indicate that moderate masking functions as an effective regularizer during contrastive physiological pre-training.

We additionally note that retrieval accuracy is highest under the 0% masking configuration. This trend is likely driven by the closer match between the training objective and the retrieval evaluation when no corruption is applied, rather than reflecting superior generalization. Retrieval accuracy should therefore be interpreted jointly with downstream task performance when comparing masking strategies.

A.5 Further Investigation of Ablation Study

A.5.1 Scaling Law of Foundation Model Parameters

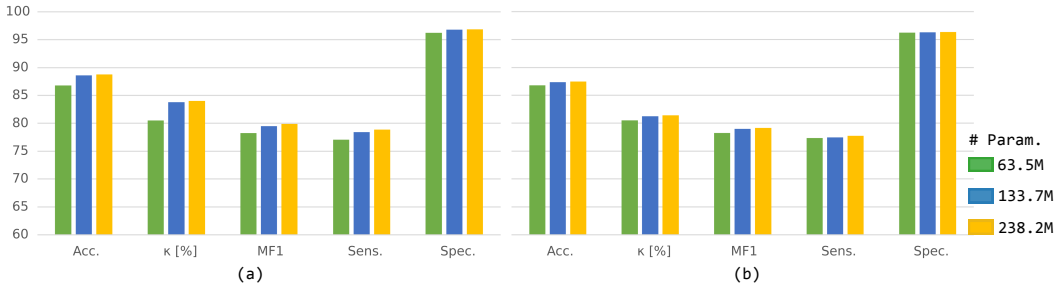

Figure 6: Performance comparison across varying model sizes (63.5M, 133.7M and 238.2M parameters) for sleep staging on **(a)** SHHS and **(b)** WSC datasets. Results demonstrate a clear scaling law, where increasing the number of parameters consistently improves Accuracy (Acc.), Cohen's Kappa (κ), Macro-F1 (MF1), Sensitivity (Sens.) and Specificity (Spec.), underscoring the effectiveness of scaling physiological foundation models in capturing complex sleep dynamics.

The scaling behavior of `sleep2vec` concerning model parameters is presented in Figure 6. The performance across two benchmark datasets, SHHS **(a)** and WSC **(b)**, consistently improves as the number of parameters increases from 63.5M to 238.2M. This improvement is evident in key metrics such as Accuracy (Acc.), Cohen's Kappa (κ), Macro-F1 (MF1), Sensitivity (Sens.) and Specificity (Spec.). Notably, the scaling effect exhibits diminishing returns, suggesting that while larger model sizes capture increasingly complex physiological patterns inherent in sleep data, the incremental gains become smaller with each parameter increase. Overall, these results affirm the robustness and scalability of the `sleep2vec` architecture, indicating its suitability for capturing detailed, multimodal physiological dynamics in sleep studies.

The training dynamics in Figure 7 and 8 indicate that increasing model capacity yields faster convergence and higher asymptotic retrieval accuracy, albeit with diminishing marginal improvements as model size grows.

A.5.2 Scaling Law of Pre-training Data Size

Table 11: Effect of pre-training data size on cross-cohort sleep staging performance (W/N1/N2/N3/REM). All medium-sized sleep2vec models are fine-tuned and evaluated on **SHHS**. **Bold numbers** denote the best performance among FMs.

	Overall Performance (↑)					Class-wise F1 (↑)					Retrieval Acc. (↑)
Data Fraction	Acc.	κ	MF1	Sens.	Spec.	Wake	N1	N2	N3	REM	
25%	87.1	0.82	77.2	76.3	96.4	93.7	37.6	86.8	77.7	89.9	0.281
50%	88.0	0.83	77.9	76.8	96.6	94.3	38.5	87.5	78.8	90.6	0.295
75%	88.4	0.83	78.7	77.1	96.7	94.6	41.2	88.0	78.3	91.1	0.340
100%	**88.6**	**0.84**	**79.5**	**78.4**	**96.8**	**94.8**	**44.1**	**88.2**	**79.2**	**91.2**	**0.368**

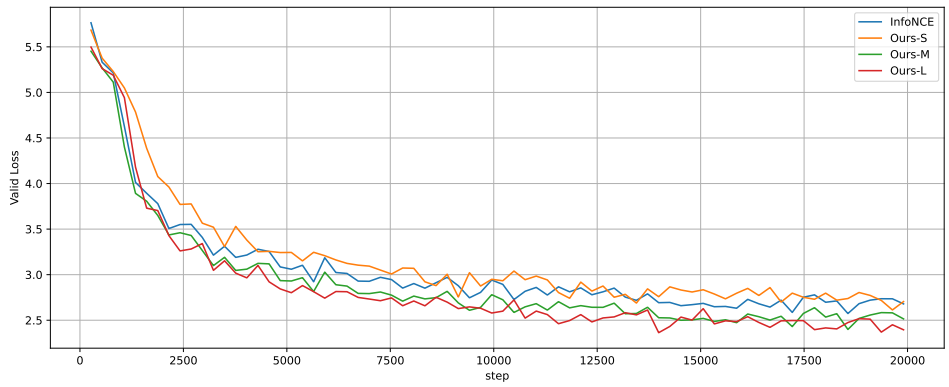

Figure 7: Validation loss curves across model scales under the vanilla InfoNCE objective and the proposed DASH-InfoNCE. Larger models (small, medium, large) consistently achieve lower validation loss and exhibit more stable convergence compared to the InfoNCE baseline, indicating increased representational capacity and more effective optimization dynamics.

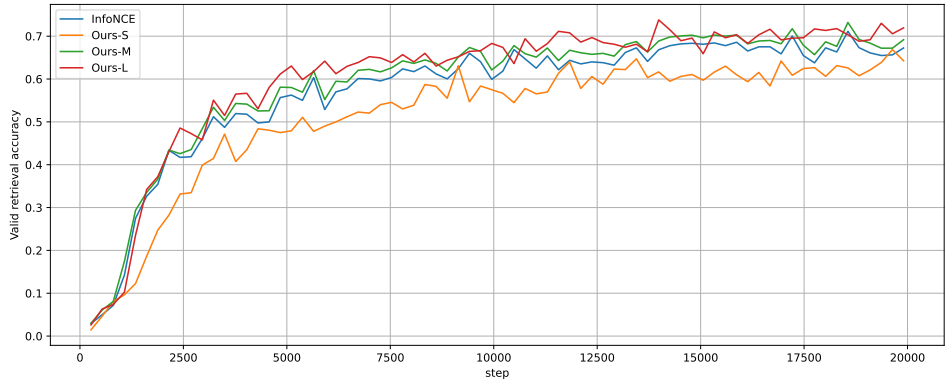

Figure 8: Validation retrieval accuracy curves across model scales under the vanilla InfoNCE objective and the proposed DASH-InfoNCE. Larger models (small, medium, large) consistently achieve higher retrieval accuracy and more stable convergence relative to the InfoNCE baseline, reflecting improved cross-modal alignment capability and stronger representation learning.

The impact of pre-training data scale on cross-cohort generalization is examined by varying the fraction of the pre-training corpus while keeping the downstream fine-tuning protocol fixed. Models are pre-trained using 25%, 50%, 75% and 100% of the available data and subsequently evaluated on an unseen cohort.

As shown in Table 11 and Table 12, increasing the amount of pre-training data leads to consistent improvements in downstream performance, particularly in Macro-F1 (MF1) and Cohen's κ. The performance improvements are most substantial when moving from low to intermediate-scale data regimes, with diminishing returns as the full dataset is utilized. This trend suggests that larger-scale physiological pre-training promotes more robust and transferable representations, which is especially valuable under distribution shift.

Collectively, these findings indicate that the proposed framework benefits notably from increased data scale and exhibits stable generalization properties across cohorts.

A.5.3 Scaling Law of Pre-Training Modality Number

The effect of pre-training modality count on downstream performance is examined by comparing four variants of the DASH-InfoNCE sleep2vec framework. The three curriculum stages differ only in the modality sets introduced during pre-training: **Stage 1** includes the most frequent and informa-

Table 12: Effect of pre-training data size on cross-cohort sleep staging performance (W/N1/N2/N3/REM). All medium-sized sleep2vec models are fine-tuned on **SHHS** and evaluated **unseen APPLES**. **Bold numbers** denote the best performance among FMs.

	Overall Performance (↑)					Class-wise F1 (↑)					Retrieval Acc. (↑)
Data Fraction	Acc.	κ	MF1	Sens.	Spec.	Wake	N1	N2	N3	REM	
25%	76.1	0.65	62.7	70.9	93.1	88.5	23.1	79.5	36.2	86.4	0.281
50%	76.8	0.66	63.0	71.7	93.3	89.5	21.6	80.0	36.2	87.7	0.295
75%	78.0	0.68	64.6	71.7	93.6	89.1	25.4	81.5	38.8	88.0	0.340
100%	**78.4**	**0.69**	**65.2**	**72.0**	**93.8**	**89.6**	**27.3**	**81.8**	**39.0**	**88.2**	**0.368**

Table 13: Effect of modality-scaling strategies on SHHS. "Single-stage" corresponds to the proposed medium sized sleep2vec model pre-trained from scratch using all available modalities. "Stage 1/2/3" implement a curriculum in which training begins with the most frequent and informative channels (EEG, RESP, and IBI in Stage 1), followed by the addition of EOG, ECG, and nasal airflow in Stage 2, and finally EMG, abdominal/thoracic belts, and SpO_2 in Stage 3, with each stage continuing from the previous checkpoint. All models are fine-tuned and evaluated on **SHHS** under identical EEG only and RESP+IBI downstream settings. **Bold numbers** indicate the best performance among FMs.

PSG Channel Set		Overall Performance (↑)					Class-wise F1 (↑)				
Inference Subset	Curriculum	Acc.	κ	MF1	Sens.	Spec.	Wake	N1	N2	N3	REM
EEG	Stage 1	86.9	0.81	77.4	76.3	96.3	93.5	41.0	86.7	77.2	88.6
	Stage 2	87.0	**0.82**	**77.7**	76.5	96.3	93.6	**42.5**	86.7	76.8	88.8
	Stage 3	87.3	**0.82**	77.2	75.7	96.4	93.7	38.9	**87.0**	77.6	**88.9**
	Single-stage	**87.4**	**0.82**	77.3	**76.6**	**96.5**	**94.2**	40.1	86.5	**77.7**	88.3
RESP+IBI	Stage 1	82.1	0.74	67.5	66.6	94.8	91.9	15.2	80.5	63.7	86.3
	Stage 2	82.2	**0.75**	66.3	66.9	94.9	91.8	6.6	80.2	**66.2**	86.6
	Stage 3	82.5	**0.75**	**68.2**	**67.3**	95.0	**92.3**	**17.1**	**80.9**	63.7	**86.8**
	Single-stage	**83.0**	**0.75**	65.9	65.8	**95.1**	86.6	5.3	**80.9**	64.3	86.6

Table 14: Effect of modality-scaling strategies. "Single-stage" corresponds to the proposed medium sized sleep2vec model pre-trained from scratch using all available modalities. "Stage 1/2/3" implement a curriculum in which training begins with the most frequent and informative channels (EEG, RESP, and IBI in Stage 1), followed by the addition of EOG, ECG, and nasal airflow in Stage 2, and finally EMG, abdominal/thoracic belts, and SpO_2 in Stage 3, with each stage continuing from the previous checkpoint. All models are fine-tuned on **SHHS** and evaluated on **unseen APPLES** under identical EEG only and RESP+IBI downstream settings. **Bold numbers** indicate the best performance among FMs.

PSG Channel Set		Overall Performance (↑)					Class-wise F1 (↑)				
Inference Subset	Curriculum	Acc.	κ	MF1	Sens.	Spec.	Wake	N1	N2	N3	REM
EEG	Stage 1	76.6	0.66	63.6	**72.8**	93.3	89.4	25.4	79.8	36.7	86.5
	Stage 2	**77.4**	**0.67**	**64.3**	72.5	**93.5**	**89.6**	**26.8**	**80.8**	37.6	86.9
	Stage 3	77.2	**0.67**	63.2	72.1	93.4	89.1	20.8	**80.8**	**37.8**	**87.3**
	Single-stage	76.7	0.66	62.5	71.9	93.3	**89.6**	19.1	80.0	37.3	86.7
RESP+IBI	Stage 1	72.3	0.60	56.8	66.4	91.9	86.7	7.1	75.3	31.0	83.8
	Stage 2	71.2	0.59	55.4	66.7	91.7	86.8	3.1	73.9	29.0	83.9
	Stage 3	72.3	0.60	57.1	**67.0**	91.9	**87.6**	7.8	74.8	30.1	**85.0**
	Single-stage	**73.2**	**0.61**	**57.8**	66.2	**92.1**	86.5	**10.2**	**76.5**	**31.5**	84.2

tive channels (EEG, RESP, and IBI); **Stage 2** resumes from the Stage 1 checkpoint and adds EOG, ECG, and nasal airflow; **Stage 3** incorporates the remaining, less frequent channels (EMG, abdominal/thoracic belts, and SpO_2). By contrast, the Single-stage model is trained from scratch using all modalities simultaneously.

All variants are fine-tuned using identical protocols on SHHS with either EEG or RESP+IBI as downstream inputs, and are evaluated both in-domain on SHHS and cross-cohort on APPLES. As reported in Table 13 and Table 14, expanding the modality set during pre-training yields small but consistent gains over the Stage 1 baseline, with the Single-stage model generally attaining the strongest or near-strongest performance. Importantly, the addition of rarer modalities in Stage 3 does not degrade performance, and all variants fall within a narrow accuracy and Macro-F1 (MF1) range. This stability indicates that sleep2vec scales predictably and robustly with respect to both modality number and modality diversity.

Moreover, as the number of pre-training modalities increases, the model retains, and often enhances, its performance when fine-tuned using only the originally available modalities, providing greater flexibility in downstream modality selection without sacrificing accuracy.

A.6 Retrieval Accuracy Matrix

The alignment quality is qualitatively validated using average recall@1 metrics computed across modalities on a test set comprising 10,109 samples. `sleep2vec` achieves a recall@1 of 36.8%, significantly surpassing the baseline with the original InfoNCE loss, which achieves 35.1%. This highlights the representational quality and discriminative capacity of the obtained embeddings.

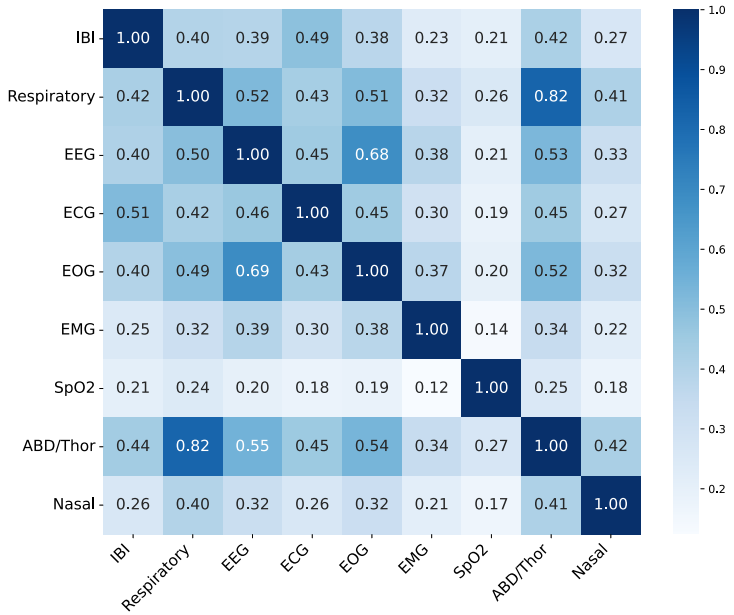

Figure 9: Recall@1 retrieval accuracy matrix of the learned representations. Rows correspond to the query modality, while columns indicate the retrieved modality.

The Recall@1 cross-modal retrieval accuracy matrix visualized in Figure 9 demonstrates distinct modality-specific alignment patterns. Notably, Respiratory and ABD/Thor exhibit exceptionally high mutual retrieval accuracy (0.82), aligning well with their known physiological coupling. EEG also demonstrates strong alignment with EOG (0.69). Conversely, modalities such as SpO_2 and EMG generally yield lower retrieval accuracies ($\approx$ 0.2–0.4), reflecting comparatively weaker physiological correlations.

A.7 Complementary Downstream Fine-tuning Results

A.7.1 Sleep Staging Configurations

For fine-tuning the transformer backbone in the sleep staging task, we utilized Low-Rank Adaptation (LoRA) to achieve parameter-efficient adaptation. Specifically, LoRA adapters were integrated into the `query`, `key` and `value` projections of every transformer layer, while keeping the original

backbone parameters frozen. Unless specified otherwise, we set the rank to $r = 8$, scaling factor $\alpha = 16$, dropout probability $p = 0.05$, without incorporating additional biases. For multimodal fine-tuning scenarios, the same set of LoRA adapters was shared across all modalities. Instead of relying on a special classification token, transformer output embeddings from the final layer at each time step were individually projected through a two-layer MLP classifier, whose hidden dimension matched that of the backbone output. Optimization was performed using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}.

Baseline models are reproduced using their original hyperparameters as reported in the corresponding publications.

A.7.2 Performance on WSC Dataset

Table 15: Performance of five-class sleep staging (W/N1/N2/N3/REM) across PSG channel sets and models on **WSC**. Reported metrics regarding overall performance including Accuracy (Acc., %), Cohen Kappa (κ), Macro-F1 (MF1, %), Sensitivity (Sens., %) and Specificity (Spec., %). Class-wise F1 (%) is also listed. Baselines reproduced by us for fair comparison are marked with †. Note that these foundation model baselines were individually pre-trained for each PSG channel subset, whereas `sleep2vec` was pre-trained only once across all modalities. "FULL CHANNELS" refers to the fixed channel configuration that each model is designed for and individually pre-trained on. Other naming conventions follow the one adopted in Table 1. Underlined numbers indicate the best overall performance within each channel set; **bold numbers** denote the best performance among FMs; **bold-underlined numbers** indicate cases where the FM surpasses specialized models.

PSG Channel Set		Overall Performance (↑)					Class-wise F1 (↑)				
Inference Subset	Model	Acc.	κ	MF1	Sens.	Spec.	W	N1	N2	N3	REM
EEG	***Foundation Model***										
	SleepFM (Thapa et al., 2024; 2025) †	84.3	0.76	73.6	72.4	95.2	90.1	40.4	**89.4**	**62.5**	85.4
	sleep2vec	**86.3**	**0.80**	**74.8**	**73.5**	**96.1**	**93.8**	**45.2**	89.1	60.1	**85.7**
IBI & RESP	***Specialized (non-FM) Model***										
	Sun et al. (2019) †	74.7	0.59	56.6	55.9	91.5	79.6	15.4	81.1	26.6	80.4
	Goldammer et al. (2022)	73.0	0.57	52.0	52.7	91.0	74.3	13.9	80.5	14.8	75.7
	Foundation Model										
	SleepFM (Thapa et al., 2024; 2025) †	77.7	0.65	56.7	57.0	92.5	83.8	10.6	83.9	24.5	81.0
	SleepFounder (Nie et al., 2025) †	79.8	0.69	65.5	64.8	93.7	86.8	28.0	**85.2**	43.5	**84.2**
	sleep2vec	**81.6**	**0.72**	**66.4**	**65.1**	**94.6**	**91.6**	**29.1**	84.0	**44.4**	82.9
EEG & EOG & EMG	***Specialized (non-FM) Model***										
	Olesen et al. (2021)	77.6	0.66	—	—	—	—	—	—	—	—
	Foundation Model										
	SleepFM (Thapa et al., 2024; 2025) †	84.5	0.77	75.1	74.9	95.4	90.0	44.3	89.3	63.8	87.9
	sleep2vec	**86.8**	**0.80**	**77.4**	**75.5**	**96.1**	**92.8**	**51.2**	**90.1**	**63.9**	**88.8**
FULL CHANNELS	***Foundation Model***										
	SleepFM (Thapa et al., 2024; 2025) †	84.6	0.77	75.4	75.0	95.4	90.0	45.2	89.4	64.1	88.1
	PFTSleep (Fox et al., 2025)	85.5	0.78	73.8	74.4	95.5	90.0	33.8	90.1	65.2	89.7
	sleep2vec (InfoNCE)	87.1	0.81	78.2	76.7	96.2	93.1	50.7	90.5	67.3	89.5
	sleep2vec	**87.3**	**0.81**	**79.0**	**77.5**	**96.3**	**93.3**	**53.8**	**90.7**	**67.5**	**89.7**

Several trends emerge from Table 15:

(i) Comprehensive research on PSG data remains limited, as existing methods commonly focus on single-channel EEG or small subsets of physiological signals. Specialized models typically provide established benchmarks, particularly in cardiorespiratory modality configurations, presenting significant evaluation standards for generalized foundation models.

(ii) Foundation models consistently exhibit strong performance, often surpassing specialized sleep staging methods. This is evident in the EEG configuration, where sleep2vec notably achieves the best overall performance among foundation models (Accuracy: 86.3%, κ: 0.80), outperforming baseline FM SleepFM (Accuracy: 84.3%, κ: 0.76).

(iii) sleep2vec consistently demonstrates superior performance across various PSG channel subsets. Specifically, in the "IBI & RESP" channel configuration, sleep2vec substantially surpasses both specialized and baseline foundation models (Accuracy: 81.6% compared to SleepFounder's 79.8% and SleepFM's 77.2%). Similarly, in the "EEG & EOG & EMG" subset, sleep2vec outperforms baseline foundation models (Accuracy: 86.8% vs. 84.5%) and considerably surpasses specialized methods (Accuracy: 86.8% *vs.* 77.6%). In the Full Channels configuration, sleep2vec achieves the highest

performance across multiple metrics (Accuracy: 87.4%, κ :0.81), underscoring the effectiveness of leveraging comprehensive modality combinations.

Table 16: Cross-cohort evaluation of five-class sleep staging (W/N1/N2/N3/REM) across PSG channel sets and models on **unseen APPLES**. Models are fine-tuned on WSC without seeing any data from APPLES during both pre-training and fine-tuning. "FULL CHANNELS" refers to the fixed channel configuration that each model is designed for and individually pre-trained on. Underlined numbers indicate the best overall performance within each channel set; **bold numbers** denote the best performance among FMs; **bold-underlined numbers** indicate cases where the FM surpasses specialized models.

PSG Channel Set		Overall Performance (↑)					Class-wise F1 (↑)				
Inference Subset	Model	Acc.	κ	MF1	Sens.	Spec.	W	N1	N2	N3	REM
IBI & RESP	***Specialized (non-FM) Model***										
	(Sun et al., 2019) †	69.9	0.54	51.6	53.3	90.4	78.6	8.9	76.2	17.1	77.2
	(Goldammer et al., 2022) †	68.6	0.52	48.6	50.3	90.2	74.3	8.1	76.7	14.6	69.5
	Foundation Model										
	SleepFM (Thapa et al., 2024; 2025) †	74.4	0.61	56.3	56.5	91.9	81.8	17.0	81.1	20.2	81.3
	SleepFounder (Nie et al., 2025) †	74.0	0.61	59.4	64.0	92.1	84.7	17.9	79.2	**31.7**	83.5
	sleep2vec (InfoNCE)	76.1	0.64	55.9	55.5	92.4	85.4	16.8	81.5	14.4	81.5
	sleep2vec	**76.4**	**0.64**	**57.7**	**57.7**	**92.7**	**85.7**	**19.2**	**81.8**	19.5	**82.5**
FULL CHANNELS	***Foundation Model***										
	SleepFM (Thapa et al., 2024; 2025) †	78.7	0.68	63.5	62.1	93.4	86.0	30.7	**85.4**	32.3	83.1
	sleep2vec (InfoNCE)	77.5	0.68	64.8	**70.5**	93.7	87.5	30.8	82.4	37.5	**86.0**
	sleep2vec	**80.1**	**0.71**	**66.3**	66.5	**94.1**	**89.1**	**33.0**	85.3	**38.9**	85.2

To further assess cross-cohort generalization, we evaluate models fine-tuned on WSC directly on the APPLES cohort, which is unseen during both pre-training and fine-tuning. A similar trend is observed in Table 16, mirroring the cross-cohort generalization results reported in Table 2.

A.7.3 INTERPRETABILITY OF CHANNEL-WISE CONTRIBUTION

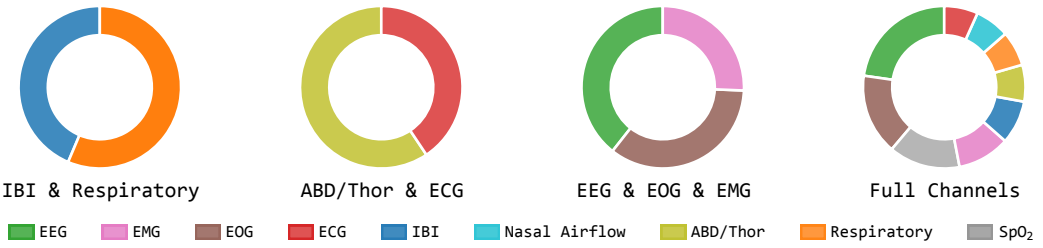

Figure 10: Visualization of modality-specific weights learned by the Gating Mechanism based feature fusion for the four SHHS sleep staging configurations in Table 1.

Beyond performance of sleep staging, the Gated scalar fusion used during fine-tuning provides a transparent, modality-level attribution of the downstream decision. Concretely, the learned scalars quantify each modality's contribution to the task-specific representation.

To illustrate, we analyze SHHS sleep staging under the four modality configurations in Table 1. Figure 10 presents the normalized fusion weights (visualized with a sharpening factor $T = 0.4$ that clarifies display while preserving relative ratios). Across tasks, EEG receives the largest weight, consistent with sleep staging being primarily annotated from EEG. EOG and EMG provide substantial complementary signal, while cardiorespiratory channels (e.g., airflow, ABD/Thor belt, IBI) carry smaller weights and SpO_2 contributes the least.

These weights are global, task-level attributions rather than per 30-second explanations, and they do not capture higher-order interactions between modalities. Nevertheless, these weights offer an interpretable summary of channel contribution that aligns with domain expectations and can inform sensor selection in channel-limited deployments.

A.7.4 Clinical Diagnosis Configurations

For clinical diagnosis tasks, the backbone parameters were kept frozen without applying LoRA adapters. Predictions were derived from the `[CLS]` token output of the final transformer layer, subsequently passed through a two-layer MLP classifier whose hidden dimension matched the backbone's output dimension. Training utilized the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}, maintaining consistency with the sleep staging experimental setup. Note that site information is not used as a covariate during the fine-tuning process.

A.8 LLM Usage

We use LLMs to correct grammatical errors.

Some elements in Figure 1 (an illustration of a participant wearing a PSG device for sleep recording, along with icons representing each physiological signal modality) and Figure 2 (participant icons in bed) were generated using LLMs.

A.9 Acknowledgment on Resource Usage

The following acknowledgment information is listed as required by the National Sleep Research Resource (NSRR) (Zhang et al., 2018) and the individual datasets.

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, "Outcomes of Sleep Disorders in Older Men," under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

This Wisconsin Sleep Cohort Study was supported by the U.S. National Institutes of Health, National Heart, Lung, and Blood Institute (R01HL62252), National Institute on Aging (R01AG036838, R01AG058680), and the National Center for Research Resources (1UL1RR025011). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).

The Apnea Positive Pressure Long-term Efficacy Study (APPLES) was supported by the National Heart, Lung, and Blood Institute (U01HL68060). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).