technologies

MDPI

Article

A Sleep Staging Method Based on Cardiopulmonary Signals Using a Unified Multimodal Model

Lin Guo [1], Yuhang Yin [1], Chen Wang [1], Hongyu Chen [1], Qinghua Cui [2] and Xiangkui Wan [1,*]

[1] School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China; 102410236@hbut.edu.cn (L.G.)
[2] School of Sports Medicine, Wuhan Sports University, Wuhan 430079, China
* Correspondence: xkwan@hbut.edu.cn

Abstract

Sleep staging based on PSG is largely confined to clinical settings, while home-based sleep monitoring often faces the challenges of insufficient unimodal information and missing modalities. Aiming to overcome these challenges, this paper proposes a unified multimodal model for sleep staging based on cardiopulmonary signals. First, a heterogeneous multi-scale feature encoder with long and short branches is adopted to adapt to the cross-modal heterogeneity of ECG and THX. It combines a Transformer encoder and a Dilated CNN to complete feature fusion and temporal modeling. Subsequently, the unified model adaptively handles flexible modality combinations by introducing global context via a modal feature alignment strategy, which is built upon a framework consisting of a bimodal global branch and unimodal dedicated branches. On the SHHS dataset, the proposed model achieved Cohen's kappa coefficients of 0.7547, 0.7121, and 0.7305 for four-stage sleep classification under ECG+THX, ECG-only, and THX-only inputs, respectively, demonstrating consistent improvements over three separately trained individual models. Furthermore, the model exhibits robust generalization performance on the P2018 external dataset and across samples with different severity levels of SDB. This work establishes a reliable algorithmic baseline for unobtrusive, long-term home sleep monitoring with missing modalities.

Keywords: sleep staging; cardiopulmonary signals; modal feature alignment; multi-scale feature encoder; home sleep monitoring

1. Introduction

Sleep is a fundamental physiological process vital for maintaining human health, with its quality being intricately linked to immune function, cognitive performance, and cardiovascular well-being [1]. Sleep staging is a core component of sleep quality assessment and the clinical diagnosis of sleep disorders. According to the criteria published by the American Academy of Sleep Medicine (AASM), human sleep is categorized into five stages: Wake, Non-Rapid Eye Movement (NREM) stages 1, 2, and 3 (N1, N2, and N3), and Rapid Eye Movement (REM) [2]. Manual sleep staging based on polysomnography (PSG) remains the clinical gold standard [3]. However, PSG requires simultaneous acquisition of multiple physiological signals, including electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG), resulting in complicated procedures, high costs, and the well-known first-night effect. These limitations restrict its applicability for long-term and naturalistic home sleep monitoring [4].

Academic Editor: Loris Nanni

Received: 5 June 2026
Revised: 9 July 2026
Accepted: 16 July 2026
Published: 17 July 2026

Driven by the demand for automated sleep staging, deep learning methods have achieved remarkable progress by replacing manual feature engineering and expert scoring [5,6]. Representative approaches based on convolutional neural networks (CNNs) [7,8], recurrent neural networks (RNNs) [9–12], and Transformer architectures [13,14] have demonstrated excellent performance, particularly when using EEG-centered PSG recordings. Nevertheless, these methods still rely on cumbersome PSG systems and therefore cannot fully satisfy the practical requirements of unobtrusive long-term home monitoring [15]. Clinical studies have shown that cardiopulmonary signals, including electrocardiogram (ECG) and respiratory activity, exhibit characteristic physiological variations across sleep stages, making them promising alternatives for portable sleep staging systems [16].

With the rapid development of wearable sensors and millimeter-wave radar, portable sleep monitoring has become increasingly feasible. However, practical home monitoring remains vulnerable to sensor detachment, body movement, environmental interference, and hardware instability, often resulting in incomplete physiological recordings or missing modalities [17–20]. Among available physiological signals, ECG and thoracic respiratory signal (THX) preserve rich cardiopulmonary information, provide relatively high signal quality, and are widely supported by existing wearable devices. Consequently, developing sleep staging algorithms that remain reliable under missing-modality conditions has become an important challenge for practical home sleep monitoring.

To improve robustness against incomplete physiological recordings, various missing modality learning strategies have recently been proposed. Representative approaches include coordinated multimodal representation learning [21], cross-modal completion through "imagine-compare" learning [22], masked modality reconstruction using autoencoders [23], and multimodal knowledge distillation [24]. Meanwhile, mature temporal modeling backbones, such as Transformer networks [14] and Dilated CNNs [8], have demonstrated strong capability for sequence representation learning. However, most existing missing-modality methods are primarily designed for homogeneous physiological modalities with strong correlations, such as multi-channel EEG. Their underlying assumption is that one modality can be reconstructed from or strongly constrained by another. This assumption may not hold for heterogeneous cardiopulmonary signals. ECG and THX originate from different physiological mechanisms and exhibit substantially different temporal dynamics and feature distributions. Direct modality reconstruction or strong feature-level alignment may therefore suppress modality-specific information and reduce representation flexibility.

To address these challenges, this study proposes a unified missing-modality learning framework specifically designed for heterogeneous cardiopulmonary sleep staging. The proposed framework adopts well-established temporal backbone architecture and centers on a unified training strategy that enables a single model to operate under arbitrary modality availability. Specifically, a dual-branch heterogeneous encoder is designed to preserve modality-specific physiological representations, while a weakly constrained modal feature alignment strategy encourages semantic consistency without enforcing feature-level equivalence. Consequently, the model can be trained once and directly deployed under dual-modality, ECG-only, or THX-only conditions without retraining, thereby reducing deployment complexity while maintaining stable sleep staging performance.

The main contributions of this work are summarized as follows.

(1) We formulate heterogeneous missing-modality sleep staging as a unified learning problem and develop a unified framework capable of supporting arbitrary combinations of ECG and thoracic respiration using a single trained model.

(2) We propose a dual-branch heterogeneous encoder together with a weakly constrained modal feature alignment strategy, enabling effective cross-modal knowledge transfer while preserving modality-specific physiological representations.
(3) Extensive experiments on two large-scale public sleep datasets demonstrate that the proposed framework achieves stable performance across different modality configurations while simplifying practical deployment for long-term home sleep monitoring.

2. Materials and Methods

2.1. Datasets and Preprocessing

To develop a sleep staging model with broad population representativeness and strong clinical generalizability, this study utilized two large-scale, publicly available medical datasets that had undergone ethical review: the Sleep Heart Health Study (SHHS) dataset and the 2018 PhysioNet Sleep Challenge (P2018) dataset [25,26]. The SHHS dataset was initiated by the National Heart, Lung, and Blood Institute (NHLBI). It has been reviewed and approved by the relevant institutional ethics committee, with all participants providing informed consent. The dataset is accessible to researchers worldwide for non-commercial purposes upon authorization. Similarly, the P2018 dataset is an open medical resource released by the PhysioNet platform. It has undergone ethical review and completed the informed consent process, remaining available for free, non-commercial research use. This study strictly complies with the data usage agreements and research ethics guidelines of both platforms. All data were de-identified, and this retrospective analysis, involving no additional human trials, was exempted from review by the host institution's ethics committee. Throughout the study, the ethical requirements of the Declaration of Helsinki were strictly followed [27].

The labeling system in this study is based on the sleep scoring criteria published by the AASM, simplifying the standard 5-category system into a 4-category model optimized for out-of-hospital monitoring: Wake, Light (N1+N2), Deep (N3), and REM [28]. To address noise and inconsistent durations in the raw data, a preprocessing workflow consisting of filtering, normalization, resampling, and interpolation was implemented to generate standardized inputs. First, to eliminate high-frequency interference and low-frequency baseline drift, a 4th-order Butterworth bandpass filter was applied: the passband for ECG was set to 0.5~15 Hz to preserve heart rate features, while the passband for THX was set to 0.1~4.0 Hz to isolate respiratory components. Second, signals were unit-normalized using the Z-score method to eliminate dimensional differences across devices. For temporal alignment, signals were segmented and resampled into 30 s epochs: ECG was resampled to 34.13 Hz (1024 samples per epoch) and THX to 8.53 Hz (256 samples per epoch). Finally, to meet the requirements for batch training, all sequences were standardized to a length of 1200 epochs (corresponding to 10 h) via zero-padding.

2.2. Network Architecture Design

As a core component of the model, the multi-scale feature encoder is designed to simultaneously capture both instantaneous physiological details and long-term rhythmic features, accounting for the cross-modal heterogeneity across different signals [29]. Given the significant differences between ECG and THX in terms of sampling rates and frequency characteristics, this study designed an encoding structure comprising a basic CNN and a long-short dual-branch CNN. For ECG, the encoder employs a deep convolutional backbone to progressively extract low-frequency fundamental features, which are subsequently split into long and short branches. The short branch utilizes small convolutional kernels ($k = 3$) to focus on transient fluctuations between adjacent heartbeats. The long branch employs large kernels ($k = 9$) to expand the receptive field, thereby capturing cross-cycle

trends in heart rate variability. For THX, given its low-frequency characteristics, the encoder reduces the depth of the backbone and adjusts the long branch's kernels to k = 7, focusing on extracting the periodic rhythms of nocturnal breathing and the contours of events such as apnea.

Sleep structure exhibits both local correlations between adjacent cycles and long-term state evolution across the entire night. Relying solely on a multi-scale feature encoder makes it difficult to fully capture the global evolutionary patterns of sleep sequences. To this end, this study draws on the temporal modeling approach proposed by Davidson et al. [30], incorporating a cascaded structure comprising a Transformer encoder and a Dilated CNN following the multi-scale encoder. Specifically, the Transformer encoder is configured with 8 attention heads and a 512-dimensional hidden layer. It utilizes a multi-head self-attention mechanism to perform cross-modal semantic aggregation of ECG and THX data, extracting a global state representation for each epoch via the classification token (CLS Token). The Dilated CNN is configured with a kernel size of 7 and a spatial dimension of 128, with dilation ratios increasing exponentially according to {1, 2, 4, 8, 16, 32}, achieving coverage of long temporal receptive fields without increasing the number of parameters. The complete network architecture is shown in Figure 1.

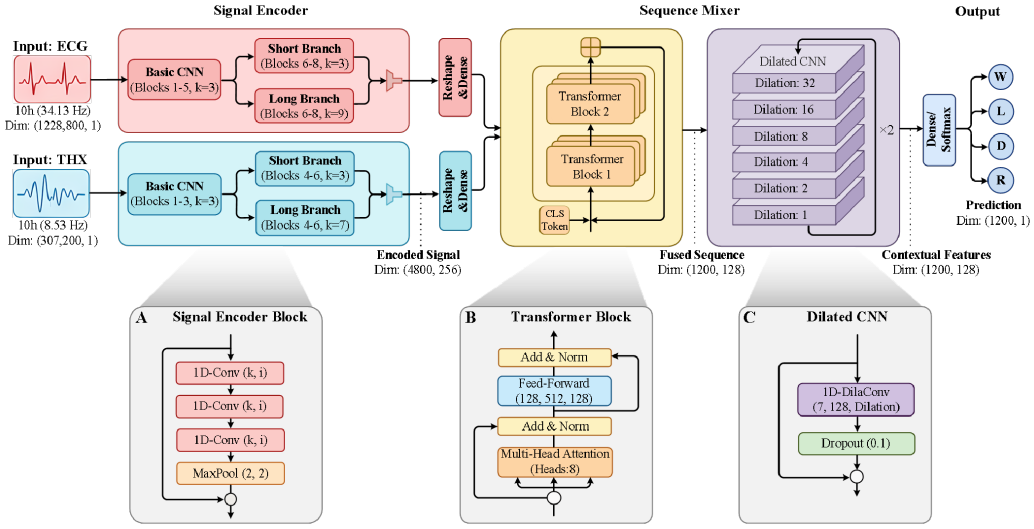

Figure 1. Architecture of the proposed network.

In the figure, i denotes the spatial dimension of the output from each layer of the multi-scale feature encoder. In the ECG encoder, i takes the values 16, 16, 32, 32, 64, 64, 128, 256, while in the THX encoder, i takes the values 16, 32, 64, 64, 128, 256. The input signals are processed into high-order feature vectors of size (4800, 256). To align with the sleep scoring rules for a 30 s epoch, the encoder's output features are first reshaped into a (1200, 1024) vector, then reduced via a fully connected layer (Dense) to (1200, 256), providing standardized feature inputs. The sequence fuser correlates the input features across the temporal dimension. Finally, contextual features are reduced to a dimension of (1200, 4), and the category with the highest probability is determined using a softmax function, with W, L, D, and R denoting Wake, Light, Deep, and REM respectively.

2.3. Model Training Framework

Formally, this framework adopts a multi-branch collaborative training paradigm inspired by knowledge distillation. However, its core mechanism is not the traditional strong-constraint distillation [31], but rather a weak-constraint modal feature alignment. Its underlying mechanism can be defined as follows: addressing the cross-modal feature space

discrepancies of heterogeneous physiological signals, it uses consistency constraints at the regularization level to guide unimodal features to progressively align with the bimodal global feature space, which contains richer complementary information. This process allows the absorption of global context to enhance feature robustness while entirely preserving the inherent discriminative characteristics of unimodal signals. Unlike traditional distillation, which forces the student model to replicate the teacher's output distribution, this strategy employs bimodal features merely as weak-supervision guidance signals, avoiding the disruption of the inherent representations of heterogeneous signals.

The framework utilizes a single unified model architecture, comprising one bimodal global guidance branch (ECG+THX, denoted as ALL) and two unimodal target branches (ECG and THX). All branches fully share the multi-scale feature encoder, the Transformer encoder, the Dilated CNN, and the same classification head; they differ only in their input modalities. Specifically, the bimodal branch is responsible for aggregating the cross-modal complementary semantics of the cardiopulmonary signals to generate baseline features rich in global context. Conversely, the unimodal branches absorb this global contextual information through weak constraints while retaining their modality-specific features.

To balance global guidance and the preservation of modality specificity, this framework employs a hybrid loss function consisting of a classification loss and a cross-modal consistency constraint loss. The total loss is calculated as shown in Equation (1):

$$L_{total} = L_{cls} + \lambda \times L_{cons} \quad (1)$$

In this study, the balancing coefficient λ is initially set to 0.5, strictly controlling the constraint strength at the regularization level. The primary rationale is that ECG is a high-frequency electrical signal, whereas THX is a low-frequency mechanical signal; thus, their feature spaces possess fundamental differences. If λ is set too high, strong consistency constraints would forcefully pull two distinct feature spaces into the same distribution, destroying the inherent discriminative information of each modality and leading to a decline in overall model performance. In contrast, weak constraints at the regularization level merely guide the unimodal features to absorb the global context while entirely preserving their modality-specific representations.

The classification loss L_{cls} employs the cross-entropy (CE) loss to simultaneously optimize the error between the predictions of the three branches and the ground-truth sleep labels. An equal weight of 1/3 is assigned to each branch to prevent performance degradation in any single branch, as calculated in Equation (2):

$$L_{cls} = \frac{1}{3}[CE(\hat{y}_{ALL}, y) + CE(\hat{y}_{ECG}, y) + CE(\hat{y}_{THX}, y)] \quad (2)$$

The cross-modal consistency constraint loss L_{cons} is the core of the weak constraint mechanism. It utilizes the mean square error (MSE) to measure the distribution discrepancy between the unimodal features and the bimodal global features. Before calculation, L2 normalization is applied to all features to eliminate scale differences, as shown in Equation (3):

$$L_{cons} = \frac{1}{2}[MSE(norm(z_{ECG}), norm(z_{ALL})) + MSE(norm(z_{THX}), norm(z_{ALL}))] \quad (3)$$

The training phase adopts an end-to-end single-pass training mode, as illustrated in Figure 2. For the same sleep sample, three types of inputs are generated: bimodal (ECG+THX), unimodal ECG, and unimodal THX. Three independent forward passes are executed on the same model instance to obtain the corresponding global features (z_{ALL}, z_{ECG}, z_{THX}) and prediction results ($\hat{y}_{ALL}$, $\hat{y}_{ECG}$, $\hat{y}_{THX}$). Based on the outputs of these three forward passes, the total loss is calculated and backpropagated to synchronously

update the shared parameters across the entire network, eliminating the need to train separate models for different modalities.

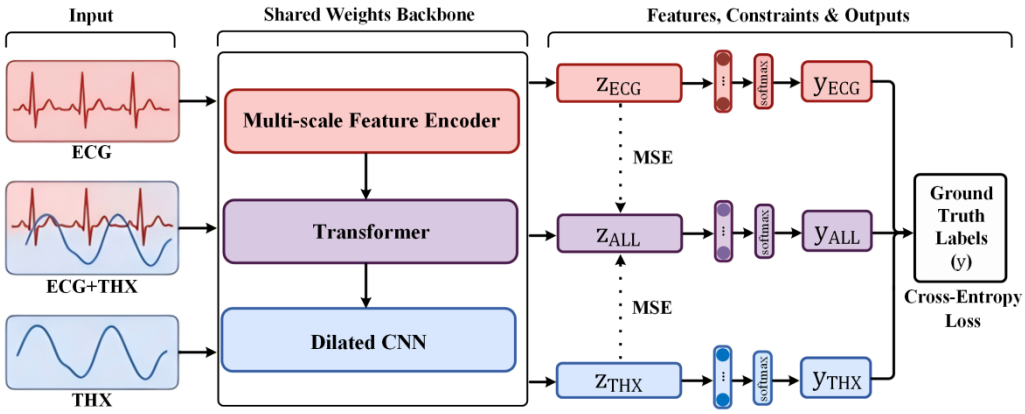

Figure 2. Overall pipeline of the proposed multi-modal training framework.

During the inference phase, the model automatically detects the number of modality channels input. If the input has 2 channels (ECG+THX), it outputs the prediction result of the bimodal branch; if the input has 1 channel, it automatically identifies it as the corresponding unimodal signal. Concurrently, the Transformer's self-attention mechanism physically masks the positions of the missing modality, ensuring that the final output features are unaffected by the absent signals, bypassing the need for additional model switching or re-inference.

2.4. Experimental Setup

The experimental platform was configured with an NVIDIA GeForce RTX 3090 GPU and implemented using the PyTorch (1.12.0+cu113) deep learning framework. During training, the batch size was set to 4 with a random seed of 42, combined with a gradient accumulation strategy, this resulted in an effective batch size of 16. The Adam optimizer was selected with an initial learning rate of 1×10^{-3} and a weight decay factor of 1×10^{-2}. The learning rate schedule employed an exponential warm-up and decay strategy, consisting of a 2000-step warm-up followed by exponential decay with a time constant τ of 6000. The maximum number of training epochs was set to 50, with an early stopping patience of 5.

To ensure data quality, this study identified ECG segments with R-wave counts outside the 10~80 range and THX segments where energy in the 0.1~0.6 Hz band accounted for less than 50% as signal quality outliers [32]. Such segments were labeled as artifacts, and low-quality samples where artifacts exceeded 15% of the total sleep time (TST) were excluded. The SHHS dataset originally contained 5596 records; after filtering, 5248 valid records were retained and randomly partitioned at the subject level (subject-independent) into training, validation, and test sets in an 8:1:1 ratio. This strict subject-independent division ensures no data leakage occurs between sets. For the P2018 dataset, 500 independent subject records were randomly selected solely for cross-dataset generalization testing and were not included in model training.

The model distinguishes single-modality inputs via dictionary signal routing and training negative-inference modality masking, masks tail zero-padding features through the embedding invalid signal masking module and Transformer padding mask, and alleviates class imbalance with standard cross-entropy loss. During loss calculation, padding bits were explicitly ignored to avoid gradient generation. This ensures that invalid segments do not interfere with the learning of sleep structures and allows the model to flexibly process

sleep records of varying durations during inference. To enhance model robustness against signal polarity variations introduced by different sensor orientations, each input signal underwent random polarity inversion with a probability of 50% during training, following the data augmentation strategy in [33]. Specifically, the entire waveform was multiplied by -1 while preserving its original temporal order and sleep-stage annotation.

Evaluation metrics included accuracy (ACC), Cohen's Kappa (kappa), and macro-F1 score (MF1), alongside per-class F1 scores. All performance metrics were calculated based on 30 s epoch-by-epoch alignment without temporal smoothing or macro-level matching, ensuring the clinical comparability and rigor of the results [34].

3. Results

3.1. Overall Performance

To evaluate the overall effectiveness of the proposed multimodal alignment approach, we analyzed the quantitative performance of our unified training framework on the held-out SHHS test set for automated sleep scoring. Two experimental configurations were designed for fair comparison under identical backbone networks:

1. Comparative method: Three separate independent models (ECG-only, THX-only, and bimodal ALL) were trained without the proposed modal feature alignment module; each model was optimized exclusively for its single input modality.
2. Proposed method: A single unified multimodal model integrated the modal feature alignment strategy, which adaptively processes three input types including unimodal ECG, unimodal THX, and bimodal ALL.

All evaluation metrics in Table 1 are accompanied by 95% percentile bootstrap confidence intervals, calculated through 2000 iterations of full-night-level resampling with replacement: each replicate randomly draws 524 full-night records from the fixed test set containing 524 independent overnight samples. This nonparametric approach requires no normality assumption, so the resulting intervals generally present naturally slight asymmetry.

Table 1. Overall staging performance of comparative and proposed models.

Method	Modality	ACC (%) (95% CI)	Kappa (95% CI)	MF1 (95% CI)
Comparative method	ALL	83.11 (82.56, 83.60)	0.7510 (0.7429, 0.7585)	0.8075 (0.7998, 0.8148)
	ECG	80.43 (79.82, 81.00)	0.7113 (0.7030, 0.7191)	0.7794 (0.7712, 0.7871)
	THX	81.55 (80.99, 82.08)	0.7271 (0.7193, 0.7344)	0.7905 (0.7826, 0.7981)
Proposed method	ALL	83.33 (82.83, 83.80)	0.7547 (0.7472, 0.7618)	0.8110 (0.8037, 0.8180)
	ECG	80.46 (79.90, 80.99)	0.7121 (0.7043, 0.7196)	0.7806 (0.7729, 0.7880)
	THX	81.72 (81.18, 82.23)	0.7305 (0.7230, 0.7376)	0.7943 (0.7867, 0.8016)

As shown in Table 1, integrating the modal feature alignment strategy delivers consistent but modest performance improvements across all three input modality conditions. The proposed unified model achieves Cohen's kappa of 0.7547 (95% CI: 0.7472, 0.7618), 0.7121 (95% CI: 0.7043, 0.7196) and 0.7305 (95% CI: 0.7230, 0.7376) for full bimodal, ECG-only and THX-only inputs respectively. Under each modality setting, the entire confidence interval of the proposed method shifts upward relative to the baseline, and the interval width is consistently narrower. This indicates that the cross-modal alignment constraint effectively reduces prediction variability and improves result reproducibility. Bimodal input uniformly outperforms both unimodal configurations across all metrics, confirming the complementary value of cardiopulmonary coupling information.

Table 2 provides a detailed per-stage F1 breakdown with individually calculated 95% bootstrap CIs. Under bimodal input, Wake and REM stages obtain high F1 scores (over 91% and 86%) with narrow intervals, reflecting stable recognition. In contrast, deep sleep achieves the lowest average F1 of 64.30% (95% CI: 63.17, 65.37) and the widest interval. This limitation arises from inherent physiological constraints: cardiopulmonary signals cannot directly reflect cortical delta waves, the gold-standard biomarker of deep sleep [9], and the low proportion of deep sleep epochs further increases estimation uncertainty.

Table 2. Per-stage F1 scores across all input modalities.

Method	Modality	Per-Class F1 Score (%) (95% CI)			
		Wake	**REM**	**Light**	**Deep**
Comparative method	ALL	91.24 (90.65, 91.77)	85.80 (85.11, 86.45)	82.45 (81.80, 83.09)	63.63 (62.48, 64.73)
	ECG	88.80 (88.15, 89.41)	81.62 (80.85, 82.34)	79.73 (79.00, 80.43)	61.53 (60.35, 62.65)
	THX	90.09 (89.45, 90.68)	84.03 (83.30, 84.72)	80.76 (80.05, 81.44)	61.36 (60.21, 62.48)
Proposed method	ALL	91.54 (90.99, 92.05)	86.20 (85.56, 86.81)	82.39 (81.78, 82.98)	64.30 (63.17, 65.37)
	ECG	89.08 (88.47, 89.65)	81.76 (81.05, 82.43)	79.69 (79.01, 80.34)	61.70 (60.56, 62.78)
	THX	90.04 (89.44, 90.60)	83.88 (83.18, 84.55)	80.83 (80.16, 81.47)	63.00 (61.89, 64.06)

Despite this epoch-wise classification limitation quantified by per-stage F1 uncertainty, the model retains reliable global sleep cycle characterization. As visualized by hypnogram reconstruction in Figure 3, even for difficult individual recordings with a subject-level Kappa of 0.7198, the model accurately recovers the macro-scale cyclic architecture of full overnight sleep, maintaining high global consistency against clinical PSG annotations.

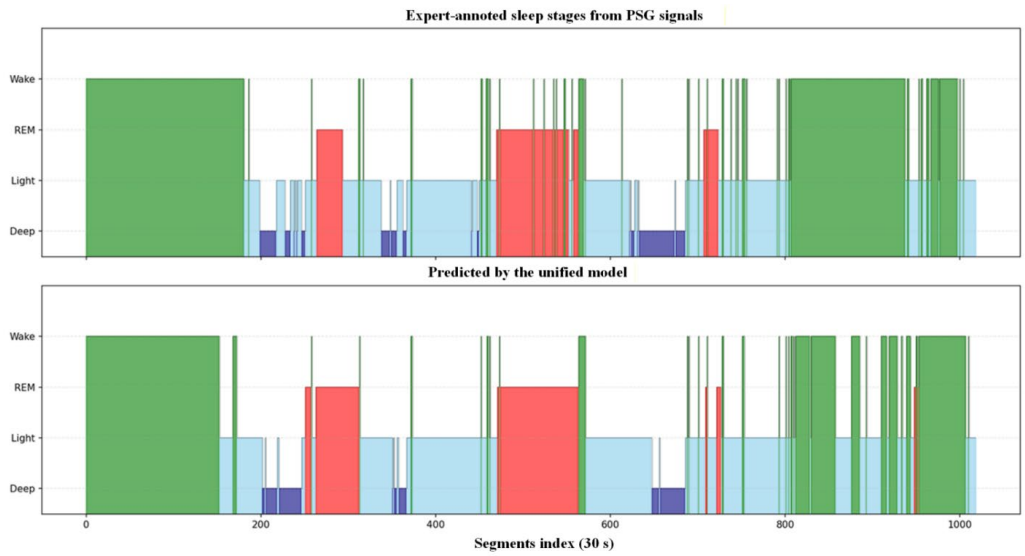

Figure 3. Example sleep hypnograms for a subject. Color blocks vertically correspond to four sleep stages: Wake, REM, Light sleep and Deep sleep.

3.2. Ablation Experiment

To verify the independent contributions and synergistic optimization effects of the core modules in this study, we conducted ablation experiments based on the convergence curves of the Kappa coefficient on the SHHS validation set. The results are illustrated in Figure 4.

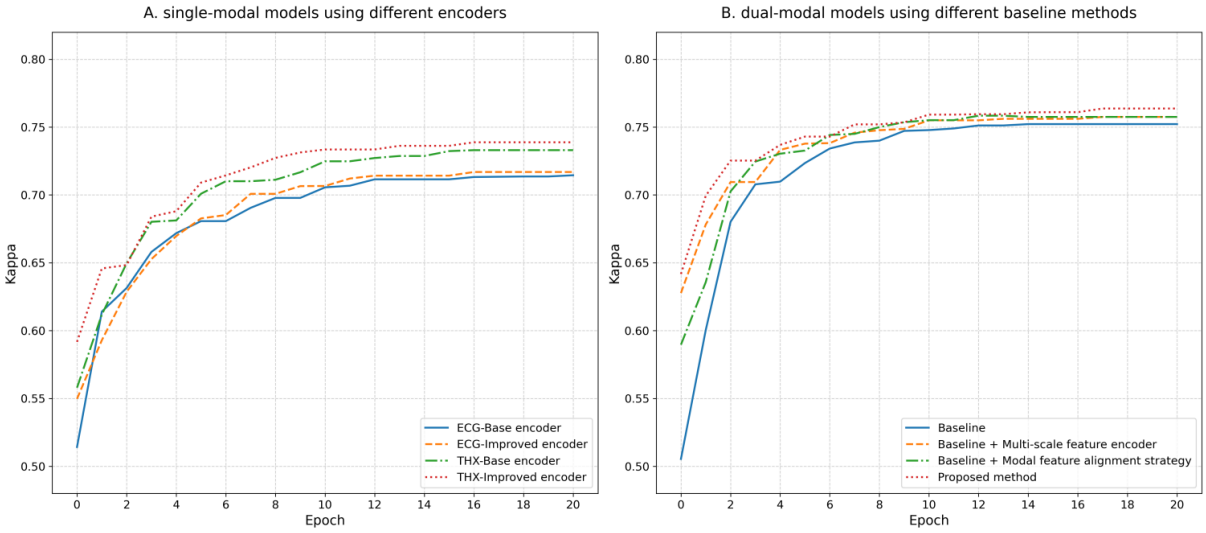

Figure 4. Iteration curves of kappa coefficient of different models on the validation set.

Figure 4A compares the performance of the proposed heterogeneous long-short branch multi-scale feature encoder with a baseline encoder (which lacks a dual-branch structure and utilizes a fixed convolution kernel size of 3) under unimodal ECG and THX signals. The results demonstrate that the multi-scale feature encoder significantly accelerates model convergence and reduces performance fluctuations during training. Simultaneously, it enhances the robustness of feature representation and the upper limit of model performance, effectively mitigating the risk of overfitting.

Figure 4B presents a performance comparison of different module combinations under a bimodal input scenario. The baseline method employs a strategy of base encoder + direct training of three independent models. Building upon this, the multi-scale feature encoder and the modal feature alignment strategy were introduced sequentially, with the results ultimately compared against the complete method proposed in this paper.

The results indicate that following the sequential introduction of these core modules, the model's convergence speed and classification performance exhibit a stepwise improvement. The proposed unified method maintains optimal validation set performance throughout the training process, showing the most pronounced advantages in both convergence efficiency and the final performance ceiling. These findings confirm that multi-scale feature encoding and the modal feature alignment strategy exhibit complementary optimization effects, steadily improving the model's overall capability in sleep scoring.

3.3. Parameter Analysis

To balance model performance with parameter scale, this study analyzed the impact of the number of dual-branch layers within the encoder on both staging accuracy and GPU memory consumption using the SHHS validation set. As shown in Table 3, model performance exhibits an initial increase followed by a decline as the number of layers increases, with performance gain gradually narrowing. Specifically, when the number of dual-branch layers increases from 0 to 3, performance improves steadily while the increase in memory consumption remains minimal. However, once the number of layers exceeds 3, performance begins to decline, and memory consumption rises significantly. Considering both classification performance and deployment efficiency, three dual-branch layers were ultimately selected for the final architecture. Additional implementation parameters and ablation experiments are provided in Appendix A.1.

Table 3. Encoder differences at different scales.

Modality	Index	Number of Branches and Results				
		0	**1**	**2**	**3**	**4**
ECG	Kappa for the validation set	0.7145	0.7162	0.7153	0.7169	0.7154
	GPU Memory (GB)	9.315	9.323	9.329	9.331	9.810
THX	Kappa for the validation set	0.7330	0.7371	0.7367	0.7388	0.7355
	GPU Memory (GB)	4.974	4.979	4.983	5.327	5.974

To evaluate the engineering robustness of the proposed training strategy, we tested model performance across different values of the loss balance coefficient λ (Table 4) on the validation set. Extreme values (λ = 0 or 100) yielded the poorest performance, confirming the necessity of feature alignment while demonstrating that excessively strong constraints disrupt the inherently heterogeneous physical spaces of ECG and THX signals. Among the tested configurations, λ = 0.5 achieved the highest overall performance and was therefore selected as optimal engineering configuration. Furthermore, overall model performance remained highly stable across λ in [0.1, 10], demonstrating that the framework is relatively insensitive to hyperparameter variations and eliminates the need for extensive, dataset-specific tuning.

Table 4. Impact of loss balance coefficient on performance.

Modality	Independent Training	Kappa for Each λ					
		0	**0.1**	**0.5**	**5**	**10**	**100**
ALL	0.7575	0.7614	0.7627	0.7637	0.7631	0.7625	0.7620
ECG	0.7169	0.7198	0.7212	0.7218	0.7214	0.7212	0.7202
THX	0.7388	0.7370	0.7391	0.7398	0.7393	0.7386	0.7384

3.4. Method Comparison

The random masking strategy has recently emerged as an effective approach for improving robustness against missing modalities in multimodal sleep staging [33]. During training, portions of the input modalities are randomly masked, enabling a single model to adapt to both complete and incomplete input conditions. To evaluate the contribution of the proposed weakly constrained modal feature alignment strategy, three training schemes were compared under identical backbone architecture, training settings, and data partitions: direct training, random masking, and the proposed modal feature alignment framework. Their robustness was evaluated under progressively increasing subject-level missing-modality ratios on the SHHS test set.

As illustrated in Figure 5, all three methods exhibit performance degradation as the proportion of missing-modality subjects increases. Nevertheless, the proposed framework consistently maintains the highest Kappa under both ECG-missing and THX-missing scenarios, with a noticeably slower performance decline than the comparison methods. The random masking strategy provides improved robustness compared with direct training by exposing the network to incomplete inputs during optimization. However, because modality interactions are learned only implicitly through random masking, its performance gradually deteriorates as the missing ratio increases. In contrast, the directly trained bimodal model relies heavily on complete paired observations and therefore exhibits the fastest performance degradation when one physiological modality becomes unavailable. These results indicate that the proposed weakly constrained modal feature alignment facilitates more effective cross-modal knowledge sharing while preserving modality-specific

representations, thereby improving robustness under progressively increasing missing-modality conditions.

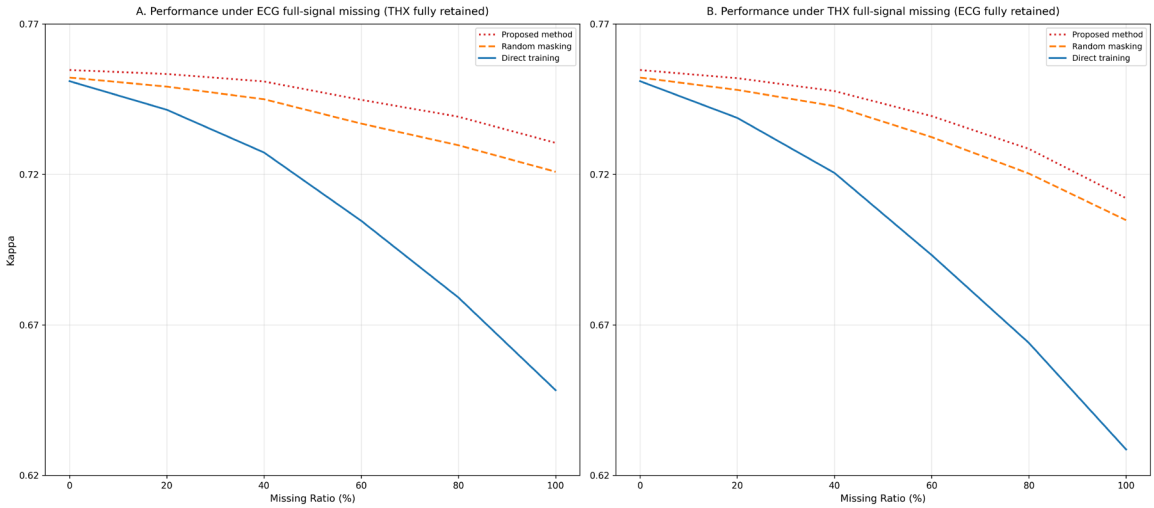

Figure 5. Performance comparison of different training strategies.

To further evaluate training stability, all three models were independently trained using six different random initialization seeds while keeping identical data partitions and training configurations. The six validation Kappa values obtained for each method were treated as independent observations for subsequent statistical analysis. As summarized in Table 5, the proposed framework achieved the highest average validation Kappa (0.7618) while converging consistently within 15–17 epochs across all random seeds. By comparison, the random masking strategy required substantially more training epochs (24–32), whereas direct training produced both the lowest average validation performance and the largest variation across different initializations. Because the absolute performance differences among the three methods were relatively modest, paired Student's t-tests were performed to evaluate whether the observed improvements were statistically reliable. The proposed framework achieved a mean validation Kappa improvement of 0.0027 over random masking (p = 0.00066, 95% CI [0.0018, 0.0036]) and 0.0063 over direct training (p = 0.00027, 95% CI [0.0045, 0.0082]). Reporting the mean paired differences together with their confidence intervals provides a more direct assessment of the practical magnitude of the observed improvements. Together with the analysis in Figure 5, these findings suggest that the proposed modal feature alignment strategy yields consistent optimization behavior across different random initializations while providing more robust performance under varying missing-modality conditions. Additional experimental results and supplementary analysis are presented in Appendix A.2.

Table 5. Results of each method with different random seeds.

Method	**Comparative Metrics**	**Random Seeds and Their Effects**					
		7	**19**	**23**	**37**	**61**	**42**
Proposed method	Number of iterations	17	15	17	17	17	17
	Validation set kappa	0.7612	0.7617	0.7617	0.7617	0.7607	0.7637
Direct training	Number of iterations	15	11	17	17	15	17
	Validation set kappa	0.7569	0.7565	0.7548	0.7524	0.7546	0.7575
Random masking	Number of iterations	24	32	26	27	28	27
	Validation set kappa	0.7597	0.7589	0.7593	0.7592	0.7580	0.7595

3.5. Generalization Validation

To validate cross-dataset generalization for practical home monitoring, the model trained on SHHS was directly evaluated on the independent P2018 dataset (Figure 6). An expected overall performance degradation of 5–7% was observed during this transition. This graceful degradation is a physically grounded phenomenon in real-world deployments, primarily stemming from hardware variations and demographic differences. Notably, the unimodal ECG accuracy dropped more significantly (6.71%) than THX. This aligns with the physiological reality that electrical cardiac signals exhibit higher inter-subject variance than mechanical respiratory rhythms. Crucially, the model maintained an overall accuracy of approximately 73–78% without catastrophic prediction failure. This confirms that the framework successfully captures generalized physiological rhythms rather than overfitting to specific device noise, demonstrating strong reliability for unconstrained environments.

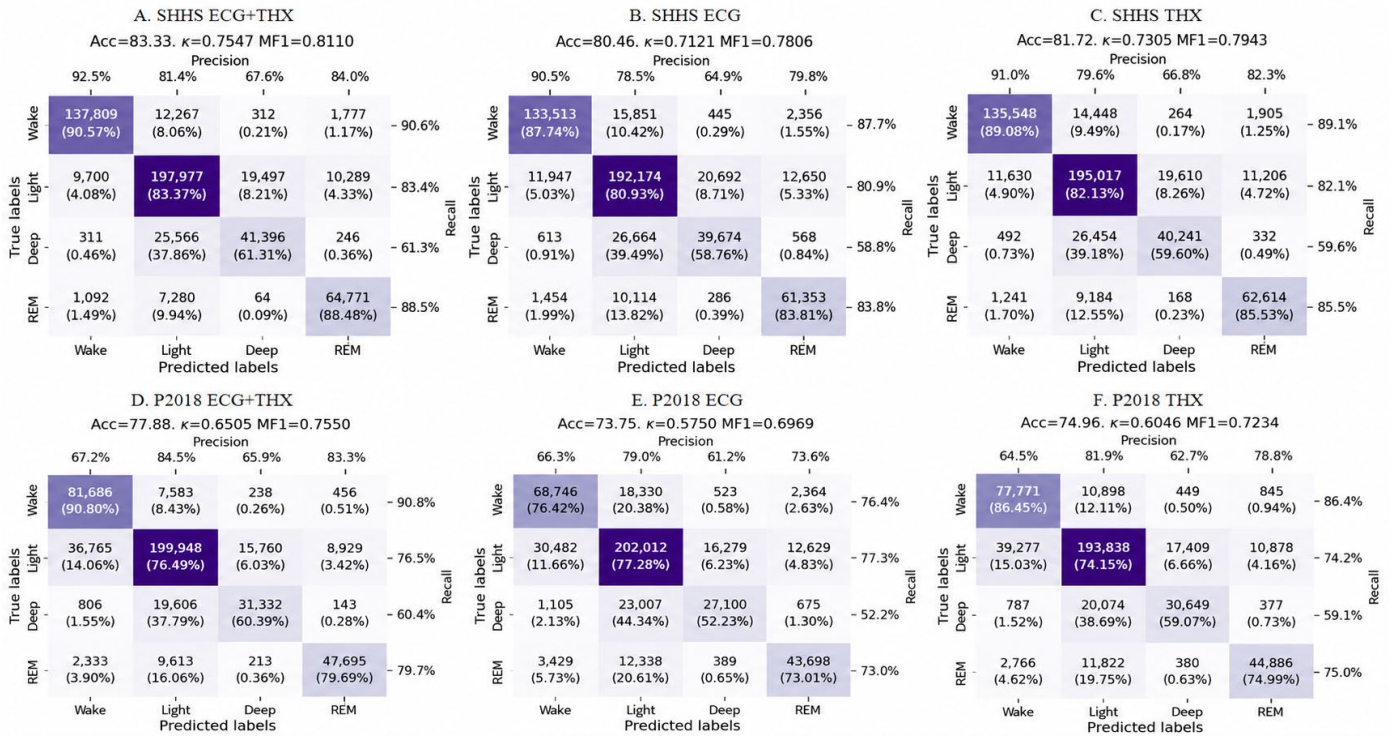

Figure 6. Confusion matrix of the model on the internal (SHHS) and external (P2018) test sets. Darker cell color indicates a larger sample count at the corresponding predicted-true label position.

Furthermore, to assess model robustness across different physiological conditions, samples were categorized into four groups according to the severity of sleep-disordered breathing (SDB) based on the Apnea–Hypopnea Index (AHI): normal (AHI < 5), mild (5 ≤ AHI < 15), moderate (15 ≤ AHI < 30), and severe (AHI ≥ 30) [35]. As shown in Figure 7, the bimodal model consistently achieved high Kappa values with relatively compact performance distributions across all SDB severity groups. The compact distributions across different SDB severity groups indicate stable model performance under heterogeneous physiological conditions. Although the unimodal ECG and THX configurations yielded slightly lower performance than the bimodal setting, both maintained stable sleep staging capability across different severity levels, with only minor fluctuations observed for THX in the severe SDB subgroup. The relatively small within-group performance variation indicates that the proposed framework adapts well to subjects with heterogeneous physiological characteristics. These subgroup analyses provide complementary evidence that the proposed model maintains robust performance across different SDB severity levels,

thereby further supporting its cross-dataset generalization capability beyond the overall evaluation metrics.

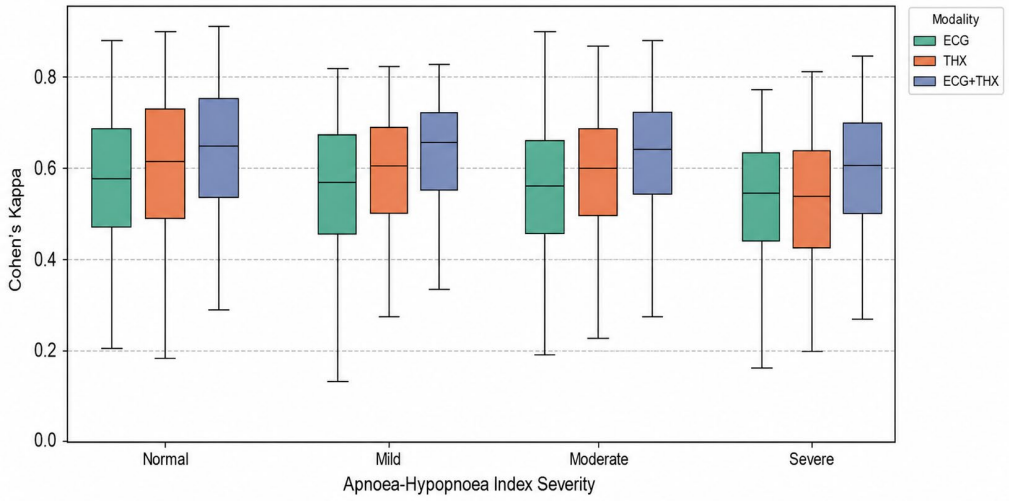

Figure 7. Model performance for different severity levels of SDB.

4. Discussion

This study addresses the missing-modality problem of cardiopulmonary signals in home sleep monitoring by developing a unified multimodal learning framework. The proposed framework combines a dual-branch heterogeneous encoder built upon well-established temporal backbone architecture with a weakly constrained modal feature alignment strategy, enabling a single model to operate under dual-modality, ECG-only, and THX-only conditions. Experimental results on more than 5000 full-night recordings from the SHHS dataset demonstrate reliable sleep staging performance across all input configurations. Compared with separately trained modality-specific models, the proposed framework consistently achieves improved overall performance while eliminating the need to maintain multiple independent models.

The observed performance improvement is numerically moderate on the large-scale SHHS dataset, which is expected for two reasons. First, the large number of training subjects allows baseline models to approach their performance ceiling, leaving limited room for further improvement in conventional evaluation metrics. Second, the primary objective of the proposed framework is not to maximize peak accuracy under complete dual-modality input, but to improve robustness when one physiological modality becomes unavailable during practical deployment. Unlike conventional missing-modality approaches that mainly rely on modality reconstruction or strong feature consistency, our framework is specifically designed for heterogeneous cardiopulmonary signals. ECG and THX originate from different physiological mechanisms and provide complementary rather than equivalent information. Therefore, enforcing strict feature-level consistency may suppress modality-specific representations. Instead, the proposed weakly constrained modal feature alignment encourages semantic consistency while preserving modality individuality, allowing the two modality-specific encoders to exchange task-relevant knowledge without sacrificing their independent discriminative capability.

This learning strategy explains the stable performance observed under missing-modality conditions. During multimodal training, each modality encoder benefits from complementary physiological information while remaining capable of independent inference. Consequently, when one modality is absent during testing, the remaining encoder has already learned more generalized sleep-stage representations through cross-modal interaction, leading to improved robustness compared with separately trained unimodal

models. This advantage is further supported by the external validation on the PhysioNet 2018 dataset, where the proposed framework maintains stable performance across different modality configurations.

The proposed framework also demonstrates favorable deployment feasibility. The final model contains approximately 4.4 million parameters with a checkpoint size of 50.2 MB. On a standard laptop CPU (AMD Ryzen 7 8845H) without GPU acceleration, the pure forward inference of a full-night recording takes only 0.86 s on average (Std: 0.05 s). When including additional overheads of model loading, data reading, inference execution, and result saving, the total processing time for one complete overnight recording is approximately 10 s, with a peak memory consumption of around 0.7 GB. These results indicate that the proposed framework does not rely on dedicated GPU hardware during inference. In practical usage, wearable sensors first collect full-night physiological signals, which can then be transmitted to a personal computer for staging analysis. Users may flexibly adopt offline manual file transfer or real-time automatic network transmission according to specific home monitoring application demands.

Table 6 provides a contextual comparison with representative sleep staging methods based on cardiopulmonary signals. Such cross-study comparisons should be interpreted cautiously because previous studies differ substantially in datasets, sleep-stage definitions, evaluation metrics, and experimental protocols. We include this comparison because most published studies focus either on EEG-based sleep staging or on single-modality cardiopulmonary signals, while publicly available implementations for unified heterogeneous missing-modality learning remain extremely limited. Therefore, Table 6 is intended to provide background rather than direct evidence of superiority. Within this context, the proposed framework achieves performance comparable to existing representative methods while additionally supporting inference under missing-modality conditions. Unlike the representative approaches summarized in Table 6, our framework enables unified deployment across dual-modality and single-modality scenarios using a single trained model, reducing deployment complexity for practical home sleep monitoring.

Table 6. Comparison with recent methods based on cardiopulmonary dual-modality signals.

Studies	Year	Methods	Stages	Kappa/ACC	Missing Modality
Sun et al. [36]	2020	CNN + LSTM	W-R-N W-R-N1-N2-N3	0.586 (Kappa) 0.750 (Kappa)	No Support
Si et al. [37]	2024	U-Net	W-R-N1-N2-N3	64.1% ACC)	No Support
Chu et al. [38]	2026	SVM (Snake-ACO)	W-R-N1N2-N3	68.9% (ACC)	No Support
Sharan et al. [39]	2024	1D-CNN + BiGRU	W-R-N1N2-N3	73.7% (ACC)	No Support
Cater et al. [40]	2024	Transformer	W-R-N1N2-N3	82.8% (ACC)	No Support
This paper	-	Unified Multimodal Model	W-R-N1N2-N3	0.7547 (Kappa) 83.33% (ACC)	Support

Despite these encouraging results, several limitations should be acknowledged. First, the proposed modal feature alignment is implemented using an MSE-based loss, which captures first-order feature consistency but may not fully characterize complex nonlinear relationships between heterogeneous physiological modalities. Second, the current framework mainly addresses complete modality absence, whereas partial signal degradation and epoch-level signal quality variation commonly encountered in home environments are not explicitly modeled. Third, cardiopulmonary signals inherently lack direct information on cortical brain activity, making the discrimination between light and deep sleep intrinsically more challenging than EEG-based approaches. Finally, although CPU inference

demonstrates practical deployment feasibility, lightweight optimization and validation on resource-constrained embedded wearable platforms remain topics for future works.

It should also be emphasized that the proposed framework is intended to support automatic sleep staging for long-term home monitoring rather than replace PSG-based clinical diagnosis. Because cardiopulmonary signals indirectly reflect sleep physiology and do not directly capture cortical neural activity, further validation against clinically relevant endpoints is required before clinical diagnostic use.

Future research will focus on developing more expressive cross-modal alignment objectives, incorporating quantitative signal quality assessment for partial signal degradation to discard low-quality recordings as fully missing samples, extending validation to larger multi-center and non-contact cardiopulmonary datasets, and optimizing the framework through lightweight compression and embedded deployment for real-world wearable sleep monitoring applications.

5. Conclusions

This work proposes a unified multimodal sleep staging framework for heterogeneous cardiopulmonary signals under missing-modality conditions. The proposed framework leverages mature temporal backbones, combined with a dual-branch heterogeneous encoder and a weakly constrained modal feature alignment strategy. This design allows one single trained model to handle dual-modality, ECG-only, and THX-only inputs without extra retraining. Extensive experiments on more than 5000 full-night recordings from the SHHS dataset demonstrate consistent performance improvements under different modality configurations, while external validation on the PhysioNet 2018 dataset further supports the robustness and cross-dataset generalization of the proposed framework. Together with the missing-ratio analysis, the proposed method provides a practical solution for robust long-term home sleep staging under incomplete physiological signal acquisition.

Although the proposed framework still has limitations in cross-modal feature modeling and handling partial signal degradation, it provides a unified training strategy for heterogeneous cardiopulmonary sleep staging with missing modalities. Future work will focus on improving cross-modal representation learning, extending the framework to non-contact physiological sensing, and validating its performance on larger multi-center datasets and practical home monitoring platforms. The proposed framework is intended to support long-term home sleep monitoring rather than replace PSG-based clinical diagnosis.

Author Contributions: Conceptualization, L.G. and X.W.; methodology, L.G.; software, L.G. and Y.Y.; validation, L.G. and C.W.; formal analysis, L.G. and H.C.; investigation, L.G. and C.W.; resources, Q.C. and X.W.; data curation, L.G. and Y.Y.; writing—original draft preparation, L.G.; writing—review and editing, Y.Y., C.W., H.C. and X.W.; visualization, L.G. and H.C.; supervision, X.W.; project administration, X.W.; funding acquisition, Q.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China under Grant No. 82427801.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this study are publicly available. Sleep physiological data were obtained from The Sleep Heart Health Study (SHHS) and the 2018 PhysioNet Sleep Challenge (P2018), accessible via (https://sleepdata.org/datasets/shhs, accessed on 30 October 2024) and (https://physionet.org/content/challenge-2018/1.0.0/, accessed on 18 September 2025), respectively. Derived data supporting the results can be provided upon reasonable request to the corresponding author.

Acknowledgments: The authors thank the creators of the SHHS and P2018 databases for making their data publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PSG	Polysomnography
AASM	American Academy of Sleep Medicine
ECG	Electrocardiogram
THX	Thoracic Respiratory Signal
NREM	Non-Rapid Eye Movement
REM	Rapid Eye Movement
CNN	Convolutional Neural Network
SDB	Sleep-Disordered Breathing
AHI	Apnea–Hypopnea Index
ACC	Accuracy
MF1	Macro-F1 Score
SHHS	Sleep Heart Health Study
P2018	2018 PhysioNet Sleep Challenge

Appendix A. Additional Results and Discussion

Appendix A.1. Preprocessing Rationale

Consistent with our earlier classification results, the overall classification accuracy follows the order: ALL > THX > ECG under simulated hardware constraints. Beyond hardware simulation, these sampling frequencies are rationally configured to balance physiological integrity and computational efficiency: the 34.13 Hz ECG down sampling fully retains the core heart rate and rhythm-related physiological components required for sleep staging without redundant high-frequency noise, while the 8.53 Hz thoracic respiratory down sampling covers all effective respiratory fluctuation characteristics associated with sleep state changes, avoiding invalid frequency band information. In addition, all recordings are standardized to 1200 epochs via zero-padding to unify model input dimensions. This uniform-length standardization eliminates performance bias caused by inconsistent overnight recording durations, ensures fair and stable feature learning across all subjects, and significantly optimizes batch training efficiency and GPU memory utilization, while the padding strategy does not introduce valid physiological interference or distort intrinsic signal feature distribution [33].

To explore the essence of performance disparities under such simulated constraints, Figure A1 visualizes the features using t-SNE. Notably, the multi-scale design (long branch) effectively enhances intra-class compactness of the learned features. The results reveal that while ECG features exhibit superior intra-class compactness due to highly discriminative heart rate patterns, they also present significant individual variability, which poses challenges for generalization. Conversely, THX features, though slightly more dispersed, reflect consistent respiratory patterns across the population, providing a stable baseline decision boundary. By fusing both modalities, the dual-modal approach yields the most robust representation and optimal overall performance.

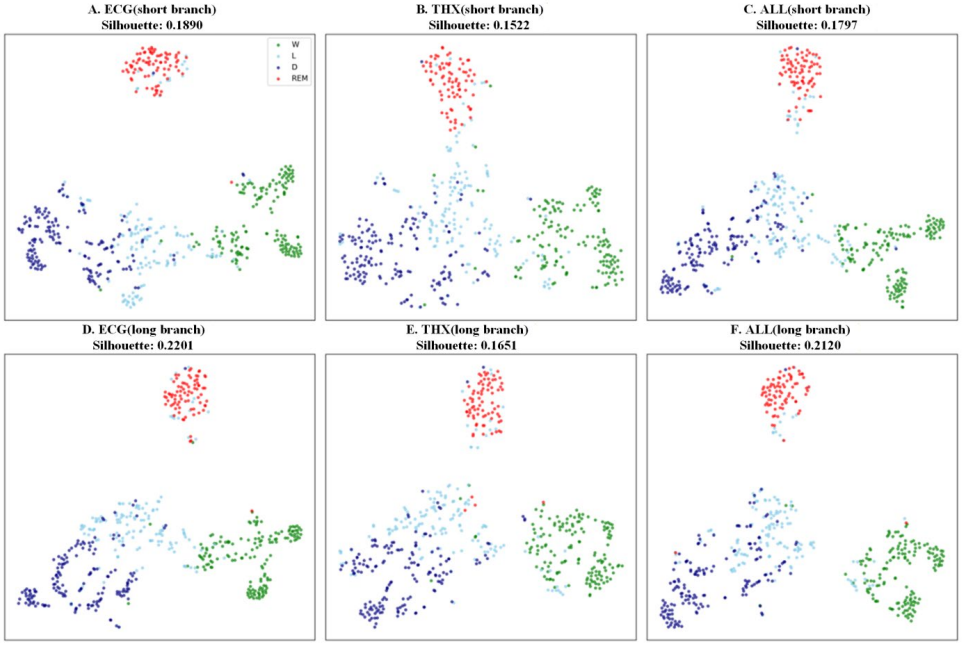

Figure A1. Visualization and silhouette coefficient analysis.

Appendix A.2. Qualitative Analysis

To qualitatively compare the intra-night prediction consistency of the three training strategies under single-modality signal loss, we present representative sleep staging hypnograms for one test subject in Figures A2–A4, color coding is consistent with Figure 3. Each figure arranges three prediction rows: full dual-modal input (ALL), ECG-only input, and THX-only input.

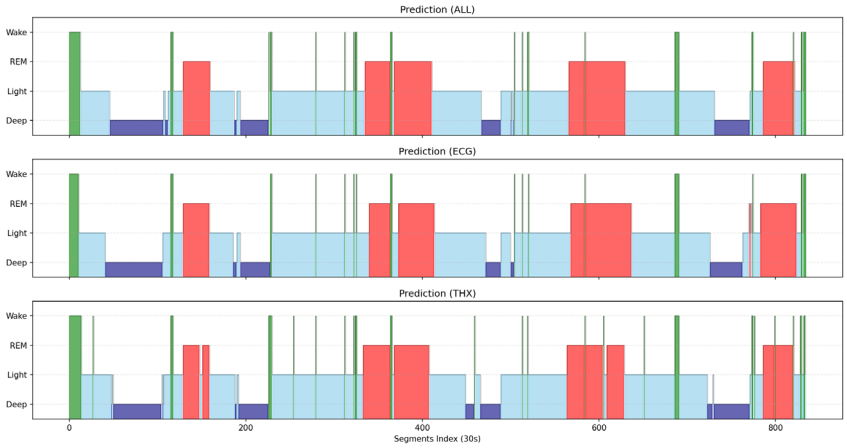

Figure A2. Sleep staging hypnograms predicted by the proposed cross-modal alignment method.

We take the prediction results of complete dual-modal signals as the reference benchmark, and evaluate how closely single-modal outputs align with this staging sequence:

1. For the proposed method (Figure A2), the staging patterns generated by ECG-only and THX-only inputs show excellent overall alignment with the full dual-modal baseline. Only sporadic scattered inconsistent epochs exist between single-modal and dual-modal outputs, with no large-scale continuous stage mismatches.
2. For the random masking baseline (Figure A3), single-modal predictions retain the general sleep cycle trend matching the dual-modal benchmark, yet more scattered inconsistent epochs appear across the whole recording relative to the proposed method.

3. For the direct training baseline (Figure A4), severe long-range mismatches emerge when only ECG signals are available. A large portion of continuous epochs produce drastically different staging results from the full dual-modal reference, demonstrating weak cross-modal transfer capacity.

These qualitative visualizations corroborate the quantitative performance trends in Figure 5: the proposed modal alignment strategy enables the model to retain highly consistent sleep staging logic even when one physiological modality is entirely absent.

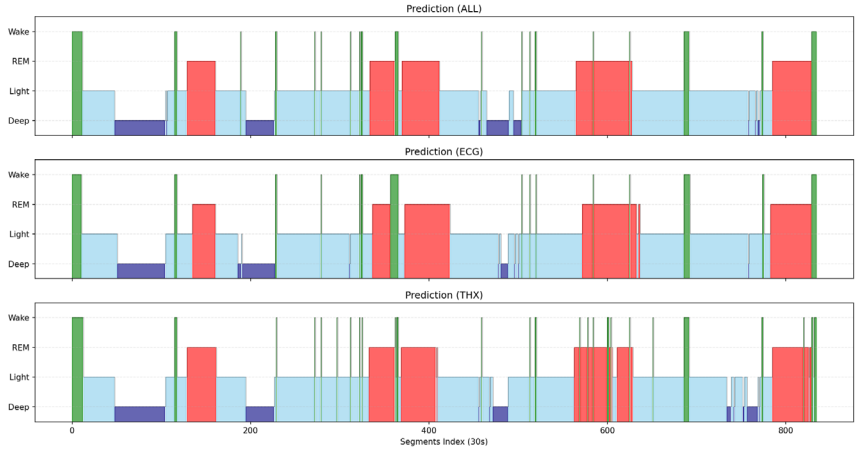

Figure A3. Sleep staging hypnograms predicted by the random masking baseline method.

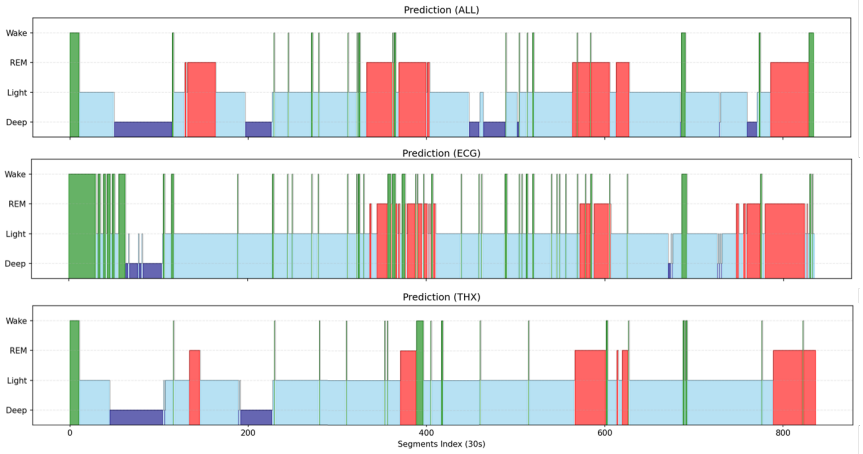

Figure A4. Sleep staging hypnograms predicted by the direct training dual-modal baseline method.

References

1. Baranwal, N.; Yu, P.K.; Siegel, N.S. Sleep physiology, pathophysiology, and sleep hygiene. *Prog. Cardiovasc. Dis.* **2023**, *77*, 59–69. [CrossRef] [PubMed]
2. American Academy of Sleep Medicine (AASM). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications (Update v2.4)*; American Academy of Sleep Medicine: Darien, IL, USA, 2017.
3. Jones, A.M.; Itti, L.; Sheth, B.R. Expert-level sleep staging using an electrocardiography-only feed-forward neural network. *Comput. Biol. Med.* **2024**, *176*, 108545. [CrossRef] [PubMed]
4. Byun, J.H.; Kim, K.T.; Moon, H.J.; Motamedi, G.K.; Cho, Y.W. The first-night effect during polysomnography, and patients' estimates of sleep quality. *Psychiatry Res.* **2019**, *274*, 27–29. [CrossRef] [PubMed]
5. Yue, H.; Chen, Z.; Guo, W.; Sun, L.; Dai, Y.; Wang, Y.; Ma, W.; Fan, X.; Wen, W.; Lei, W. Research and application of deep learning-based sleep staging: Data, modeling, validation, and clinical practice. *Sleep Med. Rev.* **2024**, *74*, 101897. [CrossRef] [PubMed]
6. Liu, P.; Qian, W.; Zhang, H.; Zhu, Y.; Hong, Q.; Li, Q.; Yao, Y. Automatic sleep stage classification using deep learning: Signals, data representation, and neural networks. *Artif. Intell. Rev.* **2024**, *57*, 301. [CrossRef]

7. Masad, I.S.; Alqudah, A.; Qazan, S. Automatic classification of sleep stages using EEG signals and convolutional neural networks. *PLoS ONE* **2024**, *19*, e0297582. [CrossRef] [PubMed]
8. Liu, G.; Wei, G.; Sun, S.; Mao, D.; Zhang, J.; Zhao, D.; Tian, X.; Wang, X.; Chen, N. MicroSleepNet: Efficient deep learning model for mobile sleep staging using dilated convolution and multi-scale fusion. *Front. Neurosci.* **2023**, *17*, 1218072. [CrossRef] [PubMed]
9. Supratak, A.; Dong, H.; Wu, C.; Guo, Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 1998–2008. [CrossRef] [PubMed]
10. Toma, T.I.; Choi, S. An end-to-end multi-channel convolutional Bi-LSTM network for sleep stage detection. *Sensors* **2023**, *23*, 4950. [CrossRef] [PubMed]
11. Wang, J.Q.; Zhao, S.; Jiang, H.T.; Zhou, Y.; Yu, Z.; Li, T.; Pan, G. CareSleepNet: A hybrid deep learning network for automatic sleep staging. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 7392–7405. [CrossRef] [PubMed]
12. Phan, H.; Lorenzen, K.P.; Heremans, E.; Chén, O.Y.; Tran, M.C.; Koch, P.; Mertins, A.; Baumert, M.; Mikke, K.B. L-SeqSleepNet: Whole-cycle long sequence modelling for automatic sleep staging. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 4748–4757. [CrossRef] [PubMed]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; NeurIPS: Long Beach, CA, USA, 2017; pp. 5998–6008.
14. Guo, Y.; Nowakowski, M.; Dai, W. FlexSleepTransformer: A transformer-based sleep staging model adaptable to varying channel counts. *Sci. Rep.* **2024**, *14*, 26312. [CrossRef] [PubMed]
15. Chen, Y.; Yuan, J.; Tang, J. A high precision vital signs detection method based on millimeter wave radar. *Sci. Rep.* **2024**, *14*, 25535. [CrossRef] [PubMed]
16. Ganglberger, W.; Krishnamurthy, P.V.; Quadri, S.A.; Tesh, R.A.; Bucklin, A.A.; Adra, N.; Westover, M.B. Sleep staging in the ICU with heart rate variability and breathing signals: An exploratory cross-sectional study using deep neural networks. *Front. Netw. Physiol.* **2023**, *3*, 1120390. [CrossRef] [PubMed]
17. Yoon, H.; Choi, S.H. Technologies for sleep monitoring at home: Wearables and nearables. *Biomed. Eng. Lett.* **2023**, *13*, 313–327. [CrossRef] [PubMed]
18. Vitazkova, D.; Foltan, E.; Kosnacova, H.; Micjan, M.; Donoval, M.; Kuzma, A.; Vavrinsky, E. Advances in respiratory monitoring: A comprehensive review of wearable and remote technologies. *Biosensors* **2024**, *14*, 90. [CrossRef] [PubMed]
19. Charlton, P.H.; Allen, J.; Bailón, R.; Baker, S.; Behar, J.A.; Chen, F.; Zhu, T. The 2023 wearable photoplethysmography roadmap. *Physiol. Meas.* **2023**, *44*, 021001. [CrossRef]
20. Fayyaz, H.; D'Souza, N.S.; Beheshti, R. Multimodal sleep apnea detection with missing or noisy modalities. *Proc. Mach. Learn. Res.* **2024**, *252*, 1–22. [CrossRef]
21. Kontras, K.; Chatzichristos, C.; Phan, H.; Suykens, J.; De Vos, M. CoRe-Sleep: A multimodal fusion framework for time series robust to imperfect modalities. *arXiv* **2023**, arXiv:2304.06485.
22. Shen, Q.; Xin, J.; Dai, B.T.; Zhang, S.; Wang, Z. *Robust Sleep Staging over Incomplete Multimodal Physiological Signals via Contrastive Imagination*; NeurIPS: Long Beach, CA, USA, 2024.
23. Kweon, Y.-S.; Shin, G.-H.; Kwak, H.-G.; Jo, H.N. Multi-signal reconstruction using masked autoencoder from EEG during polysomnography. *arXiv* **2023**, arXiv:2311.07868.
24. Xie, Z.; Liang, H.; Jia, Z. A multimodal knowledge distillation framework for sleep physiological data. In *Advanced Data Mining and Applications*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 256–265.
25. Quan, S.F.; Howard, B.V.; Iber, C.; Kiley, J.P.; Nieto, F.J.; O'Connor, G.T.; Rapoport, D.M.; Redline, S.; Robbins, J.; Samet, J.M.; et al. The Sleep Heart Health Study: Design, rationale, and methods. *Sleep* **1997**, *20*, 1077–1085. [CrossRef] [PubMed]
26. Ghassemi, M.M.; Moody, B.E.; Lehman, L.W.H.; Song, C.; Li, Q.; Sun, H.; Clifford, G.D. *You Snooze, You Win: The PhysioNet/Computing in Cardiology Challenge 2018*; PhysioNet: Cambridge, MA, USA, 2018.
27. World Medical Association. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA* **2013**, *310*, 2191–2194. [CrossRef] [PubMed]
28. Fonseca, P.; Ross, M.; Cerny, A.; Anderer, P.; van Meulen, F.; Janssen, H.; Pijpers, A.; Dujardin, S.; van Hirtum, P.; van Gilst, M.; et al. A computationally efficient algorithm for wearable sleep staging in clinical populations. *Sci. Rep.* **2023**, *13*, 9182. [CrossRef] [PubMed]
29. Zhuang, Z.; Xue, B.; An, Q.; Chu, H.; Zhang, Y.; Chen, R.; Xu, J.; Ding, N.; Cui, X.; Wang, E.; et al. Advancing sleep health equity through deep learning on large-scale nocturnal respiratory signals. *Nat. Commun.* **2025**, *16*, 9334. [CrossRef] [PubMed]
30. Davidson, S.M.; Carter, J.; Stanyer, E.C.; Sharman, R.; Roman, C.; Kyle, S.D.; Tarassenko, L. Longitudinal cardiorespiratory wearable sleep staging in the home. *Front. Neurosci.* **2026**, *20*, 1693860. [CrossRef] [PubMed]
31. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
32. Le, V.K.D.; Ho, H.B.; Karolcik, S.; Hernandez, B.; Greeff, H.; Nguyen, V.H.; Vietnam ICU Translational Applications Laboratory (VITAL) Investigators. vital_sqi: A Python package for physiological signal quality control. *Front. Physiol.* **2022**, *13*, 1020458. [CrossRef] [PubMed]

33. Carter, J.F.; Tarassenko, L. wav2sleep: A unified multi-modal approach to sleep stage classification from physiological signals. *arXiv* **2024**, arXiv:2411.04644.
34. Zhang, Y.; Cao, W.; Feng, L.; Wang, M.; Geng, T.; Zhou, J.; Gao, D. SHNN: A single-channel EEG sleep staging model based on semi-supervised learning. *Expert Syst. Appl.* **2023**, *213*, 119288. [CrossRef]
35. Korkalainen, H.; Aakko, J.; Nikkonen, S.; Kainulainen, S.; Leino, A.; Duce, B.; Leppänen, T. Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. *IEEE J. Biomed. Health Inf.* **2020**, *24*, 2073–2081.
36. Sun, H.; Ganglberger, W.; Panneerselvam, E.; Leone, M.J.; Quadri, S.A.; Goparaju, B.; Westover, M.B. Sleep staging from electrocardiography and respiration with deep learning. *Sleep* **2020**, *43*, zsz306. [CrossRef] [PubMed]
37. Si, K.; Dong, K.; Lu, J.; Zhao, L.; Xiang, W.; Li, J.; Liu, C. A U-Sleep Model for Sleep Staging Using Electrocardiography and Respiration Signals. In *Asian-Pacific Conference on Medical and Biological Engineering*; Springer Nature: Cham, Switzerland, 2024; pp. 475–482.
38. Chu, W.; Wang, C.; Yang, L.; Guo, L.; Wu, C.; Wang, B.; Wan, X. Sleep Staging Method Based on Multimodal Physiological Signals Using Snake-ACO. *Appl. Sci.* **2026**, *16*, 1316. [CrossRef]
39. Sharan, R.V.; Takeuchi, H.; Kishi, A.; Yamamoto, Y. Macro-Sleep Staging With ECG-Derived Instantaneous Heart Rate and Respiration Signals and Multi-Input 1-D CNN-BiGRU. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2535212. [CrossRef]
40. Cater, J.F.; Jorge, J.; Gibson, O.; Tarassenko, L. SleepVST: Sleep Staging from Near-Infrared Video Signals using Pre-Trained Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 12479–12489.