SLEEP2WORD: PERIPHERAL-TO-EEG TRANSLATION VIA DURATION-AWARE WORDIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating full-night electroencephalography (EEG) spectrograms from peripheral measurements could extend EEG-informed sleep analysis to recordings acquired without EEG. This task requires compact representations of signals with different temporal structures. We introduce *Sleep2Word*, a framework for translating respiration, electrocardiography (ECG), or photoplethysmography (PPG) into EEG spectrograms through duration-aware physiological words. Building on context-dependent chunking, *Sleep2Word* groups physiological embeddings into variable-duration chunks and jointly quantizes their content and duration into modality-specific wordbooks. A block-autoregressive translator then predicts target EEG word identities and occurrence-specific duration refinements within fixed-duration blocks. This design accommodates independent source and target wordbooks, limits temporal drift, and shortens the autoregressive decoding path. Trained on over 20K PSG cases across 9 datasets, *Sleep2Word* consistently outperforms existing spectrogram generation baselines in EEG spectrogram reconstruction, sleep staging, arousal detection, and survival prediction. Compressing full night sequences by up to 69.7%, the learned words yield a favorable accuracy–compute trade-off. Sleep event analysis further shows that the learned words capture physiologically plausible associations across modalities, suggesting their potential for interpretable cross-modal analysis.

1 INTRODUCTION

Electroencephalography (EEG) provides a detailed view of sleep-related brain activity, but acquiring it routinely outside a sleep laboratory remains challenging. Cardiac and respiratory measurements are more accessible and carry partial information about brain state through shared sleep physiology (Jurysta et al., 2003; Schreiner et al., 2023). We study **peripheral-to-EEG translation**: estimating a full-night EEG spectrogram from an individual peripheral source, respiration, electrocardiography (ECG), or photoplethysmography (PPG), without EEG at inference. The target is spectral structure that can be predicted from the observed peripheral signal, not exact recovery of an unobserved EEG waveform. The estimated spectrogram serves as a common intermediate output for multiple downstream analyses, rather than limiting the peripheral-to-brain mapping to a single task-specific label.

Physiology as Language (Zha et al., 2026) established the feasibility of respiration-to-sleep-EEG synthesis using continuous respiratory conditioning, a fixed grid of discrete EEG tokens, and masked generative prediction. We build on this task by investigating how both source and target signals should be represented. Fixed-duration tokenization provides a convenient temporal grid but uses the same segmentation granularity throughout a recording. Short intervals produce long full-night sequences, whereas longer intervals provide a coarser representation. Different physiological modalities also exhibit distinct temporal structures, motivating segmentations that can v*ary both within and across signals*. We therefore ask whether learned, variable-duration segments can reduce the cost of full-night translation while retaining the peripheral information needed to estimate EEG. This motivates **adaptive representations on both sides of the translation**, rather than prescribing one segment duration throughout each recording.

Adaptive segmentation makes duration part of the translation problem. The fifth respiratory token and the fifth EEG token need not describe the same recording interval, even when the underlying

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.

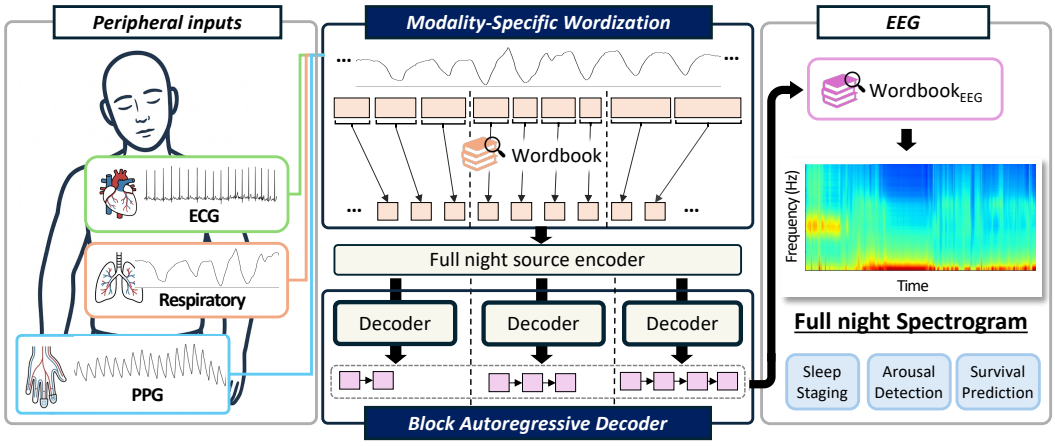


Figure 1: **Modality-specific physiological words on a shared recording timeline**. *Sleep2Word* represents peripheral and EEG signals using independently learned, variable-duration words. EEG words are generated autoregressively within block, regularizing duration drift and enabling parallel generation of different blocks.

signals are synchronized. Although source timestamps remain known, the decoder must predict both the content and duration of the target segments. In a continuous autoregressive chain, accumulated duration errors can shift later predictions along the recording. We therefore seek modality-specific segmentations whose translation remains referenced to a common physical timeline, without requiring corresponding physiological patterns to occur simultaneously.

We introduce *Sleep2Word* (Figure 1), guided by a principle: let each modality learn its own segments, but organize translation around common recording intervals. *Sleep2Word* adapts context-dependent chunking (Hwang et al., 2026) to physiological embeddings from SleepMaMi (Park et al., 2026), grouping adjacent embeddings into variable-duration chunks under reconstruction supervision. **Duration-aware wordization** then quantizes each chunk's pooled representation together with its duration into a modality-specific wordbook. A physiological word is a learned discrete prototype of chunk content and temporal extent. Chunking shortens the sequence, while wordization expresses translation as prediction over reusable segment codes. Because occurrences can differ from a word's typical duration, the decoder predicts an occurrence-specific duration adjustment.

To translate these words, a full-night source encoder first captures context across the recording. A **block-autoregressive decoder** then generates EEG words and their duration adjustments within fixed-duration *blocks*, with sequential prediction inside each block and parallel generation across blocks. The blocks organize generation in time, while the words retain adaptive segmentation within them. Compared with an uninterrupted full-night autoregressive chain, this design shortens the sequential decoding path and is intended to limit cumulative timing errors. Tokenization and generation are therefore coordinated through a common temporal partition.

Our contributions are threefold:

- **Duration-aware physiological wordization**. We adapt context-dependent chunking and jointly quantize chunk content and duration to construct modality-specific wordbooks for peripheral-to-EEG translation. The representation supports variable-duration source and target sequences without imposing matching segmentation.
- **Temporally anchored word-to-word translation**. We adapt block-autoregressive decoding to physical recording intervals, complemented by occurrence-specific duration prediction. This confines temporal drift of variable-length words within individual blocks and keeps the source and target aligned on the recording timeline.
- **Evaluation across peripheral sources and downstream tasks**. We evaluate respiration-, ECG-, and PPG-to-EEG translation across SHHS1 and MESA as well as a held-out dataset CFS. *Sleep2Word* outperforms existing baselines in spectrogram reconstruction, downstream tasks including sleep staging, arousal detection and survival prediction, while demonstrating superior computational efficiency.

2 Related Work

2.1 Cross-System Physiological Coupling and Cross-Modal Modeling

Accumulating evidence indicates that central, autonomic, and respiratory processes are coupled across multiple timescales. During sleep, respiration has been shown to modulate the timing of non-rapid-eye-movement (NREM) slow oscillations and sleep spindles, such that these events preferentially occur at specific phases of the respiratory cycle (Schreiner et al., 2023). In addition, cortical slow oscillations are temporally coupled with transient heart-rate bursts during sleep, demonstrating an interaction between central and autonomic activity (Naji et al., 2019). Together, these findings suggest that peripheral physiological signals retain information about brain states through shared physiological regulation.

Physiological machine learning exploits cross-system dependencies for representation learning, direct prediction, and signal generation. Sleep foundation models learn reusable representations from multimodal recordings (Park et al., 2026; Thapa et al., 2026), while xMAE uses masked cross-modal reconstruction as supervision for biosignal representation learning (Zhou et al., 2026). Direct prediction methods such as wav2sleep (Carter & Tarassenko, 2025) instead map physiological inputs to sleep-stage labels. Cross-modal signal generation methods includes PPG-to-ECG (Ding et al., 2025), PPG-to-respiration (Shuzan et al., 2024), and respiration-to-EEG (Zha et al., 2026) synthesis. Most closely related, Physiology as Language (Zha et al., 2026) combines continuous respiratory conditioning with a fixed temporal grid of discrete EEG codes. Instead, ***Sleep2Word*** represents both source and target through *modality-specific words*. It investigates whether this word-based compact representation can support accurate cross-modal estimation.

2.2 Adaptive Chunking and Discrete Signal Representations

Adaptive sequence models reduce fine-grained processing by forming input-dependent segments. SpaceByte (Slagle, 2024) concentrates higher-capacity processing around delimiter-like boundaries, while BLT (Pagnoni et al., 2025) forms variable-length byte patches using predictive entropy. H-Net (Hwang et al., 2026) learns context-dependent boundaries through a trainable routing module. ***Sleep2Word*** builds on this learned chunking mechanism and adapts it using physiological reconstruction and synchronization-oriented boundary regularization.

Discrete signal vocabularies have also been studied beyond language. TFM-Tokenizer (Pradeepkumar et al., 2026) learns discrete time–frequency motifs from single-channel EEG. In speech, VARSTok (Zheng et al., 2026) combines a content-code index and a discrete duration into a unified token ID. In contrast, ***Sleep2Word*** makes duration part of the quantization space, so content and temporal extent jointly determine each *word* in a fixed-size wordbook. On decoding, durations are then adjusted per occurrence rather than recovered exactly from the token ID. We study this representation in peripheral-to-EEG translation, where the learned segmentation is additionally coordinated with the recording intervals used for decoding.

2.3 Parallel Decoding and Temporal Organization

Semi-autoregressive generation reduces sequential decoding while retaining target dependencies. Wang et al. (2018) generate token groups autoregressively, with parallel prediction within each group. RecoverSAT (Ran et al., 2020) instead generates segments in parallel and tokens sequentially within each segment. ***Sleep2Word*** uses a related schedule. Fixed-duration blocks are decoded in parallel, while EEG words are predicted sequentially within each block. This keeps modality-specific segmentations aligned to a common recording timeline.

3 Method

3.1 Overview and Problem Formulation

Our objective is to estimate an EEG spectrogram from peripheral physiological signals. Let $\mathbf{x}^m \in \mathbb{R}^L$ denote a full night recording from peripheral input m, where $L = f_s N_s$, f_s is the sampling rate and N_s is the recording duration in seconds. We transform the paired target EEG signal into a

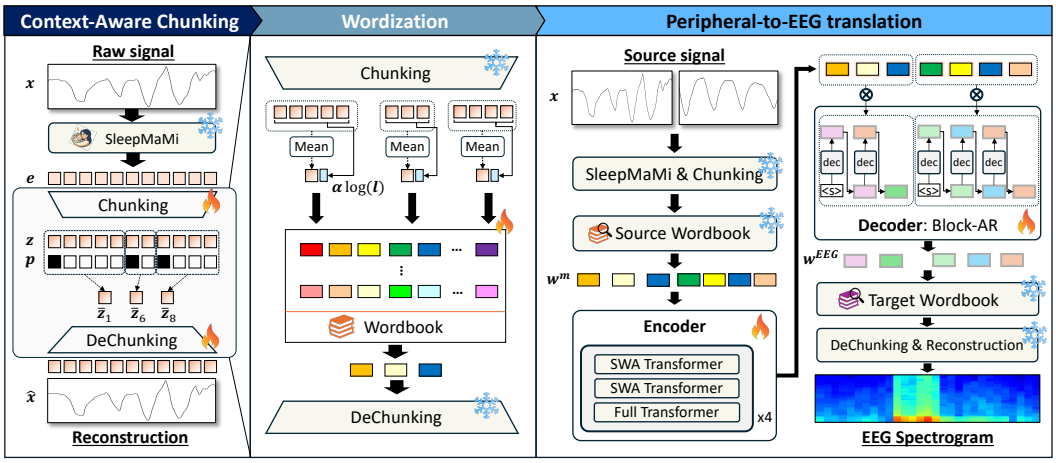


Figure 2: **Training pipeline**. *Sleep2Word* is trained in three stages. The full night sequence is segmented into context-based chunks and discretized by modality-specific wordbook. The wordbook encodes both content and duration. The full night word sequence is processed by an encoder, followed by block-autoregressive decoder for temporal consistency.

multi-taper spectrogram (Prerau et al., 2017), denoted by $\mathbf{y} \in \mathbb{R}^{T_y \times N_f}$ where T_y is the number of non-overlapping ω-second windows and N_f is the number of frequency bins (See Appendix D for details). For clarity, we omit the modality superscript where the distinction is evident.

Sleep2Word estimates the target through a compact intermediate representation. (i) *Context-dependent chunking* first groups adjacent embeddings into variable-duration segments. (ii) *Wordization* maps each segment to a reusable content-duration word. (iii) The translator then predicts EEG word identities and their occurrence-specific durations within common recording intervals, and the EEG reconstruction modules map the resulting sequence to a spectrogram. The overall pipeline is depicted in Figure 2.

Base embeddings. We construct a modality-specific wordbooks using shared embeddings produced by the micro-encoder of SleepMaMi (Park et al., 2026), a recent sleep foundation model. Whereas the original SleepMaMi forms shared representations by concatenating embeddings across all modalities, we modify its shared module to process each modality independently. Consequently, the source signal can be encoded without access to the target modality. Further details of this modification are provided in the Appendix E.

3.2 Context-Dependent Chunking

In the first stage, we learn to divide each full-night representation sequence into variable-length chunks whose boundaries depend on temporal context. To this end, we adapt the dynamic chunking module of Hwang et al. (2026) to physiological reconstruction. Given a sequence of SleepMaMi embeddings $\mathbf{e} \in \mathbb{R}^{L' \times d}$, where L' is the number of base patches in the full night recording and d is the embedding dimension, the chunking module refines the embeddings and predicts the probability p_i that base position i begins a new chunk:

$$(\mathbf{z}, \mathbf{p}) = \mathrm{Chunk}(\mathbf{e}), \quad \mathbf{z} \in \mathbb{R}^{L' \times d}, \quad \mathbf{p} \in [0,1]^{L'}. \tag{1}$$

During the forward pass, the start positions are $\{1\} \cup \{i \geq 2 \text{ and } p_i \geq 0.5\}$. Let $s_1 < \cdots < s_C$ denote the resulting start positions, with $s_{C+1} = L' + 1$. The representation of the c-th chunk is obtained by average pooling:

$$\bar{\mathbf{z}}_c = \frac{1}{l_c} \sum_{t=s_c}^{s_c+l_c-1} \mathbf{z}_t, \quad l_c = s_{c+1} - s_c. \tag{2}$$

The pair (s_c, l_c) is retained alongside each pooled representation to preserve its position and temporal length. The chunk representations are subsequently scattered back to their original positions:

$$\mathbf{u}_t = \bar{\mathbf{z}}_c \quad \text{for} \quad s_c \leq t < s_{c+1} \tag{3}$$

The resulting sequence is processed by the dechunking module and a convolutional reconstruction head, which recovers the original waveform for peripheral modalities and the corresponding spectrogram for EEG. We use mean squared error as reconstruction objective ($\mathcal{L}_{\text{recon}}$).

Because hard thresholding prevents reconstruction gradients from propagating through the boundary decisions, we apply a straight-through surrogate by replacing $\mathbf{u}_t$ with $\tilde{\mathbf{u}}_t = \mathbf{u}_t(1 + (p_{s_c} - \text{sg}[p_{s_c}]))$ where $\text{sg}[\cdot]$ denotes the stop-gradient operator. This does not influence the forward pass, but allows the gradient to reach $\mathbf{p}$ during backpropagation.

In addition to the reconstruction objective, we introduce three auxiliary losses. First, following (Hwang et al., 2026), we adopt a ratio loss ($\mathcal{L}_{\text{ratio}}$) to regulate the proportion of predicted boundaries. Second, an alignment loss ($\mathcal{L}_{\text{align}}$) anchors the refined embeddings to the original shared embeddings, preventing them from deviating substantially toward modality-specific representation space. Finally, the barrier loss ($\mathcal{L}_{\text{barrier}}$) encourages boundaries to start at the common recording intervals (B base positions). This constraint promotes periodic synchronization between the source and target sequences, facilitating the block autoregressive translation described in Section 3.4. The details of these auxiliary losses are provided in Appendix F.

The overall chunking objective is defined as follows:

$$\mathcal{L}_{\text{chunk}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{ratio}}\mathcal{L}_{\text{ratio}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{barrier}}\mathcal{L}_{\text{barrier}} \tag{4}$$

3.3 WORDIZATION

We construct a modality-specific wordbook $\mathcal{W} = \{w_k\}_{k=1}^{K}$ from the learned chunk representations. To incorporate temporal extent into each word, we augment the pooled representation with log duration: $\mathbf{z}'_c = [\bar{\mathbf{z}}_{s_c}; \alpha \log(l_c)]$ where α controls the scale of duration relative to content. We represent each wordbook entry as $w_k = [\nu_k; \eta_k]$, where $\nu_k \in \mathbb{R}^d$ represents its content and $\eta_k \in \mathbb{R}$ represents its duration component. Each chunk $\mathbf{z}'_c$ is assigned to its nearest wordbook entry following this rule:

$$k_c = \arg\min_{k \in \{1,\dots,K\}} [\|\nu_k - \bar{\mathbf{z}}_c\|_2^2 + (\alpha \log l_c - \eta_k)^2] \tag{5}$$

where k_c is the discrete word index. This formulation helps both content and duration to contribute directly to word identity. In addition, for each word, we maintain an exponential moving average μ_k of the durations of its assigned chunks for use in subsequent duration prediction.

During wordization, the chunking module is frozen. Each chunk is assigned to a word index k_c and the corresponding word embedding is passed through the dechunking module and reconstruction head to recover the original peripheral signal or EEG spectrogram. The wordization objective is

$$\mathcal{L}_{\text{Wordization}} = \mathcal{L}_{recon} + \frac{\lambda_Q}{C} \sum_{c=1}^{C} \|\text{sg}[\mathbf{z}'_c] - \mathbf{w}_{k_c}\|_2^2 \tag{6}$$

With fixed chunk representations and fixed α, we do not include the term $\|\mathbf{z}'_c - \text{sg}[\mathbf{w}_{k_c}]\|_2^2$, unlike conventional vector quantization (Van Den Oord et al., 2017).

We initialize the wordbook and calculate α using statistics from the first I batches (We use $I = 10$) produced by the frozen chunking module. The calibration of α balances the contributions of content and log-duration to word assignment, encouraging sensitivity to the chunk duration without requiring a distinct word for every possible length. More details of wordbook initialization are provided in Appendix C.

3.4 PERIPHERAL TO EEG TRANSLATION

After constructing separate wordbooks for the source and target modalities, we train an encoder-decoder model to translate a sequence of peripheral words into a sequence of EEG words, with chunking, dechunking and reconstruction modules frozen. Before target generation begins, the encoder processes the complete peripheral word sequence $\{\mathbf{w}_{k_c}^m\}_{c=1}^{C_m}$. Rotary positional embeddings Su et al. (2024) encode the original start position of each source word, retaining recording time information despite variable word durations. The encoder consists of four Transformer

blocks Vaswani et al. (2017), each containing two sliding window self-attention layers (Beltagy et al., 2020) followed by one full-attention layer. This architecture captures both local temporal patterns and full-night source context.

The decoder generates target EEG words conditioned on the encoded source sequence. Because the source and target modalities are chunked independently, their word sequences may differ in both the number and duration of their words. Naïve word-to-word decoding therefore accumulates temporal misalignment over the course of a full-night recording. To constrain this drift, we employ **Block AutoRegressive (AR)** decoding. Let $\mathbf{H}_b^m$ denote the encoded source memory associated with the b-th block. Within each block, the decoder generates EEG words autoregressively while conditioning on $\mathbf{H}_b^m$. This formulation bounds any temporal discrepancy within individual blocks and reduces the sequential decoding path relative to full-night AR generation.

Let $N_B = \lceil L'/B \rceil$ be the number of blocks and n_b the number of target words in block b. For training labels $k_{b,s}^{EEG}$ obtained from frozen EEG tokenizer, the word-identity objective is

$$\mathcal{L}_{\text{id}} = -\frac{1}{N_B}\sum_{b=1}^{N_B}\frac{1}{n_b}\sum_{s=1}^{n_b}\log p(k_{b,s}^{EEG}|k_{b,<s}^{EEG}, \mathbf{H}_b^m) \tag{7}$$

Occurrence-specific duration. In addition to word identity, the decoder predicts a duration residual $\hat{r}_{b,s}$ relative to the EMA of the selected target word. Each word represents a prototypical content-duration pattern, and the residual accommodates variation across individual occurrences. Using target tokenizer residual $r_{b,s} = l_{b,s}^{EEG} - \text{round}(\mu_{k_{b,s}})$ as supervision, the duration and overall translation loss are defined as

$$\mathcal{L}_{\text{dur}} = -\frac{1}{N_B}\sum_{b=1}^{N_B}\frac{1}{n_b}\sum_{s=1}^{n_b}\log p(r_{b,s}^{EEG}|k_{b,s}^{EEG}, \mathbf{H}_b^m), \quad \mathcal{L}_{\text{trans}} = \mathcal{L}_{\text{id}} + \lambda_{\text{dur}}\mathcal{L}_{\text{dur}} \tag{8}$$

Training. Throughout translation training, the chunking, dechunking and reconstruction modules remain frozen. Training begins with teacher-forcing after which scheduled sampling (Bengio et al., 2015) is applied with a progressively increasing probability to mitigate exposure bias.

Inference and reconstruction. During inference, the target words are generated within each block until their cumulative predicted duration reaches the corresponding block length in base positions. For block b, this target length is $T_b = \min(B, L' - (b-1)B)$. If the cumulative duration exceeds T_b, the duration of the final word is truncated. If the generated sequence length falls short, the duration of the final word is extended to fill the remaining positions. The predicted EEG word embeddings are then expanded onto the base grid using the predicted durations and processed by the frozen EEG dechunker and reconstruction head to obtain the full night EEG spectrogram.

4 Experiments

4.1 Datasets, Preprocessing and Baselines

We train on a combined set of nine publicly available sleep datasets: the Sleep Heart Health Study (SHHS1/2) (Zhang et al., 2018; Quan et al., 1997), Korea Image-based Sleep Study (KISS) (Jeong et al., 2023), Korea Video Sleep Study (KVSS) (Choi et al., 2026), Physionet 2018 (PHY) (Goldberger et al., 2000; Ghassemi et al., 2018), Multi-Ethnic Study of Atherosclerosis (MESA) (Chen et al., 2015) and Osteoporotic Fractures in Men Study (MrOS) (Blackwell et al., 2011). Overall, we use 20,036 PSG recordings (107,469 hours) for training. We use the Cleveland Family Study (CFS) (Zhang et al., 2018; Redline et al., 1995) as a held-out dataset to assess *Sleep2Word*'s generalizability to previously unseen data. Further dataset details are provided in Appendix A.

We apply standard preprocessing to all modalities. EEG signals are filtered using a 0.3–40 Hz bandpass filter, and ECG signals undergo a 0.3–60 Hz bandpass filtering. A 60 Hz notch filter is additionally applied to both EEG and ECG signals. Respiratory signals are low-pass filtered at 15 Hz. Biosignals are resampled at 100 Hz to harmonize the differing sampling rates across the diverse datasets. We use C4-A1 EEG channel and abdominal belt respiratory signal. For PPG, the raw signal is polarity-corrected, low-pass filtered at 8 Hz and clipped at ± 3 standard deviations. Before model input, all signals are z-score normalized independently within each 30-second epoch.

Table 1: **EEG spectrogram reconstruction performance**. MAE and SNR are provided across modalities and datasets. Bold values indicate the best among spectrogram-generation methods.

Dataset	Metric	Respiratory			ECG		PPG	
		Ours	PasL	C-T baseline	Ours	C-T baseline	Ours	C-T baseline
SHHS1	MAE (↓)	**0.047**	0.074	0.050	**0.048**	0.056	-	-
(Internal)	SNR (↑)	**17.35**	16.03	16.86	**17.26**	15.71	-	-
MESA	MAE (↓)	**0.056**	0.069	0.064	**0.059**	0.075	**0.057**	0.066
(Internal)	SNR (↑)	**14.38**	11.74	13.88	**14.12**	12.51	**14.23**	12.93
CFS	MAE (↓)	**0.058**	0.112	0.061	**0.065**	0.067	**0.059**	0.060
(External)	SNR (↑)	**15.26**	13.89	14.12	**14.38**	13.71	**14.69**	14.67

Table 2: **Sleep staging performance**. Best spectrogram-generation results are bolded.

Dataset	Metric	EEG	Respiratory				ECG			PPG		
		GT	Spectrogram generation			Specialized	Spectrogram generation		Specialized	Spectrogram generation		Specialized
		Spec	*Sleep2Word*	PasL	C-T baseline	wav2sleep	*Sleep2Word*	C-T baseline	wav2sleep	*Sleep2Word*	C-T baseline	wav2sleep
SHHS1	MF1	0.872	**0.757**	0.668	0.640	0.777	**0.714**	0.671	0.788	-	-	-
	Acc.	0.879	**0.766**	0.681	0.661	0.806	**0.729**	0.689	0.814	-	-	-
	Kappa	0.822	**0.658**	0.531	0.503	0.711	**0.603**	0.547	0.724	-	-	-
MESA	MF1	0.858	**0.757**	0.642	0.677	0.757	**0.699**	0.583	0.753	**0.709**	0.332	0.746
	Acc.	0.884	**0.799**	0.701	0.723	0.854	**0.741**	0.657	0.841	**0.754**	0.566	0.840
	Kappa	0.825	**0.694**	0.534	0.584	0.770	**0.609**	0.476	0.750	**0.626**	0.216	0.749
CFS	MF1	0.862	**0.743**	0.417	0.642	0.816	**0.670**	0.651	0.841	**0.679**	0.282	0.803
	Acc.	0.866	**0.748**	0.563	0.660	0.834	**0.686**	0.662	0.858	**0.688**	0.523	0.821
	Kappa	0.806	**0.639**	0.304	0.511	0.759	**0.546**	0.521	0.794	**0.550**	0.154	0.741

As a baseline, we construct a simple CNN-Transformer model for peripheral-to-EEG conversion (referred to C-T baseline, details in Appendix B). For respiratory-to-EEG conversion, we compare against the recently proposed Physiology as Language (PasL) (Zha et al., 2026). We reproduce PasL following the paper, since the source code is not released. wav2sleep (Carter & Tarassenko, 2025), a sleep staging specialized model using peripheral inputs, is also adopted to compare against raw signal based method.

Baseline split alignment. For SHHS1, MESA and CFS, we align our evaluation sets with the official wav2sleep splits to ensure comparisons are performed on the same recordings. Details of the split construction are provided in Appendix A.

4.2 EEG Reconstruction Performance

Table 1 provides the EEG spectrogram reconstruction quality of *Sleep2Word* and other baselines. Following PasL, we report the quality using two metrics: Mean Absolute Error (MAE) and Signal-to-Noise Ratio (SNR), in comparison to the ground truth spectrogram. Across all source modality inputs and datasets, *Sleep2Word* exhibits lower MAE and higher SNR compared to other baselines. Particularly, *Sleep2Word* demonstrates its robustness to the unseen dataset (CFS), while PasL shows suboptimal performance.

4.3 Downstream Performance

We evaluate whether the generated EEG spectrograms preserve information relevant to three downstream tasks: four-stage sleep staging, arousal detection, and survival prediction.

Sleep staging. Table 2 reports performance on four-stage sleep staging (Wake, N1/N2, N3, and REM) using Macro-F1, accuracy, and Cohen's kappa. For spectrogram-generation methods (*Sleep2Word*, PasL, and the C-T baseline), we froze the generation model and separately fine-tuned CNN classifier on spectrograms from the training split and evaluate it on the corresponding test split of each dataset. For ground-truth (GT) EEG spectrograms, a CNN classifier is also fine-tuned on training split and evaluated. For wav2sleep, we directly evaluate the released pretrained models.

Among the evaluated spectrogram-generation methods, *Sleep2Word* achieves the strongest sleep staging performance across the datasets and source modalities. Within *Sleep2Word*, respiration consistently yields better performance than ECG or PPG, suggesting that sleep-stage-related EEG structure can be recovered more reliably from respiratory signals in the evaluated cohorts.

Nonetheless, *Sleep2Word* generally underperforms wav2sleep and the classifier trained on GT EEG spectrograms. This comparison should be interpreted in light of their different objectives: wav2sleep

Table 3: **Arousal detection performance**. Results are provided in AUROC conditioned on non-wake epochs. Bold values indicate the best performance among spectrogram-generation methods.

Dataset	EEG	Respiratory				ECG			PPG		
	GT	Spectrogram generation			Specialized	Spectrogram generation		Specialized	Spectrogram generation		Specialized
	Spec	*Sleep2Word*	PasL	C-T baseline	wav2sleep	*Sleep2Word*	C-T baseline	wav2sleep	*Sleep2Word*	C-T baseline	wav2sleep
SHHS1	0.879	**0.799**	0.669	0.743	0.666	**0.802**	0.728	0.681	-	-	-
MESA	0.827	**0.824**	0.633	0.760	0.655	**0.779**	0.662	0.650	**0.781**	0.521	0.647
CFS	0.891	**0.777**	0.660	0.772	0.595	**0.788**	0.721	0.616	**0.757**	0.511	0.588

Table 4: **Survival prediction performance on SHHS1**. Results are reported using the C-index. Bold values indicate the best performance among spectrogram-generation methods.

Dataset	Outcome	Respiratory				ECG		
		Spectrogram generation			Specialized	Spectrogram generation		Specialized
		Sleep2Word	PasL	C-T baseline	wav2sleep	*Sleep2Word*	C-T baseline	wav2sleep
SHHS1	All-cause mortality	**0.700**	0.645	0.574	0.697	**0.736**	0.680	0.713
	CVD mortality	**0.832**	0.648	0.596	0.758	**0.826**	0.729	0.775
	Stroke	**0.824**	0.618	0.587	0.824	**0.837**	0.833	0.820

is optimized specifically for sleep staging using direct stage supervision, whereas *Sleep2Word* learns task-agnostic cross-modal generation and is evaluated using a separate downstream classifier. The remaining gap to GT EEG further suggests that some stage-relevant information is either unavailable in peripheral signals or lost during word-based generation. Despite these differences, the consistent improvement over other spectrogram-generation baselines demonstrates that *Sleep2Word* preserves more stage-discriminative EEG structure without task-specific optimization.

Arousal detection. Table 3 compares the arousal detection performance of *Sleep2Word* with those of the baseline models. To match the temporal resolution of the input spectrograms, we convert the second-level arousal annotations into epoch-level labels. An epoch is labeled as positive if it contains at least three seconds of annotated arousal activity. Using the epoch-level arousal labels, we train a linear classifier on the generated spectrogram. The arousal prediction AUROC conditioned on non-wake epochs is reported, to align with the arousal determination rule (Berry, 2012).

Sleep2Word achieves the highest AUROC among spectrogram-generation methods across datasets and modalities. In particular, it consistently outperforms wav2sleep. Since wav2sleep is trained for sleep staging, this comparison indicates that its representations carry little arousal information beyond what staging requires. This supports the use of generated spectrograms as a task-agnostic intermediate rather than a proxy for a single label.

Survival prediction. We evaluate prognostic information in generated EEG spectrograms on SHHS1 using a linear Cox head for all-cause mortality, CVD mortality, and stroke. *Sleep2Word* achieves higher C-indices than PasL and C-T across all respiratory-input outcomes, and exceeds wav2sleep for CVD mortality while remaining comparable for the other outcomes. With ECG input, *Sleep2Word* achieves higher C-indices than C-T and wav2sleep across all outcomes (Table 4).

4.4 Compute Efficiency

To assess the computational efficiency of *Sleep2Word*, we compare its cumulative training compute, measured in TFLOPs, with linear probed sleep staging accuracy on SHHS1. Here, we introduce Base2base, which directly translates a sequence of source-modality base patch embeddings $\mathbf{e}^m$, into the corresponding sequence of EEG base patch embeddings $\mathbf{e}^{EEG}$, without intermediate chunking or wordization. As shown in Figure 3, *Sleep2Word* achieves higher accuracy than the baselines at comparable or lower compute budgets. This result demonstrates the superior accuracy-compute trade-off enabled by the word-based framework.

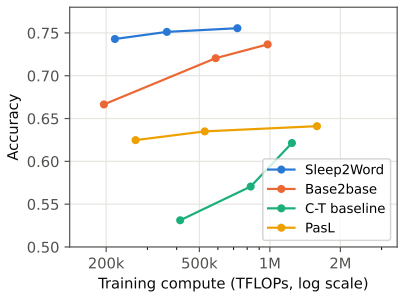
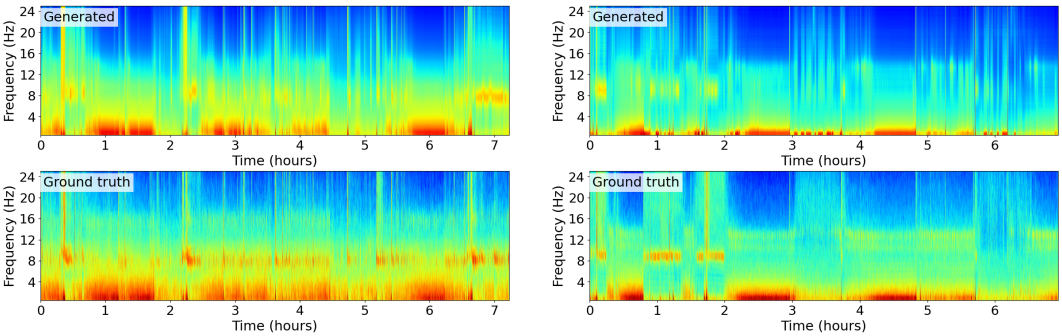


Figure 3: **Compute efficiency.** Sleep staging accuracy versus cumulative training compute. *Sleep2Word* achieves the best accuracy-compute trade-off among the compared methods

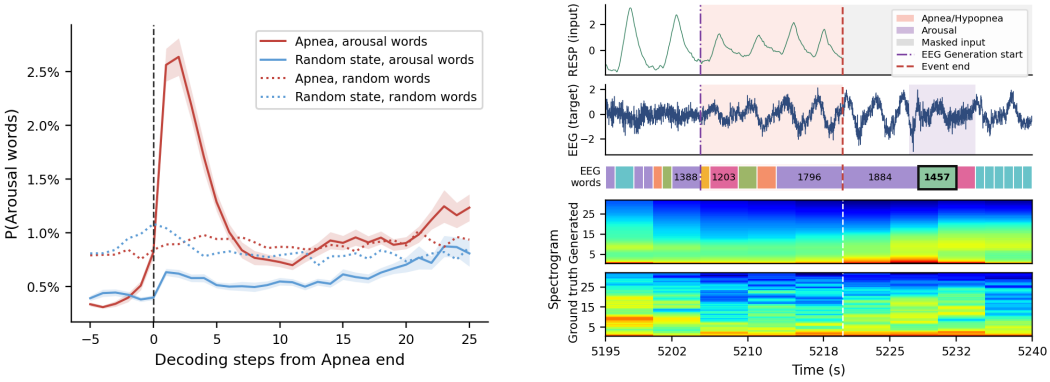
4.5 Qualitative Analysis and Ablation

Figures 4a and 4b compare representative EEG spectrograms generated from peripheral signals with their corresponding ground-truth spectrograms. The generated spectrograms reproduce broad full-



(a) Spectrogram (from Resp.): 60-year-old Male

(b) Spectrogram (from ECG): 54-year-old Male



(c) EEG arousal-word probabilities on apnea input

(d) Arousal generation example after apnea

Figure 4: **Qualitative analysis**. (a-b) Full-night comparison between the generated and ground-truth EEG spectrograms. (c) Changes in the probability of generating arousal-related EEG words after apnea termination. (d) An example showing the input respiratory signal from respiratory event, generated EEG words, and spectrogram. The bolded box represents an arousal word.

night spectral structure, including variations in low frequency, spindle frequency range, and higher frequency power, although they do not capture all fine-grained structures present in the ground-truth EEG. To examine whether the model captures cross-modal physiological relationships, we investigate apnea-arousal coupling. We identify the 20 EEG words (the top-1% of wordbook) most strongly associated with annotated arousals and provide respiratory input only until apnea termination, while allowing EEG generation to continue beyond the event. Figure 4c shows that the probability of arousal-related EEG words increases sharply after apnea termination and exceeds those under random-state conditioning. The example in Figure 4d further illustrates this pattern, suggesting that the model learns the temporal relationship between disordered breathing and subsequent cortical arousal. Together, these results highlight the potential of word-based approach for model-based exploration of cross-modal physiological associations.

Ablation results (Appendix G) confirm the contribution of each core component: loss components (Table 13), chunking and wordization (Table 14), duration modeling (Table 15), wordbook size (Table 16), target chunking ratio (Table 17) and block size (Table 18).

5 Conclusion

In this paper, we present *Sleep2Word*, a word-based framework for peripheral-to-EEG translation. By introducing context-dependent chunking, duration-aware physiological wordization, and a block-AR decoder, our model captures the distinct temporal structures of different modalities while effectively mitigating temporal drift. Our evaluations confirm that *Sleep2Word* outperforms existing baselines in EEG spectrogram reconstruction and clinical downstream tasks. Finally, the word-based architecture reduces the sequence length by 69.7%, enabling a favorable compute-accuracy trade-off. This work provides a strong foundation for expanding sleep neurophysiology analysis using more accessible peripheral measurements.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology or experiments, and implement methods. We have not used generative AI tools for other tasks with required disclosure, and they are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, summarize or analyze existing literature and edit a research paper to improve readability. We have reviewed all AI-assisted work. For example, we manually checked the ideas, sentences, existing literature suggested by AI for its factual accuracy and potential plagiarism. LLM-generated code was also verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide the model specification (Appendix B) and implementation details (Appendix C). The summary information on datasets (Appendix A) and preprocessing steps (Section 4) are also provided.

REFERENCES

Hyun Ahn, Younghoon Na, and Hyun-Woo Shin. Refining sleep-disordered breathing annotations across multiple public sleep study datasets. *Journal of Sleep Research*, 35, 12 2026. doi: 10.1111/jsr.70264.

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

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

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

Richard B Berry, Rita Brooks, Charlene Gamaldo, Susan M Harding, Robin M Lloyd, Stuart F Quan, Matthew T Troester, and Bradley V Vaughn. Aasm scoring manual updates for 2017 (version 2.4). *Journal of Clinical Sleep Medicine*, 13(5):665–666, 2017.

Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E Ensrud, Marcia L Stefanick, Alison Laffan, Katie L Stone, and Osteoporotic Fractures in Men Study Group. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *Journal of the American Geriatrics Society*, 59(12):2217–2225, 2011.

Jonathan F. Carter and Lionel Tarassenko. wav2sleep: A unified multi-modal approach to sleep stage classification from physiological signals. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 186–202. PMLR, 15–16 Dec 2025.

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

You Rim Choi, Dongik Park, Hyun Kyung Lee, Suahn Bae, Sehyun Park, Dong-Kyu Kim, Yong Min Kim, Chan-Soon Park, Hyung-Sin Kim, and Hyun-Woo Shin. Non-contact on-device detection of obstructive sleep apnea from infrared video. *Nature Communications*, 2026.

Zhengyao Ding, Yujian Hu, Ziyu Li, Yiheng Mao, Haitao Li, Dongchen Zhou, Xuesen Chu, Long Yu, Ziyi Liu, Fei Wu, et al. Ai modeling photoplethysmography to electrocardiography useful for predicting cardiovascular disease. *NPJ Digital Medicine*, 9(1):61, 2025.

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

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Sukjun Hwang, Brandon Wang, and Albert Gu. Dynamic chunking for end-to-end hierarchical sequence modeling. In *International Conference on Learning Representations*, volume 2026, pp. 149273–149313, 2026.

Jaemin Jeong, Wonhyuck Yoon, Jeong-Gun Lee, Dongyoung Kim, Yunhee Woo, Dong-Kyu Kim, and Hyun-Woo Shin. Standardized image-based polysomnography database and deep learning algorithm for sleep-stage classification. *Sleep*, 46(12):zsad242, 2023.

Fabrice Jurysta, Philippe Van De Borne, P-F Migeotte, Martine Dumont, J-P Lanquart, J-P Degaute, and Paul Linkowski. A study of the dynamic interactions between sleep eeg and heart rate variability in healthy young men. *Clinical neurophysiology*, 114(11):2146–2155, 2003.

Mohsen Naji, Giri P Krishnan, Elizabeth A McDevitt, Maxim Bazhenov, and Sara C Mednick. Coupling of autonomic and central events during sleep benefits declarative memory consolidation. *Neurobiology of learning and memory*, 157:139–150, 2019.

Artidoro Pagnoni, Ramakanth Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason E Weston, Luke Zettlemoyer, et al. Byte latent transformer: Patches scale better than tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9238–9258, 2025.

Keondo Park, Younghoon Na, You Rim Choi, Hyunwoo Ryu, Hyun-Woo Shin, and Hyung-Sin Kim. Sleepmami: A universal sleep foundation model for integrating macro- and micro-structures. In *Forty-third International Conference on Machine Learning*, 2026.

Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Tokenizing single-channel eeg with time-frequency motif learning. In *International Conference on Learning Representations*, volume 2026, pp. 36611–36637, 2026.

Michael J Prerau, Ritchie E Brown, Matt T Bianchi, Jeffrey M Ellenbogen, and Patrick L Purdon. Sleep neurophysiological dynamics through the lens of multitaper spectral analysis. *Physiology*, 32(1):60–92, 2017.

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.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. Learning to recover from multi-modality errors for non-autoregressive neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3059–3069, 2020.

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*, 151:682–687, 1995.

Thomas Schreiner, Marit Petzka, Tobias Staudigl, and Bernhard P Staresina. Respiration modulates sleep oscillations and memory reactivation in humans. *Nature Communications*, 14(1): 8351, 2023.

Md Nazmul Islam Shuzan, Moajjem Hossain Chowdhury, Saadia Binte Alam, Mamun Bin Ibne Reaz, Muhammad Salman Khan, M Murugappan, and Muhammad EH Chowdhury. Ppg2respnet: a deep learning model for respirational signal synthesis and monitoring from photoplethysmography (ppg) signal. *Physical and Engineering Sciences in Medicine*, 47(4):1705–1722, 2024.

Kevin Slagle. Spacebyte: Towards deleting tokenization from large language modeling. *Advances in Neural Information Processing Systems*, 37:124925–124950, 2024.

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

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

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chunqi Wang, Ji Zhang, and Haiqing Chen. Semi-autoregressive neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 479–488. Association for Computational Linguistics, 2018.

Kaiwen Zha, Chao Li, Hao He, Peng Cao, Tianhong Li, Ali Mirzazadeh, Ellen Zhang, Jong Woo Lee, Yoon Kim, and Dina Katabi. Physiology as language: Translating respiration to sleep eeg. *arXiv preprint arXiv:2602.00526*, 2026.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Rui-Chen Zheng, Wenrui Liu, Hui-Peng Du, Qinglin Zhang, Chong Deng, Qian Chen, Wen Wang, Yang Ai, and Zhen-Hua Ling. Say more with less: Variable-frame-rate speech tokenization via adaptive clustering and implicit duration coding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35021–35029, 2026.

Hao Zhou, Simon A. Lee, Cyrus Tanade, Keum San Chun, Juhyeon Lee, Migyeong Gwak, Megha Thukral, Justin Sung, Eugene Hwang, Mehrab Bin Morshed, Li Zhu, Viswam Nathan, Md Mahbubur Rahman, Subramaniam Venkatraman, and Sharanya Arcot Desai. Physiology-aware masked cross-modal reconstruction for biosignal representation learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=NxbSbkLNMc.

A DATASETS

A brief overview of the datasets used in this study is provided below. Table 5 summarizes the main statistics of the datasets used for pretraining, while Table 6 presents the datasets used for downstream evaluation. Table 7 reports the channels and sampling rates available for each dataset.

A.1 PHYSIONET 2018 (PHY) DATASET

The PhysioNet 2018 dataset was originally developed for the 2018 PhysioNet/CinC Challenge Goldberger et al. (2000); Ghassemi et al. (2018). It was provided by the Computational Clinical Neurophysiology Laboratory and the Clinical Data Animation Laboratory at Massachusetts General Hospital. Although the main objective of the challenge was arousal detection, the dataset also contains expert-annotated sleep stages for 994 subjects. In addition, 991 recordings were reserved as a test set, although the corresponding labels remain private. In this study, both the labeled training set and the unlabeled test set were used for training *Sleep2Word*. The original PhysioNet recordings were acquired at dataset-specific sampling rates, with most physiological signals sampled at 200 Hz. During preprocessing, the raw biosignals were resampled to 100 Hz to ensure a consistent sampling rate across datasets. Sleep stages were scored following the American Academy of Sleep Medicine (AASM) guidelines.

A.2 SLEEP HEART HEALTH STUDY (SHHS) DATASET

The Sleep Heart Health Study (SHHS) Zhang et al. (2018); Quan et al. (1997) is a multicenter cohort study organized by the National Heart, Lung, and Blood Institute. The dataset contains data collected during two separate visits. SHHS1 corresponds to the first visit, conducted between 1995 and 1998, and included 6,441 men and women aged 40 and older. SHHS2 corresponds to the second visit, carried out between 2001 and 2003, and included 3,295 participants. Polysomnography (PSG) recordings were collected at home by trained technicians and included several physiological signals, such as EEG (C3-A2 and C4-A1), two EOG channels, EMG, respiratory effort, airflow, oxygen saturation, ECG, and body position.

For our experiments, the SHHS1 dataset was divided into 3,655 subjects for training, 80 subjects for validation, and 290 subjects for the final held-out test set. For survival prediction, nine training subjects were excluded due to unavailable features or zero follow-up time, and one test subject with zero follow-up time was excluded, yielding 3,646 training, 80 validation, and 289 test subjects for analysis. The test set was restricted to subjects belonging to the common evaluation set shared between our pipeline and the wav2Sleep baseline, allowing a consistent comparison between the two models. Subjects from the original held-out test set that were not part of this common subset were kept unused. The SHHS2 dataset was used for pretraining and was not used for downstream tasks.

Table 5: Statistics of training recordings (minimum duration: 256 min). Stage counts cover all annotated 30-s epochs in retained recordings, whereas effective training duration includes only complete 256-min input windows.

Dataset	Records	Subjects	Age (years)	BMI	Sex % (Female/Male)	Stage count						Stage ratio (%)				
						W	N1	N2	N3	REM	Total	W	N1	N2	N3	REM
PHY-Train	990	990	55.2 ± 14.3	N/A	33 / 67	145,317	135,210	371,718	101,551	113,765	867,561	17	16	42	12	13
PHY-Test	989	989	54.8 ± 14.3	N/A	37 / 63	N/A						N/A				
SHHS1	3,655	3,655	63.1 ± 11.5	28.1 ± 5.2	52 / 48	737,584	136,122	1,517,256	471,476	516,169	3,378,607	22	4	45	14	15
SHHS2	2,547	2,547	67.6 ± 10.3	28.3 ± 5.0	54 / 46	682,217	106,861	1,102,583	302,150	395,181	2,588,992	26	4	43	12	15
KISS	5,675	5,675	44.6 ± 14.6	25.6 ± 4.2	21 / 79	946,226	640,532	1,516,640	584,211	648,113	4,335,722	22	15	35	13	15
KVSS	800	800	51.4 ± 14.5	26.8 ± 4.4	24 / 76	134,500	122,688	261,618	46,394	95,539	660,739	20	19	40	7	14
MrOS1	2,758	2,758	76.4 ± 5.5	27.1 ± 3.8	0 / 100	943,949	129,086	1,234,915	221,090	381,366	2,910,406	32	4	43	8	13
MrOS2	991	991	81.0 ± 4.4	26.9 ± 3.8	0 / 100	381,597	80,258	425,045	44,896	129,853	1,061,649	36	8	40	4	12
MESA	1631	1631	69.4 ± 9.1	28.6 ± 5.5	53 / 47	478,846	165,620	680,574	117,046	215,291	1,657,377	29	10	41	7	13

Table 6: The dataset statistics used for validation and test.

Dataset	Records	Subjects	Age (years)	BMI	Sex % (Female/Male)	Stage count						Stage ratio (%)				
						W	N1	N2	N3	REM	Total	W	N1	N2	N3	REM
SHHS1-Val	80	80	61.6 ± 12.0	28.0 ± 4.3	54 / 46	16,396	2,912	32,636	9,311	10,882	72,137	23	4	45	13	15
SHHS1-Test	290	290	63.9 ± 11.1	28.4 ± 4.9	49 / 51	57,933	11,558	122,037	35,274	42,105	268,907	22	4	45	13	16
MESA-Val	170	170	69.1 ± 8.6	28.9 ± 5.4	59 / 41	49,657	16,265	72,398	12,467	22,476	173,263	29	9	42	7	13
MESA-Test	62	62	68.8 ± 8.9	29.6 ± 6.0	52 / 48	18,569	6,281	24,570	5,406	8,373	63,199	29	10	39	9	13
CFS-Train	452	452	43.1 ± 19.8	32.0 ± 9.3	52 / 48	101,447	15,927	175,733	58,789	59,978	411,874	25	4	43	14	15
CFS-Val	50	50	40.4 ± 19.0	31.4 ± 8.9	54 / 46	12,319	1,751	19,718	5,879	6,099	45,766	27	4	43	13	13
CFS-Test	100	100	38.1 ± 19.3	31.6 ± 9.0	62 / 38	21,456	3,044	37,974	13,120	11,997	87,591	25	4	43	15	13

Table 7: Channels and sampling rates included in different modalities across datasets. The manufacturer of PSG recording machine is also provided. When sampling rates are same across different channels in the modality, we only write once at the top row.

Modality	KISS / KVSS		SHHS1		SHHS2		PHY		MESA		MrOS1 / 2		CFS	
	Ch.	SR (Hz)	Ch.	SR (Hz)	Ch.	SR (Hz)	Ch.	SR (Hz)	Ch.	SR (Hz)	Ch.	SR (Hz)	Ch.	SR (Hz)
EEG	C3-M2 C4-M1 O1-M2 O2-M1	200	C3-A2 C4-A1	125	C3-A2 C4-A1	125	C3-M2 C4-M1 O1-M2 O2-M1 F3-M2 F4-M1	200	C4-M1 Oz-Cz Fz-Cz	256	C3-M2 C4-M1 O1-M2 O2-M1	256	C3-A2 C4-A1	128
ECG	ECG	200	ECG	125	ECG	125	ECG	250	ECG	256	ECG1-ECG2	512	ECG1-ECG2	256/512
Respiratory	Flow Thermistor Thorax Abdomen	200	Flow	10	Flow Thorax Abdomen	10	Flow Thorax Abdomen	200	Flow Thorax Abdomen Press	32	Flow Thermistor Thorax Abdomen	64 16	Nasal Pres Airflow Thor Effort Abdo Effort	64/128 32 32 32
PPG	–	–	–	–	–	–	–	–	Pleth	256	–	–	PlethWV	128
Manufacturer	Nox, Embla		Compumedics		Compumedics		Unknown		Compumedics		Compumedics		Compumedics	

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

The Korea Image-based Sleep Study (KISS) dataset Jeong et al. (2023) is a standardized polysomnography (PSG) dataset collected between 2013 and 2020 across four sleep centers. Recordings were acquired using Embla and NOX-A1 PSG systems, resulting in a total of 10,253 recordings. Sleep stages were scored by experts according to the AASM version 2.6 guidelines Berry (2012); Berry et al. (2017). Each recording contains 21 different biosignals, including several EEG, EOG, and EMG channels, as well as respiratory and movement-related signals. The dataset is publicly available through AI Hub AIHub (2020).

The training split defined in Jeong et al. (2023) are used for training. KISS was not used for downstream evaluation or final testing.

A.4 Korea Video Sleep Study (KVSS) Dataset

The Korea Video Sleep Study (KVSS) dataset is a retrospectively constructed multicenter clinical cohort that includes synchronized infrared sleep video and polysomnography (PSG) recordings with expert annotations Choi et al. (2026). Data were collected under IRB-approved protocols at three hospitals: Chungnam National University Hospital, The Catholic University of Korea St. Vincent's Hospital, and Hallym University Hospital. The complete KVSS cohort contains 936 PSG examinations with synchronized infrared video, while 881 examinations were retained in this study after screening and quality control. The recordings were obtained during routine clinical sleep studies for various sleep-related conditions, including suspected obstructive sleep apnea, insomnia, periodic limb movement disorder, and REM sleep behavior disorder. Infrared video was recorded simultaneously with PSG to maintain temporal synchronization. Although camera position and lighting conditions differed slightly between sites, all videos were standardized to MP4 format with a resolution of 640×480 pixels at 5 fps. PSG recordings were stored in European Data Format (EDF). Sleep stages (Wake, N1, N2, N3, and REM) and AASM-defined events were annotated by certified sleep technologists and reviewed by sleep physicians. The dataset was de-identified, and expert cross-checking was performed to verify synchronization and correct obvious scoring inconsistencies. In our experiments, KVSS was used as a PSG dataset for pretraining. Only the physiological signals required by the model were processed, while the synchronized infrared video data were not used.

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

The Multi-Ethnic Study of Atherosclerosis (MESA) Chen et al. (2015) is a multicenter cohort including 6,814 adults aged between 45 and 84 years from four racial and ethnic groups (White, Black, Hispanic, and Chinese-American participants). As part of Exam 5, conducted between 2010 and 2013, the MESA Sleep study enrolled 2,237 participants who completed a single night of unattended in-home polysomnography (PSG) together with wrist actigraphy. PSG recordings were set up during evening home visits by trained staff, while sleep stages and respiratory events were scored at a centralized sleep reading center using standardized procedures. The National Sleep Research Resource (NSRR) provides the PSG recordings in European Data Format (EDF), together with XML annotation files containing sleep-stage and respiratory-event information. Respiratory-event annotations were harmonized through rule-based post-processing of the original labels to ensure consistent criteria across datasets, following the procedure described in Ahn et al. (2026). The PSG recordings include EEG (Fz-Cz, Cz-Oz, and C4-A1), EOG, EMG, ECG, nasal airflow, thoracic and abdominal respiratory effort, oxygen saturation, and body position. In our experiments, the original MESA training split was kept unchanged with 1,631 recordings. To ensure a consistent comparison with the wav2sleep baseline, the validation set was filtered to 170 recordings, while the final held-out test set was restricted to 62 recordings shared between the two evaluation pipelines. Recordings from the original test set that were not part of this common subset were kept unused.

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

The Osteoporotic Fractures in Men Sleep Study (MrOS Sleep) Blackwell et al. (2011) is a multicenter cohort of older men aged 65 years and above. It includes unattended in-home polysomnography (PSG), which was set up by trained technicians and annotated using standardized scoring procedures. The study was conducted between December 2003 and March 2005, during which 3,135 participants from the original MrOS cohort of 5,994 men completed overnight unattended PSG. The dataset available through the National Sleep Research Resource (NSRR) provides PSG recordings in European Data Format (EDF), together with XML annotation files. These recordings contain multiple physiological signals, including EEG, EOG, EMG, ECG, nasal airflow, thoracic and abdominal respiratory effort, oxygen saturation, and body position. Respiratory-event annotations were harmonized using rule-based post-processing of the original labels to maintain consistent criteria across datasets, following the procedure described in Ahn et al. (2026). In our experiments, data from both MrOS visits were included in the pretraining dataset. For MrOS1, the C3-A2 EEG channel and abdominal respiratory effort were used, resulting in 2,758 recordings. For MrOS2, the C3-M2 EEG channel and abdominal respiratory effort were used, resulting in 991 recordings. A subset of MrOS2 recordings exhibited a different signal amplitude scale and was harmonized during preprocessing.

A.7 Cleveland Family Study (CFS) Dataset

The Cleveland Family Study (CFS) (Zhang et al., 2018; Redline et al., 1995) is a family-based cohort designed to investigate familial risk factors associated with sleep apnea. The cohort includes 2,284 participants from 361 families who were followed across several visits over a period of approximately 16 years. The National Sleep Research Resource (NSRR) provides overnight polysomnography recordings, including full-night PSG data collected during Visit 5. Subjects younger than 14 years were excluded because pediatric sleep studies follow different clinical scoring criteria. The available PSG signals and corresponding sleep-stage annotations were processed using the same preprocessing pipeline applied to the other datasets.

CFS is excluded from *Sleep2Word*'s training stages. For downstream evaluation, however, we train a separate task-specific head using labels from the CFS training split and evaluate it on the held-out CFS test split. Thus, the CFS results assess the transfer of a fixed translation model to an unseen cohort under cohort-specific supervised downstream evaluation, rather than zero-shot transfer of the complete prediction pipeline. In contrast, CFS is an internal dataset for the wav2sleep baseline, as it is included among the datasets used for model training and evaluation. Therefore, while the wav2sleep results are computed on its held-out CFS test split, CFS does not represent an unseen dataset for wav2sleep, unlike in our evaluation setting.

B Model Specification

Table 8 and Table 9 summarize *Sleep2Word* and CNN-Transformer model (C-T baseline) specifications, respectively. C-T baseline is trained with the same training recipe as translation model, described in Table 11.

We also evaluated larger C-T baseline variants with increasing model capacity and window size. These variants did not converge reliably under the varied training configurations. One of the tested larger model specification and training results are provided in Table 10. This observation is limited to the settings tested and does not rule out improvements from further optimization.

C Implementation Details

We provide training hyperparameters in Table 11. The hyperparameters used for each downstream tasks are summarized in Table 12. We have trained and evaluated *Sleep2Word* and baseline methods on a single NVIDIA H200 GPU (140 GiB HBM3e) with two Intel Xeon Platinum 8462Y+ CPUs. Models are implemented in PyTorch 2.5.1 (CUDA 12.1, cuDNN 9.1) with PyTorch Lightning 2.6.0, and trained in bf16 mixed precision with distributed data parallelism. End-to-end training takes approximately 10 hours per single modality.

PPG pretraining. PPG channel is not included in SleepMaMi configuration and is only available in MESA dataset. To tackle this, we adopt two-stage pretraining strategy. First, we pretrain SleepMaMi using three modalities (EEG, ECG and Respiratory) on the full training dataset. We then introduce a PPG-specific private encoder and fine-tune only this encoder on the MESA PPG data while keeping all other modules frozen. This procedure maps PPG inputs into the common representation space learned from the other modalities, enabling alignment across all four modalities.

Sleep-stage classification. We use a CNN-TCN architecture. CNN layers extract local spectral patterns from the reconstructed EEG spectrogram. The TCN then models temporal dependencies across consecutive sleep epochs using dilated convolutions. The final features are projected to four sleep-stage classes. The model is trained for 20 epochs with weighted cross-entropy loss.

Wordbook initialization. As described in Section 3, the wordbook is initialized using data-driven approach. Wordization is disabled for the first I batches, during which chunk embeddings and their lengths are collected. The dominant chunk lengths are then identified, and chunks are allocated to these length groups in proportion to the number of chunks. Within each group, the collected chunk embeddings are clustered once using Lloyd's algorithm to initialize the corresponding *words*.

α is calibrated as the median distance between each chunk to the nearest centroid of a different length group, divided by the smallest log-duration gap between initial length groups. A smaller value may underweight duration, causing chunks with similar contents but different lengths to be assigned to the same word. In contrast, a large α may cause the duration to dominate content, making assignment primarily length-dependent. This calibration balances the contributions of content and duration to word assignment.

PasL reproduction. We reproduce PasL using the architecture and training setup described in the original paper. EEG spectrograms have a resolution of 256×512. A VQGAN tokenizer maps each spectrogram to an 8×64 grid of 512 discrete tokens using an 8,192-entry codebook. The masked Transformer uses an 8-layer encoder and decoder, 768-dimensional hidden states, and 8 attention heads. The model is trained for 500 epochs with AdamW using the masked-token cross-entropy (CE) objective. Reconstructed EEG spectrograms are obtained with the frozen VQGAN decoder. Training converges stably, although our reproduced downstream performance remains slightly below the values reported in the original paper.

D EEG Spectrogram

In *Sleep2Word*, we model the target EEG as a time-frequency representation rather than as the raw waveform. Although peripheral signals encode information coupled to brain activity, they are unlikely to support reliable recovery of the fine-grained, sample-level details of EEG waveforms. In contrast, the EEG spectrogram provides a physiologically meaningful and temporally localized

Table 8: *Sleep2Word* model specifications. D denotes the shared latent width; one base token covers 1 s of signal (30 tokens per epoch), L' is the number of base tokens and C^m is the number of words in a full night recording.

Component	Module	Configuration	Output	Notes
Frozen Base Encoder	SleepMaMi	Micro-encoder private and shared modules	$(L', 256)$	
Parameters	≈20.4M total			
Chunking	Chunk encoder	2× Mamba blocks (d_{state}=16, conv=4, expand=2, Gated MLP with ×4 expansion)	$(L', 256)$	
	Router	$W_q, W_k \in \mathbb{R}^{256\times256}$	$(L', 256)$	
Encoder	Encoder	4×(2 sliding-window Transformer + 1 Full Transformer) D=256, 8 heads, Gated MLP with ×4 expansion, RoPE at word start	$(C^m, 256)$	
Decoder	Block AR	Depth 2, Self + Cross attention per sync block id head, length head	$(C^{EEG}, \lvert\mathcal{W}\rvert + 1)$ $(C^{EEG}, 2r + 1)$	 r: Residual slot size
Dechunking	Wordbook lookup Dechunk encoder	Scatter to the base grid (STE) 2× Mamba blocks	$(T, 256)$ $(T, 256)$	
Reconstruction head	Spectrogram head	Conv1d((k, s)=$(30, 6)$) → Conv1d((k, s)=$(5, 5)$) → Linear (256→N_f)	(T_y, N_f)	Only for EEG
	Upsampling head	Conv1d((k, s)=$(30, 1)$) → Conv1d((k, s)=$(30, 1)$) → Linear (256→f_s)	$(L, 1)$ $(L = L' \times N_f)$	for peripherals
Parameters	≈25.8M total			

Table 9: CNN-Transformer baseline specifications. L is the total length of full-night vector (100 Hz × the number of seconds in full-night recording)

Component	Configuration	Output
InstanceNorm	RevIN per 30 s epoch, no affine	$(L, 1)$
Conv tokenizer	Conv1d D : 1→32→64→128→256, Depth: 1, 1, 4, 4 Kernel size: 10, 5, 1, 5 Stride is equal to kernel size for the last layer. Otherwise 1 is used.	$(L/250, 256)$
Reshape	Reshape seq from $(L/250, 256)$→$(240, 256)$ (Each seq is 20 epochs) Batch size increases by $L/(240 \times 250)$	$(240, 256)$
Encoder	(2 sliding-window Transformer + 1 Full Transformer)×4 D=256, 8 heads, Gated MLP with ×4 expansion, RoPE	$(240, 256)$
Reshape	Return to the original shape	$(L/250, 256)$
Spectrogram head	Conv1d((k, s)=$(12, 3)$) → Conv1d((k, s)=$(4, 4)$) → Linear (256→N_f)	$T_y \times N_f$
Parameters	≈14.9M total	

representation of spectral activity that is more suitable for cross-modal estimation, as suggested by Zha et al. (2026).

The EEG spectrogram is built in the following procedure. Each EEG recording is normalized by the standard deviation of each dataset's training split and partitioned into non-overlapping windows of ω seconds, from which log-power multi-taper spectra (Prerau et al., 2017) are computed over 0.5-32 Hz. Any windows carrying no usable signal (flat or non-finite, e.g., a detached electrode) are set to zero. The dB values are subsequently mapped to [0,1] using the 0.1st and 99.9th percentiles measured over the training corpus. The same percentile-based mapping is applied across datasets and splits after dataset-specific amplitude normalization. Accordingly, the resulting values represent power on a common normalized scale. We use ω of five seconds as our default setting, except the reconstruction experiment where thirty seconds is used for fair comparison to other baselines.

E MODIFICATIONS TO SLEEPMAMI ARCHITECTURE

We use the shared embeddings from SleepMaMi (Park et al., 2026), with several architectural modifications to support translation from a single source modality. In the original architecture, the shared module receives the concatenated embeddings of all available modalities. Consequently, generating shared embeddings requires access to multiple modalities at once, including the target modality - this is not permissible in our translation setting. We therefore modify the shared module to process each modality independently and align the resulting representations across modalities. This design allows the model to encode a source signal without requiring the target signal while preserving a common cross-modal representation space.

Table 10: Larger CNN-Transformer baseline specifications and training results.

Component	Configuration	Output
InstanceNorm	RevIN per 30 s epoch, no affine	$(L, 1)$
Conv tokenizer	Conv1d D : 1→32→64→128→256, Depth: 4, 8, 16 Kernel size: 10, 5, 5 Stride is equal to kernel size for the last layer. Otherwise 1 is used.	$(L/250, 256)$
Refine module	ConvNext Layer × 6	$(L/250, 256)$
Encoder	(2 sliding-window Transformer + 1 Full Transformer)×4 D=256, 8 heads, Gated MLP with ×4 expansion, RoPE	$(L/250, 256)$
Spectrogram head	Conv1d((k, s)=$(12, 3)$) → Conv1d((k, s)=$(4, 4)$) → Linear (256→N_f)	$T_y \times N_f$
Parameters	≈22.8M total	
Training results	Trained for 10 epochs, MAE: 0.0837, SNR: 11.8015, Test MF1: 0.170 on SHHS1	

Table 11: Hyperparameter settings for training.

Stage	Item	Value	Notes
Context-aware chunking	Batch size	32	PSG cases
	Optimizer	AdamW	
	β_1	0.9	
	β_2	0.99	
	Weight decay (λ)	0.01	
	Initial learning rate	2.00×10^{-3}	
	Learning rate schedule	Cosine annealing with warmup	
	Final learning rate	1.00×10^{-8}	
	Training epochs	5	
	Patch size	1 Second	
	Target chunking length	4	
	Weight for balancing ratio loss (λ_{ratio})	0.03	
	Weight for Align loss (λ^m_{align})	0.2	$m \in$ {ECG, PPG, Respiratory}, 0 for EEG
	Weight for barrier loss ($\lambda_{barrier}$)	0.01	
Wordization	Batch size	32	PSG cases
	Optimizer	AdamW	
	β_1	0.9	
	β_2	0.99	
	Weight decay (λ)	0.01	
	Initial learning rate	1.00×10^{-4}	
	Learning rate schedule	Cosine annealing with warmup	
	Final learning rate	1.00×10^{-8}	
	Training epochs	5	
	Weight for quantization loss (λ_Q)	0.5	
	Wordbook size (K)	2048	
Translation	Batch size	8	PSG cases
	Optimizer	AdamW	
	β_1	0.9	
	β_2	0.99	
	Weight decay (λ)	0.01	
	Initial learning rate	2.00×10^{-3}	
	Learning rate schedule	Cosine annealing with warmup	
	Final learning rate	1.00×10^{-8}	
	Training epochs	5	
	Block size for AR	3	Epochs (90 seconds)
	Weight for duration prediction loss (λ_{dur})	1.0	

We also modify the original leave-one-out contrastive learning (LOO-CL) objective. For each source modality, we randomly sample a subset of the remaining modalities and contrast the representations averaged over them with that of the source modality. Exposing the model to different modality subsets encourages it to capture physiological features that are consistent across diverse combinations of modalities.

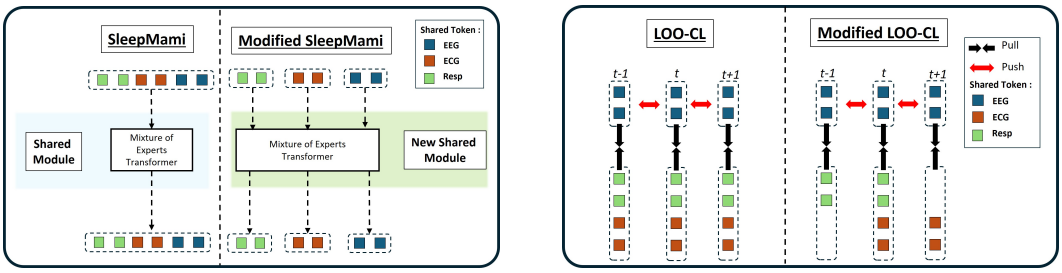
Finally, we modify the fusion module to combine the shared embeddings obtained from a randomly selected subset of modalities with the private embeddings. This modification further encourages the shared representations to encode modality-agnostic information.

F AUXILIARY LOSSES

This section provides a more detailed description of auxiliary losses used during the context-aware chunking stage.

Table 12: Training details used for downstream tasks.

Downstream task	Item	Value	Notes
Sleep staging	Model	CNN-TCN	PSG cases
	Loss	Weighted Cross Entropy	
	Optimizer	AdamW	
	Initial learning rate	3.00×10^{-4}	
	Weight decay	1.00×10^{-2}	
	LR schedule	Cosine annealing	
Arousal detection	Model	Linear ($256 \rightarrow 2$)	PSG cases
	Loss	CrossEntropy	
	Optimizer	Adam	
	Initial learning rate	1.00×10^{-2}	
	Weight decay	1.00×10^{-4}	
Survival prediction	Model	Linear ($256 \rightarrow 3$)	PSG cases
	Input	Full-night mean pooling	
	Loss	Cox PH	
	Optimizer	AdamW	
	Initial learning rate	1.00×10^{-2}	
	Weight decay	1.00×10^{-4}	
Common settings	Batch size	8	
	Training epochs	20	



(a) Shared module modification

(b) LOO-CL objective modification

Figure 5: Overview of the modifications to the SleepMaMi architecture. (a) The shared module processes each modality independently, enabling single-modality inputs while preserving a common representation space. (b) LOO-CL is modified to handle different subsets of available modalities.

Ratio loss ($\mathcal{L}_{\text{ratio}}$): We adopt this loss to regulate the proportion of predicted boundaries. The formulation follows Hwang et al. (2026), where N controls the target chunking ratio.

$$\mathcal{L}_{ratio} = \frac{N}{N-1}((N-1)FG + (1-F)(1-G)), \quad F = \frac{1}{L'}\sum_i \mathbb{1}_{p_i \geq 0.5}, \quad G = \frac{1}{L'}\sum_i p_i \tag{9}$$

Alignment loss ($\mathcal{L}_{\text{align}}$): This loss prevents the refined embeddings of chunk encoder from deviating substantially to modality-specific representation space.

$$\mathcal{L}_{align} = \sum_{c=1}^{C} \|\bar{\mathbf{z}}_c - \bar{\mathbf{e}}_c\|_2^2, \quad \bar{\mathbf{e}}_c = \frac{1}{l_c}\sum_{t=s_c}^{s_c+l_c-1} \mathbf{e}_t \tag{10}$$

Table 13: Ablation on loss components.

Index	$\mathcal{L}_{\text{Recon}}$	$\mathcal{L}_{\text{Ratio}}$	$\mathcal{L}_{\text{Align}}$	$\mathcal{L}_{\text{Barrier}}$	Acc.
1	✓	✓			0.649
2	✓	✓	✓		0.746
3	✓	✓	✓	✓	**0.753**

Barrier loss ($\mathcal{L}_{\text{barrier}}$): This loss encourages boundaries to be inserted at regular temporal intervals, thereby the resulting chunks are well placed for the block autoregressive decoder.

$$\mathcal{L}_{\text{barrier}} = -\sum_{a \in \mathcal{A}} \log(p_a), \quad \text{where} \quad \mathcal{A} = \{1 + jB : j \geq 1, 1 + jB \leq L'\} \tag{11}$$

G Ablation

We conduct ablation study to investigate the individual contribution of our core components.

Loss components. Table 13 provides the ablation results for the auxiliary loss components. We conduct this experiment using the SHHS1 training set and evaluate respiratory-to-EEG sleep staging accuracy on SHHS1 under the linear-probing protocol. Removing $\mathcal{L}_{\text{Barrier}}$ results in accuracy decrease of 0.007 (0.7 percentage points), demonstrating its contribution to maintaining temporal alignment between the source and target sequences. Removing $\mathcal{L}_{\text{Align}}$ and $\mathcal{L}_{\text{Barrier}}$ decreases accuracy by 0.104 (10.4 percentage points), suggesting that the alignment objective prevents the refined chunk representations from drifting toward modality-specific features. We do not report the configuration without $\mathcal{L}_{\text{Ratio}}$, since removing it causes the chunking mechanism to collapse and does not provide meaningful results.

Chunking mechanism. To assess the effectiveness of *context-aware chunking* and *wordization*, we compare *Sleep2Word* against two alternative methods: *Base2base* performs no chunking or wordization and directly translates the source base embeddings $\mathbf{e}^m$ into the target base embeddings $\mathbf{e}^{EEG}$. *FixedPool* replaces context-aware chunking with average pooling over non-overlapping windows of four base embeddings, matching the chunking ratio of $N = 4$. Table 14 provides respiratory-to-EEG sleep staging accuracy on SHHS1 under the linear-probing protocol, using SHHS1 training set. *Sleep2Word* outperforms Base2base by 0.022 (2.2 percentage points), demonstrating that word-level abstraction improves downstream performance. It also surpasses FixedPool by 0.020 (2.0 percentage points), showing that adaptively locating chunk boundaries preserves more task-relevant temporal structure at a comparable compression level. These gains highlight that the benefit of *Sleep2Word* arises not only from shorter sequences and improved computational efficiency, but also from learning context-aware, physiologically meaningful abstractions.

Duration awareness and duration residual prediction. Each word in *Sleep2Word* consists of a content component $\nu_k \in \mathbb{R}^d$ and a duration component $\eta_k \in \mathbb{R}$. The decoder additionally predicts an occurrence-specific duration residual to account for variations among chunks assigned to the same word. Table 15 evaluates the contributions of these two components. We conduct this experiment using the SHHS1 training set and report respiratory-to-EEG sleep staging accuracy on SHHS1 under the linear-probing protocol.

In the content-only variant, the word identity is determined solely by the content. Incorporating the duration term η_k in the wordbook enhances accuracy from 0.732 to 0.748, an increase of 0.016 (1.6 percentage points). This result suggests that explicitly distinguishing chunks with similar content but different durations improves temporal alignment during translation. Residual prediction further improves accuracy by 0.005 (0.5 percentage points), indicating that modeling deviations from the prototypical duration helps reduce residual temporal discrepancy.

Wordbook size. Table 16 provides the effect of varying K. We conduct this experiment using the full training set and evaluate respiratory-to-EEG sleep staging accuracy on SHHS1 under the linear-probing protocol, while fixing $B = 3$ and $N = 4$. Performance improves as the wordbook size increases, with the gains saturating at $K = 2048$. This suggests that a sufficiently large vocabulary is required to capture diverse patterns, while further increasing its size provides limited benefit.

Table 14: Ablation on chunking mechanism.

Item	Acc.
Base2base	0.731
FixedPool	0.733
Sleep2Word	**0.753**

Table 15: Ablation on duration modeling.

Item	Acc.
Content-only wordbook	0.732
+ Content- and duration-aware wordbook	0.748
+ Occurrence-specific residual prediction	**0.753**

Table 16: Ablation on K.

K	Acc.
256	0.710
512	0.746
1024	0.752
2048	**0.757**

Table 17: Ablation on N.

N	Acc.
2	**0.757**
4	**0.757**
8	0.740
16	0.754

Table 18: Ablation on B

B (epochs)	Acc.
3	**0.757**
5	0.746
10	0.729

Table 19: Effective compression ratio.

Modality	Compression ratio (%)
EEG	67.6
Respiratory	64.9
ECG	64.3
PPG	69.7

Target chunking ratio. Table 17 provides the results for different values of N. We conduct this experiment using the full training set and evaluate respiratory-to-EEG sleep staging accuracy on SHHS1 under the linear-probing protocol, while fixing $K = 2048$ and $B = 3$. Increasing the chunking ratio up to $N = 4$ does not harm the model performance, but larger values lead to a modest decrease in accuracy. These results suggest that the wordization mechanism can effectively compress the target sequence while preserving task-relevant information. The experiment is done under linear probing protocol using the full training set and respiratory-to-EEG sleep staging accuracy on SHHS is reported.

Block size. Table 18 provides the effect of varying B. We conduct this experiment using the full training set and evaluate respiratory-to-EEG sleep staging accuracy on SHHS1 under the linear-probing protocol, while fixing $K = 2048$ and $N = 4$. Accuracy decreases as the block size increases, potentially because temporal misalignment accumulates over longer generation spans. This result highlights the importance of introducing block boundaries at appropriate intervals to limit temporal drift.

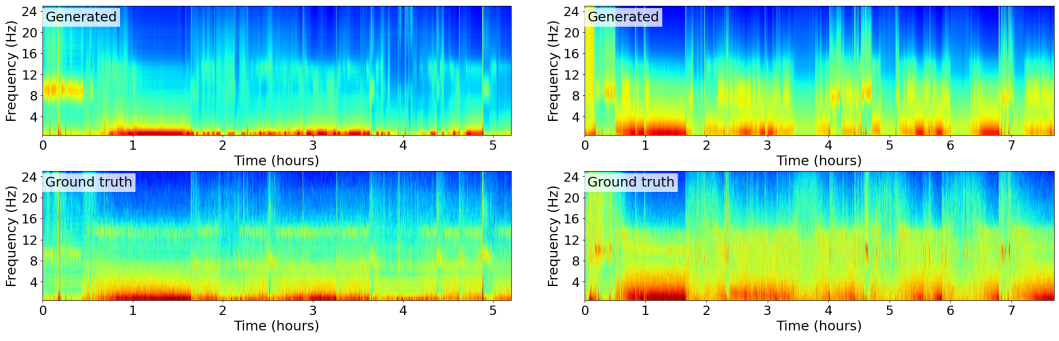
H Compression Ratio

The modality-specific wordbooks reduce sequence length by representing each variable-length chunk with a single token. We quantify this effect using pretrained checkpoints and 50 subjects from SHHS1 or MESA (PPG). Specifically, we define the effective compression ratio as $100(1 - C/L')$, where L' and C denote the number of base patches and resulting word tokens, respectively. All measurements are obtained with $N = 4$.

As presented in Table 19, the wordbooks reduce the full-night sequence length by 64.3% to 69.7%, depending on the modality. These empirical reductions are moderately below the target of 75%, implied by $N = 4$. This discrepancy arises because chunk boundaries and lengths are adaptively determined based on their context.

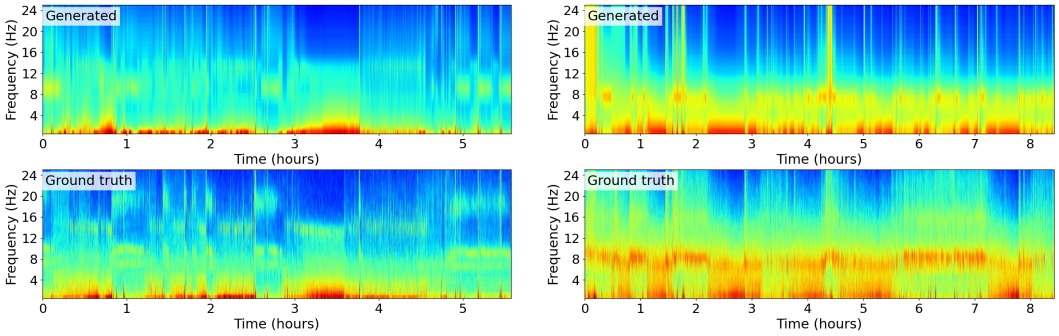
I GENERATED SPECTROGRAM SAMPLES.

We provide selected samples of generated spectrograms using respiratory, ECG and PPG, respectively in Figure 6, 7, and 8.



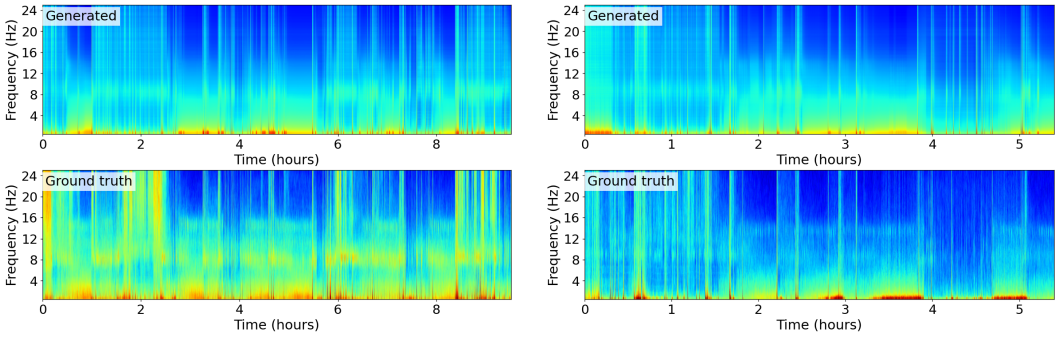
(a) Spectrogram : 64-year-old Female (b) Spectrogram : 65-year-old Female

Figure 6: Generated spectrograms from respiratory.



(a) Spectrogram : 46-year-old Male (b) Spectrogram : 79-year-old Female

Figure 7: Generated spectrograms from ECG.



(a) Spectrogram : 55-year-old Male (b) Spectrogram : 65-year-old Male

Figure 8: Generated spectrograms from PPG.