SEMPSG: A SEMANTIC CHANNEL-AWARE FOUNDATION MODEL FOR POLYSOMNOGRAPHY ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Polysomnography (PSG) integrates multiple physiological signals to provide a comprehensive characterization of human sleep. The growing availability of large-scale PSG recordings creates an opportunity to develop unified time series models across diverse applications. Existing studies often overlook the structured semantics encoded in heterogeneous PSG channel identities. Meanwhile, multi-scale temporal dynamics and entangled intra- and inter-modality dependencies pose challenges for unified PSG representation learning. In this paper, we propose **SemPSG**, a Semantic channel-aware foundation model for heterogeneous PSG analysis. SemPSG learns representations from time series and image views while incorporating the channel semantics. Notably, it introduces a channel semantics tokenizer that encodes physiological information from channel identities. A semantic-aware time-series and image architecture jointly captures long-range temporal dynamics, time-frequency patterns, and local waveform morphology, alleviating the trade-off across temporal scales. Finally, the intra- and inter-modality encoders explicitly model channel-level dependencies within each modality and cross physiological interactions. We evaluate SemPSG on six sleep and health-related tasks, experimental results demonstrate SemPSG outperforms general-purpose and PSG-specific foundation models, achieving a 9.6 percentage-point gain for sleep-disordered breathing classification, while demonstrating transferability across heterogeneous PSG and channel configurations. The code is available at `https://anonymous.4open.science/r/SemPSG`.

1 INTRODUCTION

Polysomnography (PSG) provides a comprehensive view of human sleep by continuously capturing multiple physiological signals, including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), and respiratory signals (Troester et al., 2023; Lee et al., 2025). Jointly, these signals capture brain activity, eye movements, muscle activity, cardiac activity, and respiratory dynamics (Krugliakova et al., 2026), respectively. The synchronized recordings form multi-physiological time series that support diverse downstream analyses (Ma et al., 2025). However, conventional PSG models are commonly developed for individual tasks, such as sleep staging or event detection, and rely on task-specific annotations (Phan et al., 2021; Wang et al., 2024; Kjaer et al., 2026). The growing availability of large-scale PSG recordings creates an opportunity to develop unified time series models that learn reusable physiological representations and transfer them across diverse applications (Thapa et al., 2024; Shuai et al., 2026).

Recent progress in time series foundation models has demonstrated the effectiveness of self-supervised pretraining for learning transferable representations across domains (Nie et al., 2023; Liu et al., 2024; Fu et al., 2026). However, PSG differs from conventional time series. As illustrated in Figure 1 (a), PSG comprises multiple physiological modalities with distinct sleep-related temporal dynamics and complex interactions across physiological systems, making direct adaptation of general-purpose foundation models non-trivial. These characteristics have motivated the development of foundation models specifically tailored to sleep- related physiological data (Xu et al., 2026). For example, SleepFM explores multimodal contrastive pretraining across physiological signals (Thapa et al., 2026), while SleepMaMi employs hierarchical temporal representation learning (Park et al., 2026). Other studies investigate cross-modal alignment across heterogeneous nocturnal biosignals (Yuan et al., 2026) and unified time-frequency modeling (Huang et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

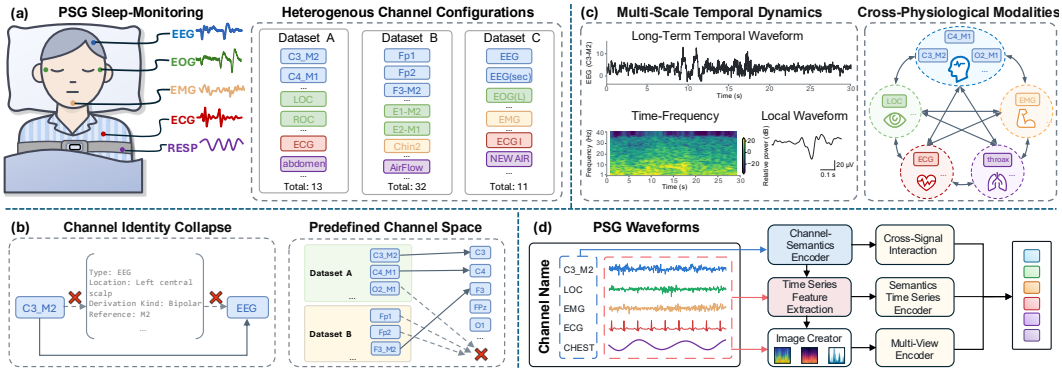

Figure 1: **Motivation.** (a) An example of PSG monitoring. (b) Limitations of conventional PSG. (c) PSG data characteristics. (d) The pipeline of SemPSG.

Despite these advances, existing approaches still struggle to learn generalizable representations across heterogeneous PSG recordings due to several challenges. The first challenge lies in *channel semantics heterogeneity*. PSG channels vary substantially across datasets from different acquisition systems, as illustrated in Figure 1 (a). One major issue is channel identity collapse, where distinct channels within the same physiological modality become indistinguishable when represented only by coarse modality labels, as shown in the left panel of Figure 1 (b). For example, *C3-M2* is not merely an EEG channel: *C3* specifies the recording location over the left central scalp region, whereas *M2* specifies the reference electrode over the right mastoid. Moreover, *C3-M2* and *M2-C3* contain the same electrode identities but opposite derivation directions, resulting in reversed waveform. Furthermore, predefined channel spaces preserve identities by assigning them to a fixed vocabulary of canonical channels, as shown in the right part of Figure 1 (b). It is difficult to accommodate unseen channels arising from different derivation configurations and naming conventions.

The second challenge is *temporal dynamics trade-off*. As shown in the left panel of Figure 1 (c), PSG signals exhibit distinct temporal dynamics at multiple scales: long-term temporal waveform characterizes the evolution of physiological states over extended periods, whereas time-frequency patterns reveal spectral characteristics; Meanwhile, local waveform morphology captures transient physiological events. These temporal dynamics provide complementary information for sleep analysis, but remains difficulties to balance for effective temporal representations. The third challenge is *intra- and inter-modality dependency entanglement*. As illustrated in the right part of Figure 1 (c), PSG contains dependencies both within and across physiological modalities. Existing approaches often emphasize inter-modality interactions while underexploring intra-modality channel relationships (Thapa et al., 2026; Huang et al., 2026; Park et al., 2026), causing channel-level correlations to become entangled with modality-level interactions.

In this paper, we propose **SemPSG**, a Semantic channel-aware foundation model for PSG analysis. As illustrated in Figure 1 (d), SemPSG learns representations from time series and image views while incorporating the channel semantics. To address the channel semantic heterogeneity, a channel semantics tokenizer is designed to represent each channel through the structured physiological semantics encoded in its identity. These semantic representations condition both waveform feature learning and channel aggregation, preserving channel-specific information without collapsing channels into coarse modality labels and restricting them to a predefined channel space. To alleviate the temporal dynamics trade-off, the semantic-aware time-series module captures long-range temporal dynamics from physiological waveforms, while the time-frequency and morphology branch extracts spectral patterns and local waveform morphology through Wavelet Transform (CWT) scalograms, Short-Time Fourier Transform (STFT) spectrograms, and morphology envelopes. The three-category views are then aggregated using learned sample-dependent weights. To disentangle the intra- and inter-modality dependencies, an intra-modality encoder first captures relationships among channels within each physiological modality, followed by semantic channel aggregation that forms modality-level representations. An inter-modality encoder then models interactions across different physiological systems, preventing channel-level correlations from being directly mixed with modality-level dependencies.

Overall, our main contributions are summarized as follows:

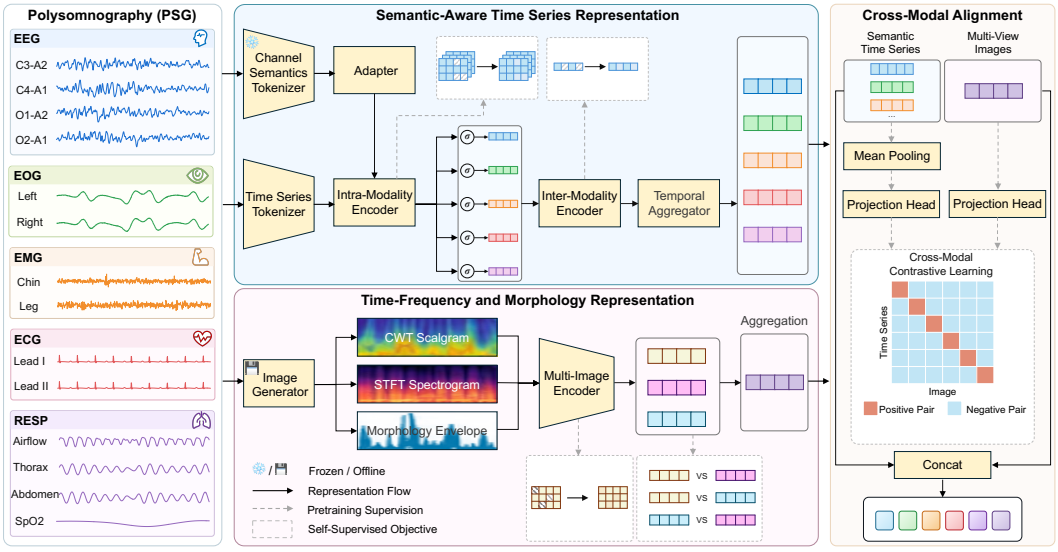

Figure 2: **Overall Framework of SemPSG.**

(i) We propose **SemPSG**, a Semantic channel-aware foundation model for heterogeneous PSG analysis. SemPSG introduces a channel semantics tokenizer that transforms dataset-specific channel identities into structured physiological representations, establishing a unified semantic space for knowledge sharing across heterogeneous channel configurations and acquisition systems.

(ii) SemPSG learns a unified representation from semantic time series and image views. The semantic-aware time series module hierarchically models intra- and inter-modality dependencies, disentangling channel-level correlations from modality-level interactions, while the multi-scale temporal module balances the long-range dynamics, spectral, and local waveform patterns.

(iii) Experiments conducted on six downstream tasks covering diverse sleep and health outcomes. SemPSG consistently outperforms general-purpose and PSG-specific foundation models, achieving a 9.6 percentage-point gain for sleep-disordered breathing classification, while demonstrating transferability across heterogeneous PSG datasets and channel configurations.

2 METHODOLOGY

SemPSG learns unified PSG representations from two views: a semantic-aware time series view and a time-frequency and morphology view, as shown in Figure 2. In the semantic time-series view, PSG signals are tokenized into temporal representations, while channel identities are encoded into semantic tokens that capture the structured information associated with each PSG channel. The fused semantic time-series embeddings are learned with intra- and inter-modality reconstruction objectives used during pretraining. In the image view, each physiological signal is transformed into Continuous Wavelet Transform (CWT) scalograms, Short-Time Fourier Transform (STFT) spectrograms, and morphology envelopes to capture complementary time-frequency and waveform characteristics. Finally, semantic, intra- and inter-modal, and multi-scale temporal representations are aligned in a shared space, enabling the model to capture both the semantics and physiological patterns of PSG.

2.1 SEMANTIC-AWARE TIME-SERIES ENCODING

To incorporate channel identity into physiological dependency modeling, the time-series branch combines channel-semantic conditioning with successive intra- and inter-modality encoding. It maps the waveforms $\mathbf{X}^{(t)}$, channel names $\mathcal{N}^{(t)}$, and validity mask $A^{(t)}$ to window-level representations $\{\mathbf{u}_m^{(t)}\}_{m=1}^{M}$, retaining a separate output token for each physiological modality.

Channel Semantics Tokenizer. Raw channel names provide recording identity cues that can inform how physiological signals are modeled. To this end, we develop a channel semantics Tokenizer

consists of byte-level embeddings, a convolutional stem, Transformer blocks, and padding-masked pooling followed by layer normalization. For channel c of modality m, its name $n_{m,c}^{(t)}$ is converted into a UTF-8 byte-token sequence $\mathbf{b}_{m,c}^{(t)} = (b_{m,c,1}^{(t)}, ..., b_{m,c,L_{m,c}}^{(t)})$, where $L_{m,c}$ denotes the sequence length, each byte is mapped to a token in $\{1, ..., 256\}$, and 0 is reserved for padding. This byte level representation preserves the original lexical structure, including case, whitespace, punctuation and byte order. The convolutional stem and Transformer blocks encode the byte sequence into contextual states $\mathbf{H}_{m,c}^{(t)} \in \mathbb{R}^{L_{m,c} \times D_n}$. Then, the encoder E_{name} provides the channel-name representation:

$$\mathbf{h}_{m,c}^{(t)} = \mathrm{LN}\left(\mathrm{Mean}_{valid}\left(E_{name}\left(\mathbf{b}_{m,c}^{(t)}\right)\right)\right) \in \mathbb{R}^{D_n}. \tag{1}$$

To endow $\mathbf{h}_{m,c}^{(t)}$ with physiological semantics, we pretrain the Channel Semantics Tokenizer using structured channel annotations. Specifically, auxiliary heads predict semantic type, derivation, anatomical location, and sensor information, while an alias-contrastive objective aligns different channel names sharing the same semantic signature. For ordered derivations, direction-aware hard negatives further distinguish reversed channel orders. After pretraining, the auxiliary heads are discarded, the tokenizer is frozen, and a lightweight adapter learns to incorporate the fixed semantic representations into waveform modeling:

$$\mathbf{s}_{m,c}^{(t)} = \mathrm{GELU}\left(\mathbf{W}_{\mathrm{adp}}\,\mathrm{LN}\left(h_{m,c}^{(t)}\right) + \mathbf{b}_{\mathrm{adp}}\right), \tag{2}$$

where $\mathbf{s}_{m,c}^{(t)} \in \mathbb{R}^D$ is trainable semantic conditioning for intra-modality and channel aggregation.

Intra-Modality Encoder. To model dependencies among channels within each modality, we jointly encode channel and temporal context under semantic conditioning. Each valid channel $\mathbf{X}_{m,c,:}^{(t)}$ is divided into P non-overlapping patches $\{\mathbf{x}_{m,c,p}^{(t)}\}_{p=1}^{P}$, which are mapped by a shared convolutional tokenizer to physiological tokens $\mathbf{v}_{m,c,p}^{(t)} = T_{\mathrm{ts}}(\mathbf{x}_{m,c,p}^{(t)}) \in \mathbb{R}^D$. The resulting tokens are then processed by Transformer blocks with semantic modulation and joint channel-temporal self-attention. Channel semantics guide how these physiological tokens interact by conditioning their hidden representation before self-attention. The semantic representation $\mathbf{s}_{m,c}^{(t)}$ is projected into feature-wise modulation parameters $\gamma_{m,c}^{(t)}$ and $\beta_{m,c}^{(t)}$ that are shared across all temporal patches of channel c:

$$[\gamma_{m,c}^{(t)}, \beta_{m,c}^{(t)}] = \mathbf{W}_{\mathrm{sem}}\mathbf{s}_{m,c}^{(t)} + \mathbf{b}_{\mathrm{sem}}, \tilde{\mathbf{z}}_{m,c,p}^{(t)} \qquad = \left(1 + \gamma_{m,c}^{(t)}\right) \odot \mathrm{LN}\left(\mathbf{z}_{m,c,p}^{(t)}\right) + \beta_{m,c}^{(t)}, \tag{3}$$

where $\mathbf{z}_{m,c,p}^{(t)}$ denotes the hidden representation of patch p. Self-attention is performed on the modulated representations. Applying the shared encoder separately to each modality yields contextualized physiological tokens $\{\mathbf{z}_{m,c,p}^{(t)}\}_{c,p}$.

To recover local waveform information from intra-modality context, we mask selected valid physiological tokens while preserving their channel semantics. A prediction head then reconstructs the masked token representations as $\hat{\mathbf{v}}i = R\mathrm{intra}(\mathbf{z}_i)$, with the objective

$$\mathcal{L}_{\mathrm{intra}} = \frac{1}{|\Omega_{\mathrm{intra}}|} \sum_{i \in \Omega_{\mathrm{intra}}} [1 - \cos(\hat{\mathbf{v}}_i, \mathrm{sg}(\mathbf{v}_i))], \tag{4}$$

where Ω_{intra} denotes the set of valid masked positions, $\mathbf{v}_i$ is the tokenizer output before masking, and $\mathrm{sg}(\cdot)$ stops gradients through the reconstruction target.

After intra-modality encoding, a semantic-aware gate aggregates channel representations into modality-level temporal sequences. For each patch p, the gate score for channel c combines its contextual token $\mathbf{z}_{m,c,p}^{(t)}$ with the projected channel-semantic representation:

$$e_{m,c,p}^{(t)} = \mathbf{w}^\top \left(\mathrm{LN}\left(\mathbf{z}_{m,c,p}^{(t)}\right) + \mathbf{W}_s\mathbf{s}_{m,c}^{(t)}\right). \tag{5}$$

The scores are normalized over the valid channels using the channel validity mask $A^{(t)}$, and the contextualized representations are aggregated as

$$\alpha_{m,c,p}^{(t)} = \frac{A_{m,c}^{(t)} \exp(e_{m,c,p}^{(t)})}{\sum_{c'=1}^{C} A_{m,c'}^{(t)} \exp(e_{m,c',p}^{(t)})}, \mathbf{f}_{m,p}^{(t)} = \sum_c \alpha_{m,c,p}^{(t)} \mathbf{z}_{m,c,p}^{(t)}, \tag{6}$$

where $\mathbf{f}_{m,p}^{(t)} \in \mathbb{R}^D$ integrates valid channels while retaining the modality and temporal indices.

Inter-Modality Encoder. To capture dependencies between physiological modalities, we design inter-modality interaction among the aggregated modality-level representation $\mathbf{f}_{m,p}^{(t)}$ at each temporal patch. The Inter-Modality Encoder consists of learnable modality embeddings and a Transformer. For each modality m, a learnable modality embedding $\mathbf{e}_m \in \mathbb{R}^D$ is added to $\mathbf{f}_{m,p}^{(t)}$ to indicate its modality identity. Then, the tokens at p are jointly encoded:

$$\mathbf{G}_p^{(t)} = E_{\text{inter}}\left(\left[\mathbf{f}_{1,p}^{(t)} + \mathbf{e}_1, ..., \mathbf{f}_{M,p}^{(t)} + \mathbf{e}_M\right]\right) \in \mathbb{R}^{M \times D}. \tag{7}$$

The resulting $\mathbf{g}_{m,p}^{(t)} \in \mathbb{R}^D$ denotes the contextualized representation of modality m at temporal position p, while missing modalities are excluded from attention.

During pretraining, we further leverage recovery from inter-modality context by masking selected valid modality–temporal representations. For each masked position i, a prediction head reconstructs the corresponding channel-aggregated representation as $\hat{\mathbf{f}}_i = R_{\text{inter}}(\mathbf{g}_i)$, with the objective

$$\mathcal{L}_{\text{inter}} = \frac{1}{|\Omega_{\text{inter}}|} \sum_{i \in \Omega_{\text{inter}}} [1 - \cos(\hat{\mathbf{f}}_i, \text{sg}(\mathbf{f}_i))], \tag{8}$$

where Ω_{inter} denotes the set of valid masked modality-temporal positions.

To obtain window-level representations while preserving modality-specific information, a temporal aggregator independently summarizes each sequence $\mathbf{g}^{(t)}m, pp = 1^P$. The temporal weights are:

$$a_{m,p}^{(t)} = \text{softmax}_p\left(\mathbf{w}_{\text{temp}}^\top \text{LN}\left(\mathbf{g}_{m,p}^{(t)}\right)\right), \tag{9}$$

and obtain the corresponding window-level modality representation as $\mathbf{u}_m^{(t)} = \sum_{p=1}^P a_{m,p}^{(t)} \mathbf{g}_{m,p}^{(t)}$.

2.2 Time-Frequency and Morphology Encoding

Raw PSG waveforms exhibit heterogeneous temporal characteristics, including multi-scale oscillations, localized spectral variations, and waveform morphology. To explicitly capture these complementary structures, we construct three image views for each PSG window: a CWT scalogram for scale-dependent oscillatory patterns, an STFT spectrogram for localized spectral content, and a morphology map for local time-domain waveform characteristics. CWT and STFT provide complementary time-frequency descriptions (Moca et al., 2021), while the morphology view preserves waveform information beyond spectral transforms.

Temporal Images Generation. To convert multichannel PSG signals into consistently organized image inputs, the Image Creator consists of channel-wise feature transformations and modality-wise spatial composition. For each valid channel c of modality m, we construct a feature map

$$\mathbf{Y}_{v,m,c}^{(t)} = \mathcal{T}_v\left(\mathbf{X}_{m,c,:}^{(t)}\right), v \in \{\text{cwt}, \text{stft}, \text{morph}\}, \tag{10}$$

where $\mathcal{T}_v$ denotes the corresponding transformation. The CWT and STFT views apply a Morlet wavelet and a Hann analysis window, respectively, followed by the log-power mapping $\log(1+|\cdot|^2)$.

To accommodate varying channel configurations while preserving a fixed modality layout, valid channels within each modality are aggregated using nonlearned weights derived from window-level signal statistics. CWT maps are normalized and resized before aggregation, whereas STFT and morphology maps are processed after aggregation. This yields a modality-specific image strip $\mathbf{J}_{v,m}^{(t)} \in [0,1]^{H \times W}$, and the complete view is formed by vertically stacking the M modality strips:

$$I_v^{(t)} = \text{Concat}_{\text{height}}\left(\mathbf{J}_{v,1}^{(t)}, ..., \mathbf{J}_{v,M}^{(t)}\right) \in [0,1]^{(MH) \times W}. \tag{11}$$

Missing modalities retain blank strips and are marked as invalid by spatial validity masks. All transformations are performed offline, and the image set $\mathcal{I}^{(t)} = \{I_{\text{cwt}}^{(t)}, I_{\text{stft}}^{(t)}, I_{\text{morph}}^{(t)}\}$, together with its validity masks, serves as the input to the Multi-Image Encoder.

Multi-Image Encoder. To accommodate these differences while learning shared contextual features, the Multi-Image Encoder consists of view-specific convolutional stems and a parameter-shared Transformer. Each stem ϕ_v maps its input into spatial tokens $\mathbf{R}_v^{(t)} = \phi_v\left(I_v^{(t)}\right) \in \mathbb{R}^{Q\times D}$, where Q is the number of spatial patches. A learnable view embedding $\mathbf{e}_v \in \mathbb{R}^D$ is added to preserve view identity, after which the tokens are passed through a Transformer f_{img} shared across the three views. We denote the resulting contextualized representation as $\mathbf{H}_v^{(t)} = f_{\text{img}}(\mathbf{R}_v^{(t)} + \mathbf{e}_v) \in \mathbb{R}^{Q\times D}$.

To preserve local structure, we apply masked latent reconstruction to the Multi-Image Encoder. Valid patches are masked before encoding and replaced with a learnable mask token, while a prediction head reconstructs their target latents. For position q in view v, the target is $\mathbf{y}^{(t)}v, q = \mathbf{R}^{(t)}v, q + \mathbf{e}_v$, obtained from the unmasked image using the same view-specific stem under stop-gradient:

$$\mathcal{L}_{\text{mim}} = \frac{1}{|\Omega_{\text{img}}|} \sum_{i\in\Omega_{\text{img}}} [1 - \cos(\hat{\mathbf{y}}_{v,q}, \text{sg}(\mathbf{y}_{v,q}))], \tag{12}$$

where Ω_{img} contains the valid masked positions, $\hat{\mathbf{y}}_{v,q}$ denotes the prediction.

Local reconstruction captures view-specific spatial structure but does not explicitly align different transformations of the same PSG window. We therefore introduce cross-view contrastive learning at the representation level. The contextualized tokens are averaged over valid spatial positions:

$$\mathbf{r}_v^{(t)} = \frac{1}{|\mathcal{Q}_v^t|} \sum_{q\in\mathcal{Q}_v^t} \mathbf{H}_{v,q}^{(t)}, \tag{13}$$

where $\mathcal{Q}_v^t$ denotes the valid spatial positions in view v. Representations from the same PSG window form positive pairs across views, while other eligible windows in the batch provide negatives. For a view pair (v, v'), let N denote the number of samples for which both views are valid. The directional contrastive loss and its symmetric aggregation are

$$\mathcal{L}_{v\to v'} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{r}_{v,i}, \mathbf{r}_{v',i})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{r}_{v,i}, \mathbf{r}_{v',j})/\tau)}, \mathcal{L}_{\text{view}} \quad = \frac{1}{\mathcal{P}} \sum_{(v,v')\in\mathcal{P}} \frac{\mathcal{L}_{v\to v'} + \mathcal{L}_{v'\to v}}{2}, \tag{14}$$

where $\mathbf{r}_{v,i}$ denotes the view-level representation of sample i, $\text{sim}(\cdot,\cdot)$ is cosine similarity between L_2-normalized representations, τ is the temperature, and $\mathcal{P} = \{(\text{cwt}, \text{stft}), (\text{cwt}, \text{morph}), (\text{stft}, \text{morph})\}$ contains the three complementary view pairs.

After cross-view representation learning, the view-level representations are integrated into a single image representation using learned sample-dependent weights:

$$a_v^{(t)} = \text{softmax}_{v\in\mathcal{V}^{(t)}}\, \mathbf{w}_{\text{img}}^\top \text{LN}(\mathbf{r}_v^{(t)}),\ \mathbf{z}_{\text{img}}^{(t)} = \sum_{v\in\mathcal{V}^{(t)}} \alpha_v^{(t)}\mathbf{r}_v^{(t)} \in \mathbb{R}^D, \tag{15}$$

where $\mathcal{V}^{(t)}$ denotes the set of valid views and $\mathbf{w}_{\text{img}} \in \mathbb{R}^D$ is a shared learnable scoring vector.

2.3 Cross-Modal Alignment: Bridging Time-Series and Image Representations

While multi-image views captures temporal dynamics, the semantic time-series view captures waveform dynamics and modality structure. To align these two views, we average valid modality representations only along the alignment pathway and project the pooled time-series and aggregated image representations into a shared contrastive space:

$$\bar{\mathbf{u}}^{(t)} = \frac{\sum_m^M a_m^{(t)}\mathbf{u}_m^{(t)}}{\sum_m^M a_m^{(t)}}, \mathbf{q}_{\text{ts}}^{(t)} = g_{\text{ts}}\left(\bar{\mathbf{u}}^{(t)}\right), \mathbf{q}_{\text{img}}^{(t)} = g_{\text{img}}\left(\mathbf{z}_{\text{img}}^{(t)}\right), \tag{16}$$

where $a_m^{(t)}$ indicates the validity of modality m, and g_{ts} and g_{img} are modality projection heads.

For N paired windows, same-windows form positives and others form negatives. We use cosine similarity between L_2-normalized projections to define the cross-modal contrastive objective:

$$\mathcal{L}_{\text{ts}\to\text{img}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{q}_{\text{ts},i}, \mathbf{q}_{\text{img},i})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{q}_{\text{ts},i}, \mathbf{q}_{\text{img},j})/\tau)}, \mathcal{L}_{\text{align}} \quad = \frac{\mathcal{L}_{\text{ts}\to\text{img}} + \mathcal{L}_{\text{img}\to\text{ts}}}{2}. \tag{17}$$

The final representation is $Z^{(t)} = [\mathbf{u}_1^{(t)}, \ldots, \mathbf{u}_M^{(t)}, \mathbf{z}_{\text{img}}^{(t)}] \in \mathbb{R}^{(M+1)\times D}$, formed by stacking the unprojected time-series and image representations. Combining cross-modal alignment with the reconstruction and cross-view objectives, the overall pretraining loss is:

$$\mathcal{L} = \mathcal{L}_{\text{intra}} + \lambda_{\text{inter}}\mathcal{L}_{\text{inter}} + \lambda_{\text{mim}}\mathcal{L}_{\text{mim}} + \mathcal{L}_{\text{view}} + \mathcal{L}_{\text{align}}, \tag{18}$$

where λ_{inter} is progressively increased during pretraining to strengthen cross-modal reconstruction, λ_{mim} controls the contribution of masked image reconstruction.

3 Experiments

We evaluate six downstream tasks, with the four task results in Sections 3.2– 3.6, two tasks results are reported in Appendix A.8.1– A.8.2, and additional results are provided in Appendix A.8.3– A.10.

3.1 Dataset Usage

As shown in Figure 3, we construct a heterogeneous PSG corpus from 10 cohorts, comprising 37,082 participants, 49,207 recordings, and 409,462 hours of sleep recordings. *Pretraining:* SemPSG is pretrained on nine cohorts: SHHS (Quan et al., 1997), MrOS (Blackwell et al., 2011), WSC (Young et al., 2009), MESA (Chen et al., 2015), APPLES (Quan et al., 2011), CCSHS (Rosen et al., 2003), PN2018 (Ghassemi et al., 2018), HSP-S0001 (Li et al., 2026), and iSLEEPS (Maiti et al., 2026). *For cohorts also used downstream, only training subjects are used for pretraining to prevent leakage. Downstream Evaluation:* We evaluate on SHHS, APPLES, and CFS. SHHS and APPLES assess within-cohort transfer, while CFS (Redline et al., 1995) is excluded from pretraining and serves as an external cohort for cross-dataset generalization. For SHHS and CFS, we follow the partition protocol of SleepMaMi (Park et al., 2026).

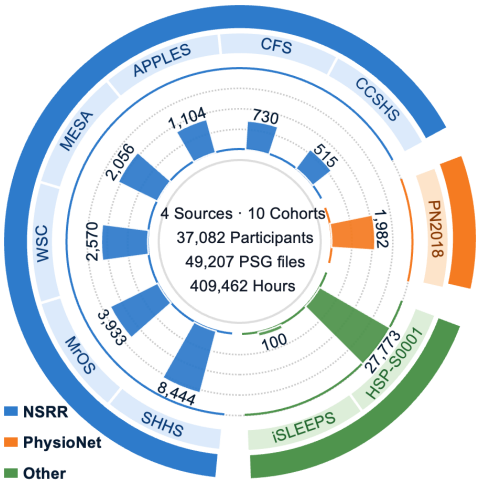

Figure 3: Statistics of the PSG datasets.

3.2 Sleep Staging

Sleep staging is formulated as a five-class classification task over 30-s PSG epochs. We evaluate both in-cohort generalization on SHHS1 and cross-dataset generalization on the unseen CFS cohort. As shown in Table 1, SemPSG consistently outperforms other sleep foundation models across both datasets and all metrics, with the largest gain on SHHS1, improving Kappa by 5.2 percentage points over SleepMaMi. General-purpose time-series models remain competitive, with Zeus performing particularly well on CFS, indicating that generic temporal representations capture useful staging information. SemPSG further improves performance by explicitly modeling intra-modality channel dependencies and inter-modality physiological interactions.

3.3 Sleep-Disordered Breathing Analysis

We evaluate sleep-disordered breathing (SDB) using two downstream tasks, SDB classification and apnea–hypopnea index (AHI) regression. SDB classification identifies normal and disordered breathing on a second-by-second basis, whereas AHI regression estimates overall respiratory disturbance severity at the recording level. As shown in Tables 2, SemPSG performs strongly on the SDB classification, which improves Macro-F1 by 9.6 percentage points over the second-best model SleepMaMi. For the experimental results of AHI regression, please see Appendix A.8.1

3.4 Disease Prediction

We evaluate whether the learned PSG representations transfer to downstream disease prediction. Based on the outcomes available in SHHS1, we consider six cardiovascular and cerebrovascular

Table 1: Sleep staging performance on SHHS1 and CFS. All metrics are reported as percentages.

Dataset	Category	Model	Accuracy ↑	Kappa ↑	Macro-F1 ↑
SHHS1	General-purpose TSFM	MOMENT	79.4	70.0	65.6
		UniTS	64.2	53.3	59.2
		Zeus	79.2	<u>70.5</u>	69.5
	EEG Foundation Model	LaBraM	70.2	55.7	54.1
	Sleep Foundation Model	SleepFM	69.7	55.9	56.3
		SleepMaMi	<u>81.9</u>	70.0	<u>74.1</u>
		SemPSG	**82.4**	**75.2**	**74.6**
CFS	General-purpose TSFM	MOMENT	71.4	56.7	52.4
		UniTS	64.1	53.3	58.1
		Zeus	<u>80.8</u>	<u>72.9</u>	69.7
	EEG Foundation Model	LaBraM	77.5	67.3	61.8
	Sleep Foundation Model	SleepFM	71.1	57.5	57.4
		SleepMaMi	80.7	<u>72.9</u>	<u>71.2</u>
		SemPSG	**81.9**	**74.8**	**73.6**

Table 2: Sleep-disordered breathing classification performance on SHHS1 and CFS.

Dataset	Category	Model	Accuracy ↑	Macro-F1 ↑
SHHS1	General-purpose TSFM	MOMENT	73.4	33.4
		UniTS	**88.2**	48.8
	EEG Foundation Model	LaBraM	50.2	20.4
	Sleep Foundation Model	SleepFM	77.5	39.4
		SleepMaMi	77.3	<u>60.6</u>
		SemPSG	<u>83.5</u>	**70.2**
CFS	General-purpose TSFM	MOMENT	74.2	38.5
		UniTS	<u>85.5</u>	40.0
	EEG Foundation Model	LaBraM	47.0	41.7
	Sleep Foundation Model	SleepFM	82.1	46.6
		SleepMaMi	79.9	<u>66.1</u>
		SemPSG	**86.5**	**69.4**

endpoints: angina, CVD death, CHF, CHD death, MI, and stroke. Performance is evaluated using the concordance index (C-index). As shown in Table 3, SemPSG achieves the highest C-index for four of the six outcomes, including CVD death, CHF, CHD death, and stroke. The largest gains are observed for CHD death and CVD death, where SemPSG outperforms the second-best model, SleepFM, by 7.2% and 5.7%, respectively. Its leading performance on CHF and stroke further indicates that the gains extend across multiple cardiovascular and cerebrovascular endpoints.

3.5 Cognition and Affective Assessment

We assess cognitive and affective information in the learned PSG representations via HAM-D and WASI score prediction on APPLES. As shown in Table 4, SemPSG achieves the lowest MAE on both tasks, reducing MAE by 6.0% for HAM-D and 6.5% for WASI over the second-best SleepFM. These results demonstrate that SemPSG preserves information relevant to both affective symptoms and cognitive function beyond conventional sleep outcomes

3.6 Modality Semantics Visualization

We visualize the representation of channel semantics tokenizer to explain the semantics of PSG. We assess the effect of channel semantics on channel fusion by shuffling channel-name assignments while keeping waveforms unchanged. As shown in Figure 5 (right), the gating pat-

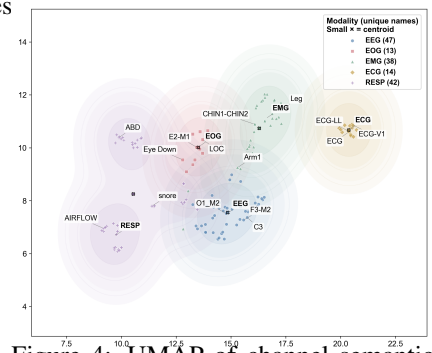

Figure 4: UMAP of channel semantic embeddings across modalities.

Table 3: Disease prediction performance measured on SHHS1.

Category	Models	Disease Outcomes					
		Angina	CVD Death	CHF	CHD Death	MI	Stroke
		C-Index ↑					
General-purpose TSFM	MOMENT	0.676	0.698	0.692	0.689	0.584	0.609
	UniTS	0.593	0.640	0.636	0.667	0.579	0.620
Sleep Foundation Model	SleepFM	0.632	0.791	0.764	0.781	0.636	0.729
	SleepMaMi	**0.778**	0.788	0.793	0.776	**0.662**	0.718
	SemPSG	0.727	**0.836**	**0.798**	**0.837**	0.660	**0.751**

Table 4: Cognition and affective assessment performance on APPLES.

Dataset	Category	Model	HAM-D MAE ↓	WASI MAE ↓
APPLES	General-purpose TSFM	MOMENT	3.15	9.12
		UniTS	3.13	8.86
	Sleep Foundation Model	SleepFM	3.02	8.56
		SleepMaMi	3.05	9.88
		SemPSG	**2.84**	**8.00**

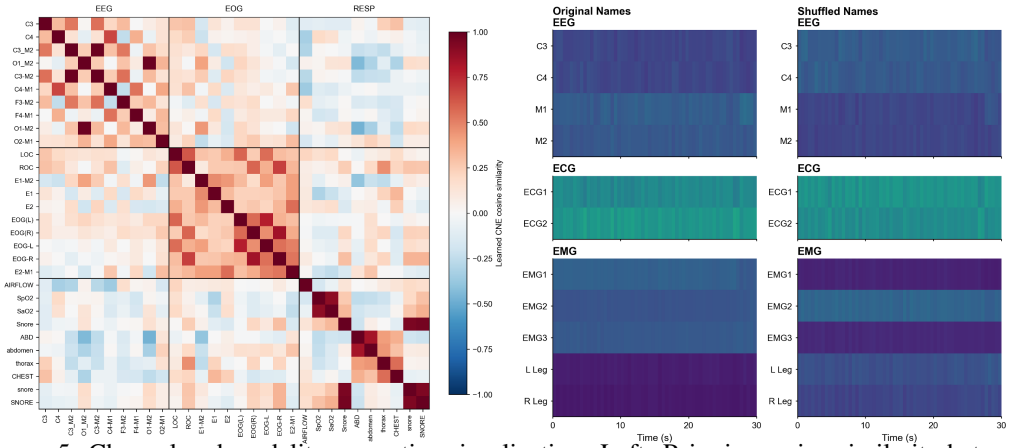

Figure 5: Channel and modality semantics visualization. Left: Pairwise cosine similarity between learned channel-semantic embeddings. Right: Learned channel-gating weights over a 30-s window for EEG, ECG, and EMG under the original and shuffled channel-name assignments.

terns change for EEG, ECG, and EMG, demonstrating that channel-name semantics directly influence intra-modality channel weighting. Figure 5 (left) shows that channels within the same physiological modality generally exhibit higher semantic similarity, while Figure 4 reveals finer-grained intra-modality structure. For example, respiratory channels form distinct subregions rather than collapsing into a single cluster, indicating that the learned space captures both modality-level and channel-specific semantics.

4 Conclusion

In this work, we present **SemPSG**, a semantic-aware foundation model for learning generalizable representations from heterogeneous PSG recordings. SemPSG integrates channel semantics, captures both intra- and inter-modality dependencies, and aligns waveform and image representations in a shared representation space. Experiments across a broad range of downstream tasks show that the learned representations transfer effectively across different prediction settings, supporting the use of SemPSG as a general-purpose foundation model for sleep and physiological signal analysis.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and with limited implementation and code-related tasks, including code editing and debugging. The research methodology, model architecture, experimental design, and interpretation of results were developed by the authors.

ETHICS STATEMENT

This study uses de-identified PSG recordings from publicly available datasets and involves no new human-subject data collection. Ethical approval and informed consent for the original studies were obtained by the respective institutions, and all datasets were used in accordance with their access and usage requirements.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, data preprocessing, pretraining objectives, training settings, and downstream evaluation protocols to support reproducibility. Key implementation details, including data splits, optimization settings, and loss configurations, are specified throughout the paper. Code is provided in both Abstract and Supplementary Material.

REFERENCES

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.

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.

Mingyue Cheng, Xiaoyu Tao, Zhiding Liu, Qi Liu, Hao Zhang, Rujiao Zhang, and Enhong Chen. Timemae: Self-supervised representations of time series with decoupled masked autoencoders. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pp. 498–508, 2026.

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

Yisong Fu, Zezhi Shao, Chengqing Yu, Yujie Li, Yongjun Xu, Xueqi Cheng, and Fei Wang. Zeus: Towards tuning-free foundation model for time series analysis. In *ICML*, 2026.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. UniTS: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37:140589–140631, 2024.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pp. 1–4. IEEE, 2018.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 16115–16152, 2024.

Weixuan Huang, Yan Wang, Hanrong Cheng, Wei Xu, Tingyue Li, Xiuwen Wu, Hui Xu, Pan Liao, Zaixu Cui, Qihong Zou, et al. A unified time-frequency foundation model for sleep decoding. *Nature Communications*, 17(1):1198, 2026.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, 2024.

Magnus Ruud Kjaer, Umaer Hanif, Andreas Brink-Kjaer, Mads Olsen, Oliver Sum-Ping, Oscar Carrillo, Scott A Sands, Susan Redline, Katie L Stone, Poul Jennum, et al. Expert-level probabilistic breathing event detector informs phenotyping of sleep apnea. *Nature Communications*, 17(1): 2548, 2026.

Elena Krugliakova, Friederike Breuer, Nico Adelhöfer, Alejandra Alonso, Luciana Besedovsky, Keith Murphy, Emma Peters, Karolina Raczek, Björn Rasch, Leila Salvesen, Sophia Snipes, Sarah Schoch, Thomas Schreiner, Rick Wassing, Til Ole Bergmann, and Martin Dresler. Hacking the functions of sleep: noninvasive approaches to stimulate sleep neurophysiology. *Physiological Reviews*, 106(2):675–749, 2026.

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.

Qichen Li, Shenghan Wen, Haoqi Sun, Wolfgang Ganglberger, Ayush Tripathi, Niels Turley, Samuel Waters, Arnav Gupta, Aditya Gupta, Manohar Ghanta, et al. The human sleep project: A multicenter clinical polysomnography dataset across the human lifespan. *Sleep*, pp. zsag215, 2026.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Forty-first International Conference on Machine Learning*, 2024.

Jiaming Ma, Qihe Huang, Haofeng Ma, Guanjun Wang, Sheng Huang, Zhengyang Zhou, Pengkun Wang, Xu Wang, Binwu Wang, and Yang Wang. PHAT: Modeling period heterogeneity for multivariate time series forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 135659–135685, 2026.

Shuo Ma, Yingwei Zhang, Yiqiang Chen, Hualei Wang, Yuan Jin, Wei Zhang, and Ziyu Jia. Sleepsmc: Ubiquitous sleep staging via supervised multimodal coordination. In *International Conference on Learning Representations*, volume 2025, pp. 24170–24191, 2025.

Suvadeep Maiti, Shivam Kumar Sharma, S Mythirayee, Srijithesh Rajendran, and Raju S Bapi. Polysomnography dataset for sleep analysis in ischemic stroke patients. *Scientific Data*, 13(1): 421, 2026.

Vasile V Moca, Harald Bârzan, Adriana Nagy-Dăbâcan, and Raul C Mureșan. Time-frequency super-resolution with superlets. *Nature communications*, 12(1):337, 2021.

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

Keondo Park, Younghoon Na, Yourim Choi, Hyunwoo Ryu, Hyun-Woo Shin, and Hyung-Sin Kim. Sleepmami: A universal sleep foundation model for integrating macro- and micro-structures. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

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.

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, et al. The association between obstructive sleep apnea and neurocognitive performance—the apnea positive pressure long-term efficacy study (apples). *Sleep*, 34(3):303–314, 2011.

Susan Redline, Peter V Tishler, Tor D Tosteson, John Williamson, Kenneth Kump, Ilene Browner, Veronica Ferrette, and Patrick Krejci. The familial aggregation of obstructive sleep apnea. *American journal of respiratory and critical care medicine*, 151(3_Part_1):682–687, 1995.

Carol L Rosen, Emma K Larkin, H Lester Kirchner, Judith L Emancipator, Sarah F Bivins, Susan A Surovec, Richard J Martin, and Susan Redline. Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity. *Journal of Pediatrics*, 142(4):383–389, 2003.

Zitao Shuai, Zongzhe Xu, David Yang, Wei Wang, and Yuzhe Yang. OSF: On pre-training and scaling of sleep foundation models. In *Forty-third International Conference on Machine Learning*, 2026.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Zou. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg, and respiratory signals. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48019–48037. PMLR, 2024.

Rahul Thapa, Magnus Ruud Kjaer, Bryan He, Ian Covert, Hyatt Moore, IV, Umaer Hanif, Gauri Ganjoo, M. Brandon Westover, Poul Jennum, Andreas Brink-Kjaer, et al. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, pp. 1–11, 2026.

M.M. Troester, S.F. Quan, American Academy of Sleep Medicine, and R.B. Berry. *The AASM Manual for the Scoring of Sleep and Associated Events, Version 3*. American Academy Of Sleep Medicine, 2023. ISBN 9780970613714.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 265–273, 2024.

Zongzhe Xu, Zitao Shuai, Eideen Mozaffari, Ravi S Aysola, Rajesh Kumar, and Yuzhe Yang. SleepLM: Natural-language intelligence for human sleep. In *Forty-Third International Conference on Machine Learning*, 2026.

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.

Weixuan Yuan, Zengrui Jin, Yichen Wang, Donglin Xie, Ziyi Ye, Chao Zhang, et al. sleep2vec: Unified cross-modal alignment for heterogeneous nocturnal biosignals. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8980–8987, 2022.

Zhe Zhao, Pengkun Wang, Haibin Wen, Shuang Wang, Liheng Yu, and Yang Wang. STEM-LTS: Integrating semantic-temporal dynamics in llm-driven time series analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22858–22866, 2025.

A Appendix

A.1 Related Work

Time-Series Representation Learning aims to extract generalizable features from sequential data that can be transferred across downstream tasks. To reduce the dependence on task-specific annotations, self-supervised learning has become a widely adopted paradigm, with representative approaches based on temporal-view consistency (Yue et al., 2022; Ma et al., 2026) and masked reconstruction (Dong et al., 2023; Cheng et al., 2026; Zhao et al., 2025). More recently, general-purpose time series foundation models have further scaled representation learning to heterogeneous datasets

and tasks. For example, MOMENT (Goswami et al., 2024) learns generalizable representations through large-scale masked pretraining, while UniTS (Gao et al., 2024) adopts a unified architecture across different time series tasks. These models demonstrate the potential of large-scale pretraining in general-purpose time series analysis, but they are not specifically designed to capture the structured physiological information associated with PSG channels.

Foundation Models for Polysomnography Analysis have recently emerged to exploit the heterogeneous physiological signals and temporally structured sleep patterns in PSG. For instance, SleepFM (Thapa et al., 2026) learns representations through contrastive learning across physiological signals, while SleepMami (Park et al., 2026) models sleep patterns at multiple temporal scales. Other studies have explored representation learning across heterogeneous nocturnal biosignals (Yuan et al., 2026) and channel-adaptive time-frequency modeling (Huang et al., 2026), further broadening the range of PSG signal settings considered. Despite these advances, existing approaches do not explicitly represent PSG channel identity through its structured physiological attributes, such as signal type, recording location, and reference configuration. This limits their ability to exploit semantics among channels and transfer knowledge across heterogeneous channel configurations.

A.2 Preliminaries

PSG Signals. A PSG (Polysomnography) signal $\mathcal{R}$ contains synchronized physiological time-series signals from five modalities $\mathcal{M} = \{EEG, EOG, EMG, ECG, RESP\}$. We divide each recording into non-overlapping 30s windows $\{\mathcal{W}^{(t)}\}_{t=1}^{T}$. For the t-th window, the time series signals are represented as $\mathbf{X}^{(t)} \in \mathbb{R}^{M\times C\times S}$, where $M = 5$, C is the padded channel dimension, and S is the number of samples per channel. The corresponding raw channel names are denoted by $\mathcal{N}^{(t)} = \{n_{m,c}^{(t)}\}$, with the channel validity indicated by $A^{(t)} \in \{0,1\}^{M\times C}$. Each window is also associated with $V = 3$ image views $\mathcal{I}^{(t)} = \{\mathbf{I}_v^{(t)}\}_{v=1}^{V}$ derived from the same physiological signals. Thus, each PSG window is represented as $\mathcal{W}^{(t)} = (\mathbf{X}^{(t)}, \mathcal{N}^{(t)}, \mathcal{I}^{(t)}, A^{(t)})$.

Problem Formulation. Given a PSG window $\mathcal{W}^t$, our goal is to learn a foundation model f_θ that maps its time series signals, channel name, and image views to a unified representation, $Z^{(t)} = f_\theta(\mathcal{W}^{(t)}) = [\mathbf{u}_1^{(t)}, ..., \mathbf{u}_M^{(t)}, \mathbf{z}_{img}^{(t)}]$, where $Z^{(t)} \in \mathbb{R}^{(M+1)\times D}$, $\mathbf{u}_m^{(t)} \in \mathbb{R}^D$ denotes the representation of physiological modality m, and $\mathbf{z}_{img}^{(t)} \in \mathbb{R}^D$ denotes the image representation. The foundation model is pretrained on unlabeled PSG windows with self-supervised objectives that capture local physiological structure and promote consistency across time-series and image representations.

A.3 Physiological Modalities and Channel Organization in PSG

PSG provides a multimodal view of sleep by jointly recording neurophysiological and cardiorespiratory activity. We organize the recorded signals into five physiological modalities: electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), and respiration (RESP).

Neurophysiological Signals. EEG, EOG, and EMG jointly characterize brain activity, eye movements, and muscle tone during sleep, providing complementary information about neurophysiological states.

Cardiorespiratory Signals. ECG captures cardiac activity and autonomic variations, while RESP may include airflow, thoracic and abdominal effort, and SpO_2, jointly characterizing respiratory dynamics during sleep.

Channel Organization. Each modality may contain multiple channels with different recording locations, derivations, and sensor characteristics. We retain these channel-level distinctions within a consistent modality space, motivating the modeling of both intra-modality channel dependencies and inter-modality physiological interactions.

A.4 Dataset Details

We pretrain SemPSG on multiple PSG cohorts collected from diverse populations and recording settings to expose the model to substantial variation in physiological signals, participant character-

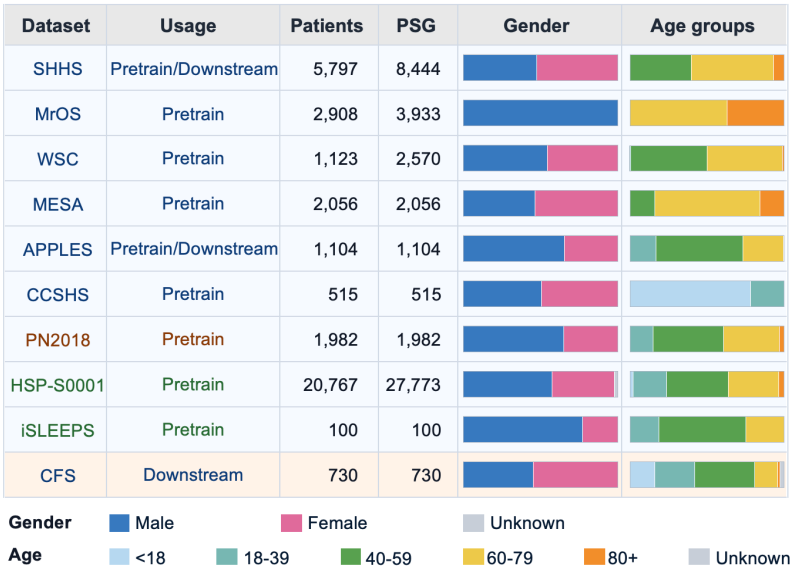

Figure 6: **Statistics and demographic characteristics of the PSG datasets used in this study.** For each cohort, we report its usage in pretraining and downstream evaluation, the number of patients and PSG recordings, and the distributions of gender and age groups. SHHS and APPLES are used for both pretraining and downstream evaluation, CFS is reserved exclusively for downstream evaluation, and the remaining cohorts are used for pretraining. The colored bars visualize the demographic composition of each cohort.

istics, and channel configurations during representation learning. As summarized in Figure 6, the pretraining corpus includes SHHS (Quan et al., 1997), MrOS (Blackwell et al., 2011), WSC (Young et al., 2009), MESA (Chen et al., 2015), APPLES (Quan et al., 2011), CCSHS (Rosen et al., 2003), PN2018 (Ghassemi et al., 2018), HSP-S0001 (Li et al., 2026), and iSLEEPS (Maiti et al., 2026). These cohorts differ markedly in scale, demographic composition, clinical characteristics, acquisition environments, and PSG channel configurations SHHS and APPLES contribute to both pretraining and downstream evaluation, with evaluation subjects excluded from representation learning to prevent information leakage. The remaining cohorts are used exclusively for pretraining, further broadening the diversity of the training distribution. For downstream evaluation, we additionally include CFS (Redline et al., 1995), which is completely excluded from pretraining and therefore serves as an unseen cohort for assessing cross-dataset generalization. Below, we describe the characteristics of each cohort and its specific role in our experimental setup.

- **Sleep Heart Health Study (SHHS)**. SHHS is a large-scale cohort study containing overnight PSG recordings together with extensive demographic and clinical information. The recordings provide multiple physiological signals covering neurophysiological, cardiac, and respiratory activity. In this study, SHHS is included in the pretraining corpus and is also used for downstream evaluation. Subjects assigned to downstream test sets are kept separate from the recordings used for representation learning.
- **MrOS Sleep Study (MrOS)**. The MrOS Sleep Study is an ancillary study of the Osteoporotic Fractures in Men cohort and contains overnight PSG recordings from community-dwelling older men. The sleep assessment includes unattended home PSG with neurophysiological, cardiac, and respiratory measurements, providing recordings from an older population and a home-based acquisition setting. We include MrOS as part of the pretraining corpus.
- **Multi-Ethnic Study of Atherosclerosis (MESA)**. MESA includes overnight sleep recordings collected as part of a population-based study of cardiovascular health. Its PSG recordings contain signals from multiple physiological modalities and provide additional variation in participant characteristics and recording configurations. We include MESA as part of the pretraining corpus.

- **Wisconsin Sleep Cohort (WSC)**. WSC is a longitudinal observational study established to investigate the prevalence, causes, and consequences of sleep disorders in adults. It includes repeated overnight in-laboratory PSG recordings collected from community-dwelling participants, together with extensive demographic and clinical information. The cohort therefore contributes longitudinal PSG data and additional variation in participant characteristics and recording sessions. We include WSC as part of the pretraining corpus.

- **Apnea Positive Pressure Long-term Efficacy Study (APPLES)**. APPLES is a multicenter clinical study of obstructive sleep apnea originally conducted to investigate the long-term effects of continuous positive airway pressure therapy. The dataset contains diagnostic PSG recordings together with demographic, clinical, and neurocognitive information from participants with suspected or diagnosed obstructive sleep apnea. In this study, APPLES is included in the pretraining corpus and is also used for downstream evaluation. Subjects assigned to downstream test sets are excluded from pretraining to maintain separation between representation learning and evaluation.

- **Cleveland Children's Sleep and Health Study (CCSHS)**. CCSHS is a population-based cohort established to investigate sleep and sleep-disordered breathing in children and adolescents. The available TREC visit includes full overnight laboratory PSG recordings with multiple physiological signals, introducing a younger population and a laboratory-based recording setting into the pretraining data. We include CCSHS as part of the pretraining corpus.

- **PhysioNet/Computing in Cardiology Challenge 2018 (PN2018)**. PN2018 contains overnight PSG recordings collected from subjects monitored in the Massachusetts General Hospital sleep laboratory for the diagnosis of sleep disorders. The recordings include multiple physiological signals together with expert annotations of sleep stages and sleep-related arousals. We use the physiological recordings for self-supervised pretraining, without relying on the accompanying task annotations.

- **Human Sleep Project (HSP).** The Human Sleep Project is a large-scale clinical sleep database containing overnight PSG recordings collected from patients undergoing sleep assessment. In this study, we use the HSP-S0001 subset, which provides multichannel physiological recordings from a clinical sleep population and further expands the diversity of recording conditions and channel configurations represented in the pretraining corpus. HSP-S0001 is used exclusively for pretraining.

- **iSLEEPS.** iSLEEPS is a clinical PSG dataset collected from patients with ischemic stroke at the National Institute of Mental Health and Neurosciences in India. It contains full-night PSG recordings acquired in a clinical setting, together with expert sleep-stage annotations, respiratory events, and associated demographic and clinical information. The recordings capture multiple physiological signals across neurophysiological, cardiac, and respiratory modalities, providing PSG data from a population with underlying neurological conditions. Compared with population-based sleep cohorts, iSLEEPS introduces additional variation in participant characteristics, disease status, and recording conditions. We include iSLEEPS exclusively in the pretraining corpus to broaden the physiological and clinical diversity encountered during representation learning and to expose SemPSG to heterogeneous PSG patterns.

- **Cleveland Family Study (CFS).** CFS is a family-based cohort established to investigate the familial aggregation of sleep apnea and related physiological characteristics. The study includes participants from a broad age range, with full overnight PSG available from the later study visit. In our experiments, CFS is used exclusively for downstream evaluation and is not included in the pretraining corpus, ensuring that its recordings remain unseen during representation learning.

A.5 ALGORITHMS

Algorithm 1 Channel semantics tokenizer Pretraining

procedure TRAINSEMANTICS($\mathcal{D}_{\text{name}}, \mathcal{R}$)
  Initialize encoder T_ϕ and semantic heads C_ψ
  for each batch (N, Y, A) **do**
    $h \leftarrow T_\phi(N)$; $z \leftarrow \text{Normalize}(h)$
    $\mathcal{L}_{\text{sem}} \leftarrow \text{SEMANTICLOSS}(C_\psi(h), Y, A)$ ▷ Attribute supervision
    $\mathcal{L}_{\text{alias}} \leftarrow \text{ALIASCONTRAST}(z, \mathcal{R})$ ▷ Alias consistency
    $\mathcal{L}_{\text{dir}} \leftarrow \text{DIRECTIONLOSS}(T_\phi, N, \mathcal{R})$ ▷ Ordered derivations
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{alias}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}}$
    Update (ϕ, ψ) to minimize $\mathcal{L}$
  end for
  return validation-selected T_ϕ
end procedure

Algorithm 2 SemPSG Pretraining Overview

procedure PRETRAINPSGFM($\mathcal{D}, T_\phi$)
  Freeze T_ϕ; initialize trainable parameters θ
  Prepare CWT, STFT, and morphology views
  for each batch (X, N, I, A) at step s **do**
    $(U, \mathcal{L}_{\text{intra}}, \mathcal{L}_{\text{inter}}) \leftarrow \text{ENCODESIGNALS}(X, N, A)$
    $(R, z_{\text{img}}, \mathcal{L}_{\text{mim}}) \leftarrow \text{ENCODEIMAGES}(I, A)$
    $\mathcal{L}_{\text{view}} \leftarrow \text{VIEWCONTRAST}(R)$ ▷ Across image views
    $\mathcal{L}_{\text{align}} \leftarrow \text{ALIGN}(U, z_{\text{img}}, A)$ ▷ Signals and images
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{intra}} + \lambda_{\text{inter}}(s)\mathcal{L}_{\text{inter}} + \lambda_{\text{mim}}\mathcal{L}_{\text{mim}} + \mathcal{L}_{\text{view}} + \mathcal{L}_{\text{align}}$
    Update θ to minimize $\mathcal{L}$
  end for
  return f_θ
end procedure

function ENCODESIGNALS(X, N, A)
  $E \leftarrow \text{Adapter}(T_\phi(N))$; $V \leftarrow \text{TOKENIZE}(X)$
  $(H, \mathcal{L}_{\text{intra}}) \leftarrow \text{MASKEDINTRA}(V, E, A)$ ▷ Within modalities
  $F \leftarrow \text{SEMANTICPOOL}(H, E, A)$ ▷ Across channels
  $(G, \mathcal{L}_{\text{inter}}) \leftarrow \text{MASKEDINTER}(F, A)$ ▷ Across modalities
  $U \leftarrow \text{TEMPORALPOOL}(G, A)$
  return $(U, \mathcal{L}_{\text{intra}}, \mathcal{L}_{\text{inter}})$
end function

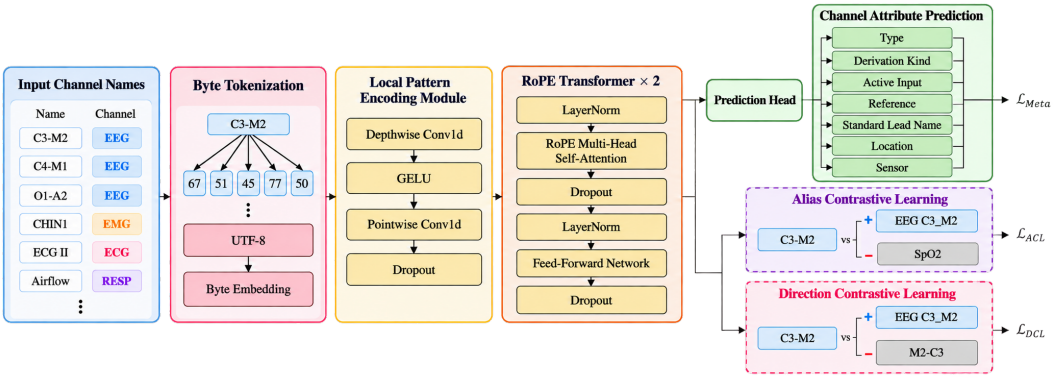

Figure 7: Channel semantics tokenizer training process.

A.6 BASELINES

We compare SemPSG with representative pretrained models from three categories, including general-purpose time-series foundation models, sleep-specific foundation models, and EEG foundation models where applicable. The general-purpose baselines include UniTS (Gao et al., 2024), MOMENT (Goswami et al., 2024), and Zeus (Fu et al., 2026), while the sleep-specific baselines include SleepFM (Thapa et al., 2026) and SleepMaMi (Park et al., 2026). LaBraM (Jiang et al., 2024) is additionally included as an EEG foundation model for tasks where it is applicable. Since not all pretrained models support every downstream setting, the exact set of baselines varies across tasks. All methods are evaluated using the same downstream data splits and task-specific evaluation protocols.

- **UniTS**. UniTS is a general-purpose time-series model designed to handle multiple predictive and generative tasks within a unified architecture. It introduces task tokenization to represent different task specifications and employs a shared Transformer backbone to learn transferable representations from heterogeneous time-series datasets spanning different domains, sampling rates, and temporal scales. The pretrained model can be adapted to downstream tasks such as classification, forecasting, imputation, and anomaly detection. In our implementation, we load the released UniTS x128 pretrained backbone and keep it frozen, using ten fixed zero prompt tokens and masked mean pooling over valid chunks, patches, and channels to obtain 128-dimensional representations for downstream evaluation.
- **MOMENT**. MOMENT is a family of general-purpose time-series foundation models pretrained on the Time Series Pile, a large collection of publicly available time-series datasets from diverse domains. Its pretraining framework is designed to learn transferable temporal representations that can be adapted to a range of downstream tasks under limited supervision, including classification, forecasting, imputation, and anomaly detection. We load the pretrained MOMENT-base checkpoint through the official package in embedding mode, freeze the encoder, and apply masked mean pooling over valid chunks, patches, and channels to obtain 768-dimensional representations for each 30-second epoch.
- **Zeus**. Zeus is a general-purpose time-series foundation model developed for tuning-free adaptation across heterogeneous analysis tasks. It combines point-wise tokenization with a multi-scale Transformer organized in a U-shaped hierarchy and introduces multi-objective temporal masking to capture temporal information required by different tasks within a shared pretrained model. This design enables the same model to perform multiple downstream tasks without task-specific fine-tuning. We use the official pretrained checkpoint as a frozen feature extractor, concatenating temporally pooled features from its five hidden-state scales and averaging across signal chunks and channels under our downstream protocol.
- **LaBraM**. LaBraM is an EEG foundation model designed to learn transferable representations across heterogeneous EEG datasets. It partitions multichannel EEG signals into channel-wise patches and uses a vector-quantized neural tokenizer to convert continuous EEG segments into discrete neural codes, after which a Transformer is pretrained to recover the codes of masked EEG patches. Because LaBraM is specifically designed for EEG rather than complete PSG recordings, we include it only for downstream tasks in which the required EEG inputs are available.
- **SleepFM**. SleepFM is a multimodal sleep foundation model pretrained on large-scale PSG recordings spanning EEG, ECG, EMG, and respiratory signals. It adopts a channel-agnostic architecture together with leave-one-out contrastive learning to learn across heterogeneous PSG configurations, encouraging representations from one physiological modality to align with information from the remaining modalities. The pretrained representations are designed to support transfer across sleep-related and clinical prediction tasks. We load the released SleepFM checkpoint with its accompanying model configuration and freeze the encoder, applying its spatial and temporal attention pooling within each modality and concatenating the resulting modality representations into a 512-dimensional vector.
- **SleepMaMi**. SleepMaMi is a sleep foundation model that jointly captures short-term physiological patterns and long-range sleep structure through a hierarchical dual-encoder architecture. Its Micro-Encoder models fine-grained biosignal characteristics using masked au-

toencoding and multimodal contrastive learning, while its Macro-Encoder captures longer-term sleep dynamics and incorporates demographic-guided contrastive learning. This hierarchical design provides representations at both local physiological and extended temporal scales. We use the official pretrained checkpoint with frozen parameters, following its channel and filtering conventions while adapting inference to complete nights and concatenating macro- and micro-encoder features into 3328-dimensional epoch representations.

A.7 IMPLEMENTATION

SemPSG is pretrained on 8 $\times$ Ascend NPUs using AdamW with a learning rate of 10^{-4} and a global batch size of 64. The pretrained encoder is frozen for downstream evaluation, where task-specific linear heads are trained.

Input processing. PSG recordings are processed as 30-s windows at 128 Hz, yielding 3,840 samples per channel. Continuous signals are normalized per channel at the recording level, without additional window-level normalization or clipping. Oxygen saturation values are scaled by 0.01. Each waveform is divided into patches of 64 samples, resulting in 60 temporal tokens per channel. Channels are organized into five modality groups: EEG, EOG, EMG, ECG, and RESP. Validity masks identify missing channels and modalities.

Model configuration. The intra-modality and inter-modality encoders each use two Transformer layers with a hidden dimension of 128, four attention heads, a feed-forward dimension of 2,048, and a dropout rate of 0.1. The separately pretrained channel-name encoder remains frozen throughout PSG pretraining, while the semantic adapter is trainable. The image branch receives three grayscale views of size 160×160, each arranged into five modality strips of height 32. The CWT view uses real Morlet wavelets with 32 geometrically spaced scales from 1 to 128. The STFT view uses a 256-sample Hann window with a hop size of 64, and the morphology view uses 80 temporal bins. Images are generated offline. The image encoder uses view-specific stems of width 64, 16×16 patches, and a shared two-layer Transformer with the same hidden dimension, attention head count, feed-forward dimension, and dropout as the waveform encoders. The final window representation consists of six 128-dimensional tokens, one for each physiological modality and one for the image branch. Invalid output tokens are zeroed, and their validity masks are retained.

Pretraining. SemPSG is pretrained on eight Ascend NPUs in FP32 using AdamW with a learning rate of 10^{-4} and weight decay of 0.01. The batch size is eight windows per device, giving a global batch size of 64 without gradient accumulation. We use 2,000 warmup steps followed by cosine learning-rate decay and clip the gradient norm at 1.0. Training runs for 80,245 optimizer steps with random seed 42. The masking probabilities for channel-level temporal tokens, modality-level temporal tokens, and image patches are 0.35, 0.25, and 0.40, respectively. Masking is applied only to valid tokens. The contrastive temperature is 0.07. Validation is performed every 5,000 steps using at most 20 global batches of 64 windows. For checkpoint selection, the validation weights for intra-modality reconstruction, inter-modality reconstruction, image reconstruction, cross-view contrast, and waveform-image alignment are $(1, 1, \lambda_{\mathrm{mim}}, 1, 1)$, respectively.

Downstream adaptation. The entire pretrained encoder is frozen during downstream evaluation, and only task-specific prediction heads are trained on cached representations. For window-level linear probes, the six output tokens are concatenated into a 768-dimensional feature vector.

A.8 ADDITIONAL EXPERIMENTAL RESULTS

Figure 8 provides an overall comparison across six downstream task groups, including sleep staging, sleep-disordered breathing analysis, AHI regression, disease prediction, cognition and emotion assessment, and demographic estimation. For visualization, all metrics are normalized to a common scale, with MAE-based metrics inverted so that larger values consistently indicate better performance. SemPSG exhibits strong and balanced performance across most task categories, with its curve remaining close to the outer boundary of the radar chart. In contrast, competing models tend to perform strongly on particular tasks but show larger variations across different physiological and clinical outcomes. This comparison highlights the broad transferability of SemPSG representations across heterogeneous downstream settings.

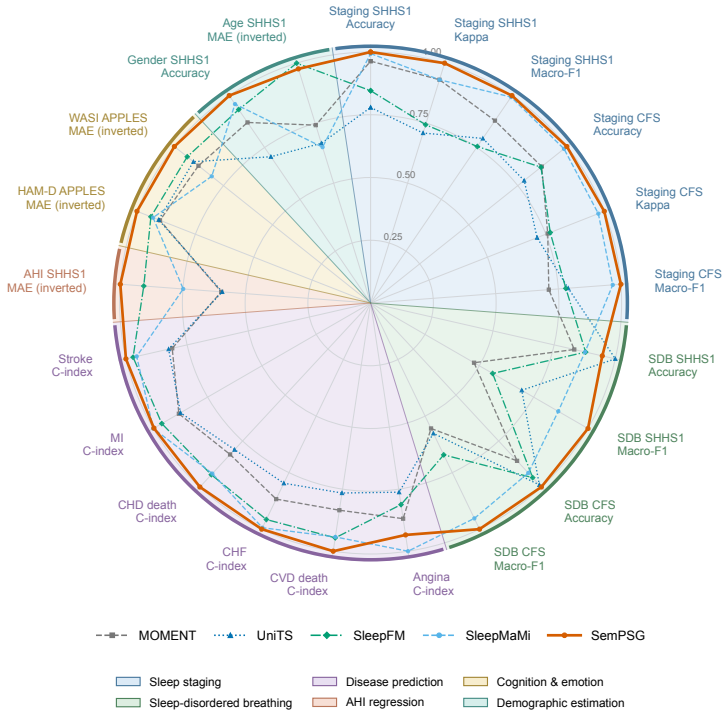

Figure 8: **Overall downstream performance of SemPSG and competing foundation models.** Each axis represents a normalized evaluation metric from one of six downstream task groups. MAE-based metrics are inverted such that higher values consistently indicate better performance. Colored sectors denote different task categories.

Table 5: AHI regression performance on SHHS1.

Category	Model	MAE ↓
General-purpose TSFM	MOMENT	10.54
	UniTS	10.61
Sleep Foundation Model	SleepFM	<u>6.94</u>
	SleepMaMi	8.40
	SemPSG	**6.29**

A.8.1 Sleep-Disordered Breathing Analysis: AHI Regression

Another type of sleep-disordered breathing analysis is apnea–hypopnea index (AHI) regression, SemPSG achieves a 9.4% lower MAE than the second-best model SleepFM, indicating more accurate estimation of recording-level respiratory disturbance severity. These results show that the learned PSG representations capture information relevant to both fine-grained breathing abnormalities and overall respiratory burden. By preserving modality-specific representations while modeling intra-modality channel dependencies and inter-modality interactions, SemPSG captures respiratory patterns alongside physiological responses distributed across other PSG signals, supporting both second-by-second SDB classification and recording-level AHI estimation.

A.8.2 Demographic Estimation

We evaluate whether the learned PSG representations encode demographic information through gender classification and age regression. As shown in Table 6, SemPSG achieves the highest accuracy for gender classification and the second-lowest MAE for age estimation. The results indicate that

Table 6: Demographic estimation performance on SHHS1.

Dataset	Category	Model	Gender Accuracy ↑	Age MAE ↓
SHHS1	General-purpose TSFM	MOMENT	72.7	8.28
		UniTS	58.9	9.22
	Sleep Foundation Model	SleepFM	77.9	**6.14**
		SleepMaMi*	<u>80.1</u>	9.45
		SemPSG	**83.5**	<u>6.29</u>

* SleepMaMi is evaluated using only its Micro-Encoder for both gender classification and age regression, as its Macro-Encoder is pretrained with the demographic contrastive objective.

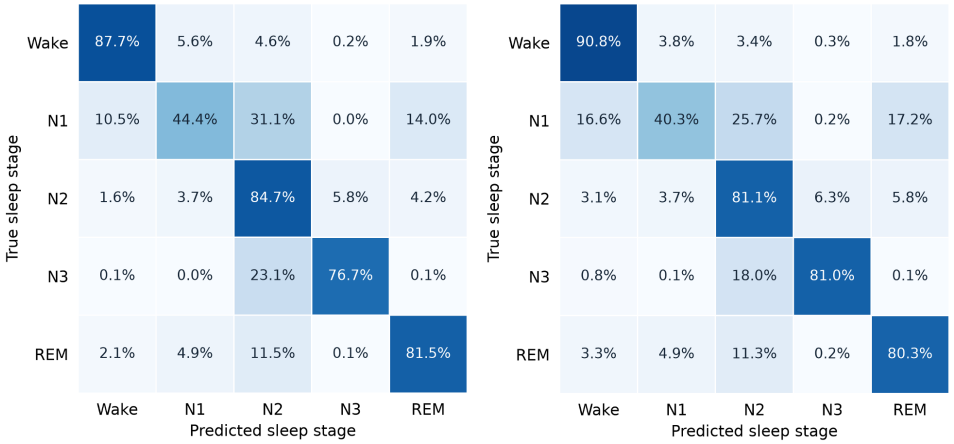

Figure 9: Row-normalized confusion matrices for sleep staging using frozen SemPSG representations with validation-selected linear probes on the SHHS-1 (left) and CFS (right) test sets. Rows indicate ground-truth stages and columns indicate predicted stages. Each entry reports the percentage of epochs within the corresponding ground-truth stage; diagonal entries therefore represent per-stage recall.

the learned representations retain information associated with demographic characteristics such as gender and age.

A.8.3 Sleep Staging

Figure 9 provides a stage-wise analysis of SemPSG's sleep-staging predictions. Both datasets exhibit predominantly diagonal confusion matrices, with recall exceeding 76% for Wake, N2, N3, and REM. Wake achieves the highest recall on SHHS-1 and CFS (87.7% and 90.8%, respectively), while N2 and REM maintain recall above 80% on both datasets. N3 recall is 76.7% on SHHS-1 and 81.0% on CFS, with most N3 errors assigned to N2 (23.1% and 18.0%, respectively). The reverse confusion is substantially smaller: only 5.8% and 6.3% of true N2 epochs are classified as N3.

A.9 Channel Semantics and Modality Analysis

Figure 10 and figure 11 extend the semantic visualizations in the main text to all five modality groups. Figure 10 presents pairwise cosine similarities between learned channel-name embeddings. The matrix exhibits modality-related structure without forming uniformly similar blocks. ECG channels show relatively high within-group similarity, whereas EEG, EOG, and EMG display more heterogeneous patterns, with stronger similarities among particular channel subsets. Respiratory-related channels also form distinct subgroups rather than a single homogeneous block. These observations indicate that the learned embeddings capture distinctions within the broad modality categories.

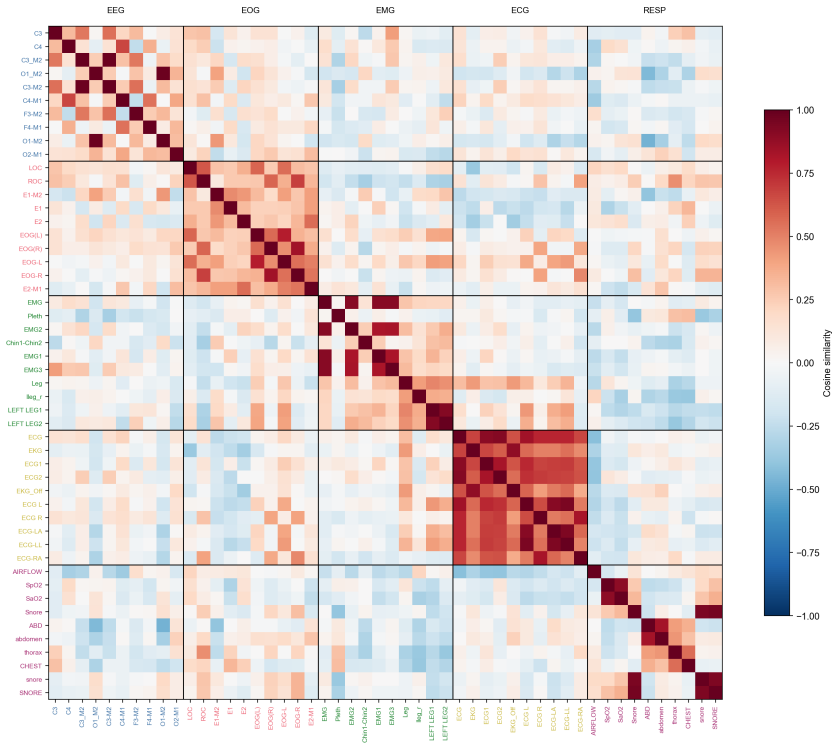

Figure 10: Full pairwise cosine similarity matrix of learned channel-name embeddings across EEG, EOG, EMG, ECG, and respiratory-related channels, extending the subset shown in the main text.

Figure 11 examines how channel-name assignments affect channel aggregation while keeping the waveform inputs fixed. Shuffling the names changes the relative gating weights of several channels, most visibly within EEG, EMG, and the respiratory-related group. For example, the contrast between the EEG channel rows changes, and the relative weighting of the EMG and leg channels is redistributed. EOG and ECG show comparatively small visual changes in this window, indicating that sensitivity to name reassignment varies across modality groups. The differences primarily concern channel-wise weighting, while the weights remain relatively stable over time within the displayed segment.

Together, these visualizations connect the structure of the channel-name embedding space with the model's use of channel metadata during aggregation. The shuffling comparison demonstrates that channel weighting depends on the assigned names for the illustrated input. It does not, by itself, establish the contribution of this dependence to downstream performance; the gating weights characterize model behavior rather than physiological channel importance.

A.10 Analysis of Pretrained Representations

Figure 12 visualizes the representations extracted from three pretrained models with their parameters frozen. SleepMaMi and SleepFM exhibit substantial overlap among sleep stages, although some regions show stage-specific concentrations. In SleepMaMi, Wake epochs extend toward the right of the main group, while the remaining stages are largely intermingled. SleepFM produces a more elongated distribution, with local concentrations of Wake and N3 but considerable overlap throughout the projection.

SemPSG exhibits a more visible stage-related organization: Wake epochs concentrate toward the upper region, N3 epochs are more prevalent in the lower region, and N2 occupies much of the intervening space. N1 and REM remain interspersed with other stages, indicating that the representation does not yield fully separated stage clusters. This pattern suggests that SemPSG captures sleep-stage structure before downstream task-specific training, while retaining overlap between stages.

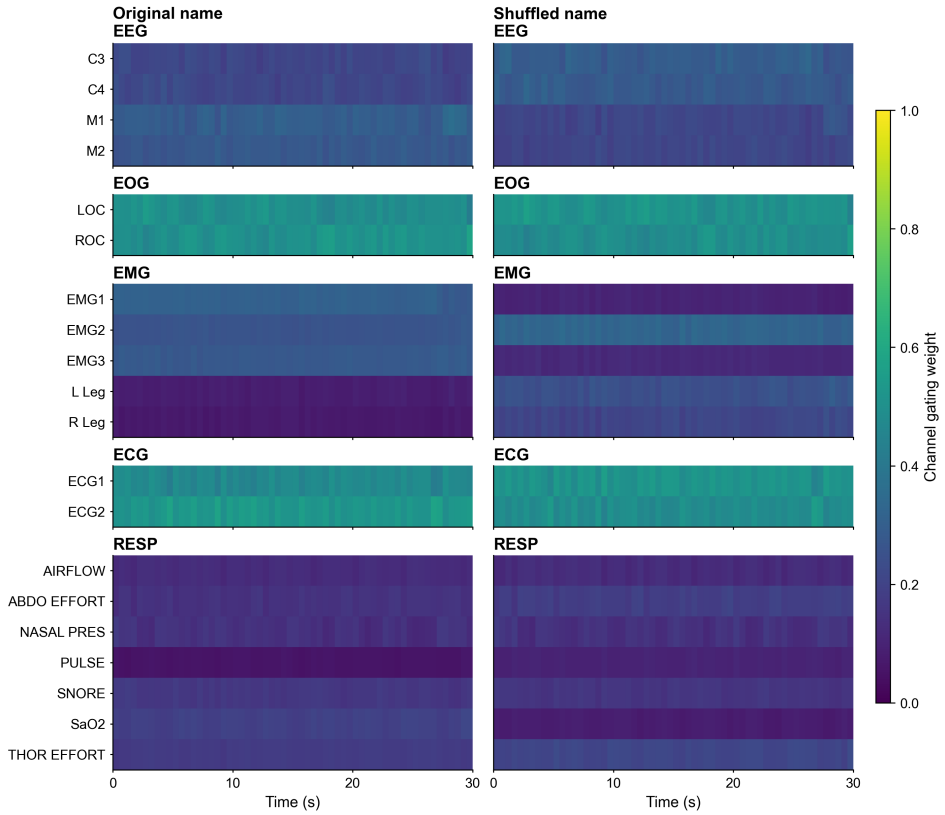

Figure 11: Learned channel-gating weights across all five modality groups over a 30-s window under the original (left) and shuffled (right) channel-name assignments, extending the visualization in the main text.

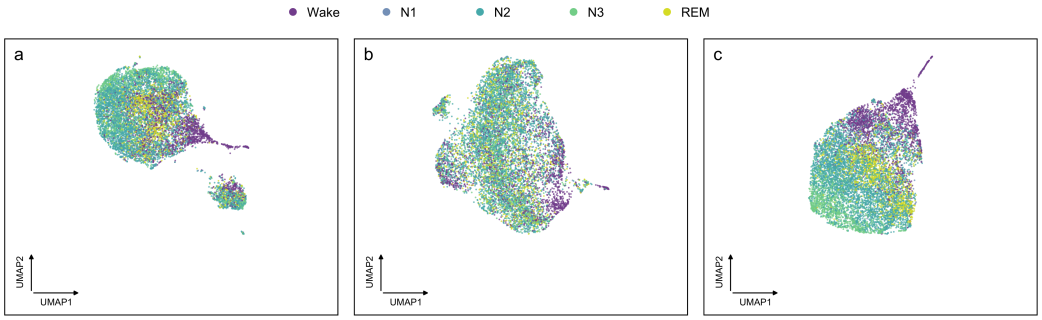

Figure 12: UMAP visualizations of frozen pretrained representations from (a) SleepMaMi, (b) SleepFM, and (c) SemPSG on SHHS-1. Each point represents a sleep epoch and is colored by its ground-truth sleep stage: Wake, N1, N2, N3, or REM. The projections provide a qualitative view of how sleep-stage information is organized in the pretrained representation spaces.