SLEEPXA: COMPOSING PRETRAINED NEURAL AND CARDIAC FOUNDATION MODELS FOR SLEEP REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Physiological signals recorded during sleep provide synchronized views of neural, cardiac, respiratory, and other bodily systems, while preserving modality-specific information. Existing EEG and ECG foundation models already capture rich but distinct physiological structure, motivating an alternative to training a multimodal sleep model from scratch: composing existing pretrained models. We introduce SleepXA, a framework that uses synchronized polysomnography (PSG) to compose independently pretrained neural and cardiac foundation models. SleepXA combines pretrained EEG and ECG encoders with lightweight encoders for additional PSG modalities, and learns shared and modality-specific representations through sleep-guided alignment. We pretrain SleepXA on 9,230 overnight PSG recordings from three large cohorts and evaluate it on 27 downstream tasks spanning sleep representation, clinical and disease prediction, and daytime physiological transfer. SleepXA ranks among the top two methods on 22 of 27 tasks. Controlled experiments further show consistent benefits across sensor configurations and EEG–ECG backbone choices, while pretrained initialization outperforms architecture-matched training from scratch and SleepXA surpasses alternative multimodal composition strategies. These results suggest that synchronized multimodal physiology can serve as a natural interface for composing existing physiological foundation models, yielding reusable representations that extend beyond sleep to clinical and daytime physiological settings.

1 INTRODUCTION

Sleep is a fundamental physiological process supporting cognition, metabolism, cardiovascular function, and mental well-being (Consortium et al., 2026). Clinical sleep assessment commonly relies on Polysomnography (PSG), which synchronously records multiple physiological signals, including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), respiratory signals, and oxygen saturation (SpO_2) (Berry et al., 2017). Together, these modalities provide complementary yet coupled views of neural, cardiac, respiratory, and other physiological processes during sleep, making PSG a natural substrate for learning representations relevant to both sleep analysis and broader health assessment.

Recent sleep foundation models (FMs) have demonstrated the value of large-scale multimodal PSG pretraining for transferable sleep representations (Thapa et al., 2026; Huang et al., 2026; Yuan et al., 2026; Shuai et al., 2026; Hou et al., 2026). Most existing approaches, however, treat heterogeneous PSG modalities largely as parallel inputs for joint encoding or fusion, and they are mostly trained from scratch. However, pretraining such a sleep foundation model requires the availability of well-curated and paired multimodal signals, at a scale and diversity that are prohibitively difficult to obtain, raising concerns about the representational breadth of the resulting models. Moreover, these sleep FMs are primarily optimized for shared sleep related information across signals, often without proper mechanisms to optimize modality-specific information.

On the other hand, large-scale pretraining has advanced general physiological representation learning (Gu et al., 2025), and has resulted in a series of state-of-the-art and open-sourced foundation models on the specific modalities belonging to PSG, such as cardiac (Li et al., 2025; Gu et al., 2026) and neural (Jiang et al., 2024; Wang et al., 2025) signals.

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.

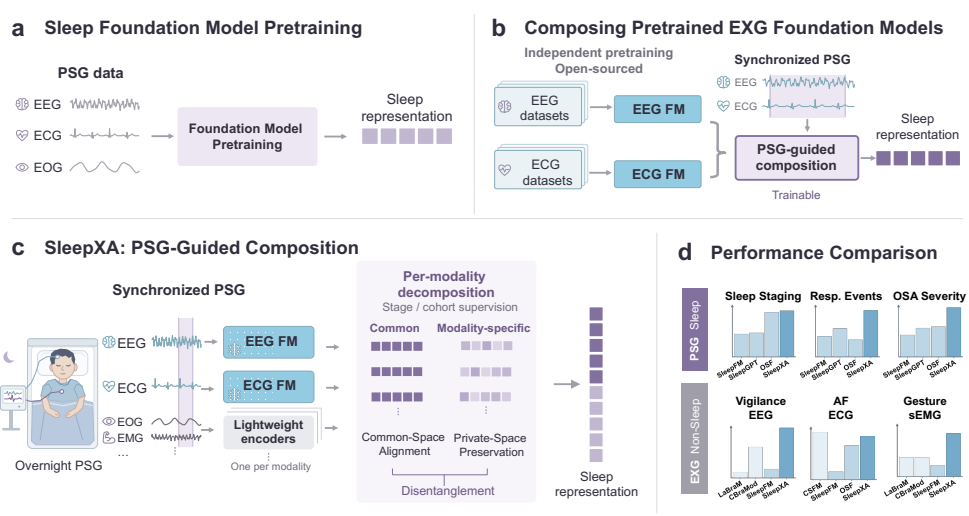


Figure 1: **Overview of SleepXA. a.** Conventional sleep FMs are pretrained on polysomnography (PSG) from scratch. **b.** Independently pretrained, open-sourced EEG and ECG FMs can be composed for sleep representation learning. **c.** SleepXA aligns common representations, preserves modality-specific information, and reduces redundancy between modalities. **d.** Representative evaluation spans sleep and non-sleep tasks (normalized by panel, higher is better; details in Section 4.2).

This motivates a complementary alternative to sleep representation learning: *Can we reuse and compose existing pretrained modality-specific FMs to extract good representations of sleep physiology?*

However, such composition is non-trivial for the following three reasons.

(i) First, existing neural and cardiac FMs are pretrained independently, and mostly on non-sleep datasets. They may produce representations that are not naturally aligned due to signal semantics, temporal scales, training objectives, etc., thus making direct concatenation or fusion suboptimal.

(ii) Second, different modalities may contain shared yet complementary information. Naïvely suppressing them into a common modality-invariant representation space may remove useful modality-specific information that is not captured by other modalities.

(iii) Third, the available modalities during sleep monitoring may vary substantially across settings, from full PSG to reduced clinical montages and wearable devices, making it challenging to learn representations that remain robust under partial modality availability.

To address these challenges, we present **SleepXA**, a compositional framework that integrates independently pretrained modality-specific foundation models for sleep representation learning. SleepXA combines pretrained EEG and ECG FMs, with lightweight encoders for additional PSG modalities, and learns to separate shared *common* representation from modality-*specific* information, with three fundamental modules (common space alignment, modality-specific preservation, and common-specific disentanglement) as detailed in the Methods Section.

We study this compositional paradigm at scale by pretraining SleepXA on 9,230 overnight PSG recordings from three large cohorts and evaluating it on 27 downstream tasks across 15 datasets, spanning general sleep representation, clinical and longitudinal prediction, and cross-context physiological transfer beyond sleep. SleepXA ranks first on 17 tasks and within the top two on 22.

In particular, extensive experiments demonstrate that SleepXA is not merely a "better sleep model", but also preserves broader modality-specific physiological knowledge inherited from the original FMs, while extending their capabilities to respiratory, cardiac, longitudinal, and out-of-sleep tasks. We further examine how pretrained initialization, EEG and ECG backbone choice, multimodal integration, sensor availability, and data-efficient adaptation shape these capabilities.

2 Related Work

Sleep Foundation Models. Large-scale PSG pretraining has become an effective paradigm for transferable sleep representation learning. Existing approaches explore complementary directions, including cross-modal contrastive learning in SleepFM (Thapa et al., 2026) and sleep2vec (Yuan

et al., 2026), time-frequency and multi-objective pretraining in SleepGPT (Huang et al., 2026), robust pretraining and scaling across channel configurations in OSF (Shuai et al., 2026), and physiologically structured CNS-ANS modeling in Omni-Sleep (Hou et al., 2026). These studies primarily learn multimodal physiological representations with a dedicated sleep-pretraining pipeline. Differently, SleepXA investigates whether transferable physiology knowledge already embedded in independently pretrained foundation models can be extracted and composed to facilitate sleep representation learning.

Physiological Foundation Models. Large-scale foundation models have also been developed for individual physiological domains. For neural signals, LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), BrainOmni (Xiao et al., 2025), and REVE (El Ouahidi et al., 2025) learn transferable representations from EEG or EEG/MEG recordings. For cardiac signals, ECGFounder (Li et al., 2025) and CSFM (Gu et al., 2026) learn reusable representations from large-scale ECG or ECG/PPG corpora. These models provide rich modality-specific physiology knowledge, but are pretrained independently and remain largely confined to their original physiological domains. Rather than simplistic late fusion to leverage these FMs in sleep PSG, SleepXA treats these as composable blocks and introduces a series of modules to align, preserve, and merge their representations for sleep representation learning.

Multimodal Fusion and Alignment. Multimodal representations can be connected through direct feature fusion, cross-attention, or contrastive alignment (Tsai et al., 2019; Radford et al., 2021). Recent physiological studies further demonstrate cross-modal knowledge transfer: AnyPPG uses a pretrained ECG encoder to guide PPG representation learning (Nie et al., 2026), while MindMix aligns EEG with pretrained acoustic representations (Liu et al., 2026). Common-specific factorization has also been explored to separate shared and modality-dependent information in multimodal learning (Qian et al., 2026; Meng et al., 2026). However, it remains unclear how multiple modality-specific FMs can be aligned into a shared and reusable representation. This is particularly challenging in sleep PSG, where modality configurations vary substantially across datasets/deployment settings.

3 METHOD

3.1 MOTIVATION AND OVERVIEW

SleepXA addresses the problem of composing physiological foundation models that were pretrained independently on different modalities and datasets. In our implementation, EEG and ECG are encoded by pretrained foundation models, while the remaining PSG modalities are represented using lightweight modality-specific encoders.

Following standard sleep-scoring conventions, a PSG recording can be segmented into consecutive 30-s epochs (Berry et al., 2017). For epoch i and modality m, we define $\mathbf{h}_i^m = E_m(\mathbf{x}_i^m)$, where $\mathbf{x}_i^m$ denotes the 30-s input of modality m.

For simplicity but without loss of generality, we use E_m to denote the modality encoder in all cases: for EEG and ECG, E_m is initialized from a pretrained FM (with the first several layers frozen, details in Appendix A.2), while for the remaining PSG modalities, E_m denotes a lightweight encoder trained during SleepXA pretraining. Meanwhile, we denote their corresponding sleep-stage label as y_i and cohort (or dataset) identity as d_i.

Directly combining these representations $\mathbf{h}_i^m$ is suboptimal for two reasons. First, independently pretrained encoders produce heterogeneous representation spaces that are not naturally aligned. Second, forcing all modalities into a single shared space may suppress physiological information that may be specific to individual modalities.

SleepXA therefore decomposes each representation into

$$\mathbf{c}_i^m = F_{\text{com}}^m(\mathbf{h}_i^m), \qquad \mathbf{s}_i^m = F_{\text{spec}}^m(\mathbf{h}_i^m), \tag{1}$$

where $\mathbf{h}_i^m$, $\mathbf{c}_i^m$, and $\mathbf{s}_i^m$ denote its encoded common and specific representations, from projector F_{com}^m and F_{spec}^m respectively.

Below, we introduce three key modules, common-space alignment, modality-specific preservation, and common-specific disentanglement, as shown in Figure 2, to optimize the learning of $\mathbf{c}_i^m$ and $\mathbf{s}_i^m$.

3.2 COMMON-SPACE ALIGNMENT

The first challenge is to align representations from heterogeneous modality encoders. Because PSG modalities are synchronously recorded, signals within the same 30-s epoch provide different phys-

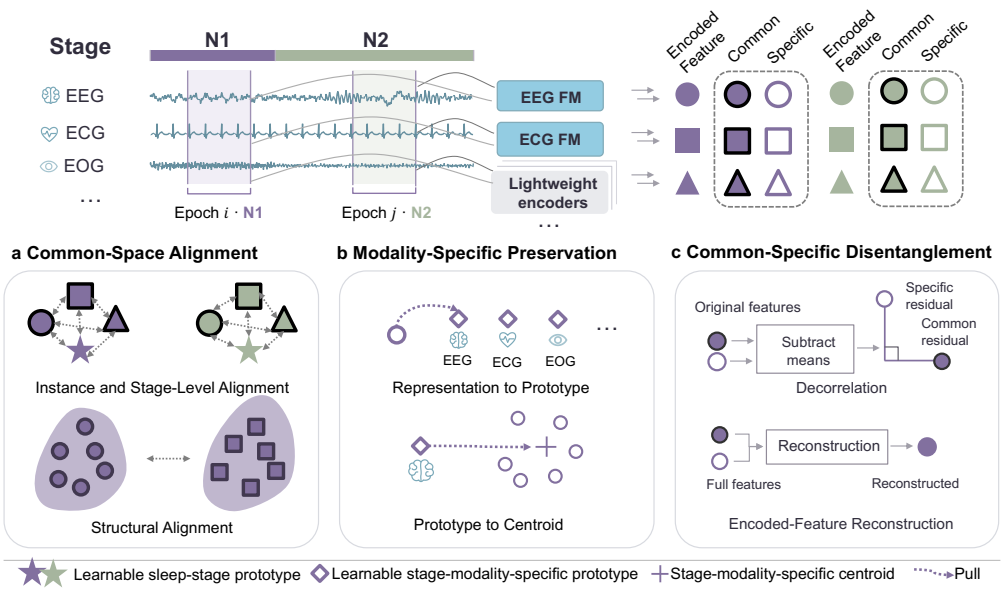


Figure 2: **Sleep-guided cross-modality common-specific learning in SleepXA. a.** Common-space alignment combines instance- and stage-level alignment with distribution structural alignment within each cohort. **b.** Specific preservation matches each feature to its corresponding same-stage-modality prototypes while matching prototypes to their corresponding centroids. **c.** Disentanglement reduces redundancy between the two components while retaining complementary physiology.

iological views of the same underlying state. We use this correspondence to organize the common representations at complementary levels, while discouraging them from encoding cohort identity.

Instance and sleep-stage alignment. We first align common representations $\mathbf{c}_i^m$ at two hierarchical levels, as indicated in Equation 2. At the instance level, synchronized modality pairs are pulled together in a shared representation space. At the semantic level, representations from the same sleep stage y_i are forced to be aligned to corresponding modality-invariant stage prototypes $\mathbf{p}_{y_i}$, which are learnable, similar to Cui et al. (2021). Both objectives are implemented using contrastive learning, as detailed in Appendix A.6.

$$\mathcal{L}_{\text{sem}} = \underbrace{\mathcal{L}_{\text{inst}}(\mathbf{c}_i^m)}_{\text{instance level}} + \alpha_{\text{stage}} \underbrace{\mathcal{L}_{\text{stage}}(\mathbf{c}_i^m, \mathbf{p}_{y_i})}_{\text{sleep-stage level}}. \tag{2}$$

Structural alignment. The above semantic alignment introduces alignment at both individual and stage level. However, at the global distribution level, it is also expected that for the same sleep stage, the distribution of different modalities, such as $\{\mathbf{c}_i^m : m = \text{EEG}, y_i = N_1\}$ and $\{\mathbf{c}_i^m : m = \text{ECG}, y_i = N_1\}$, are expected to exhibit similar feature distributions. This is analogous to previous research pursuing distribution alignment across semantically similar categories (Wang et al., 2023; Sun & Saenko, 2016). To realize this, we perform conditional distribution alignment across modalities within each cohort-stage group.

$$\mathcal{L}_{\text{struc}} = \mathbb{E}_{d,y} \sum_{\substack{\{m,n\} \subseteq \mathcal{M} \\ m \neq n}} D_{\text{struct}} \left(\{\mathbf{c}_i^m\}_{d,y}, \{\mathbf{c}_i^n\}_{d,y} \right). \tag{3}$$

where d and y denote cohort and sleep stage, respectively, and D_{struct} measures the discrepancy between the conditional representation structures of two modalities. Please refer to Appendix A.6 for implementation details.

The resulting common-space objective is $\mathcal{L}_{\text{com}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{struc}}$.

3.3 Modality-Specific Preservation

While the common branch F_{com}^m promotes cross-modal consistency, the specific branch F_{spec}^m is supposed to retain physiological information unique to each individual modality. We achieve this by encouraging modality-specific representations to remain distinguishable within each sleep stage. This

is different from naively enforcing global modality discriminability, which may conflate modality-specific variation with differences arising from sleep stage.

To implement this, we introduce learnable prototypes $\mathbf{q}_y^m$ for each modality m and sleep stage y. For a specific representation $\mathbf{s}_i^m$ with stage label y_i, we encourage it to be closer to the prototype of its source modality than to those of other modalities under the same sleep stage. We further anchor each prototype to the corresponding modality distribution using its stage-conditioned centroid. The resulting objective is

$$\mathcal{L}_{\text{spec}} = \underbrace{\mathcal{L}_{\text{proto}}^{\rightarrow}\left(\{\mathbf{s}_i^m, y_i\}, \{\mathbf{q}_y^m\}\right)}_{\text{representation-to-prototype}} + \underbrace{\mathcal{L}_{\text{proto}}^{\leftarrow}\left(\{\mathbf{q}_y^m\}, \{\bar{\mathbf{s}}_{d,y}^m\}\right)}_{\text{prototype-to-centroid}}. \tag{4}$$

We implement both terms using contrastive objectives. Here, $\bar{\mathbf{s}}_{d,y}^m$ denotes the centroid of modality m within cohort d and sleep stage y. Because not all modalities are available in every cohort, we construct these centroids separately within each cohort-stage condition, using only the modalities observed in that cohort. The first term preserves modality identity within each sleep stage, while the second keeps the learned prototypes representative of their corresponding modality distributions. Detailed formulations are provided in Appendix A.6.

3.4 Common-Specific Disentanglement

Although the common and specific branches are optimized for different purposes, they may still encode redundant information. For example, both $\mathbf{c}_i^m$ and $\mathbf{s}_i^m$ may vary with sleep stage, such as between N_2 and REM, and this is not necessarily undesirable. The problem is when they continue to encode the same variation even within the same sleep stage and cohort. We therefore avoid enforcing global orthogonality, which could remove meaningful stage- or cohort-related information shared by both branches.

We therefore measure redundancy relative to each cohort-stage condition (refer to Appendix A.4.8 for an ablation of this constraint). Specifically, we remove the variation explained by the corresponding context and define

$$\Delta\mathbf{c}_i^m = \mathbf{c}_i^m - \bar{\mathbf{c}}_{d,y}^m, \qquad \Delta\mathbf{s}_i^m = \mathbf{s}_i^m - \bar{\mathbf{s}}_{d,y}^m, \tag{5}$$

We then decorrelate only the remaining residual variation:

$$\mathcal{L}_{\text{dis}} = \underbrace{\mathbb{E}_{i,m}\left[\cos^2\left(\Delta\mathbf{c}_i^m, \Delta\mathbf{s}_i^m\right)\right]}_{\text{conditional decorrelation}} + \lambda_{\text{rec}} \underbrace{\mathbb{E}_{i,m}\left[\left\|\mathbf{h}_i^m - R_m([\mathbf{c}_i^m; \mathbf{s}_i^m])\right\|_2^2\right]}_{\text{representation reconstruction}}. \tag{6}$$

Here, $\cos(\cdot,\cdot)$ denotes cosine similarity between the two residual representations, while the latter reconstruction, implemented by a lightweight MLP R_m, prevents the common-specific decomposition from discarding information contained in the original encoded representation. Together, these objectives reduce redundancy between the common and specific branches without removing meaningful variation associated with sleep stage or cohort.

3.5 Overall Objective and Downstream Representation

Overall training objective. SleepXA combines three complementary modules introduced above. The overall training objective is realized as below,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{com}} + \lambda_{\text{spec}}\mathcal{L}_{\text{spec}} + \lambda_{\text{dis}}\mathcal{L}_{\text{dis}}. \tag{7}$$

Further details of hyperparameter settings are available in Appendix A.6.

Downstream representation. For downstream tasks, the common and modality-specific representations from the available modalities are combined to form the task representation, with the representation pathway adapted according to the downstream evaluation setting. For linear probing, we average the available common features and concatenate the result with the modality-specific features. For fine-tuning, task-adaptive attention separately fuses the common and specific features across available modalities.

4 Experiments and Results

4.1 Experimental Setup

Pre-training and downstream data. We perform multimodal alignment pre-training on 9,230 overnight PSG recordings from SHHS, MESA, and CFS (Quan et al., 1997; Chen et al., 2015; Red-

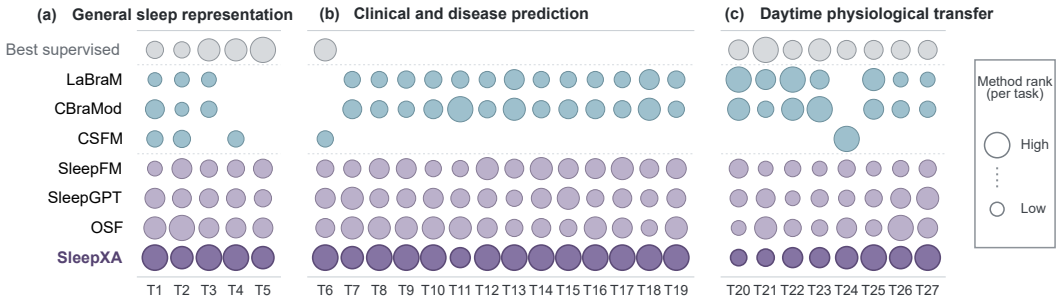


Figure 3: **Cross-task performance landscape.** Results span general sleep representation, clinical and longitudinal prediction, and daytime physiological transfer. Larger bubbles indicate better task-wise ranks. Gray, blue, lavender, and dark-purple denote supervised references, physiological foundation models, sleep foundation models, and SleepXA, respectively. Task definitions for T1–T27 are provided in Appendix A.1.

line et al., 1995). Downstream evaluation spans 27 tasks across held-out Test SHHS1/2 and seven external sleep cohorts (APPLES, WSC, SleepEDF20, HMC, ISRUC-1, MASS-SS2, and CAP) (Quan et al., 2011; Young et al., 2009; Kemp et al., 2000; Alvarez-Estevez & Rijsman, 2021; Khalighi et al., 2016; O'reilly et al., 2014; Terzano et al., 2001), together with seven non-sleep physiological datasets covering EEG, ECG, EOG, and EMG (Mumtaz et al., 2017; Miltiadous et al., 2023; Zyma et al., 2019; Tangermann et al., 2012; Clifford et al., 2017; Zúñiga-Castillo et al., 2025; Zheng & Lu, 2017). We evaluate SleepXA across these 15 datasets on 27 downstream tasks in three groups: *general sleep representation*, *clinical and longitudinal prediction*, and *sleep-to-non-sleep physiological transfer*. Complete dataset statistics, split construction, leakage control, and task definitions are provided in Appendix A.1.

Evaluation and baselines. Foundation models are evaluated under linear probing (LP) and full fine-tuning (FT), using cross-subject splits whenever participant identifiers are available. We compare SleepXA against three groups of baselines: task-specific supervised models; physiological foundation models, including LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025) for EEG and CSFM (Gu et al., 2026) for ECG; and sleep foundation models, including SleepFM (Thapa et al., 2026), SleepGPT (Huang et al., 2026), and OSF (Shuai et al., 2026). Unless otherwise specified, SleepXA uses LaBraM and CSFM as its pretrained EEG and ECG encoders, respectively. All methods use the same splits and metrics, with supported input modalities. Detailed optimization settings and task-specific evaluation protocols are provided in Appendices A.2 and A.3.

4.2 Performance across Downstream Tasks

Figure 3 summarizes performance across the 27-task benchmark. To emphasize pretrained representation quality, we use an LP-first ranking protocol, with full details provided in Appendix A.3. **SleepXA ranks first on 17 of 27 tasks and within the top two on 22**, demonstrating competitive performance across all three evaluation groups. The advantage is particularly consistent for general sleep representation and clinical prediction, while meaningful transfer is also retained on cross-context physiological tasks involving EEG, ECG, and multimodal inputs. Together, these results highlight the strength and broad transferability of SleepXA representations across diverse downstream settings.

General sleep representation. We first evaluate whether SleepXA captures general sleep-related physiology at multiple prediction granularity levels. At the subject level, SleepXA achieves the strongest age-estimation performance under both LP and FT, while achieving the best sex-classification performance after FT (Fig. 4, Appendix A.4). We next evaluate epoch-level sleep staging as a direct test of cross-cohort transfer. On Test SHHS1/2 and the two external cohorts APPLES and WSC, SleepXA achieves the best performance among the evaluated foundation models under LP, reaching accuracy scores of 84.07%, 77.63%, and 85.16%, respectively (Table 1). FT further improves performance on Test SHHS1/2 and WSC while maintaining comparable performance on APPLES, indicating that the representation is both directly transferable and adaptable to downstream supervision. Further sleep staging evaluations are provided in Appendix A.4.1. Finally, SleepXA also shows strong transfer to fine-grained respiratory-event (Table A.6) and spindle detection (Appendix A.4.1), suggesting that its representation captures broader sleep physiology beyond sleep-stage classification.

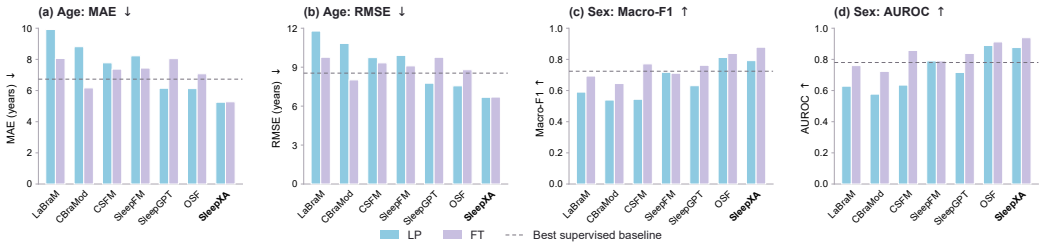


Figure 4: **Age and sex prediction.** Performance of foundation models under LP and FT. Age prediction is evaluated by MAE and RMSE (↓), while sex prediction is evaluated by Macro-F1 and AUROC (↑). Dashed lines denote the best supervised baselines.

Table 1: **Sleep staging on three representative cohorts.** Acc and MF1 are percentages; all metrics are higher-is-better. **Bold** and underlined values indicate the **best** and second-best results.

Model	Channels	Test SHHS1/2			APPLES			WSC		
		Acc.	MF1	κ	Acc.	MF1	κ	Acc.	MF1	κ
Task-specific supervised models										
DeepSleepNet (Supratak et al., 2017)	E	80.10	72.66	0.728	75.12	67.87	0.658	81.89	74.70	0.732
SleePyCo (Lee et al., 2024)	E	82.12	74.62	0.753	76.57	65.92	0.667	83.23	75.78	0.748
U-Time (Perslev et al., 2019)	E	79.10	70.58	0.712	76.26	67.80	0.665	83.11	75.26	0.745
AttnSleep (Eldele et al., 2021)	E	74.98	63.59	0.650	73.42	58.00	0.616	80.48	72.82	0.712
XSleepNet (Phan et al., 2021)	E/O/M	84.41	77.38	0.783	76.81	70.20	0.678	84.70	77.47	0.769
ASTGSleep (Chen et al., 2025)	Full	83.81	77.08	0.777	76.88	70.25	0.690	**85.97**	**78.25**	**0.782**
XSleepFusion (Hu et al., 2025c)	Full	82.44	73.95	0.756	75.99	68.09	0.665	83.01	75.98	0.747
SleepXViT (Lee et al., 2025)	Full	83.89	71.54	0.767	76.41	65.12	0.662	83.40	73.49	0.741
Foundation models: linear probing (LP)										
LaBraM (Jiang et al., 2024)	E	74.65	61.57	0.639	67.40	59.85	0.556	72.35	60.40	0.577
CBraMod (Wang et al., 2025)	E	76.52	64.44	0.667	69.57	63.62	0.588	73.18	62.01	0.594
SleepFM (Thapa et al., 2026)	Full	75.22	68.05	0.667	70.89	62.70	0.600	77.40	67.97	0.658
SleepGPT (Huang et al., 2026)	Full	77.91	69.75	0.697	72.24	63.92	0.615	79.51	68.35	0.681
OSF (Shuai et al., 2026)	Full	83.14	74.42	0.764	75.50	67.30	0.660	83.52	75.61	0.794
SleepXA (Ours)	Full	84.07	76.24	0.780	77.63	**71.13**	**0.692**	85.16	76.09	0.765
Foundation models: full fine-tuning (FT)										
LaBraM (Jiang et al., 2024)	E	83.43	72.97	0.765	75.56	67.57	0.660	79.73	70.04	0.692
CBraMod (Wang et al., 2025)	E	83.91	74.76	0.774	76.60	68.05	0.671	81.36	71.03	0.712
SleepFM (Thapa et al., 2026)	Full	74.51	67.73	0.658	72.09	62.95	0.610	68.73	63.79	0.562
SleepGPT (Huang et al., 2026)	Full	82.13	74.37	0.753	73.06	66.51	0.632	82.46	71.54	0.722
OSF (Shuai et al., 2026)	Full	84.09	75.58	0.782	76.72	68.20	0.670	83.18	74.97	0.742
SleepXA (Ours)	Full	**87.04**	**78.71**	**0.815**	**77.84**	70.51	0.690	85.95	77.61	0.780

E: EEG; E/O/M: EEG+EOG+EMG; Full: full experimental PSG configuration.

Clinical and longitudinal prediction. We next assess whether SleepXA demonstrates transferrable capability to predict clinically meaningful subject-level outcomes. For sleep-disordered breathing, we evaluate both continuous AHI estimation and four-level OSA severity classification on Test SHHS1/2 and WSC. As shown in Table A.8, SleepXA shows strong cross-cohort transfer already under LP and further improves with FT, reaching MAEs of 4.200 on Test SHHS1/2 and 3.120 on WSC. These results indicate that the learned representation supports both continuous apnea-burden estimation and clinically defined OSA severity assessment. More broadly, across 12 frozen-representation clinical tasks, including six current disease diagnoses, two longitudinal sleep outcomes, and four five-year cardiovascular outcomes, SleepXA achieves the highest mean AUROC among the evaluated foundation models on 11 tasks, including five of six current diagnoses and all six longitudinal outcomes (Tables A.9 and A.10). More details are in Appendices A.4.2 and A.4.3.

Daytime physiological transfer. We further examine whether representations learned from nocturnal PSG transfer beyond the sleep setting. As shown in Fig. 5, these results highlight an underexplored property of PSG-guided physiology FM composition: *nocturnal PSG appears to facilitate learning whole-body physiological structure, yielding knowledge that remains reusable beyond sleep itself.* (i) First, SleepXA supports decoding brain states that reflect health conditions and physiological function beyond sleep. It transfers to multiple tasks, such as Major depressive disorder (MDD) and Alzheimer's disease (AD) (Fig. 5 a-b). (ii) Second, SleepXA remains competitive even when the downstream task is semantically distant from sleep. On EEG based mental arithmetic recogni-

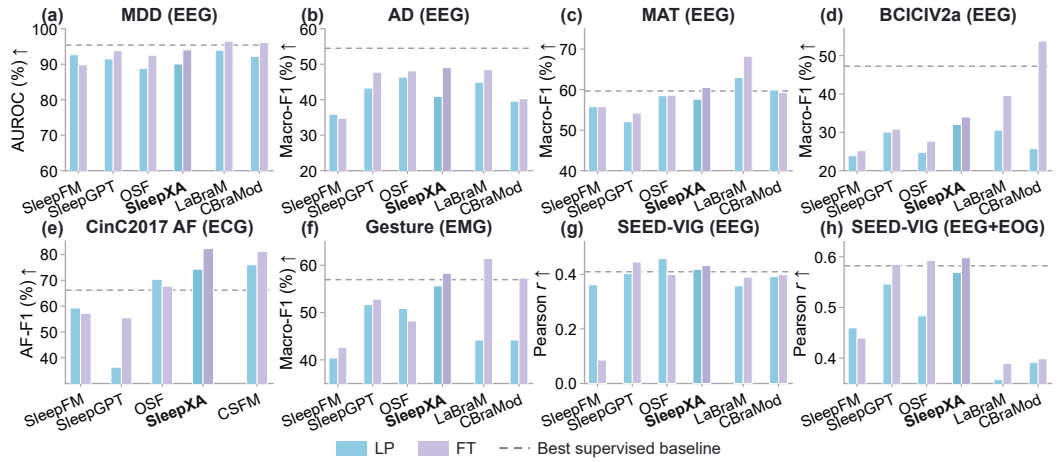


Figure 5: **Cross-context physiological transfer beyond nocturnal sleep.** Performance across eight downstream settings involving EEG, ECG, EOG, and surface EMG. Foundation models are evaluated under LP and FT, and dashed lines denote the best supervised baseline for each setting.

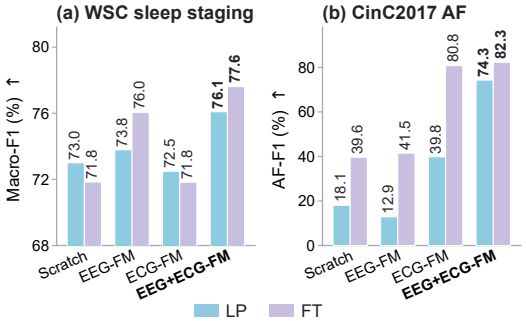


Figure 6: **Effect of physiological foundation-model initialization.** Comparison on (a) WSC sleep staging and (b) CinC2017 AF recognition.

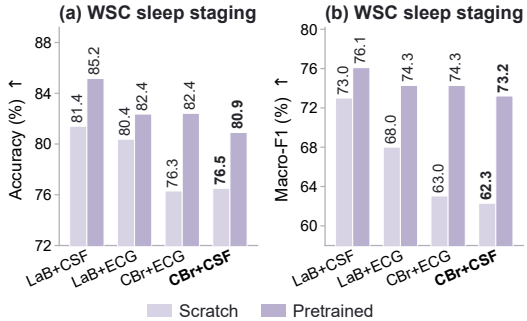


Figure 7: **Scratch vs pretrained initialization across FM combinations.** LaB:LaBraM, CBr:CBraMod, CSF:CSFM, ECG:ECGFounder.

tion (MAT) and motor-imagery decoding, as well as ECG-based atrial-fibrillation (AF) recognition, SleepXA ranks consistently among the top two sleep FMs, demonstrating the effective use of the pretrained modality-specific encoders beyond sleep-related semantics (Fig. 5 c-e). (iii) Third, SleepXA supports both cross-modality transfer and multimodal integration. It performs competitively on EMG-based gesture recognition and EEG-only vigilance estimation, and achieves the best linear-probing performance when EEG and EOG are jointly used, supporting the utility of the learned representations when modalities are used individually or combined (Fig. 5 f-h). These results support the value of PSG-based multimodal alignment for organizing complementary modality-specific information. See Appendix A.4.4 for further discussion.

4.3 Analysis of Model Composition

Effect of physiological foundation-model initialization. To assess the contribution of pretrained physiological knowledge during composition, we investigate different EEG/ECG encoder initialization strategies while keeping all other settings fixed. As shown in Fig. 6, joint EEG+ECG foundation-model initialization performs best on both WSC sleep staging (full PSG) and CinC2017 AF recognition (ECG), supporting the benefit of composing complementary pretrained neural and cardiac knowledge through synchronized PSG alignment. Complete results are provided in Appendix A.4.5.

Robustness to physiological foundation-model backbones. To test the FM-agnostic of our composition, we test all 2×2 combinations of two EEG foundation models (LaBraM and CBraMod) and two ECG foundation models (CSFM and ECGFounder (Li et al., 2025)). Across all four pairs, pretrained initialization outperforms matched scratch training by 1.98-6.12 points in accuracy on WSC sleep staging (Fig. 7; Appendix A.4.6), showing that the gains are not specific to the default LaBraM+CSFM configuration.

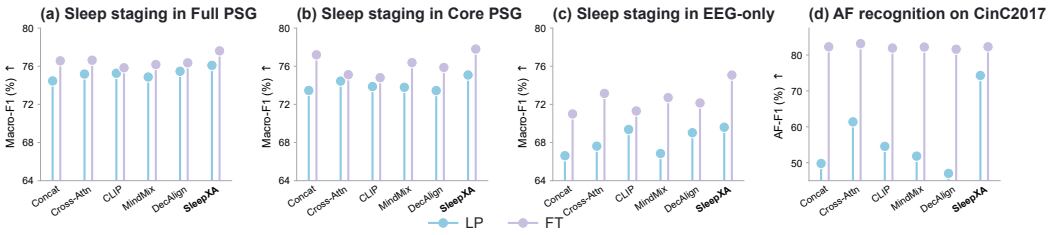


Figure 8: **Comparison of multimodal integration strategies.** WSC sleep staging under (a) Full, (b) Core, and (c) EEG-only inputs, and (d) CinC2017 AF recognition under LP and FT.

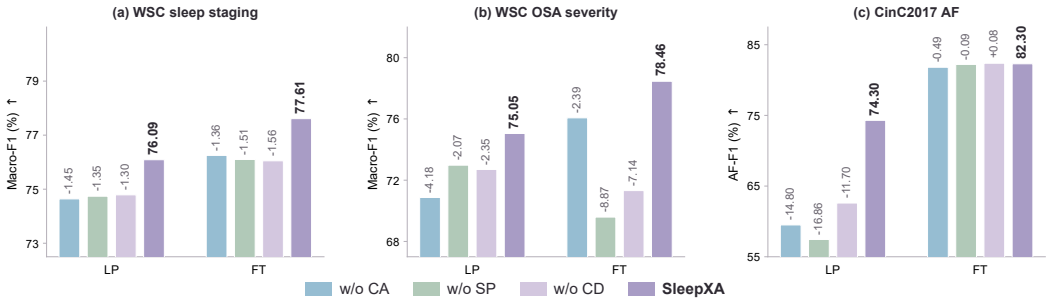


Figure 9: **Ablation of the major alignment objectives.** Effects of removing CA, SP, or CD on (a) WSC sleep staging, (b) WSC OSA severity classification, and (c) CinC2017 AF recognition. Annotations report the change relative to the original SleepXA model under the same LP/FT setting.

Comparison with multimodal integration strategies. We compare SleepXA with representative fusion and alignment alternatives in four configurations: WSC sleep staging with full-PSG, core-PSG (EEG+EOG+EMG), or EEG-only inputs, and ECG AF recognition on CinC2017. SleepXA performs best across all WSC settings and shows a clear LP advantage on CinC2017, while FT largely narrows the gap (Fig. 8). These results suggest that SleepXA's alignment improves the quality of frozen representations by preserving and organizing complementary pretrained modality-specific knowledge. More details are provided in Appendix A.4.7.

Ablation of major alignment objectives. We independently remove condition-aware alignment (CA), stage-conditioned specific preservation (SP), and condition-relative disentanglement (CD). As shown in Fig. 9, removing any module degrades WSC sleep staging and OSA severity prediction under both LP and FT. On CinC2017, alignment improves LP but adds little after FT, as FT may unlock AF-relevant knowledge already encoded in the pretrained ECG FM. Together, these results demonstrate each module's complementary contribution, with their larger effect under LP indicating that they primarily improve pretrained representation quality (Appendix A.4.8).

Robustness to modality and supervision constraints. Finally, we evaluate robustness to reduced sensor availability and limited downstream supervision. SleepXA remains effective with reduced PSG inputs, including EEG-only input. We further compare single-domain SHHS, matched multi-domain, and full multi-domain pre-training under different downstream labeling ratios. Full multi-domain pre-training most consistently benefits sleep-domain transfer, whereas gains on more distant tasks are more dependent on downstream supervision and domain shift. Details are in Appendix A.5.

5 Conclusion

We introduced SleepXA, a compositional framework that uses synchronized PSG to integrate independently pretrained physiological foundation models. SleepXA combines an asymmetric architecture with conditional common-specific alignment to capture cross-modal common structure while preserving modality-specific physiology. Across 27 downstream tasks spanning sleep representation, clinical prediction, and daytime physiological transfer, SleepXA demonstrates strong transfer across cohorts, sensor configurations, and adaptation settings. Controlled comparisons and ablations further show that the gains arise not only from pretrained physiological knowledge, but also from how heterogeneous pretrained representations are composed. These results support a compositional alternative to training multimodal physiological foundation models from scratch: synchronized PSG can provide a common observational basis for aligning existing physiological foundation models into transferable multisystem representations.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and the preparation of visual materials. Specifically, large language models (LLMs) were used to polish the language of initial manuscript drafts, including improving grammar, clarity, and presentation, without altering the underlying scientific claims or conclusions. We also used generative AI tools to create illustrative visual elements that were subsequently incorporated as components of the manuscript figures, such as the sleep-scene illustration shown in Figure 1.

All AI-assisted text was reviewed and revised by the authors. AI-generated visual elements were also manually inspected, selected, and integrated into the final figures by the authors to ensure that they accurately represented the intended concepts and did not introduce misleading scientific information. The experimental results, quantitative analyses, and scientific conclusions reported in this work were determined and verified by the authors.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study conducts secondary analysis of existing de-identified physiological and polysomnographic datasets obtained from established public or controlled-access research repositories. No new participants were recruited, no new human-subject data were collected, and no intervention or direct interaction with participants was conducted as part of this work. Ethical approval, informed consent, and participant privacy protections for the original data collection were obtained by the institutions responsible for the respective studies. We accessed and used all datasets in accordance with their licenses, data-use agreements, and repository-specific access requirements, including the National Sleep Research Resource (NSRR), PhysioNet, and other official data repositories. No attempt was made to re-identify individual participants.

REPRODUCIBILITY STATEMENT

To support reproducibility, the appendix describes the dataset preprocessing, subject-level data splits, evaluation protocols, implementation details, and key training settings used in our experiments.

REFERENCES

Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one*, 16(8):e0256111, 2021.

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). *Journal of clinical sleep medicine*, 13(5):665–666, 2017.

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.

Xiaoyu Chen, Yiyuan Zhang, Qiangqiang Chen, Ligang Zhou, Hongyu Chen, Huijuan Wu, Yunxia Xu, Kun Chen, Bin Yin, Wei Chen, et al. Astgsleep: Attention-based spatial–temporal graph network for sleep staging. *IEEE Transactions on Instrumentation and Measurement*, 74:1–14, 2025.

Gari D Clifford, Chengyu Liu, Benjamin Moody, Li-wei H Lehman, Ikaro Silva, Qiao Li, Alistair E Johnson, and Roger G Mark. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 computing in cardiology (CinC)*, pp. 1–4. IEEE, 2017.

MULTI Consortium, Cliodhna Kate O'Toole, Zhiyuan Song, Filippos Anagnostakis, Zhijian Yang, Ye Ella Tian, Michael R Duggan, Chunrui Zou, Yue Leng, Yi Cai, et al. Sleep chart of biological ageing clocks in middle and late life. *Nature*, pp. 1–11, 2026.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 695–704. IEEE, 2021.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg-adapting to any setup with large-scale pretraining on 25,000 subjects. *Advances in Neural Information Processing Systems*, 38, 2025.

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.

Xiao Gu, Yuxuan Shu, Jinpei Han, Yuxuan Liu, Zhangdaihong Liu, James Anibal, Veer Sangha, Edward Phillips, Bradley Segal, Hang Yuan, et al. Foundation models for biosignals: A survey. 2025.

Xiao Gu, Wei Tang, Jinpei Han, Veer Sangha, Fenglin Liu, Shreyank N Gowda, Antonio H Ribeiro, Patrick Schwab, Kim Branson, Lei Clifton, et al. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, 8(2):220–233, 2026.

Zhoujie Hou, Song Wang, Kexin Lou, Mo Wang, Chen Wei, and Quanying Liu. Omni-sleep: A sleep foundation model via hierarchical contrastive learning of cns–ans dynamic. *arXiv preprint arXiv:2607.07720*, 2026.

Shuaicong Hu, Wenjie Cai, Tijie Gao, and Mingjie Wang. A hybrid transformer model for obstructive sleep apnea detection based on self-attention mechanism using single-lead ecg. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.

Shuaicong Hu, Jian Liu, Yanan Wang, Cong Fu, Jichu Zhu, Huan Yu, and Cuiwei Yang. Transparent artificial intelligence-enabled interpretable and interactive sleep apnea assessment across flexible monitoring scenarios. *Nature Communications*, 16(1):7548, 2025a.

Shuaicong Hu, Yanan Wang, Jian Liu, Zhaoqiang Cui, Cuiwei Yang, Zhifeng Yao, and Junbo Ge. Ipct-net: Parallel information bottleneck modality fusion network for obstructive sleep apnea diagnosis. *Neural Networks*, 181:106836, 2025b.

Shuaicong Hu, Yanan Wang, Jian Liu, and Cuiwei Yang. Xsleepfusion: A dual-stage information bottleneck fusion framework for interpretable multimodal sleep analysis. *Information Fusion*, 123:103275, 2025c.

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*, volume 2024, 2024.

Lars Kaulen, Justus TC Schwabedal, Jules Schneider, Philipp Ritter, and Stephan Bialonski. Advanced sleep spindle identification with neural networks. *Scientific reports*, 12(1):7686, 2022.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016.

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.

Seongju Lee, Yeonguk Yu, Seunghyeok Back, Hogeon Seo, and Kyoobin Lee. Sleepyco: Automatic sleep scoring with feature pyramid and contrastive learning. *Expert Systems with Applications*, 240:122551, 2024.

Jun Li, Aaron D Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *Nejm ai*, 2(7):AIoa2401033, 2025.

Rui Liu, Zhige Chen, Shu Peng, Wenlong You, Zhi-An Huang, Jibin Wu, and Kay Chen Tan. Mindmix: A multimodal foundation model for auditory perception decoding via deep neural-acoustic alignment. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Chunlei Meng, Jiabin Luo, Pengbin Feng, Zhenglin Yan, Chengyin Hu, Zhongxue Gan, and Chun Ouyang. Mitigating shared-private branch imbalance via dual-branch rebalancing for multimodal sentiment analysis. *arXiv preprint arXiv:2604.25179*, 2026.

Andreas Miltiadous, Katerina D Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G Tsalikakis, Pantelis Angelidis, Markos G Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp eeg recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine eeg. *Data*, 8(6):95, 2023.

Wajid Mumtaz, Likun Xia, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, Muhammad Hussain, and Aamir Saeed Malik. Electroencephalogram (eeg)-based computer-aided technique to diagnose major depressive disorder (mdd). *Biomedical Signal Processing and Control*, 31:108–115, 2017.

Guangkun Nie, Xiaocheng Fang, Gongzheng Tang, Yujie Xiao, Jun Li, Bo Liu, Hongyan Li, and Shenda Hong. Anyppg: An ecg-guided ppg foundation model trained on over 100,000 hours of recordings for holistic health profiling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11717–11727, 2026.

Christian O'reilly, Nadia Gosselin, Julie Carrier, and Tore Nielsen. Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of sleep research*, 23(6):628–635, 2014.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602. IEEE, 2022.

Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *Advances in neural information processing systems*, 32, 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.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, 2026.

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.

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020.

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

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

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, 2017.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

Nicolás I Tapia and Pablo A Estévez. Red: Deep recurrent neural networks for sleep eeg event detection. In *2020 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Mario Giovanni Terzano, Liborio Parrino, Adriano Sherieri, Ronald Chervin, Sudhansu Chokroverty, Christian Guilleminault, Max Hirshkowitz, Mark Mahowald, Harvey Moldofsky, Agostino Rosa, et al. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep. *Sleep medicine*, 2(6):537–553, 2001.

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*, 32(2):752–762, 2026.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, 2025.

Runqi Wang, Hao Zheng, Xiaoyue Duan, Jianzhuang Liu, Yuning Lu, Tian Wang, Songcen Xu, and Baochang Zhang. Few-shot learning with visual distribution calibration and cross-modal distribution alignment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23445–23454. IEEE, 2023.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified eeg and meg signals. *Advances in Neural Information Processing Systems*, 38, 2025.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=c2LZyTyddi.

Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI*, 2(1):e46769, 2023b.

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, and Xuesong Chen. sleep2vec: Unified cross-modal alignment for heterogeneous nocturnal biosignals. *arXiv preprint arXiv:2602.13857*, 2026.

Wei-Long Zheng and Bao-Liang Lu. A multimodal approach to estimating vigilance using eeg and forehead eog. *Journal of neural engineering*, 14(2):026017, 2017.

Cinthya Alejandra Zúñiga-Castillo, Víctor Alejandro Anaya-Mosqueda, Natalia Margarita Rendón-Caballero, Marcos Aviles, José M Álvarez-Alvarado, Roberto Augusto Gómez-Loenzo, and Juvenal Rodríguez-Reséndiz. Surfaceemg datasets for hand gesture recognition under constant and three-level force conditions. *Data*, 10(12):194, 2025.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A Appendix

A.1 Datasets and Evaluation Protocols

We evaluate SleepXA using a broad downstream benchmark spanning general sleep representation, clinical and disease prediction, and daytime physiological transfer. Table A.1 summarizes the dataset-level configuration, including signal modalities, sample sizes, and evaluation protocols, while Table A.2 provides the definitions and prediction targets of the 27 downstream tasks.

For multimodal alignment pre-training, we use three large synchronized polysomnography (PSG) cohorts: the Sleep Heart Health Study (SHHS; Visits 1 and 2) (Quan et al., 1997), the Multi-Ethnic Study of Atherosclerosis (MESA) (Chen et al., 2015), and the Cleveland Family Study (CFS) (Redline et al., 1995). To prevent subject leakage across longitudinal SHHS visits, we first hold out Test SHHS1/2, comprising 1,000 participants with PSG recordings available at both SHHS Visits 1 and 2 (2,000 recordings in total), from all multimodal pre-training. After this exclusion, alignment pre-training uses 9,230 overnight PSG recordings. Test SHHS1/2 is reserved exclusively for downstream evaluation and provides the held-out SHHS population used for demographic prediction, sleep analysis, OSA assessment, and current and longitudinal clinical tasks.

We further evaluate generalization on multiple external sleep cohorts that are never used during alignment pre-training. These include APPLES (Quan et al., 2011), the Wisconsin Sleep Cohort (WSC) (Young et al., 2009), SleepEDF20 (Kemp et al., 2000), HMC (Alvarez-Estevez & Rijsman, 2021), ISRUC-1 (Khalighi et al., 2016), MASS-SS2 (O'reilly et al., 2014), and the CAP Sleep Database (Terzano et al., 2001). Together, these cohorts provide complementary evaluations of sleep staging, respiratory-event detection, OSA severity estimation, sleep-spindle detection, and sleep-disorder classification under different populations, recording configurations, and cohort sizes.

Daytime physiological transfer is evaluated on seven additional physiological-signal datasets, none of which are used during alignment pre-training. These comprise major depressive disorder (MDD) diagnosis from EEG (Mumtaz et al., 2017), dementia classification including Alzheimer's disease (AD) from resting-state EEG using OpenNeuro ds004504 (Miltiadous et al., 2023), mental-arithmetic (MAT) recognition from EEG (Zyma et al., 2019), motor-imagery classification using BCI Competition IV-2a (Tangermann et al., 2012), atrial-fibrillation (AF) recognition from single-lead ECG using the PhysioNet/Computing in Cardiology 2017 Challenge dataset (Clifford et al., 2017), constant-force hand-gesture recognition from surface EMG using DS1 (Zúñiga-Castillo et al., 2025), and vigilance estimation using SEED-VIG (Zheng & Lu, 2017). SEED-VIG is evaluated under both EEG-only and EEG+EOG settings, yielding eight transfer settings across the seven datasets. We use *daytime physiological transfer* as a compact umbrella term for physiological downstream tasks outside nocturnal PSG settings, including awake-state, ambulatory, and task-specific recordings; the term does not imply that every recording, such as those in CinC 2017, was explicitly acquired during daytime hours.

The downstream benchmark is organized into three groups. **General Sleep Representation Evaluation** probes fundamental sleep representations through demographic prediction, sleep staging, apnea/hypopnea detection, and sleep-spindle detection, using both Test SHHS1/2 and external sleep cohorts. **Clinical and Disease Prediction** evaluates OSA severity, sleep-disorder classification, current disease status, and longitudinal prediction of subsequent sleep-related and cardiovascular outcomes. Finally, **Daytime Physiological Transfer** evaluates cross-context transfer from overnight PSG to independent non-sleep physiological tasks across EEG, ECG, EOG, and surface-EMG settings. When participant identifiers are available, we split the data by participant, keeping all recordings and sessions from the same participant in the same split or fold. Larger cohorts use fixed 8:1:1 train/validation/test splits, while smaller cohorts are evaluated using participant-level k-fold cross-validation.

A.2 Implementation Details

Pre-training modalities and signal preprocessing. SleepXA performs multimodal alignment over seven physiological modalities commonly recorded in overnight polysomnography (PSG): electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), airflow, abdominal respiratory effort (ABD), and pulse oximetry (SpO_2). All PSG signals are temporally synchronized and segmented into non-overlapping 30-s epochs. EEG,

Table A.1: **Dataset configuration for alignment pre-training and downstream evaluation.** Task IDs correspond to Table A.2.

Dataset	Task ID(s)	Signal	Records/Nights	Subjects	Split
Alignment pre-training					
SHHS-1/2	–	PSG	6,444	4,797	5,798 / 646
MESA	–	PSG	2,056	2,056	1,850 / 206
CFS	–	PSG	730	730	657 / 73
Downstream evaluation					
Test SHHS1/2	T1–T4, T6	PSG	2,000	1,000	8:1:1
APPLES	T3–T4	PSG	1,096	1,096	8:1:1
WSC	T3–T4, T6	PSG	2,534 / 2,570	1,121 / 1,123	8:1:1
SleepEDF20	T3	PSG	39	20	5-fold
ISRUC-1	T3	PSG	100	100	5-fold
HMC	T3	PSG	151	151	5-fold
MASS-SS2	T5	EEG	19	19	5-fold
CAP	T7	PSG	108	108	5-fold
Test SHHS1/2	T8–T13	PSG	2,000	1,000	5-fold
Test SHHS1 $\rightarrow$ SHHS2	T14–T15	PSG	1,000 pairs	1,000	5-fold
Test SHHS2	T16–T19	PSG	1,000	1,000	5-fold
MDD	T20	EEG	64	64	5-fold
OpenNeuro ds004504	T21	EEG	88	88	5-fold
MAT	T22	EEG	–	36	5-fold
BCI IV-2a	T23	EEG	–	9	5-fold
CinC 2017[†]	T24	ECG	8,528	–	5-fold
sEMG-DS1	T25	sEMG	2,000	20	5-fold
SEED-VIG	T26–T27	EEG / EEG+EOG	–	21	5-fold

Notes. Pre-training splits are reported as train/validation. Downstream splits are participant-level unless otherwise noted. For WSC, sample sizes differ slightly across tasks because of task-specific label availability and preprocessing, and are reported for T3–T4 / T6, respectively. [†]CinC 2017 uses class-stratified record-level five-fold cross-validation because reliable participant identifiers are unavailable.

EOG, EMG, and ECG are first standardized to 100 Hz, airflow and ABD to 10 Hz, and SpO_2 to 1 Hz; EEG and ECG are subsequently resampled to the native sampling rates required by their pretrained backbones. Each signal is normalized independently by channel, with extreme normalized values clipped to reduce the influence of outliers. Because signal availability varies across cohorts and recordings, unavailable modalities are excluded from the corresponding alignment terms using explicit availability masks rather than being imputed from other signals.

Modality encoders. The EEG and ECG branches are initialized from pretrained LaBraM and CSFM encoders, respectively. The remaining modalities, including EOG, EMG, airflow, ABD, and SpO_2, are processed by independently parameterized lightweight 1D CNN encoders, respectively (Table A.3). Each lightweight encoder consists of three convolutional layers, $\mathrm{Conv1D}(1, 32, k = 25, s = 5)$, $\mathrm{Conv1D}(32, 64, k = 15, s = 3)$, and $\mathrm{Conv1D}(64, 128, k = 9, s = 2)$, followed by global average pooling to produce a 128-dimensional representation. Temporal padding is applied where required for low-rate inputs.

Optimization and hyperparameters. All optimization settings are summarized here for reproducibility. Multimodal alignment pre-training is performed for up to 50 epochs using AdamW with a learning rate of 3×10^{-5}, weight decay 1×10^{-4}, linear learning-rate warm-up, and cosine decay. Training uses BF16 mixed precision on four NVIDIA H200 GPUs with a batch size of 512 per GPU, corresponding to a global batch size of 2,048. During alignment pre-training, the lower layers of the pretrained LaBraM and CSFM encoders remain frozen, while their upper layers are progressively unfrozen. The lightweight modality encoders and alignment modules are optimized throughout pre-training. Validation is performed periodically according to optimization steps, and the checkpoint with the lowest validation alignment loss is retained. Early stopping uses a patience of 10 validation evaluations. For downstream adaptation, LP uses a learning rate of 1×10^{-4} for

Table A.2: **Overview of the 27 downstream tasks.** Task IDs correspond to those used in the cross-task performance summary.

ID	Task	Type	Primary metric(s)
GENERAL SLEEP REPRESENTATION			
T1	Age prediction	R	MAE, RMSE, R^2
T2	Sex prediction	C2	Acc., MF1, AUROC
T3	Sleep staging	C5	Acc., MF1, κ
T4	Apnea/hypopnea detection	C2	Acc., AUROC, AUPRC
T5	Sleep-spindle detection	D	Precision, Recall, F1
CLINICAL AND DISEASE PREDICTION			
T6	AHI estimation / OSA severity	R+C4	MAE, RMSE, R^2; Acc., MF1
T7	Sleep-disorder classification	C2	MF1, AUROC
T8	COPD	C2	AUROC
T9	Emphysema	C2	AUROC
T10	Clinical depression	C2	AUROC
T11	Depressive affect	C2	AUROC
T12	Heart failure	C2	AUROC
T13	Stroke	C2	AUROC
T14	AHI worsening	C2	AUROC
T15	Incident hypersomnia	C2	AUROC
T16	Composite CVD	C2	AUROC
T17	Incident heart failure	C2	AUROC
T18	Incident stroke	C2	AUROC
T19	CVD death	C2	AUROC
DAYTIME PHYSIOLOGICAL TRANSFER			
T20	MDD diagnosis	C2	AUROC
T21	Dementia classification	C3	MF1
T22	Mental-arithmetic recognition	C2	MF1
T23	Motor-imagery classification	C4	MF1
T24	Atrial-fibrillation recognition	C4	Ch-F1, AF-F1, AUROC
T25	Hand-gesture recognition	C5	MF1
T26	Vigilance estimation (EEG)	R	Pearson r
T27	Multimodal vigilance estimation (EEG+EOG)	R	Pearson r

Notes. R denotes regression, Ck denotes k-class classification, and D denotes event detection. T6 jointly evaluates continuous AHI estimation and four-level OSA severity classification. T24 is a four-class ECG rhythm classification task (Normal, AF, Other, and Noisy). Ch-F1 is the official CinC 2017 score averaged over Normal, AF, and Other; AF-F1 and AUROC evaluate AF recognition. T26 and T27 share the same vigilance target but differ in input modality: T26 uses EEG only, whereas T27 uses EEG+EOG to assess multimodal fusion capability in downstream transfer specifically.

the prediction head. FT uses learning rates of 1×10^{-4} and 1×10^{-6} for the prediction head and backbone, respectively, corresponding to a backbone learning-rate scale of 0.01. Both settings use AdamW with weight decay 1×10^{-4}. Downstream models are generally trained for up to 50 epochs with a batch size of 32 and an early-stopping patience of 4. Downstream experiments are also conducted using NVIDIA H200 GPUs. We typically use a fixed random seed of 42 for both alignment pre-training and downstream evaluation, with repeated-seed experiments explicitly indicated where applicable.

A.3 BASELINES AND DOWNSTREAM EVALUATION PROTOCOL

Baseline methods We organize the compared methods into three groups.

(1) **Non-foundation models** include task-specific supervised architectures, general-purpose time-series models, and conventional self-supervised representation-learning methods. We select baselines according to the characteristics of each task, prioritizing strong domain-specific models with

Table A.3: **Pre-training provenance of inherited EEG and ECG foundation encoders.** LaBraM and CSFM have no cohort overlap with our downstream evaluation datasets.

Model	Signal	Original Pre-training Corpus	Pre-training Scale	Overlap
LaBraM	EEG	20 heterogeneous EEG datasets spanning BCI, emotion, resting-state, and clinical EEG, including the SEED and TUH series, together with other public and self-collected EEG datasets.	2,534.78 h	**None**
CSFM	ECG/PPG	MIMIC-III-WDB, MIMIC-IV-ECG, and CODE-Full.	$\sim$1.7M individuals	**None**

Note. "Overlap" denotes direct reuse of the same source cohort or recordings between foundation-model pretraining and our downstream evaluation.

public implementations when available and using a unified set of general-purpose baselines otherwise.

The general-purpose baselines are selected following representative benchmark suites used in LaBraM (Jiang et al., 2024), BrainOmni (Xiao et al., 2025), BIOT (Yang et al., 2023a) and CSFM (Gu et al., 2026). Depending on the input modality and task, these include ST-Transformer architectures for EEG (Song et al., 2021), CNN–Transformer models (Peh et al., 2022), ResNet1D-based architectures commonly used for ECG analysis (Ribeiro et al., 2020; Gu et al., 2026), general time-series models including TimesNet (Wu et al., 2023) and iTransformer (Liu et al., 2024), and ContraWR (Yang et al., 2023b) as a conventional self-supervised EEG representation-learning baseline.

For sleep-related tasks, we additionally include representative domain-specific supervised methods. For sleep staging, we compare against DeepSleepNet (Supratak et al., 2017), SleePyCo (Lee et al., 2024), U-Time (Perslev et al., 2019), AttnSleep (Eldele et al., 2021), XSleepNet (Phan et al., 2021), ASTGSleep (Chen et al., 2025), XSleepFusion (Hu et al., 2025c), and SleepXViT (Lee et al., 2025). For respiratory event detection, we include Hybrid Transformer (Hu et al., 2022), XSleepFusion (Hu et al., 2025c), and IPCT-Net (Hu et al., 2025b). For OSA severity estimation, because publicly available OSA-specific baselines are limited, we use the unified general baseline set above and additionally include TSDNet (Hu et al., 2025a) as an OSA-specific supervised method. For sleep spindle detection on MASS-SS2, we compare against RED-Time (Tapia & Estévez, 2020) and SUMO (Kaulen et al., 2022).

(2) **Physiological foundation models** include LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025) for EEG-based evaluation, and CSFM (Gu et al., 2026) for ECG-based evaluation. These baselines assess whether the proposed multimodal alignment provides additional benefits beyond large-scale unimodal physiological pretraining.

(3) **Sleep foundation models** include SleepFM (Thapa et al., 2026), SleepGPT (Huang et al., 2026), and OSF (Shuai et al., 2026). We select these three representative sleep foundation models because they cover recent and complementary approaches to sleep representation learning while providing publicly available pretrained weights and sufficiently documented downstream evaluation pipelines for reproducible comparison. Our goal is not to exhaustively enumerate all existing sleep foundation models, but to establish a representative and reproducible comparison set spanning different pretraining paradigms and input configurations. Whenever applicable, we use their officially released checkpoints, supported input modalities, preprocessing pipelines, and downstream adaptation protocols.

Common downstream evaluation protocol. We evaluate pretrained foundation models using linear probing (LP) and full fine-tuning (FT). In LP, the pretrained backbone is frozen, and only a task-specific linear prediction head is trained. In FT, the prediction head and the backbone layers selected by the fine-tuning schedule are optimized jointly. Epoch-level tasks use individual epoch representations, whereas whole-night and subject-level tasks aggregate epoch representations according to their task protocols.

Cross-task ranking protocol. To emphasize pretrained representation quality, foundation models are ranked primarily by LP performance, which evaluates downstream utility without modifying the pretrained backbone; FT is used only to break LP ties. A supervised baseline is ranked ahead of a foundation model only when it surpasses both its LP and FT performance, so that task-specific supervised training is considered superior only when it also exceeds the fully adapted foundation model. For each task, ranking is based on its primary evaluation metric, such as Macro-F1 for classification tasks, AUROC for disease prediction tasks, and the corresponding regression metric (e.g., MAE or Pearson's r) for continuous prediction tasks.

For models with public implementations or pretrained checkpoints, we follow their documented configurations and adapt inputs as required. Downstream signals undergo dataset-specific preprocessing, including channel selection, filtering, resampling, normalization, and sample construction where appropriate, following established evaluation protocols from prior foundation-model studies (Jiang et al., 2024; Xiao et al., 2025), and the preprocessing recommendations provided with each dataset. Within each task, compared methods use the same splits, prediction targets, and evaluation metrics. Input modalities follow each baseline's supported configuration; for the sEMG task, EEG foundation models are used as the closest available baselines due to the lack of comparable pretrained sEMG foundation models.

Evaluation metrics. We use task-specific evaluation metrics following commonly adopted protocols for the corresponding benchmarks and prior literature. For sleep staging, each 30-s epoch is classified into five AASM stages: Wake, N1, N2, N3, and REM, following the AASM scoring manual (Berry et al., 2017). We report accuracy (Acc), macro-F1 (MF1), and Cohen's κ, which measure overall accuracy, class-balanced performance, and agreement beyond chance, respectively (Supratak et al., 2017; Phan et al., 2021; Eldele et al., 2021). For respiratory-event detection, we report area under the precision–recall curve (AUPRC) and area under the receiver operating characteristic curve (AUROC), which are informative under class imbalance, together with accuracy for comparison with prior sleep-event detection studies. For obstructive sleep apnea (OSA) assessment, the apnea–hypopnea index (AHI) is evaluated both as a continuous clinical measure and through conventional severity categories defined by AHI thresholds (Berry et al., 2017). Accordingly, we report mean absolute error (MAE), root mean squared error (RMSE), and R^2 for AHI regression, together with accuracy and macro-F1 for OSA severity classification (Hu et al., 2025a).

For other disease and clinical classification tasks, we primarily use AUROC for imbalanced binary outcomes and MF1 for multiclass classification benchmarks. Regression and continuous physiological prediction tasks use the corresponding standard regression or correlation metrics. Metrics for all 27 downstream tasks are also summarized in Table A.2.

A.4 Additional Results and Analyses

A.4.1 Detailed Results for General Sleep Representation

We provide additional analyses for the tasks summarized in the main text. Results already reported in full in the main text are omitted to avoid repetition.

Age and Sex Prediction. Table A.4 reports the complete results for subject-level age and sex prediction on Test SHHS1/2. For age estimation, SleepXA with full PSG achieves the best performance under both LP and FT. Its LP results (MAE 5.250, RMSE 6.680, $R^2 = 0.584$) are comparable to, and slightly better than, the corresponding FT results, indicating that age-related information is already well represented before downstream adaptation. For sex classification, OSF performs best under LP, while SleepXA with full PSG achieves the strongest FT accuracy and MF1 and remains competitive in AUROC. Several baselines show modest degradation after FT, suggesting that end-to-end adaptation can be sensitive to model architecture and task-specific optimization, whereas SleepXA remains comparatively stable across LP and FT.

Sleep Stage Representation Visualization. To complement the quantitative sleep-staging results in the main text, we qualitatively examine the organization of the learned representations across Test SHHS1/2, APPLES, and WSC. Figure A.1 visualizes epoch-level embeddings using UMAP under both LP and FT. Across the three cohorts, the representations exhibit consistent stage-dependent organization while retaining overlap between neighboring sleep states, particularly around transitional stages such as N1. After full fine-tuning, the embedding geometry becomes more structured with respect to the downstream staging labels. These visualizations provide a qualitative complement to

Table A.4: **Age prediction and sex classification on Test SHHS1/2.** Lower values are better for MAE and RMSE, while higher values are better for R^2, Acc, MF1, and AUROC. **Bold** and underlined values indicate the **best** and second-best foundation-model results within each training setting.

Model	Channels	Age prediction			Sex classification		
		MAE ↓	RMSE ↓	R^2 ↑	Acc. ↑	MF1 ↑	AUROC ↑
General non-foundation models							
TimesNet (Wu et al., 2023)	Full	7.565	9.328	0.268	0.725	0.724	0.780
CNN-Transformer (Peh et al., 2022)	Full	7.063	8.964	0.324	0.490	0.329	0.494
ResNet1D-50 (Ribeiro et al., 2020)	Full	7.375	9.642	0.218	0.585	0.584	0.579
ST-Transformer (Song et al., 2021)	Full	7.005	9.163	0.294	0.610	0.600	0.690
ContraWR (Yang et al., 2023b)	Full	6.731	8.541	0.386	0.675	0.674	0.718
iTransformer (Liu et al., 2024)	Full	8.292	10.312	0.106	0.570	0.570	0.606
Foundation models: linear probing (LP)							
LaBraM (Jiang et al., 2024)	E	9.917	11.781	-0.292	0.590	0.590	0.628
CBraMod (Wang et al., 2025)	E	8.813	10.827	0.014	0.540	0.539	0.577
CSFM (Gu et al., 2026)	C	7.777	9.735	0.118	0.561	0.544	0.635
SleepFM (Thapa et al., 2026)	Full	8.227	9.906	0.086	0.717	0.717	0.790
SleepGPT (Huang et al., 2026)	Full	6.146	7.765	0.410	0.636	0.631	0.716
OSF (Shuai et al., 2026)	Full	6.124	7.561	0.468	0.813	0.812	0.889
SleepXA (Ours)	Full	**5.250**	**6.680**	**0.584**	0.793	0.792	0.876
Foundation models: full fine-tuning (FT)							
LaBraM (Jiang et al., 2024)	E	8.059	9.763	0.112	0.699	0.693	0.760
CBraMod (Wang et al., 2025)	E	6.168	8.018	0.459	0.645	0.645	0.723
CSFM (Gu et al., 2026)	C	7.374	9.330	0.189	0.772	0.771	0.857
SleepFM (Thapa et al., 2026)	Full	7.442	9.098	0.229	0.711	0.711	0.790
SleepGPT (Huang et al., 2026)	Full	8.047	9.760	0.113	0.762	0.762	0.838
OSF (Shuai et al., 2026)	Full	7.075	8.801	0.279	0.839	0.838	0.912
SleepXA (Ours)	Full	5.277	6.697	0.582	**0.879**	**0.878**	**0.939**

E: EEG; C: ECG; E/O/M: EEG+EOG+EMG; Full: full experimental PSG configuration.

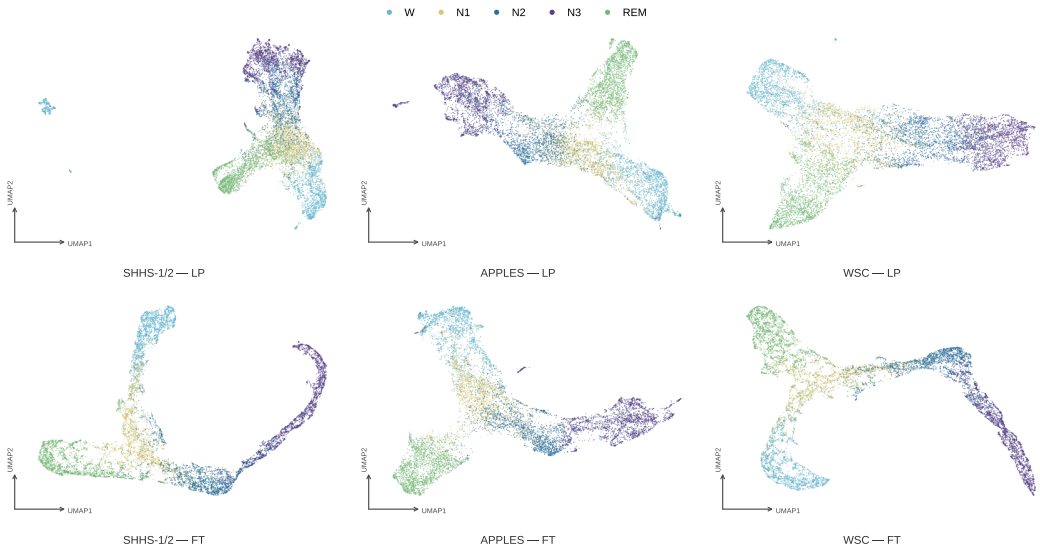


Figure A.1: **UMAP visualization of sleep-stage representations.** Epoch-level representations from Test SHHS1/2, APPLES, and WSC are shown under LP (top) and FT (bottom). Points are colored according to the five sleep stages: Wake, N1, N2, N3, and REM.

the quantitative cross-cohort results and suggest that the learned sleep-stage structure is preserved across independent cohorts.

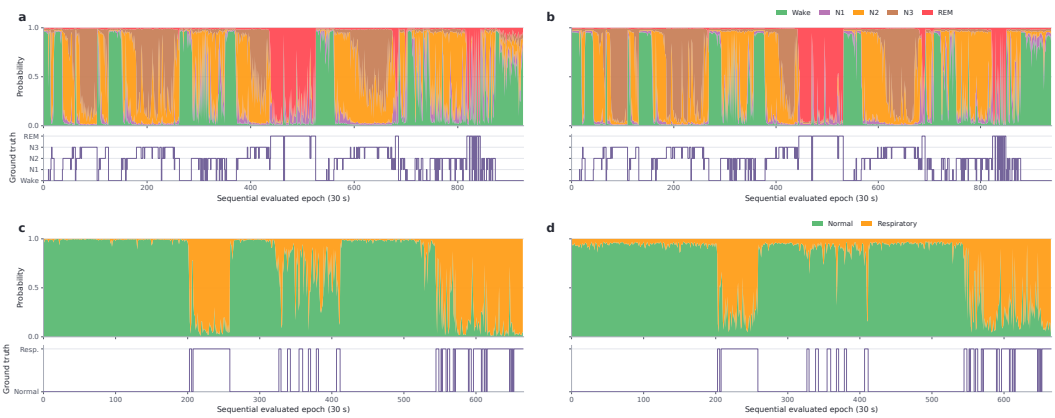


Figure A.2: **Illustrative whole-night predictions on a randomly selected WSC test recording.** Recording ID: `39201`. Panels show sleep staging under (a) LP and (b) FT, and respiratory-event detection under (c) LP and (d) FT. Prediction probabilities are shown with the corresponding ground-truth labels. Unevaluated epochs are removed and the remaining epochs are sequentially reindexed.

Sleep Staging on Small Cohorts. Beyond the large-cohort evaluations in the main text, we further assess transfer to three widely used small-cohort sleep-staging benchmarks: SleepEDF20 (Kemp et al., 2000), HMC (Alvarez-Estevez & Rijsman, 2021), and ISRUC-1 (Khalighi et al., 2016). All experiments use subject-level five-fold cross-validation to evaluate generalization to unseen participants in these small cohorts. For supervised baselines, we include DeepSleepNet as a classical single-channel sleep-staging model (Supratak et al., 2017), XSleepNet (Phan et al., 2021)as a representative multichannel sleep-staging architecture, and ASTGSleep as a full-channel graph-based fusion model (Chen et al., 2025), providing complementary comparisons with the evaluated sleep foundation models. As shown in Table A.5, SleepXA remains competitive under LP and achieves strong performance after FT across all three cohorts. In particular, FT reaches 86.76% accuracy, 81.50% Macro-F1, and $\kappa = 0.816$ on SleepEDF20; 77.44%, 76.28%, and 0.708 on HMC; and 79.07%, 77.10%, and 0.732 on ISRUC-1, respectively. These results extend the large-cohort findings in the main text and show that the learned representation remains transferable under smaller downstream cohorts, subject-level cross-validation, and different sensor configurations.

Fig. A.2(a,b) provides a complementary whole-night view of model behavior. We randomly select one recording from the WSC test set and visualize predictions across the whole night for both sleep staging and respiratory-event detection under LP and FT. For sleep staging, LP already captures the overall whole-night sleep architecture well, while FT produces more confident and temporally coherent predictions in some fragmented or transition-heavy periods.

Respiratory Event Detection. We further evaluate whether the learned representations capture epoch-level respiratory abnormalities across cohorts. Unlike subject-level AHI estimation and OSA severity classification, this task probes temporally localized respiratory events within the overnight recording. Table A.6 reports results on Test SHHS1/2, APPLES, and WSC.

Under LP, SleepXA achieves the strongest overall performance among the evaluated foundation models across all three cohorts. It reaches 83.44% accuracy, 0.763 AUPRC, and 0.879 AUROC on Test SHHS1/2; 79.61%, 0.793, and 0.881 on APPLES; and 90.73%, 0.790, and 0.938 on WSC. These results indicate that respiratory-event information is accessible from the frozen representation and transfers across independent sleep cohorts. FT further improves performance on Test SHHS1/2 and APPLES and maintains strong performance on WSC, showing that the pretrained representation also supports task-specific adaptation. For consistency with the sleep-staging visualization, Fig. A.2(c,d) shows whole-night respiratory-event probabilities under LP and FT for the same randomly selected WSC test recording.

Sleep Spindle Detection. We evaluate single-channel spindle detection on MASS-SS2 using C3 EEG and report FT results only, following the SleepGPT evaluation protocol (Huang et al., 2026).

Table A.5: **Sleep staging on small cohorts under subject-level five-fold cross-validation.** Acc. and MF1 are percentages; all metrics are higher-is-better. **Bold** and underlined values indicate the **best** and second-best results.

Model	Channels	Sleep-EDF-20			HMC			ISRUC-1		
		Acc.	MF1	κ	Acc.	MF1	κ	Acc.	MF1	κ
Task-specific supervised models										
DeepSleepNet (Supratak et al., 2017)	E	81.90	76.60	0.760	63.06	59.32	0.521	69.00	65.22	0.604
XSleepNet (Phan et al., 2021)	E/O/M	86.30	80.60	0.813	75.76	73.41	0.684	80.07	78.74	0.743
ASTGSleep (Chen et al., 2025)	Full	86.50	80.90	0.814	74.88	72.66	0.673	78.50	77.40	0.724
Foundation models: linear probing (LP)										
SleepFM (Thapa et al., 2026)	Full	76.27	69.89	0.683	66.53	64.82	0.570	72.21	70.98	0.647
SleepGPT (Huang et al., 2026)	Full	81.62	76.48	0.753	74.18	72.25	0.665	71.44	68.62	0.634
OSF (Shuai et al., 2026)	Full	82.18	77.23	0.760	75.56	74.50	0.686	75.62	73.25	0.687
SleepXA (Ours)	Full	82.32	77.09	0.761	76.69	74.83	0.698	78.63	**77.76**	0.727
Foundation models: full fine-tuning (FT)										
SleepFM (Thapa et al., 2026)	Full	79.93	73.39	0.728	72.44	69.73	0.638	74.18	72.26	0.670
SleepGPT (Huang et al., 2026)	Full	83.57	78.40	0.777	77.17	75.90	0.704	74.23	73.11	0.670
OSF (Shuai et al., 2026)	Full	83.79	79.27	0.781	76.53	75.21	0.695	78.26	77.02	0.722
SleepXA (Ours)	Full	**86.76**	**81.50**	**0.816**	**77.44**	**76.28**	**0.708**	**79.07**	77.10	**0.732**

E: EEG; E/O/M: EEG+EOG+EMG; Full: full experimental PSG configuration.

Table A.6: **Respiratory event detection performance on three representative cohorts.** Acc. is reported in percentage, while AUPRC and AUROC are reported on a 0–1 scale. Higher is better for all metrics.

Model	Channels	Test SHHS1/2			APPLES			WSC		
		Acc.	AUP	AUR	Acc.	AUP	AUR	Acc.	AUP	AUR
Task-specific supervised models										
Hybrid Transformer (Hu et al., 2022)	C	58.52	0.329	0.601	56.89	0.368	0.542	78.05	0.262	0.634
XSleepFusion (Hu et al., 2025c)	Full	81.13	0.758	0.878	81.02	0.777	0.871	**92.26**	**0.845**	**0.955**
IPCT-Net (Hu et al., 2025b)	Full	83.54	0.763	0.881	80.58	0.771	0.864	90.77	0.834	0.931
Foundation models: linear probing (LP)										
CSFM (Gu et al., 2026)	C	67.89	0.340	0.610	47.21	0.382	0.551	74.83	0.246	0.624
SleepFM (Thapa et al., 2026)	Full	77.35	0.608	0.795	77.44	0.754	0.852	89.74	0.740	0.924
SleepGPT (Huang et al., 2026)	Full	72.70	0.659	0.820	75.77	0.747	0.847	89.98	0.755	0.922
OSF (Shuai et al., 2026)	Full	83.07	0.760	0.876	78.78	0.775	0.868	89.28	0.733	0.915
SleepXA (Ours)	Full	83.44	0.763	0.879	79.61	**0.793**	0.881	90.73	0.790	0.938
Foundation models: full fine-tuning (FT)										
CSFM (Gu et al., 2026)	C	65.93	0.389	0.664	64.73	0.353	0.500	77.73	0.323	0.719
SleepFM (Thapa et al., 2026)	Full	57.21	0.505	0.747	76.51	0.714	0.820	87.12	0.627	0.856
SleepGPT (Huang et al., 2026)	Full	72.74	0.635	0.804	78.34	0.776	0.870	90.73	0.832	0.935
OSF (Shuai et al., 2026)	Full	80.01	0.715	0.856	80.38	0.771	0.868	90.29	0.775	0.931
SleepXA (Ours)	Full	**84.25**	**0.772**	**0.884**	**81.10**	0.791	**0.882**	91.07	0.839	0.936

C: ECG; Full: full experimental PSG configuration. AUP and AUR denote AUPRC and AUROC, respectively. Respiratory-event detection is performed on 30-s epochs; an epoch is labeled positive if at least 10 s of the epoch overlaps with an annotated apnea or hypopnea event.

As shown in Table A.7, SleepXA achieves the highest precision among the evaluated foundation-model methods for both expert annotations and obtains F1 scores of 0.649 and 0.757 for E1 and E2, respectively. Because SleepGPT explicitly models time–frequency structure during pre-training, it provides a particularly relevant comparison for spindle detection. SleepXA improves F1 over SleepGPT by 3.9 and 1.6 percentage points for E1 and E2, respectively. Compared with task-specific detectors, SleepXA surpasses SUMO (Kaulen et al., 2022) in F1 despite not being specifically designed for spindle detection, while remaining below RED-Time (Tapia & Estévez, 2020) due to its comparatively higher precision but lower recall.

Table A.7: **Sleep spindle detection on MASS-SS2 using C3 EEG.** E1 and E2 denote the two expert annotations. **Bold** and underlined values indicate the **best** and second-best results across all methods in each column, respectively, based on unrounded scores. Higher is better for all metrics.

Model	E1			E2		
	Precision	Recall	F1	Precision	Recall	F1
Task-specific supervised models						
RED-Time (Tapia & Estévez, 2020)	0.583	**0.866**	**0.681**	0.744	**0.841**	**0.783**
SUMO (Kaulen et al., 2022)	0.527	0.864	0.641	0.694	0.807	0.726
Foundation-model-based methods: full fine-tuning (FT)						
SleepFM (Thapa et al., 2026)	0.529	0.781	0.621	0.702	0.801	0.741
SleepGPT (Huang et al., 2026)	0.673	0.603	0.610	0.702	0.801	0.741
OSF (Shuai et al., 2026)	0.544	0.853	0.648	0.710	0.824	0.756
SleepXA (ours)	**0.679**	0.724	0.649	**0.782**	0.742	0.757

Precision, recall, and F1 are averaged independently across folds.

Table A.8: **Performance comparison on AHI estimation and OSA severity classification across two representative cohorts.**

Model	Channels	Test SHHS1/2					WSC				
		MAE ↓	RMSE ↓	R^2 ↑	Acc. (%) ↑	MF1 (%) ↑	MAE ↓	RMSE ↓	R^2 ↑	Acc. (%) ↑	MF1 (%) ↑
General non-foundation models											
TimesNet (Wu et al., 2023)	Full	5.319	7.424	0.751	66.50	64.36	3.276	5.136	0.911	78.31	75.47
CNN-Transformer (Peh et al., 2022)	Full	10.473	15.436	-0.076	39.50	14.16	10.317	17.915	-0.088	27.71	10.85
ResNet1D-50 (Ribeiro et al., 2020)	Full	4.833	7.197	0.766	69.50	69.73	4.461	8.467	0.757	75.10	69.98
ST-Transformer (Song et al., 2021)	Full	4.746	6.893	0.785	67.50	67.54	4.546	6.950	0.836	69.88	67.26
ContraWR (Yang et al., 2023b)	Full	4.829	7.253	0.763	66.00	64.31	3.359	6.542	0.855	77.91	73.14
iTransformer (Liu et al., 2024)	Full	6.192	8.552	0.670	59.00	57.24	4.499	8.536	0.753	74.70	70.01
Task-specific supervised models											
TSDNet (Hu et al., 2025a)	Full	4.363	6.171	0.828	70.00	69.50	3.234	6.116	0.873	80.32	75.64
Foundation models: linear probing (LP)											
CSFM (Gu et al., 2026)	C	10.531	15.008	-0.017	35.50	19.75	9.935	17.155	0.003	35.34	23.93
SleepFM (Thapa et al., 2026)	Full	6.085	8.781	0.652	56.50	45.98	4.349	7.504	0.809	67.47	63.89
SleepGPT (Huang et al., 2026)	Full	5.692	8.088	0.705	59.00	53.37	4.260	7.082	0.830	71.89	66.72
OSF (Shuai et al., 2026)	Full	5.239	7.235	0.764	66.50	65.34	4.571	7.354	0.817	69.88	67.29
SleepXA (Ours)	Full	4.526	6.671	0.799	68.50	68.35	3.441	6.341	0.864	79.12	75.05
Foundation models: full fine-tuning (FT)											
CSFM (Gu et al., 2026)	C	10.407	15.312	-0.059	36.00	20.23	9.641	16.885	0.034	38.55	25.84
SleepFM (Thapa et al., 2026)	Full	4.846	6.984	0.780	59.50	59.21	4.178	7.325	0.818	74.71	67.35
SleepGPT (Huang et al., 2026)	Full	4.596	6.751	0.794	69.00	68.37	3.731	**5.464**	**0.899**	79.14	76.28
OSF (Shuai et al., 2026)	Full	4.694	6.424	0.814	63.00	60.61	3.731	5.948	0.880	77.51	72.69
SleepXA (Ours)	Full	**4.200**	**5.920**	**0.842**	**73.50**	**73.32**	**3.120**	6.008	0.878	**82.73**	**78.46**

OSA severity is categorized into normal, mild, moderate, and severe groups according to the reference AHI.

A.4.2 Detailed Clinical and Disease Evaluation

This section provides additional task definitions and evaluation details for the clinical benchmarks summarized in the main text. The evaluations cover OSA severity assessment, heterogeneous sleep-disorder classification, current disease diagnosis, and longitudinal clinical outcome prediction.

AHI Estimation and OSA Severity Classification. We evaluate continuous apnea–hypopnea index (AHI) estimation together with four-level obstructive sleep apnea (OSA) severity classification on Test SHHS1/2 and WSC. OSA severity is defined using the conventional AHI thresholds: normal (< 5 events/h), mild (5–< 15), moderate (15–< 30), and severe (≥ 30 events/h) (Berry et al., 2017). Predicted severity is obtained by applying the same thresholds to the estimated AHI. This joint evaluation therefore assesses both continuous estimation of respiratory burden and clinically defined severity stratification. The complete comparison with task-specific supervised models and foundation models under LP and FT is reported in Table A.8. The results complement the epoch-level respiratory-event evaluation by examining whether the learned representations can support subject-level estimation of overnight disease burden and clinically meaningful severity.

Detailed OSA Prediction Analysis. Beyond aggregate metrics, we further analyze the prediction behavior of SleepXA on Test SHHS1/2 and WSC under both LP and FT. Figure A.3 provides four complementary views. First, the four-level confusion matrices characterize the pattern of errors among normal, mild, moderate, and severe OSA. Second, predicted-versus-reference AHI plots examine the correspondence between estimated and observed respiratory burden across its continuous

range. Third, Bland–Altman analyses provide an additional view of agreement and systematic prediction bias. We additionally examine performance across sex, age, and BMI subgroups. This analysis is particularly relevant because demographic information, including age, is itself represented in the learned sleep representation, as demonstrated by the demographic-prediction experiments. The subgroup analysis therefore serves as a diagnostic robustness check for whether OSA performance is concentrated in a narrow demographic stratum. Across the considered subgroups, the model retains useful severity classification performance under both LP and FT. This subgroup analysis is intended to characterize how performance varies across clinically relevant participant groups.

Heterogeneous Sleep-Disorder Generalization. We further evaluate whether the learned representations transfer beyond sleep-disordered breathing to heterogeneous sleep pathologies using the CAP Sleep Database (Terzano et al., 2001). The cohort includes multiple disorder categories, including nocturnal frontal lobe epilepsy (NFLE), REM sleep behavior disorder (RBD), periodic leg movements (PLM), insomnia, narcolepsy, and other sleep disorders. We formulate a subject-level binary classification task that distinguishes participants with sleep disorders from healthy controls and use subject-level five-fold cross-validation because of the limited cohort size.

We evaluate two sensing configurations: full multimodal PSG and a unified central-EEG setting. The full-PSG setting examines whether the multisystem representation transfers to heterogeneous pathological conditions, whereas the EEG-only setting tests whether disease-related information remains accessible under substantially restricted sensing. As shown in Fig. A.4, SleepXA maintains effective transfer across both settings. With full PSG, the frozen representation reaches 0.798 Macro-F1, indicating that disorder-related information is already accessible without adapting the pretrained backbone. The central-EEG setting provides a complementary test of transfer when only a restricted subset of the original PSG signals is available.

A.4.3 Current and Longitudinal Clinical Prediction

Task definitions. We consider 12 subject-level clinical prediction tasks spanning current disease status and longitudinal outcomes. Current disease diagnosis includes six conditions across three clinical domains: respiratory disorders (COPD and emphysema), mental-health conditions (clinical depression and depressive affect), and cardiovascular conditions (heart failure and stroke) (Quan et al., 1997; Hou et al., 2026). Whole-night recordings from Test SHHS1 and Test SHHS2 are used for these evaluations, with all recordings from the same participant assigned to the same fold.

Longitudinal sleep prediction uses Test SHHS1 recordings to predict AHI worsening and incident hypersomnia observed at the Test SHHS2 follow-up visit. We further use Test SHHS2 as the index examination to predict four five-year cardiovascular outcomes: a composite CVD endpoint and three specific outcomes, namely heart failure, stroke, and cardiovascular death. Participants without sufficient follow-up for the corresponding prediction horizon are excluded. The composite CVD endpoint is defined as the occurrence of angina, heart failure, myocardial infarction, stroke, coronary revascularization, or cardiovascular death within five years after SHHS2.

Evaluation protocol. Given the limited number of positive participants and substantial class imbalance in several endpoints, we evaluate these tasks using frozen pretrained representations. Epoch-level representations are aggregated over each overnight recording using their mean and standard deviation, followed by feature standardization and an ℓ_2-regularized logistic-regression probe. Within each cross-validation fold, model hyperparameters are selected using validation data drawn only from the training partition, while the held-out fold remains untouched for final evaluation. We use participant-level five-fold cross-validation, with paired SHHS1 and SHHS2 visits from the same participant retained in the same fold.

Current disease diagnosis. Table A.9 reports frozen-representation performance for the six current disease diagnoses. SleepXA achieves the highest mean AUROC among the evaluated foundation models on five of the six tasks, spanning respiratory, mental-health, and cardiovascular conditions. The only exception is depressive affect, for which CBraMod achieves the highest mean AUROC.

Longitudinal outcome prediction. Table A.10 reports results for two cross-visit sleep outcomes and four five-year cardiovascular outcomes. SleepXA achieves the highest mean AUROC among the evaluated foundation models on all six longitudinal tasks. These results complement the current-

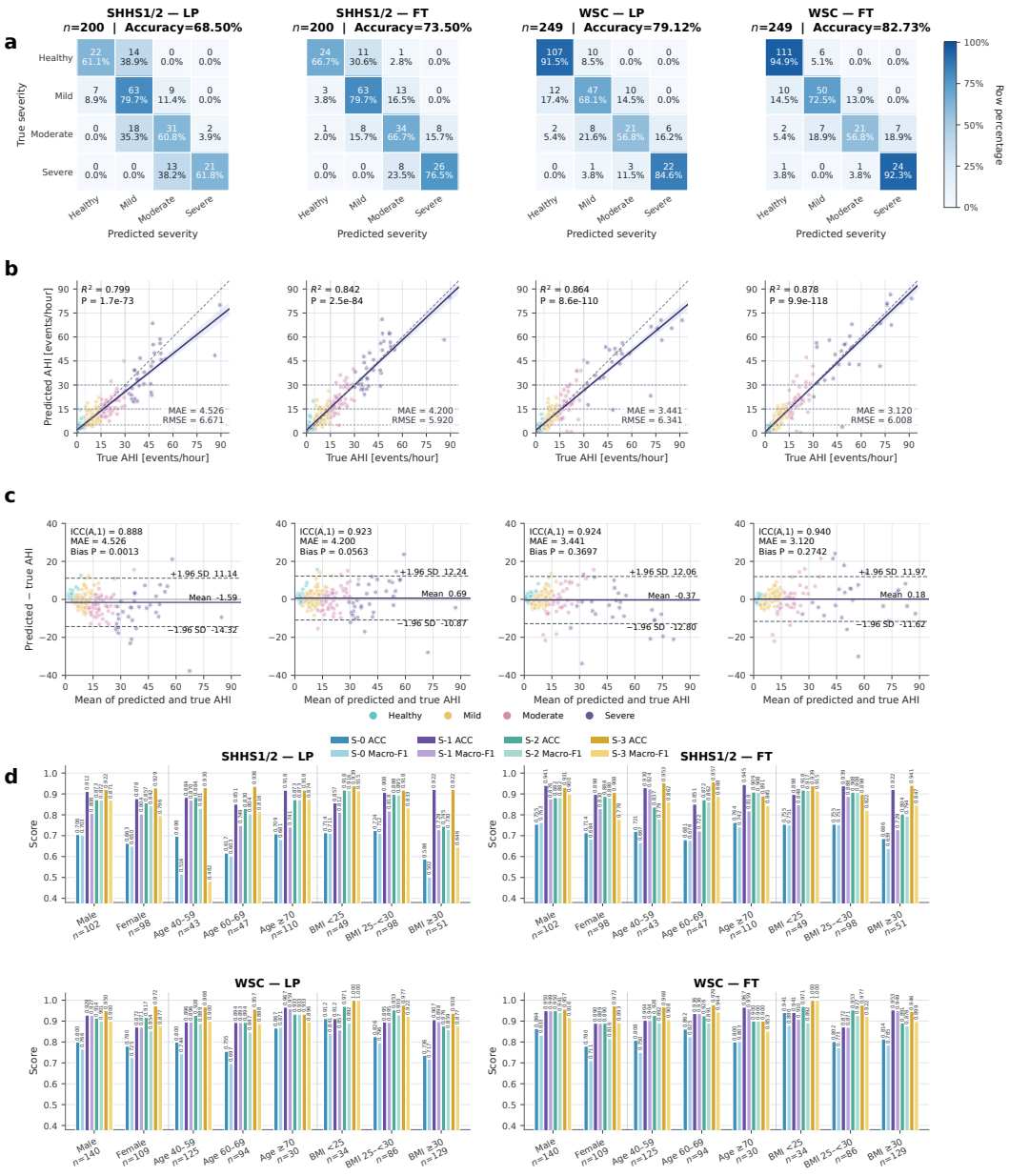


Figure A.3: **Detailed analysis of AHI estimation and OSA severity prediction.** Results are shown for Test SHHS1/2 and WSC under linear probing (LP) and full fine-tuning (FT). (a) Four-level OSA severity confusion matrices for normal, mild, moderate, and severe OSA; (b) predicted versus reference AHI with regression statistics; (c) Bland–Altman agreement analysis between predicted and reference AHI; and d) subgroup performance across sex, age, and BMI strata, with S-0 to S-3 denoting normal, mild, moderate, and severe OSA, respectively. The subgroup analysis is included as a robustness diagnostic to examine whether predictive performance is concentrated in particular demographic groups.

disease evaluation by testing whether information retained in the frozen representation is also relevant to subsequent changes in sleep-related phenotypes and longer-term cardiovascular outcomes.

A.4.4 Discussion of Daytime Physiological Transfer

The daytime transfer results show that representations learned from nocturnal PSG can remain useful beyond sleep-specific tasks. Sleep foundation models exhibit meaningful transfer to awake-state

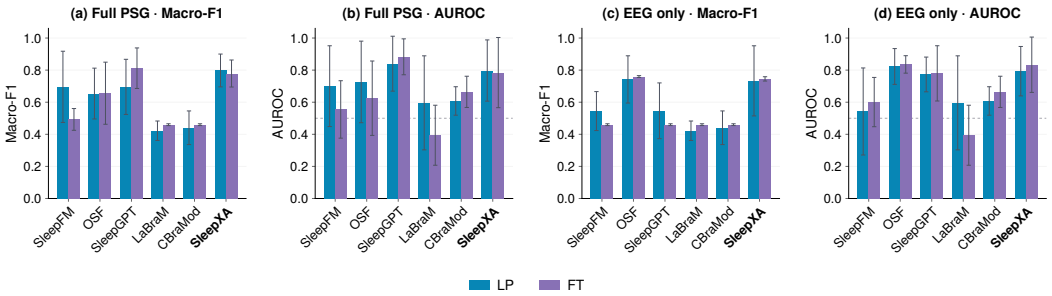


Figure A.4: **Sleep-disorder classification on the CAP Sleep Database.** Subject-level binary classification of heterogeneous sleep disorders versus healthy controls under full PSG (a–b) and central EEG-only (c–d) configurations. Results are reported as mean and standard deviation across five subject-level folds.

Table A.9: **Linear-probing performance for current disease diagnosis.**

Method	Respiratory		Mental Health		Cardiovascular	
	COPD	Emphysema	Clin. Depression	Depressive Affect	Heart Failure	Stroke
LaBraM (Jiang et al., 2024)	0.469±0.061	0.553±0.025	0.487±0.034	0.561±0.077	0.599±0.051	0.573±0.051
CBraMod (Wang et al., 2025)	0.502±0.120	0.548±0.054	0.502±0.037	**0.629±0.064**	0.618±0.048	0.582±0.043
SleepFM (Thapa et al., 2026)	0.553±0.087	0.596±0.097	0.452±0.021	0.544±0.064	0.704±0.073	0.555±0.028
SleepGPT (Huang et al., 2026)	0.549±0.052	0.566±0.076	0.520±0.044	0.580±0.073	0.643±0.050	0.524±0.051
OSF (Shuai et al., 2026)	0.642±0.055	0.636±0.044	0.535±0.028	0.593±0.068	0.679±0.039	0.545±0.026
SleepXA (ours)	**0.682±0.046**	**0.678±0.046**	**0.547±0.031**	0.591±0.069	**0.709±0.057**	**0.604±0.026**

EEG tasks, including MDD (Mumtaz et al., 2017), Alzheimer's disease (Miltiadous et al., 2023), mental-arithmetic recognition (Zyma et al., 2019), motor imagery (Tangermann et al., 2012), and vigilance estimation (Zheng & Lu, 2017). This suggests that **sleep pre-training captures neural structure that remains relevant across behavioral states rather than being restricted to sleep-stage information**.

SleepXA further shows strong transfer to tasks that differ more substantially from nocturnal sleep, including ECG-based atrial-fibrillation recognition and surface-EMG gesture classification. The SEED-VIG evaluation provides a complementary test of multimodal transfer: using EEG+EOG improves vigilance estimation over EEG alone, indicating that cross-modal complementarity remains useful in an awake-state task. These results are consistent with the compositional design of SleepXA, which inherits modality-specific knowledge from pretrained physiological encoders and organizes it through sleep-guided common-specific alignment. The larger differences observed under LP further suggest that this advantage is already present in the pretrained representation, while FT can partially reduce differences between pre-training strategies. Overall, the results support nocturnal PSG as a useful context for **learning reusable multisystem physiological representations beyond sleep**.

A.4.5 Effect of Physiological Foundation-Model Initialization

We first examine whether the performance of SleepXA benefits from physiological knowledge inherited from pretrained EEG and ECG foundation models. All variants use the same SleepXA architecture, multimodal inputs, alignment objectives, pre-training data, downstream splits, and optimization settings, differing only in EEG and ECG encoder initialization. *Scratch* randomly initializes both encoders while keeping the remaining architecture and training configuration unchanged; *EEG-FM* and *ECG-FM* initialize only the corresponding encoder from LaBraM and CSFM, respectively; and *EEG+ECG-FM* initializes both encoders from their pretrained checkpoints. This controlled comparison isolates the contribution of inherited physiological knowledge.

Sleep staging. Table A.11 reports the complete WSC sleep staging results. Under LP, joint EEG+ECG foundation-model initialization improves accuracy from 81.40% with Scratch to 85.16%, MF1 from 73.00% to 76.09%, and Cohen's κ from 0.717 to 0.765. The same trend remains after FT, where the jointly initialized model reaches 85.95% accuracy, 77.61% MF1, and $\kappa = 0.780$. The class-wise confusion matrices in Fig. A.5 further show improvements across multiple sleep stages rather than only the dominant N2 class. Notably, ECG initialization improves REM

Table A.10: **Longitudinal sleep-related progression and five-year cardiovascular outcome prediction.** Longitudinal sleep outcomes use SHHS1 representations to predict changes observed at SHHS2; the cardiovascular outcomes use SHHS2 as the index examination for five-year prediction.

Method	Longitudinal Sleep Outcomes		5-Year Cardiovascular Outcomes			
	AHI Worsening	Incident Hypersomnia	Composite CVD outcome	Heart Failure	Stroke	CVD Death
LaBraM (Jiang et al., 2024)	0.513±0.042	0.493±0.107	0.603±0.017	0.648±0.048	0.555±0.050	0.675±0.051
CBraMod (Wang et al., 2025)	0.549±0.024	0.556±0.070	0.612±0.051	0.665±0.053	0.608±0.043	0.669±0.076
SleepFM (Thapa et al., 2026)	0.594±0.026	0.559±0.060	0.634±0.044	0.695±0.046	0.540±0.062	0.695±0.032
SleepGPT (Huang et al., 2026)	0.567±0.036	0.585±0.058	0.605±0.025	0.671±0.038	0.516±0.074	0.691±0.075
OSF (Shuai et al., 2026)	0.553±0.041	0.552±0.047	0.641±0.034	0.692±0.040	0.495±0.039	0.740±0.038
SleepXA (ours)	**0.614±0.055**	**0.595±0.110**	**0.643±0.014**	**0.706±0.032**	**0.613±0.048**	**0.757±0.068**

Table A.11: **Controlled initialization study on WSC sleep staging.** All variants share the same architecture and training configuration, differing only in the initialization of the EEG and ECG encoders. All metrics except Cohen's κ are reported in percentages. **Bold** and underlined values denote the **best** and second-best results within LP or FT, respectively.

Protocol	Initialization	Acc.	MF1	κ	W F1	N1 F1	N2 F1	N3 F1	REM F1
LP	Scratch	81.40	73.00	0.717	85.61	46.07	88.34	63.33	81.63
	EEG-FM	82.36	73.79	0.727	85.89	**50.75**	88.58	62.64	81.11
	ECG-FM	80.77	72.49	0.710	84.38	46.05	87.44	62.44	82.16
	SleepXA	**85.16**	**76.09**	**0.765**	**87.27**	49.32	**91.04**	**65.28**	**87.52**
FT	Scratch	79.93	71.84	0.698	84.28	**54.76**	86.54	58.46	75.17
	EEG-FM	83.67	76.04	0.748	86.91	51.27	89.52	**73.20**	79.30
	ECG-FM	81.19	71.83	0.712	84.62	44.82	86.68	60.23	82.82
	SleepXA	**85.95**	**77.61**	**0.780**	**88.73**	51.81	**91.23**	67.02	**89.25**

recognition, consistent with the distinct autonomic and cardiac dynamics observed during REM sleep (Berry et al., 2017).

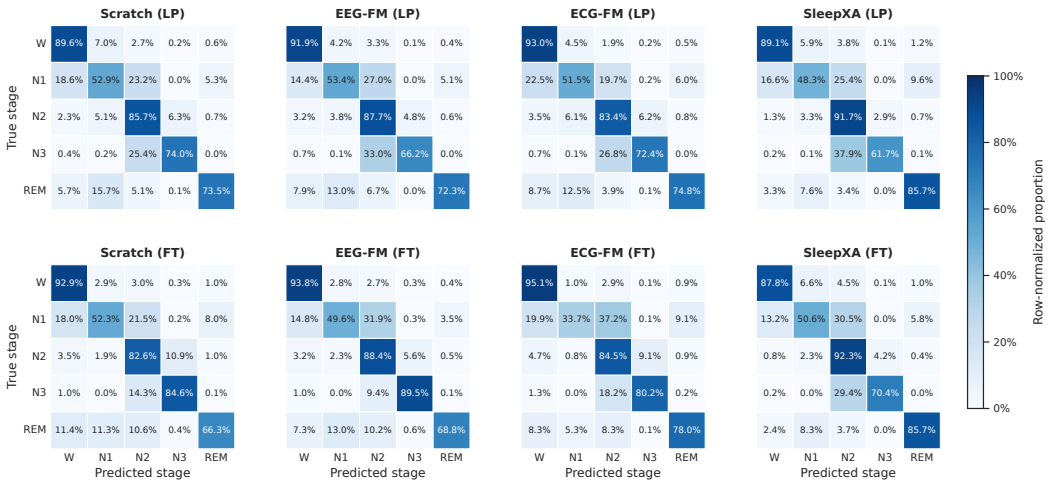


Figure A.5: **Class-wise analysis of foundation-model initialization on sleep staging.** Row-normalized confusion matrices are shown for four initialization settings under linear probing (LP, top row) and full fine-tuning (FT, bottom row). Rows denote ground-truth stages and columns denote predicted stages. Each cell reports the percentage within the corresponding ground-truth class.

Cross-context ECG transfer. We further evaluate the same initialization variants on CinC2017 atrial-fibrillation recognition, transferring representations learned from nocturnal multimodal PSG to an independent single-lead ECG task. As shown in Table A.12, ECG-FM initialization already improves AF-F1 from 0.181 with Scratch to 0.398 under LP. Joint EEG+ECG initialization further

Table A.12: **Controlled initialization study on the CinC2017 AF task.** Results are reported as mean ± standard deviation across test folds to assess the stability of different initialization strategies. **Bold** and underlined values denote the **best** and second-best results within LP or FT, respectively.

Protocol	Initialization	Ch-F1 ↑	AF-F1 ↑	AF-AUROC ↑
LP	Scratch	0.346 ± 0.005	0.181 ± 0.035	0.603 ± 0.033
	EEG-FM	0.349 ± 0.011	0.129 ± 0.034	0.602 ± 0.039
	ECG-FM	0.569 ± 0.027	0.398 ± 0.030	0.808 ± 0.033
	SleepXA	$\mathbf{0.732 \pm 0.004}$	$\mathbf{0.743 \pm 0.019}$	$\mathbf{0.912 \pm 0.027}$
FT	Scratch	0.446 ± 0.020	0.396 ± 0.026	0.846 ± 0.015
	EEG-FM	0.475 ± 0.040	0.415 ± 0.050	0.856 ± 0.022
	ECG-FM	0.812 ± 0.010	0.808 ± 0.014	$\mathbf{0.984 \pm 0.002}$
	SleepXA	$\mathbf{0.834 \pm 0.013}$	$\mathbf{0.823 \pm 0.022}$	0.984 ± 0.003

increases AF-F1 to 0.743 and Ch-F1 to 0.732. After FT, the differences become smaller, although the jointly initialized model retains the highest Ch-F1 and AF-F1.

The pronounced LP improvement is particularly informative because the pretrained representation remains frozen during downstream adaptation. These results indicate that PSG-based composition preserves the modality-specific cardiac knowledge inherited from the ECG foundation model, while integrating it with complementary neural representations.

A.4.6 Robustness to Physiological Foundation-Model Backbones

Building on the controlled initialization study above, we examine whether the benefit of pretrained physiological initialization generalizes across different constituent foundation models. We consider LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025) as EEG backbones and CSFM (Gu et al., 2026) and ECGFounder (Li et al., 2025) as ECG backbones, resulting in four EEG–ECG backbone combinations. For each pair, pretrained initialization is compared with an architecture-matched Scratch counterpart under the same pre-training and evaluation protocol. All experiments use WSC sleep staging under LP with the same fixed random seed.

As shown in Table A.13, pretrained initialization improves over Scratch for all four backbone combinations. The gains range from 1.98 to 6.12 percentage points in accuracy and from 3.09 to 11.24 points in MF1. The default LaBraM+CSFM configuration reaches 85.16% accuracy, 76.09% MF1, and $\kappa = 0.765$, while the alternative backbone combinations also benefit consistently from pretrained initialization. The class-wise results show that these gains are not confined to the dominant N2 stage. For example, with CBraMod+ECGFounder, N1 F1 increases from 28.02% to 47.11% and REM F1 from 60.90% to 85.35%, compared with a smaller change in N2 F1 from 86.01% to 88.51%. Pronounced improvements in REM recognition are also observed across the other backbone combinations.

These consistent gains across heterogeneous EEG and ECG backbone pairs show that SleepXA is not tied to the default LaBraM+CSFM configuration, although the absolute performance still depends on the quality and compatibility of the constituent pretrained representations.

A.4.7 Detailed Comparison of Multimodal Integration Strategies

We provide the complete LP and FT results for the multimodal integration comparison on WSC sleep staging and CinC2017 atrial-fibrillation recognition. The WSC experiments evaluate robustness across different PSG sensor configurations, while CinC2017 examines whether the learned representations transfer beyond the sleep domain to an independent ECG recognition task.

We compare SleepXA with representative alternatives spanning direct feature fusion, attention-based interaction, contrastive alignment, and common-specific representation learning. Specifically, we consider concatenation followed by an MLP fusion module, cross-attention fusion (Tsai et al., 2019), CLIP-style contrastive alignment (Radford et al., 2021), MindMix-style alignment (Liu et al., 2026), and DecAlign-style alignment (Qian et al., 2026). All strategies produce a fixed-dimensional downstream representation and are evaluated using the same prediction head and downstream protocol. The physiological encoders, pre-training data, downstream splits, and optimization settings are otherwise kept consistent.

Table A.13: **Robustness to different EEG and ECG foundation-model backbones on WSC sleep staging.** Δ denotes the improvement of pretrained initialization over Scratch; bold indicates the best result within each setting.

EEG	ECG	Initialization	Overall performance			Class-wise F1					Improvement	
			Acc.	MF1	κ	W	N1	N2	N3	REM	Δ Acc.	Δ MF1
LaBraM	CSFM	Scratch	**81.40**	**73.00**	**0.717**	**85.61**	**46.07**	88.34	**63.33**	**81.63**	–	–
		Pretrained	**85.16**	**76.09**	**0.765**	**87.27**	**49.32**	**91.04**	**65.28**	**87.52**	+3.76	+3.09
LaBraM	ECGFounder	Scratch	80.38	68.01	0.692	84.37	32.37	**88.48**	61.59	73.27	–	–
		Pretrained	82.36	74.28	0.733	86.39	47.38	88.51	64.17	84.93	+1.98	+6.27
CBraMod	ECGFounder	Scratch	76.30	63.04	0.627	80.94	28.02	86.01	59.35	60.90	–	–
		Pretrained	82.42	74.28	0.735	86.63	47.11	88.51	63.79	85.35	+6.12	+11.24
CBraMod	CSFM	Scratch	76.51	62.30	0.625	81.21	27.60	86.29	56.37	60.02	–	–
		Pretrained	80.91	73.21	0.714	86.20	47.53	87.58	62.25	82.48	+4.40	+10.91

Note. Scratch and Pretrained use identical backbone architectures and differ only in whether the EEG and ECG encoders are initialized from their corresponding foundation-model checkpoints. LaBraM+CSFM is the default SleepXA backbone combination. All downstream results are obtained under LP.

Sleep staging under different sensor configurations. We first evaluate the integration strategies on WSC sleep staging under three input configurations. *Full* uses all seven PSG modalities considered in this study. *Core* uses EEG, EOG, and EMG; and *Wearable* is restricted to EEG only. This comparison tests whether an integration strategy learned from multimodal PSG remains effective as the available physiological inputs are progressively restricted.

As shown in Table A.14, SleepXA achieves the strongest overall performance across the three sensor configurations under both LP and FT. The advantage is particularly consistent under LP, where the pretrained representation remains frozen. For Full PSG, SleepXA reaches an MF1 of 76.09%, compared with 75.47% for the strongest alternative. The corresponding MF1 values are 75.08% under the Core configuration and 69.60% under EEG-only input. The advantage therefore persists as the available sensing configuration is reduced, suggesting that the learned representation does not depend on access to the complete PSG montage. After FT, SleepXA also achieves the highest MF1 across all three configurations, although the gaps to several alternative integration strategies become smaller. This pattern indicates that the choice of multimodal alignment has its clearest effect on the quality of the pretrained frozen representation, while downstream end-to-end optimization can partially compensate for differences introduced during multimodal pre-training.

Cross-context ECG transfer. We next evaluate the different multimodal fusion and alignment strategies on the CinC2017 AF task (Table A.15). Unlike WSC sleep staging, this setting transfers representations learned from nocturnal multimodal PSG to an independent single-lead ECG task outside the sleep context, providing a more stringent test of whether different integration strategies preserve transferable modality-specific physiological information.

Under LP, SleepXA achieves the highest Ch-F1 (73.18%) and AF-F1 (74.30%), exceeding the strongest alternative by 12.38 and 12.92 percentage points, respectively. Cross-Attention obtains a higher AF-AUROC (93.92% versus 91.15% for SleepXA), showing that the advantage of SleepXA is more pronounced for F1-based rhythm discrimination than for AUROC. After FT, the differences among the integration strategies become substantially smaller. Cross-Attention achieves the highest Challenge-F1 and AF-F1, while SleepXA remains close to the best result across all three metrics. The contrast between LP and FT is informative: differences in multimodal pre-training are most apparent when the learned representation is frozen, whereas sufficiently strong task-specific supervision can substantially reduce these differences during end-to-end adaptation. Taken together with the WSC results, these experiments suggest that the conditional common–specific alignment used by SleepXA primarily improves the quality and transferability of the pretrained representation. Its benefits persist under reduced sensor availability within the sleep domain and are especially apparent when transferring to a physiologically distinct single-modality task.

A.4.8 Detailed Ablation of Major Alignment Objectives

The main text summarizes the effects of the three major SleepXA alignment mechanisms using representative metrics. Here, we report the complete LP and FT results for WSC sleep staging, WSC

Table A.14: **Comparison of integration strategies across input configurations on WSC sleep staging. Bold** and underlined values denote the **best** and second-best results within each configuration and adaptation setting.

Method	Config.	Linear Probing			Full Fine-tuning		
		Acc.	MF1	κ	Acc.	MF1	κ
SleepXA (Ours)	Full	**85.16**	**76.09**	**0.765**	**85.95**	**77.61**	**0.780**
MindMix-style Alignment	Full	83.43	74.87	0.744	83.86	76.18	0.754
Cross-Attention Fusion	Full	83.74	75.18	0.748	84.57	76.63	0.763
Concatenation + MLP Fusion	Full	83.75	74.46	0.747	85.16	76.57	0.769
DecAlign-style Alignment	Full	83.56	75.47	0.746	84.60	76.36	0.765
CLIP-style Contrastive Alignment	Full	84.24	75.26	0.756	83.26	75.84	0.754
SleepXA (Ours)	Core	**84.46**	**75.08**	**0.755**	**85.50**	**77.80**	**0.778**
MindMix-style Alignment	Core	82.68	73.79	0.740	83.11	76.38	0.747
Cross-Attention Fusion	Core	83.09	74.44	0.738	83.87	75.11	0.748
Concatenation + MLP Fusion	Core	83.11	73.46	0.737	84.71	77.20	0.767
DecAlign-style Alignment	Core	82.82	73.45	0.731	83.67	75.87	0.759
CLIP-style Contrastive Alignment	Core	82.89	73.87	0.732	82.99	74.80	0.741
SleepXA (Ours)	Wearable	**81.02**	**69.60**	**0.701**	**83.41**	**75.07**	**0.747**
MindMix-style Alignment	Wearable	78.18	66.85	0.663	82.00	72.71	0.721
Cross-Attention Fusion	Wearable	78.16	67.62	0.662	82.37	73.15	0.729
Concatenation + MLP Fusion	Wearable	77.79	66.62	0.658	80.24	71.00	0.717
DecAlign-style Alignment	Wearable	79.98	69.03	0.694	81.18	72.15	0.724
CLIP-style Contrastive Alignment	Wearable	80.19	69.37	0.697	80.42	71.31	0.717

Table A.15: **Cross-context transfer on CinC2017 AF task.** Results are reported as percentages. Best and second-best results are shown in bold and underlined, respectively.

Method	Linear Probing			Full Fine-tuning		
	Ch-F1	AF-F1	AF-AUROC	Ch-F1	AF-F1	AF-AUROC
SleepXA (Ours)	**73.18**	**74.30**	91.15	83.35	82.30	98.40
MindMix-style Alignment	54.35	51.83	84.90	83.29	82.22	98.32
Cross-Attention Fusion	60.80	61.38	**93.92**	**83.73**	**83.17**	**98.41**
Concatenation + MLP Fusion	55.97	49.79	84.49	83.22	82.29	**98.41**
DecAlign-style Alignment	54.07	47.02	85.78	82.24	81.61	98.23
CLIP-style Contrastive Alignment	57.93	54.55	90.92	82.47	81.94	98.38

AHI/OSA prediction, and CinC2017 atrial-fibrillation recognition. These tasks respectively probe general sleep representation, multisystem clinical phenotyping, and cross-context single-modality transfer. For each experiment, we independently remove *condition-aware alignment* (CA), *stage-conditioned specific preservation* (SP), or *condition-relative disentanglement* (CD), while keeping the model architecture, pre-training data, downstream split, and optimization protocol unchanged. The complete SleepXA model uses all three mechanisms. Specifically, *w/o CA* retains only instance-level InfoNCE in the common alignment objective. *w/o SP* removes the specific-preservation objective while retaining the specific branch and all other training objectives. *w/o CD* removes the conditional residual-decoupling objective while retaining all other training objectives.

WSC sleep staging. Table A.16 reports the complete sleep-staging results. Removing any of the three mechanisms reduces accuracy, Macro-F1, and Cohen's κ under both LP and FT. Under LP, the complete model reaches an MF1 of 76.09%, compared with 74.64%, 74.74%, and 74.79% without CA, SP, and CD, respectively. The same pattern persists after FT, where SleepXA achieves 77.61% MF1. The relatively consistent degradation across the three ablations supports their complementary roles in organizing the learned sleep representation.

WSC AHI estimation and OSA severity classification. The effect of the alignment mechanisms becomes more pronounced for the clinical OSA task (Table A.17). Under LP, the complete model achieves the best results across all regression and classification metrics, including an MAE of 3.441, R^2 of 0.864, and a severity MF1 of 75.05%. After FT, removing SP or CD results in larger degra-

Table A.16: **Ablation of major objectives on WSC sleep staging.** Best results within each adaptation setting are shown in **bold**.

Mode	Variant	Acc. (%) ↑	MF1 (%) ↑	κ ↑
LP	w/o CA	83.90	74.64	0.749
	w/o SP	84.20	74.74	0.750
	w/o CD	83.98	74.79	0.751
	SleepXA	**85.16**	**76.09**	**0.765**
FT	w/o CA	85.02	76.25	0.765
	w/o SP	85.10	76.10	0.766
	w/o CD	84.87	76.05	0.763
	SleepXA	**85.95**	**77.61**	**0.780**

Table A.17: **Ablation of major objectives on WSC AHI estimation and OSA severity classification.** Best results within each adaptation setting are shown in **bold**.

Mode	Variant	MAE ↓	RMSE ↓	R^2 ↑	Acc. (%) ↑	MF1 (%) ↑
LP	w/o CA	3.623	6.858	0.841	77.11	70.87
	w/o SP	3.791	6.842	0.841	78.71	72.98
	w/o CD	3.756	6.921	0.838	77.51	72.70
	SleepXA	**3.441**	**6.341**	**0.864**	**79.12**	**75.05**
FT	w/o CA	3.361	6.430	0.860	81.53	76.07
	w/o SP	4.124	7.239	0.822	75.50	69.59
	w/o CD	3.775	7.115	0.828	78.31	71.32
	SleepXA	**3.120**	**6.008**	**0.878**	**82.73**	**78.46**

dation: MF1 decreases from 78.46% to 69.59% and 71.32%, respectively, while the corresponding regression errors also increase. Both variants also obtain lower severity MF1 after FT than under LP. One possible explanation is that SP and CD help preserve and organize modality-specific information relevant to OSA, including SpO_2 derived desaturation patterns and respiratory cues. Removing these constraints during pre-training may weaken the effective use of such information during subsequent adaptation, potentially contributing to the degradation observed after FT.

CinC2017 AF recognition. Table A.18 further evaluates the same mechanisms under cross-context transfer from nocturnal multimodal PSG to single-lead ECG. Under LP, the complete SleepXA objective provides a clear advantage in F1-based recognition, reaching a Ch-F1 of 73.18% and an AF-F1 of 74.30%. Removing CA, SP, or CD reduces AF-F1 to 59.50%, 57.44%, and 62.60%, respectively. The highest LP AF-AUROC is obtained without CD, showing that the benefit is not uniform across all metrics. After FT, however, the differences in F1 become very small. The variant without CD marginally exceeds the complete model (82.38% versus 82.30% AF-F1), with similarly small differences in Challenge-F1 and AF-AUROC between these two variants. On this ECG-only task, task-specific fine-tuning may adapt the ECG representation sufficiently to compensate for differences introduced during multimodal pre-training. This offers a possible explanation for the larger LP gaps and the similar F1 performance after FT, suggesting that the benefit of these objectives for AF-F1 is more evident when the pretrained representation remains frozen.

Across the three representative tasks, the complete objective shows the most consistent advantage under LP. These benefits persist after full fine-tuning for WSC sleep staging and OSA prediction, whereas the F1 differences in AF recognition largely diminish. These findings support the contributions of CA, SP, and CD to learning transferable representations.

A.5 Robustness under Restricted Modalities and Supervision

We further evaluate SleepXA under two practical constraints that are not fully explored in the main text: restricted modality availability and limited downstream supervision. These experiments examine whether the learned representation remains useful when the downstream sensor configuration differs from the full PSG setting or when only a small number of labeled subjects are available.

Table A.18: **Ablation of major objectives on CinC2017 AF recognition.** Ch-F1 and AF-F1 denote the F1 scores for the challenge metric and atrial fibrillation class, respectively. Best results within each adaptation setting are shown in **bold**.

Mode	Variant	Ch-F1 (%) ↑	AF-F1 (%) ↑	AF-AUROC (%) ↑
LP	w/o CA	58.79	59.50	90.83
	w/o SP	59.26	57.44	90.85
	w/o CD	62.70	62.60	**92.37**
	SleepXA	**73.18**	**74.30**	91.15
FT	w/o CA	82.60	81.81	97.43
	w/o SP	82.76	82.21	97.48
	w/o CD	**83.53**	**82.38**	**98.43**
	SleepXA	83.35	82.30	98.40

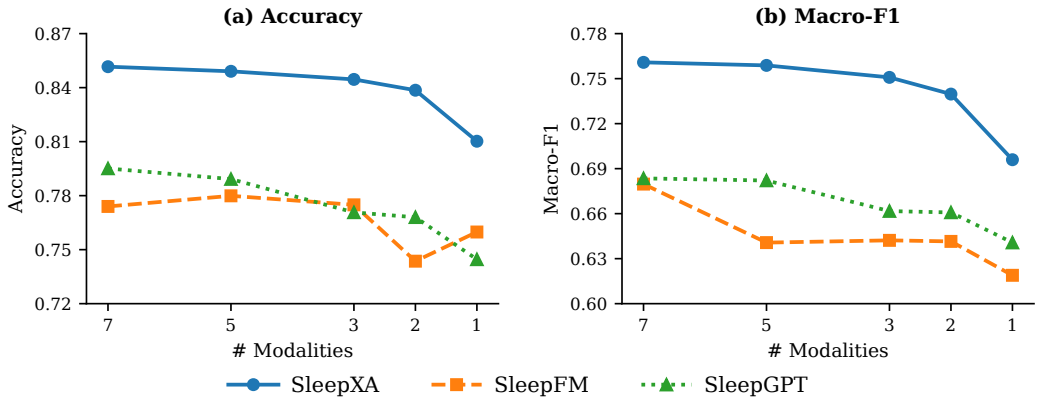


Figure A.6: **Robustness to restricted modality availability.** WSC sleep staging under LP with progressively reduced physiological inputs, from the full seven-modality PSG configuration to EEG-only evaluation.

A.5.1 Robustness to Restricted Modality Availability

We examine the effect of reducing the number of available physiological modalities at downstream evaluation. Experiments are conducted on WSC sleep staging under LP so that the pretrained backbone remains fixed, and the comparison primarily reflects the quality of the learned representations. Starting from the full seven-modality PSG configuration (EEG, EOG, EMG, ECG, airflow, ABD, and SpO_2), we progressively reduce the available inputs down to an EEG-only setting. For comparison, we include SleepFM and SleepGPT, which can be evaluated under the same modality configurations as SleepXA, enabling a controlled comparison as sensor availability is reduced. As shown in Fig. A.6, SleepXA maintains strong performance across the evaluated configurations, with gradual degradation as modalities are removed. The representation remains effective even under substantially restricted sensor availability, including the EEG-only setting. These results indicate that SleepXA can exploit complementary multimodal information when available without requiring a fixed full-PSG configuration at downstream evaluation.

A.5.2 Pre-training Scale under Limited Downstream Supervision

We next investigate how the scale and cohort diversity of sleep-domain pre-training interact with the amount of labeled downstream data. All experiments use LP. For each downstream task, the original validation and test partitions are kept fixed while the number of labeled training subjects is varied. We consider five labeling ratios: 5%, 10%, 25%, 50%, and 100%. For WSC, these correspond to 45, 90, 225, 450, and 899 labeled training subjects, respectively, with 111 subjects each in the validation and test sets. Each experiment is repeated using five downstream random seeds $\{40, 41, 42, 43, 44\}$. For each seed, subjects are randomly ordered once, and progressively larger training subsets are constructed from the same ordering, ensuring that the smaller labeled subsets are nested within the larger ones.

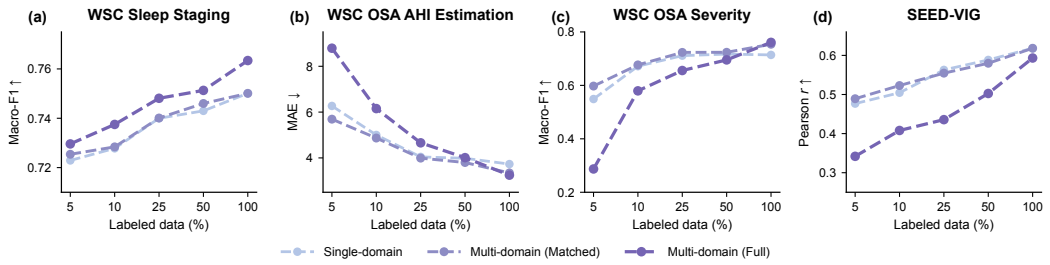


Figure A.7: **Effect of pre-training diversity and scale under limited downstream supervision** LP performance using 5%, 10%, 25%, 50%, and 100% of the labeled downstream training subjects under single-domain, matched multi-domain, and full multi-domain pre-training. (a) WSC sleep staging, evaluated by Macro-F1; (b) WSC AHI estimation, evaluated by MAE; (c) WSC OSA severity classification, evaluated by Macro-F1; and (d) SEED-VIG vigilance estimation, evaluated by Pearson's r.

We compare three pre-training configurations: *Single-domain* uses the SHHS pre-training cohort; *Multi-domain (Matched)* uses SHHS, MESA, and CFS while matching the total number of pre-training participants to the single-domain setting; and *Multi-domain (Full)* uses all available SHHS, MESA, and CFS pre-training data. The pretrained checkpoint is fixed within each configuration, and downstream seeds affect only subject subsampling and prediction-head optimization. Results are summarized as mean±standard deviation across the five downstream seeds. For SEED-VIG, performance is first averaged across the five subject-level folds within each seed and then aggregated across seeds.

Figure A.7 reveals a task-dependent effect of increasing the scale and diversity of nocturnal PSG pre-training. For WSC sleep staging, full multi-domain pre-training consistently provides the strongest transfer across labeling ratios, with the benefit already visible in the low-label regime. For AHI estimation and OSA severity classification, the effect is more dependent on downstream supervision: the matched multi-domain configuration is often stronger when labels are scarce, whereas the full multi-domain representation becomes more competitive as supervision increases. A different trend is observed for SEED-VIG, where increasing sleep-domain pre-training scale does not produce monotonic improvements. Together, these results suggest that scaling heterogeneous PSG pre-training most reliably benefits downstream tasks that remain closely aligned with nocturnal sleep physiology. For more distant clinical or daytime transfer settings, the usefulness of additional pre-training scale depends more strongly on the amount of task-specific supervision and the degree of domain shift.

A.6 Additional Methodological Details

Instance and sleep-stage alignment. The two terms in Eq. 2 implement contrastive alignment at the instance and sleep-stage levels. For $\mathcal{L}_{\text{inst}}$, synchronized representations from two available modalities of the same epoch form a positive pair, while eligible epochs from other participants in the same cohort serve as negatives. The objective is computed bidirectionally between each valid modality pair.

For $\mathcal{L}_{\text{stage}}$, we maintain learnable modality-invariant prototypes $\mathbf{p}_y$ for the five sleep stages. Each common representation is contrasted against these prototypes, with the prototype of its sleep stage as the positive. In the reverse direction, each prototype is contrasted against the corresponding batch stage centroids computed within each cohort and modality. These two directions are combined to obtain $\mathcal{L}_{\text{stage}}$.

$$\mathcal{L}_{\text{inst}} = -\mathbb{E}_{i,m,n} \log \frac{\exp!\left(\text{sim}(\mathbf{c}_i^m, \mathbf{c}_i^n)/\tau\right)}{\exp!\left(\text{sim}(\mathbf{c}_i^m, \mathbf{c}_i^n)/\tau\right) + \sum_{j \in \mathcal{N}_i^n} \exp!\left(\text{sim}(\mathbf{c}_i^m, \mathbf{c}_j^n)/\tau\right)}. \tag{8}$$

where $\mathcal{N}_i^n$ denotes eligible negatives from other participants in the same cohort. The objective is applied bidirectionally across valid modality pairs.

Structural alignment. To implement Eq. 3, we group common representations by cohort and sleep stage, and compare each modality pair using epochs in which both modalities are available. For each valid group, D_{struct} combines the cosine distance between paired representations with differences in their feature-wise means and standard deviations. All comparisons are computed on normalized common representations and aggregated across valid modality pairs, stages, and cohorts.

For a valid cohort–stage group $\mathcal{G}_{d,y}^{m,n}$, we define

$$D_{\text{pair}} = \mathbb{E}_{i \in \mathcal{G}_{d,y}^{m,n}} \left[1 - \cos(\mathbf{u}_i^m, \mathbf{u}_i^n)\right], \; D_{\text{struct}} \quad = \frac{1}{2}\left(D_{\text{pair}} + D_{\text{stat}}\right), \tag{9}$$

where $\mathbf{u}_i^m$ denotes the parameter-free LayerNorm of $\mathbf{c}_i^m$, and D_{stat} measures discrepancies in the feature-wise means and standard deviations between the two conditional representation distributions.

Modality-specific preservation. For each cohort and sleep stage, we consider the modalities observed in that group and normalize their specific representations and learnable prototypes. In the representation-to-prototype direction, each $\mathbf{s}_i^m$ is contrasted against the prototypes of the observed modalities for stage y_i, with $\mathbf{q}_{y_i}^m$ as its positive target. In the prototype-to-centroid direction, we first compute a centroid for each observed modality within the cohort-stage group. Each prototype is then contrasted against these centroids, with the centroid of its own modality as the positive target. Both directions use temperature-scaled cross-entropy and are applied only when at least two modalities are observed in the group. The resulting terms are aggregated across valid stages and cohorts. We set $\alpha_{\text{stage}} = 0.5$, $\lambda_{\text{spec}} = \lambda_{\text{dis}} = 0.1$, and $\lambda_{\text{rec}} = 1$ throughout.