

SleepLM: Natural-Language Intelligence for Human Sleep

Zongzhe Xu¹, Zitao Shuai¹, Eideen Mozaffari¹, Ravi S. Aysola¹, Rajesh Kumar¹, Yuzhe Yang^{1†}

¹University of California, Los Angeles

We present SleepLM, a family of sleep-language foundation models that enable human sleep alignment, interpretation, and interaction with natural language. Despite the critical role of sleep, learning-based sleep analysis systems operate in closed label spaces (e.g., predefined stages or events) and fail to describe, query, or generalize to novel sleep phenomena. SleepLM bridges natural language and multimodal polysomnography, enabling language-grounded representations of sleep physiology. To support this alignment, we introduce a multilevel sleep caption generation pipeline that enables the curation of the *first* large-scale sleep-text dataset, comprising over 100K hours of data from more than 10,000 individuals. Furthermore, we present a unified pretraining objective that combines contrastive alignment, caption generation, and signal reconstruction to better capture physiological fidelity and cross-modal interactions. Extensive experiments on real-world sleep understanding tasks verify that SleepLM outperforms state-of-the-art in zero-shot and few-shot learning, cross-modal retrieval, and sleep captioning. Importantly, SleepLM also exhibits intriguing capabilities including language-guided event localization, targeted insight generation, and zero-shot generalization to unseen tasks.

Code: <https://github.com/yang-ai-lab/SleepLM>

Website: <https://yang-ai-lab.github.io/SleepLM>

1. Introduction

Humans move between two distinct states: the *waking* life, structured by perception and language; and *sleep*, expressed through dense and continuous physiology. Sleep is not simply rest; it is a dynamic orchestration shaped by interactions among brain oscillations, cardiac regulation, and respiratory rhythms [7, 10, 18]. Polysomnography (PSG) captures this narrative through synchronized channels (e.g., EEG, ECG, EMG, EOG, respiration), offering a high-resolution view into human health and encoding rich biomarkers for cardiovascular, neurological, and metabolic function [1, 39, 41].

As with waking experience, making sense of sleep requires a mapping from *physiology* to *language*. The same PSG pattern can be described from low-level signal shifts (e.g., “*EEG delta power increases and heart rate slows*”) to high-level sleep states (e.g., “*transition into deep sleep*”). Learning this translation and alignment is the key to interpreting and interacting with high-frequency sleep physiology through actionable language descriptions, which is essential for democratizing sleep health [15], enabling personalized monitoring [26] and advancing clinical analysis beyond fixed categorization.

Yet, current computational methods do not meet this need. On one hand, existing machine learning models for sleep are predominantly *discriminative* and confined to *closed* label spaces (e.g., sleep stages or events) without the capacity for open-ended description or reasoning [5, 27, 30]. On the other hand, while Large Language Models (LLMs) excel at generative tasks, they are inherently ill-equipped to handle the high-dimensional, continuous nature of physiological data [25]: Fig. 1

[†]Correspondence to: yuzhey@ucla.edu.

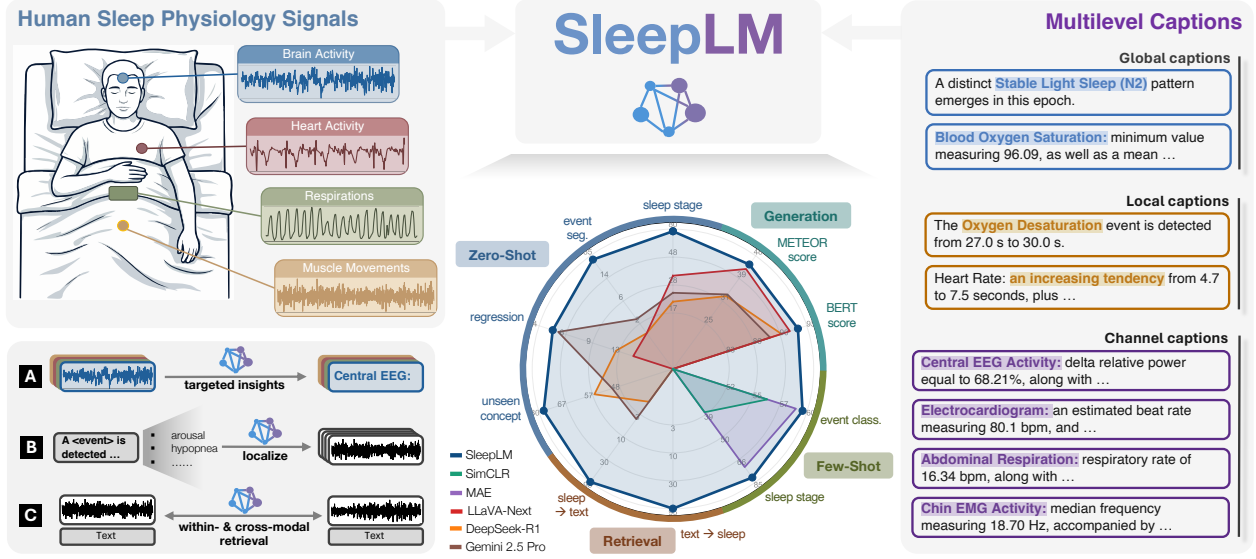


Figure 1 | **Sleep-language foundation models (SleepLM)**. We present a comprehensive study using over 100K hours of multimodal sleep PSG data from over 10,000 individuals. We design a multi-level captioning pipeline that captures PSG information at different temporal and semantic granularities. Across a wide range of downstream tasks and evaluation settings, SleepLM consistently outperforms state-of-the-art LLMs and VLMs. In addition to its predictive capabilities, SleepLM also enables: (A) targeted and controlled insights generation, (B) language-guided event localization, and (C) within- and cross-modal zero-shot retrieval (details in Sec. 5).

shows the poor zero-shot performance of Gemini 2.5 Pro [14] and DeepSeek-R1 [21] (details in Sec. 5), leaving the rich narrative of sleep largely inaccessible to modern generative AI.

To fill the gap, we present SleepLM, to our knowledge, the first family of sleep-language foundation models that unlocks meaningful interpretation of raw sleep signals and enable novel sleep applications through natural language. The efficacy of SleepLM represents a paradigm shift from *categorizing* sleep to *describing* and *interacting* with it, enabled by three key innovations: ❶ A hierarchical, automated captioning pipeline that generates structural descriptions for each sleep epoch, spanning global semantic summaries, local fine-grained details, and channel-specific characteristics; ❷ The curation of the *largest* paired sleep-text dataset to date, comprising over 100,000 hours of data from more than 10,000 individuals, enabling the learning of robust, generalized representations of human sleep; and ❸ ReCoCa, a novel and generic multimodal pretraining architecture that utilizes a compound objective (contrastive alignment, caption generation, and signal reconstruction) to support scalable learning of joint language and physiological time-series data.

To rigorously evaluate SleepLM, we benchmark it against state-of-the-art (SOTA) methods across diverse, real-world sleep understanding tasks. Extensive experiments verify that SleepLM not only achieves superior performance on established tasks, but also enables new capabilities such as controllable insight generation and zero-shot generalization to novel clinical concepts. Our contributions are as follows:

- We introduce a multilevel captioning pipeline for sleep data, yielding the largest sleep-text dataset to date with over 100,000 hours of data from over 10,000 people.
- We design SleepLM, the first family of sleep-language foundation models that enables diverse sleep capabilities and interaction through natural language.
- We present ReCoCa, a unified and generic multimodal pretraining architecture for joint learning

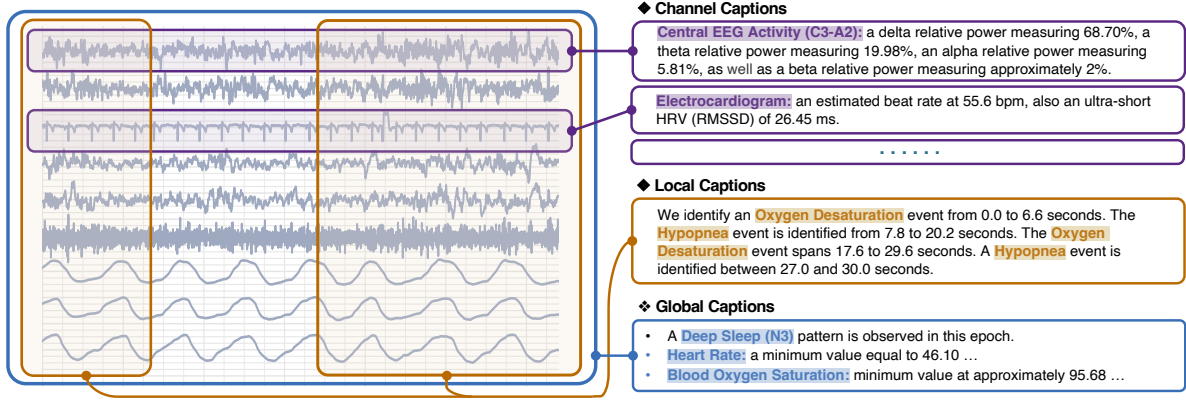


Figure 2 | **Multilevel sleep captioning pipeline.** We generate three complementary levels of captions from each PSG window: (1) **Channel** captions summarize modality-specific clinically relevant statistical features commonly used in manual scoring; (2) **Local** captions capture temporal semantics such as transient morphological changes and sleep event onsets and durations; (3) **Global** captions describe high-level physiological states such as sleep stage and overall cardiac and respiratory conditions. Example captions are provided in Appendix E.1.

over language and physiological time-series data.

- Extensive experiments across real-world sleep tasks verify the superiority of SleepLM against SOTA methods, and reveal emergent capabilities including controllable insight generation and generalization to unseen concepts.

2. Related Work

Sleep Foundation Models. Sleep research has begun adopting large-scale pretraining for physiological signals, producing strong representations for discriminative tasks such as sleep staging and event detection [11, 35, 47]. Current methods focus on self-supervised learning for biological signals, including inter-channel contrastive learning [37], masked reconstruction [29], and predictive coding for long-range sleep structure [19]. However, existing sleep foundation models remain unimodal and optimized for classification, limiting their ability to generate descriptions or handle new concepts. In contrast, our work aligns raw PSG with *natural language* to support open-ended description, controllable reporting, and zero-shot generalization beyond fixed label sets.

Time Series Foundation Models. Time series modeling has advanced with foundation models that generalize across many downstream settings, including weather forecasting and financial data [3, 13, 40]. Recent work also studies how to represent temporal structure effectively, including tokenization and parameterization [28], pretraining objectives [45], architectural choices [16], and embedding strategies [4]. Although physiological signals share the same temporal form, they differ in sampling rate, recording length, and fixed channel structure. In contrast, we adapt these time-series design principles to PSG by explicitly accounting for multi-channel physiology and long, dense recordings, and by pairing signals with language for interpretability and interaction.

Multimodal Language Models. Connecting non-text modalities with language models is now central in multimodal learning. In vision, CLIP [32] showed that contrastive alignment between images and text enables strong zero-shot transfer, and CoCa [43] combined contrastive learning with captioning to support both retrieval and generation. These alignment ideas have since been extended to temporal data: OpenTSLM [23] adapts language-modality alignment to general time series, while SensorLM [48] pairs wearable signals with automatically generated text to enable natural language interaction with sensor data. We extend this language-signal alignment framework to multi-channel PSG by

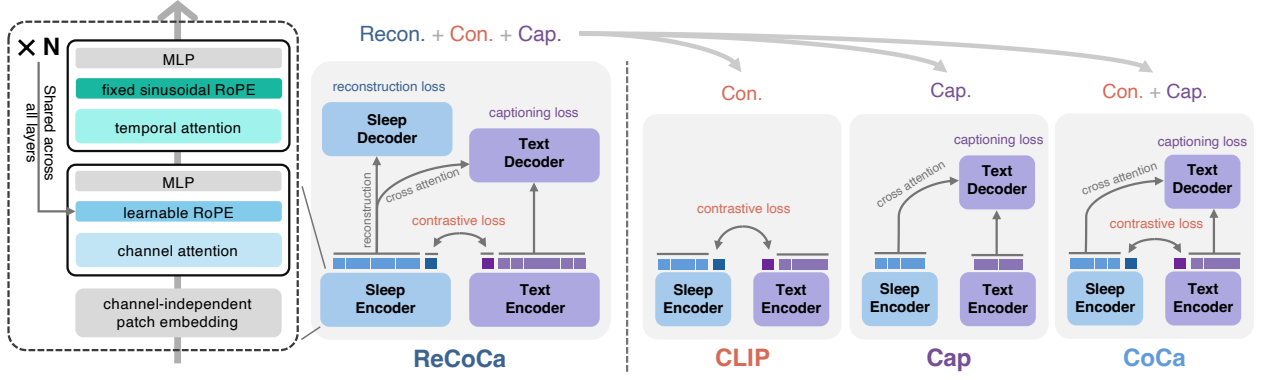


Figure 3 | **The SleepLM architecture, pretraining objectives, and variants.** We introduce ReCoCa, a generic sleep-language pretraining framework that jointly optimizes signal reconstruction, contrastive alignment, and caption generation for multi-channel PSG. By enabling or disabling components, ReCoCa yields common formulations (e.g., CLIP, Cap, CoCa) as special cases.

introducing a sleep-specific captioning pipeline, a large paired sleep-text dataset, and a multimodal pretraining architecture designed for dense physiological time series.

3. Human Sleep Captioning at Scale

PSG Data and Processing. We source our primary pretraining corpus from the National Sleep Research Resource [46]. Specifically, we aggregate five large-scale datasets: SHHS [31], MrOS [9], CFS [33], [34], and WSC [42].

For preprocessing, continuous recordings are segmented into non-overlapping 30-second epochs, following standard clinical sleep scoring conventions [2]. We use a standardized 12-channel montage grouped by physiological modality: ❶ *Brain (EEG/EOG)*: C3–A2, C4–A1, E1–A2, E2–A1. ❷ *Respiration*: Thorax, Abdominal effort, Airflow. ❸ *Cardiac*: ECG, Heart Rate, SpO2. ❹ *Somatic*: Chin EMG, Body Position. Detailed preprocessing steps are provided in Appendix B.

Multilevel Sleep Caption Generation. Foundation models benefit from dense, structured supervision, yet most sleep datasets provide only a single coarse label per 30-second epoch (e.g., stage label such as “N2”), creating a strong information bottleneck [8]. To address this, we introduce a **multilevel** captioning pipeline that produces hierarchical, text supervision for each PSG epoch. The captions are organized at three complementary granularities:

Channel Captions: These captions describe signal morphology that is directly observable in the input channels. For each modality (e.g., EEG), we extract clinically used features from standard scoring practice, such as EEG band power and respiratory periodicity and variability. We then render these features into diverse linguistic templates, promoting robustness to language variation while keeping the supervision tightly grounded to the measured signals.

Local Captions: This level captures within-epoch temporal structure and event localization. We apply trend and peak detection to identify transient changes such as heart-rate accelerations and oxygen denaturation. Beyond event presence, we also indicate onset and offset timestamps within each epoch (Fig. 2), providing fine-grained temporal supervision to enable both event *identification* and *localization*.

Global Captions: At the highest level, we integrate semantic summaries of the epoch’s holistic state, including sleep stage (e.g., “N2”, “REM”) and global autonomic descriptors. Notably, we derive some of

Table 1 | Zero-shot classification and regression. We compare SleepLM with fine-tuned VLMs and state-of-the-art LLMs across a broad set of tasks. Results are averaged over the SHHS (internal), MrOS (internal), and CFS (external) evaluation sets. Sleep-event IoU and balanced accuracy are averaged over {Central Apnea, Hypopnea, Oxygen Desaturation, Arousal}. Heart rate MAE is averaged over {Mean, Min, Max}, and SpO₂ MAE over {Min, Mean}. Channel statistics MAE is averaged over a diverse set of per-channel statistics (Appendix B.2). Detailed task setups and additional results are provided in Appendix C and D.5.

Model	Sleep Stage		Sleep Event		Heart Rate		SpO ₂		Channel Stats
	AUC [↑]	BAcc [↑]	IoU [↑]	BAcc [↑]	MAE [↓]	Recall [↑]	MAE [↓]	Recall [↑]	SMAPE [↓]
<i>Fine-tuned VLMs:</i>									
Qwen3-VL-8B-Instruct [6]	<u>70.2</u>	<u>51.3</u>	<u>13.5</u>	<u>59.8</u>	14.25	22.2	2.73	18.4	<u>34.62</u>
LLaVA-Next [24]	58.9	33.7	7.4	54.0	14.67	22.2	2.58	<u>18.8</u>	36.34
<i>State-of-the-art LLMs:</i>									
DeepSeek-R1 [21]	50.9	20.8	1.6	50.8	9.42	<u>27.3</u>	1.88	9.4	-
Gemini 2.5 Pro [14]	52.2	24.5	2.8	51.7	<u>2.01</u>	10.7	<u>2.20</u>	12.1	-
SleepLM	85.4	76.9	30.4	74.3	1.97	35.8	2.24	39.1	3.15
Gains	+15.2	+25.6	+15.9	+14.5	+0.04	+8.5	-0.36	+20.3	+21.47

these descriptors from withheld signals (e.g., average heart rate and baseline SpO₂) that are excluded from the model input, encouraging inference of these consequences from the remaining biosignals and acting as a masked-prediction proxy task during pretraining.

Large-Scale Pretraining Sleep-Text Dataset. Building on the aggregated PSG corpus and the multilevel captioning pipeline, we construct the *first* large-scale paired sleep-text dataset. In total, the dataset comprises over 100,000 hours of PSG data spanning over 12,000 recording nights from more than 10,000 individuals. Detailed cohort statistics and dataset splits are provided in Appendix B.

4. SleepLM

SleepLM introduces a generic sleep-language pretraining framework for learning joint representations of sleep PSG and text. SleepLM is trained with a compound objective that combines contrastive alignment, caption generation, and signal reconstruction. By enabling or disabling components, this framework instantiates common formulations (e.g., CLIP, Cap, CoCa) as special cases.

Reconstructive Contrastive Captioner (ReCoCa). We present ReCoCa, a unified multimodal pretraining architecture designed for dense physiological time series and forms the default instantiation of the SleepLM family (Fig. 3):

Channel-Specific Sleep Encoder. PSG channels are not interchangeable: they follow a fixed montage with modality-specific meaning (e.g., EEG vs. EOG vs. respiratory effort). To capture this structure, ReCoCa first applies a channel-independent patch embedding so that each sensor’s local morphology is encoded before cross-channel mixing. The resulting tokens are processed by interleaved temporal-attention and channel-attention blocks. Temporal attention uses sinusoidal RoPE [36] along time, while channel attention uses a shared learnable RoPE along the sensor dimension, allowing the model to represent montage topology and stable inter-channel relationships.

Signal Reconstruction Decoder. Text supervision is sparse relative to the information presented in PSG. If trained only to align with captions, the encoder may discard waveform details that are not explicitly described. ReCoCa therefore includes a lightweight reconstruction decoder that predicts the original signal patches from the encoder latents, encouraging physiologically grounded representations and acting as a regularizer during multimodal pretraining.

Modality-Conditioned Text Decoder. PSG descriptions can target different physiological systems, and generating a single caption that covers all channels can be inefficient. We therefore group captions into four systems, $\mathcal{M} = \{\text{Brain}, \text{Respiratory}, \text{Cardiac}, \text{Somatic}\}$. During training and inference, we sample a target system $m \in \mathcal{M}$ and prepend a learnable token $[m]$ to the decoder input. This conditioning guides the decoder to focus on one system at a time, enabling *controllable* and *targeted* caption generation.

Pretraining Objectives. ReCoCa uses a composite loss function that combines *contrastive*, *reconstruction*, and *generative* objectives. Given a batch of N paired examples $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n \in [N]}$, where $\mathbf{x}$ denotes a PSG epoch and $\mathbf{y}$ denotes its caption, let the sleep encoder produce a pooled embedding $\mathbf{s}_i$ (e.g., a CLS token) and the text encoder produce a pooled embedding $\mathbf{v}_i$. We first employ a symmetric InfoNCE contrastive objective [12]:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N} \left(\underbrace{\sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{s}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1, j \neq i}^N \exp(\text{sim}(\mathbf{s}_i, \mathbf{v}_j)/\tau)}}_{\text{sleep-to-text}} + \underbrace{\sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{s}_i)/\tau)}{\sum_{j=1, j \neq i}^N \exp(\text{sim}(\mathbf{v}_i, \mathbf{s}_j)/\tau)}}_{\text{text-to-sleep}} \right).$$

where $\text{sim}(\cdot, \cdot)$ is a similarity function between embeddings, and τ denotes the temperature parameter. To further preserve signal fidelity, we reconstruct the input from the sleep encoder representations. Let $\hat{\mathbf{x}}$ denote the reconstructed sleep epoch, we minimize a mean-squared error:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2.$$

For caption generation, the multimodal text decoder conditions on the sleep embeddings and a modality token $[m]$. Using an autoregressive factorization, the captioning loss is:

$$\mathcal{L}_{\text{cap}} = -\sum_{t=1}^T \log \mathbb{P}_{\theta}(\mathbf{y}_t \mid \mathbf{y}_{<t}, \mathbf{x}, [m]).$$

The SleepLM Family. The final objective is a weighted combination $\mathcal{L}_{\text{ReCoCa}} = \lambda_{\text{con}} \cdot \mathcal{L}_{\text{con}} + \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} + \lambda_{\text{cap}} \cdot \mathcal{L}_{\text{cap}}$, which yields a family of sleep-language models by varying $(\lambda_{\text{con}}, \lambda_{\text{rec}}, \lambda_{\text{cap}})$. For instance, setting $\lambda_{\text{rec}} = 0$ and $\lambda_{\text{cap}} = 0$ yields a CLIP-style dual encoder (Fig. 3). The full ReCoCa configuration uses all three losses and is our default model; we ablate these variants in Sec. 5.3.

5. Experiments and Results

Datasets. We use the pretraining splits of SHHS, MrOS, and CCSHS as the primary training corpora, and keep CFS and WSC fully held out. For zero-shot evaluation, we use the validation splits of SHHS and MrOS together with CFS to assess both internal and external generalization. Due to its higher prevalence of rare apnea subtypes, WSC is reserved for few-shot learning and unseen-concept generalization experiments. Further details are provided in Appendix B.

Baselines. For zero-shot and generative tasks, we compare against two classes of models. ❶ *Fine-tuned multimodal LLMs*: we adapt leading open-source vision-language models by replacing the vision encoder with our pretrained sleep encoder, then fine-tune the full system on our sleep-text corpus. We evaluate Qwen3-VL-8B-Instruct [6] and LLaVA-Next [24] as representative strong open-source backbones. ❷ *Proprietary LLMs*: we evaluate strong commercial models, including Gemini 2.5 Pro [14] and DeepSeek-R1 [21], by providing PSG as tabular time-series input with task-specific prompts. For few-shot tasks, we compare against SOTA self-supervised learning methods, MAE [22] and SimCLR

[12]. These models are pretrained on the same PSG corpus as all other methods for a controlled comparison. Training details and prompts for zero-shot tasks are provided in Appendix A and E.2.

Metrics. For zero-shot classification, we report area under the ROC curve (AUROC), and balanced accuracy (BAcc). For zero-shot event localization, we report intersection-over-union (IoU). We report both symmetric mean absolute percentage error (sMAPE) and mean absolute error (MAE) for regression. For few-shot learning, we report AUROC across {5, 10, 20, 50} samples per class. Cross-modal retrieval is evaluated using Recall@1 and Recall@5 (R@K). A detailed settings by task is provided in Appendix C.

Unless otherwise stated, we report the main results using the base ReCoCa configuration, denoted as SleepLM-B. We study the effects of scaling model components in Sec. 5.2.

5.1. Main Results

Zero-Shot Recognition. We evaluate SleepLM in a zero-shot setting across four task categories: ❶ five-class sleep staging, ❷ sleep event localization, ❸ implicit physiological inference (HR/SpO₂), and ❹ explicit signal grounding (channel statistics). As shown in Table 1, SleepLM consistently outperforms both proprietary LLMs and fine-tuned VLMs across all zero-shot tasks. The baselines exhibit distinct failure modes. Proprietary LLMs, even when given tabular PSG descriptors, produce near-random predictions for sleep stages and event identification, yet perform relatively well on HR and SpO₂ regression. This suggests that strong LLMs can extract and manipulate explicit numeric summaries, but struggle to map low-level signal descriptors into higher-level sleep and event states. In contrast, fine-tuned VLMs are consistently suboptimal, suggesting that simply swapping in a sleep encoder does not yield effective multimodal fusion for dense physiological time series.

Zero-Shot Cross Modal Retrieval. We evaluate cross-modal alignment via retrieval in both directions (signal-to-text and text-to-signal), reporting R@1 and R@5. As shown in Table 2, SleepLM substantially outperforms LLM baselines, which are only slightly above random in this dense retrieval setting.

On a 100-sample validation subset, SleepLM achieves near-perfect accuracy. On the full 2,000-sample validation set, where LLM baselines are not feasible due to context-length constraints, SleepLM still maintains strong retrieval performance. These results confirm that our pretraining objective learns a precise mapping between physiological states and their language descriptions. Later in Sec. 5.2, we further demonstrate that SleepLM learns a continuous embedding space that supports practical retrieval beyond exact matches, enabling the search of semantically related data clusters.

Generalization to Unseen Concepts. A key property of foundation models is the ability to generalize to concepts that are not explicitly observed during training. We evaluate SleepLM on two held-out clinical events: “Mixed Apnea” and “Obstructive Apnea”. These labels are absent from the training vocabulary, requiring the model to extrapolate from related physiological patterns (e.g., “Central Apnea”). As Table 3 confirms,

Table 2 | **Zero-shot cross-modal retrieval.** We evaluate text-to-signal (**top**) and signal-to-text (**bottom**) retrieval for SleepLM and LLM baselines. “-” indicates that the task is infeasible for an LLM baseline due to context limits. Full results are in Appendix D.6.

Model	n=100		n=2000	
	R@1 [†]	R@5 [†]	R@1 [†]	R@5 [†]
<i>Text → Signal Retrieval</i>				
DeepSeek-R1 [21]	-	-	-	-
Gemini 2.5 Pro [14]	-	-	-	-
SleepLM	94.3	99.0	82.5	91.9
<i>Signal → Text Retrieval</i>				
DeepSeek-R1 [21]	1.7	5.7	-	-
Gemini 2.5 Pro [14]	4.0	10.7	-	-
SleepLM	94.3	99.7	80.9	91.5

while LLM baselines perform at chance, SleepLM achieves approximately 80% F1 and accuracy on these unseen events. To probe the source of this behavior, we visualize the embedding space with UMAP (Fig. 4). We observe that SleepLM places mixed and “Obstructive Apnea” near “Central Apnea” (a seen concept), while separating them from physiologically distinct events such as “Oxygen Desaturation”. This structure suggests that SleepLM learns a latent concept that spans both seen and unseen variants in the aligned text and physiology spaces.

Few-Shot Learning. We evaluate the transferability of SleepLM by isolating the sleep encoder and comparing it with SOTA SSL baselines and a supervised ViT model [17]. We freeze encoder backbones and train a linear probe using a small labeled samples per class (i.e., {1, 5, 10, 20, 50}) on the held-out WSC cohort. As shown in Fig. 6, SleepLM consistently outperforms both SSL and supervised baselines on sleep staging. With only 50 samples per class, SleepLM reaches approximately 0.90 AUC, indicating strong data efficiency. Beyond general pretraining recipes, we further validate our approach by comparing SleepLM against specialized domain architectures. We conduct few-shot linear probing runs and full finetuning runs against diverse types of domain specific models: (1) *Supervised sleep stage models* [20, 30], (2) *Sleep FMs* [37], and (3) *General time series FMs* [4]. Specifically, to ensure a fair comparison, the domain-specific models are trained on the same data splits as SleepLM. We then isolated SleepLM’s sleep encoder and evaluate all models via linear probing and full fine-tuning on the completely unseen CFS dataset. Notably, for Chronos-2, we fine-tune its public checkpoint to properly assess its capacity as a general time-series foundation model. Ultimately, SleepLM demonstrates highly competitive performance across all tasks and settings. Together, these findings indicate that the semantic structure learned through caption-based supervision produces more discriminative and transferable features than standard reconstruction or invariance objectives, consistently matching or exceeding prior domain-specific state-of-the-art methods. Detailed results are provided in Appendix D.4.

5.2. Analyses

Sleep Caption Generation. We assess the generation quality of SleepLM by comparing its captions with ground-truth text and with captions from Gemini 2.5 Pro. As confirmed in Fig. 5, SleepLM

Table 3 | **Zero-shot generalization to unseen concepts.** We report performance on two held-out respiratory event classification tasks, where SleepLM remains robust across both settings.

Model	Mixed Apnea		Obstructive Apnea	
	F1 $\uparrow$	BAcc $\uparrow$	F1 $\uparrow$	BAcc $\uparrow$
Gemini 2.5 Pro	33.4	48.3	15.6	51.4
SleepLM	78.2	81.8	79.7	77.1

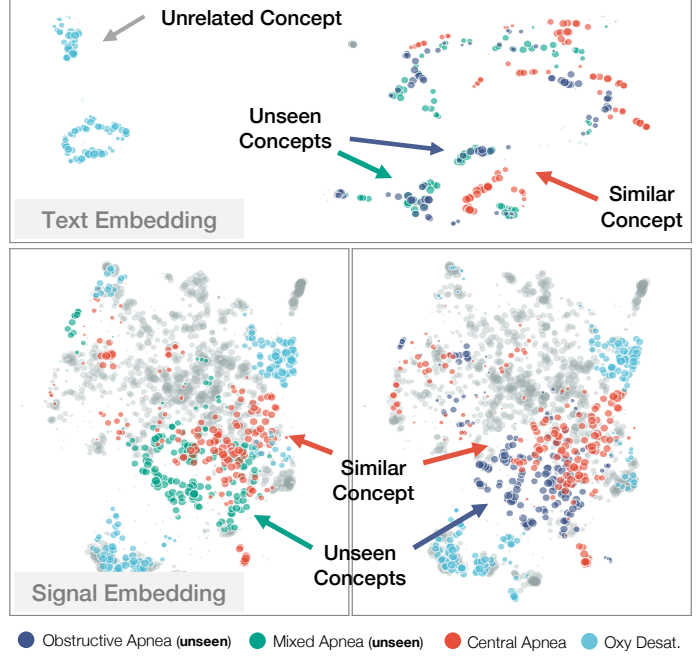


Figure 4 | **Zero-shot generalization analysis of SleepLM.** We visualize text and signal embeddings with UMAP as a case study of zero-shot concept transfer. SleepLM is capable of clustering previously unseen concepts to semantically related seen concepts.

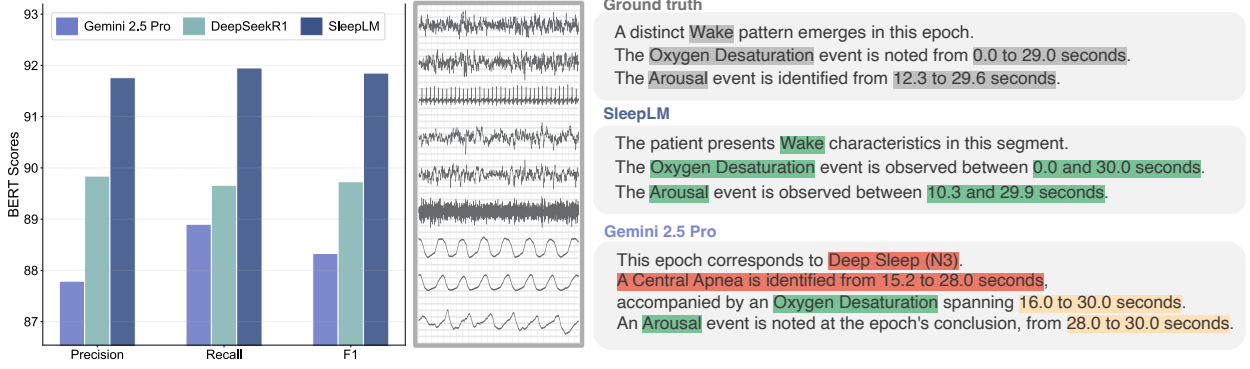


Figure 5 | **Sleep caption generation.** SleepLM captures both high-level sleep stages and fine-grained localized events, while LLM baselines often fail to recognize or localize. Additional examples, including fine-tuned VLM outputs, are provided in Appendix D.9. The generation scores are provided in Table 20.

produces concise, clinically accurate descriptions that capture both sleep stage and the timing of localized events. In contrast, Gemini 2.5 Pro frequently introduces incorrect associations and fails to reflect the underlying signal morphology.

Localization Sensitivity. One key capability is whether the model captures *when* an event occurs, rather than only its presence. To test this, we run a controlled perturbation study (Fig. 8). We select an epoch containing a ground-truth event (e.g., hypopnea) and construct synthetic captions that are identical except for their timestamp intervals. We then compute the cosine similarity between the fixed signal embedding and the text embeddings of these temporally shifted captions. The similarity shows a strong linear relationship with timestamp IoU: it peaks near the correct alignment and decreases as overlap diminishes. This indicates that SleepLM learns temporally grounded representations that are sensitive to fine-grained localization in a zero-shot setting. Additional examples are provided in Appendix E.5.

Embedding Space Continuity. To further visualize the structure of the learned latent space, we analyze retrieval results for a single PSG epoch. Fig. 7 shows the top three retrieved captions from the pool, ranked by decreasing similarity. For clarity, we display only the global sleep stage and event descriptions. We observe a smooth semantic gradient: higher similarity scores consistently return captions that match the query physiology, while lower scores correspond to physiologically distinct states. This indicates that SleepLM learns a continuous and semantically meaningful manifold in which embedding distance reflects physiological similarity, enabling data exploration based on semantic proximity rather than exact concept matches. Additional examples are provided in Appendix E.4.

Scaling Behavior (Appendix D.1 & D.2): We study how SleepLM scales with both model size and data diversity. For model scaling, we train three variants, SleepLM-T (38M), SleepLM-S (180M), and SleepLM-B (410M). As shown in Table 9, performance improves consistently as parameter count increases, with particularly clear gains in channel-statistics regression and cross-modal retrieval, and no evidence of saturation at the base scale. Model specifications are summarized in Table 10. For data scaling, we compare single-source pretraining (SHHS only) with multi-source pretraining (SHHS + MrOS + CCSHS). Multi-source training improves all metrics and even outperforms the single-source baseline on

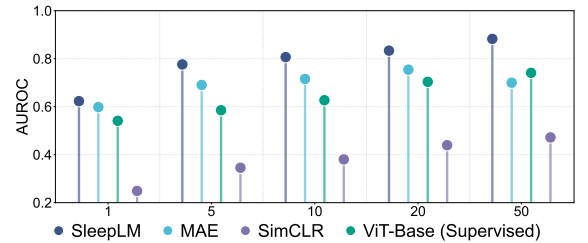


Figure 6 | **Few-shot to downstream tasks.** We compare SleepLM with SSL and supervised baselines under varying numbers of labeled samples per class. SleepLM shows higher data efficiency and stronger performance across all shot regimes.

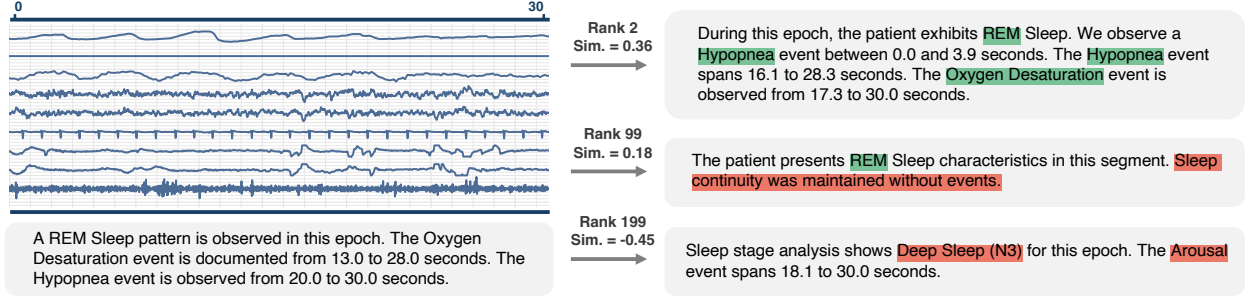


Figure 7 | **Semantic retrieval continuity of SleepLM.** Retrieval results show a smooth semantic gradient: high-similarity captions (green) match the query’s physiological state, while lower-scoring results (red) correspond to distinct conditions. This indicates that embedding distance of SleepLM reflects physiological similarity.

the internal SHHS evaluation set, indicating that our captioning pipeline provides stable supervision across cohorts and that added source diversity strengthens, rather than harms, generalization.

Full-Night Reporting (Appendix D.8). To connect with clinical practice, we aggregate epoch-level predictions across full-night recordings to derive standard diagnostic metrics, including the apnea-hypopnea index (AHI) and wake after sleep onset (WASO). We randomly select five SHHS subjects and run SleepLM in a sliding-window manner over each night, then summarize the fine-grained outputs into full-night statistics (details in Appendix D.8). SleepLM shows strong concordance with manual scoring and remains stable across thousands of epochs, whereas fine-tuned VLM baselines exhibit drift over long sequences. Fig. 19 shows an example report constructed from these derived statistics, illustrating that SleepLM can translate longitudinal PSG into actionable, physician-oriented summaries.

5.3. Ablation Studies

ReCoCa vs. other SleepLM variants. We compare ReCoCa with other pretraining formulations that can be instantiated within the SleepLM framework. As described in Sec. 4, we obtain CLIP-, Cap-, and CoCa-style variants by toggling the contrastive ($\mathcal{L}_{\text{con}}$) and captioning and captioning ($\mathcal{L}_{\text{cap}}$) objectives. To ablate our channel-specific sleep encoder, we also replace it with a standard encoder backbone consisting of convolutional feature extraction followed by a temporal transformer [11]. All models are matched to roughly the same parameter size.

Comparing these variants in Table 4 reports representative metrics across tasks and shows that ReCoCa consistently performs best across most classification, regression, and event-localization evaluations, highlighting the benefit of our sleep-specific architectural choices.

Sleep reconstruction as regularization (Appendix D.3). We hypothesize that the mismatch between dense PSG and sparse text can cause the encoder to discard fine-grained morphology under text-only supervision. To test this, we ablate the reconstruction objective by setting $\mathcal{L}_{\text{rec}} = 0$. As shown in Appendix D.3, removing reconstruction consistently degrades performance on discriminative tasks, supporting the role of reconstruction as a regularizer that preserves physiologically meaningful details.

Table 4 | **SleepLM variant ablations on representative tasks.** The full ReCoCa configuration consistently performs best, highlighting the benefit of our sleep-specific design choices. Complete results are provided in Appendix D.

Arch	Stage Acc [↑]	Event IoU [↑]	T2S R@1 [↑]	sMAPE [↓]
Cap [38]	72.0	<u>29.2</u>	-	<u>7.20</u>
CLIP [32]	74.6	-	74.3	-
CoCa [43]	<u>76.0</u>	27.4	<u>77.5</u>	7.34
ReCoCa	76.9	30.4	82.5	3.15

Multilevel caption supervision (Appendix D.3). An important design of our captioning pipeline is the integration of both low-level grounding (channel/local) and high-level summaries (global). To assess the value of low-level supervision, we compare the full-caption model (channel + local + global) with an ablation trained only on global captions. Results in Appendix D.3 demonstrate that the model trained with multilevel supervision consistently outperforms other baselines, including on high-level tasks such as sleep staging and event classification. This suggests that learning grounded waveform descriptors strengthens the representations used to infer broader physiological states.

6. Discussion

Limitations. While promising, SleepLM is a research prototype and is not clinically validated for diagnosis, treatment, or medical decision-making. In addition, our study focuses on five PSG cohorts curated from NSRR; further work may extend the data coverage to assess robustness to broader clinical variability, devices, and patient populations.

Conclusion. We present SleepLM, the first family of sleep-language foundation models that unlock human sleep understanding through natural language. By curating the first large-scale sleep-text dataset and designing a unified pretraining objective ReCoCa, we support joint learning of language and physiological time-series at scale. We verify that SleepLM achieves strong performance across diverse tasks while enabling new capabilities such as language-guided localization and generalization to unseen concepts.

Acknowledgments

We gratefully acknowledge the support by Amazon Science Hub and UCLA DataX. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

References

- [1] Prince Nii Ossah Addo, Paddington T Mundagowa, Longgang Zhao, Mufaro Kanyangarara, Monique J Brown, and Jihong Liu. Associations between sleep duration, sleep disturbance and cardiovascular disease biomarkers among adults in the united states. *BMC Public Health*, 24(1):947, 2024.

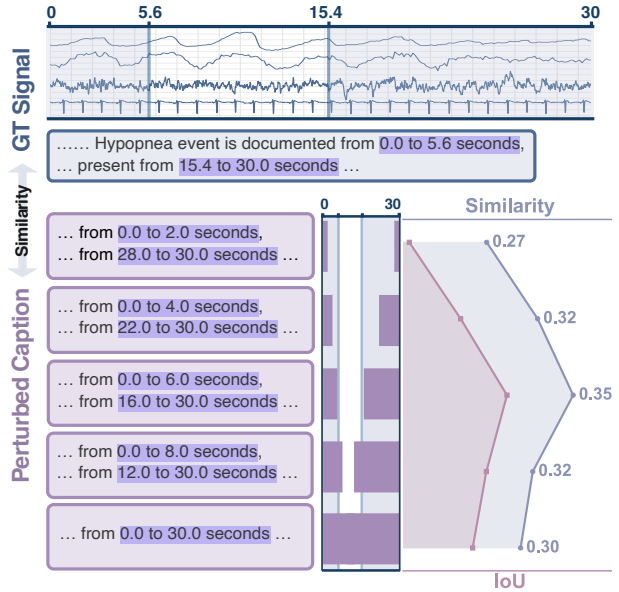


Figure 8 | **Localization sensitivity of SleepLM.** Given a fixed signal embedding and its caption, we progressively shift the event timestamp in the caption and compare the resulting text embeddings to the signal embedding. Embedding similarity increases with the IoU between the ground-truth and perturbed timestamps, peaking near the correct alignment.

- [2] American Academy of Sleep Medicine. The AASM manual for the scoring of sleep and associated events. <https://aasm.org/clinical-resources/scoring-manual/>, 2025. Accessed: 2026-01-25.
- [3] Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.
- [4] Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. Expert Certification.
- [5] Mahsa Bahrami and Mohamad Forouzanfar. Sleep apnea detection from single-lead ecg: A comprehensive analysis of machine learning and deep learning algorithms. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [7] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, Carole Marcus, Bradley V Vaughn, et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176(2012):7, 2012.
- [8] Erhan Bilal, Matheus Lima Diniz Araujo, Kristen L Beck, Catherine M Heinzinger, Samer Ghosn, Carl Y Saab, Nancy Foldvary Schaefer, Jeffrey L Rogers, and Reena Mehra. A foundation model for sleep-based risk stratification and clinical outcomes. *Research Square*, pages rs–3, 2025.
- [9] 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.
- [10] Andreas Brink-Kjaer, Eileen B Leary, Haoqi Sun, M Brandon Westover, Katie L Stone, Paul E Peppard, Nancy E Lane, Peggy M Cawthon, Susan Redline, Poul Jennum, et al. Age estimation from sleep studies using deep learning predicts life expectancy. *NPJ digital medicine*, 5(1):103, 2022.
- [11] Jonathan F Carter and Lionel Tarassenko. wav2sleep: A unified multi-modal approach to sleep stage classification from physiological signals. *arXiv preprint arXiv:2411.04644*, 2024.
- [12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [13] Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, et al. This time is different: An observability perspective on time series foundation models. *arXiv preprint arXiv:2505.14766*, 2025.

- [14] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- [15] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. Towards a personal health large language model. *arXiv preprint arXiv:2406.06474*, 2024.
- [16] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [18] Ada Eban-Rothschild, Lior Appelbaum, and Luis de Lecea. Neuronal mechanisms for sleep/wake regulation and modulatory drive. *Neuropsychopharmacology*, 43(5):937–952, 2018.
- [19] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Self-supervised learning for label-efficient sleep stage classification: A comprehensive evaluation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1333–1342, 2023.
- [20] Antoine Guillot and Valentin Thorey. Robustsleepnet: Transfer learning for automated sleep staging at scale. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1441–1451, 2021.
- [21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [23] Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A Xu, Winnie Chow, Martin Maritsch, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, et al. Opentslm: Time-series language models for reasoning over multivariate medical text-and time-series data. *arXiv preprint arXiv:2510.02410*, 2025.
- [24] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.
- [25] Sirui Li, Shuhan Xiao, Mihir Joshi, Ahmed Metwally, Daniel McDuff, Wei Wang, and Yuzhe Yang. Hearts: Benchmarking llm reasoning on health time series. *arXiv preprint arXiv:2603.06638*, 2026.
- [26] Maria P Mogavero, Giuseppe Lanza, Oliviero Bruni, Luigi Ferini-Strambi, Alessandro Silvani, Ugo Faraguna, and Raffaele Ferri. Beyond the sleep lab: A narrative review of wearable sleep monitoring. *Bioengineering*, 12(11):1191, 2025.

- [27] Guangkun Nie, Xuesong Chen, Yichen Wang, Jingxu Chen, Yunhan Shi, Jianwen Zhong, Jie Shi, Chun-feng Liu, Bei Huang, Yaping Liu, et al. A zero-burden sleep foundation model built on cardiorespiratory signals from 800,000+ hours of multi-ethnic sleep recordings. *medRxiv*, pages 2025–09, 2025.
- [28] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.
- [29] Saurav Raj Pandey, Aaqib Saeed, and Harlin Lee. Pedsleepmae: Generative model for multimodal pediatric sleep signals. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–8. IEEE, 2024.
- [30] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-sleep: resilient high-frequency sleep staging. *NPJ digital medicine*, 4(1):72, 2021.
- [31] 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.
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [33] Susan Redline, Peter V Tishler, Tor D Tosteson, John Williamson, Kenneth Kump, Ilene Browner, Veronica Ferrette, and Patrick Krejci. The familial aggregation of obstructive sleep apnea. *American journal of respiratory and critical care medicine*, 151(3):682–687, 1995.
- [34] Carol L Rosen, Emma K Larkin, H Lester Kirchner, Judith L Emancipator, Sarah F Bivins, Susan A Surovec, Richard J Martin, and Susan Redline. Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity. *The Journal of pediatrics*, 142(4):383–389, 2003.
- [35] Zitao Shuai, Zongzhe Xu, David Yang, Wei Wang, and Yuzhe Yang. Osf: On pre-training and scaling of sleep foundation models. *arXiv preprint arXiv:2603.00190*, 2026.
- [36] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [37] Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Zou. SleepFM: Multi-modal representation learning for sleep across brain activity, ECG and respiratory signals. In *Forty-first International Conference on Machine Learning*, 2024.
- [38] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36:46830–46855, 2023.
- [39] Raphael Vallat, Vyoma D Shah, and Matthew P Walker. Coordinated human sleeping brainwaves map peripheral body glucose homeostasis. *Cell Reports Medicine*, 4(7), 2023.
- [40] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *International Conference on Machine Learning*, 2024.

- [41] Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R Junna, et al. Artificial intelligence-enabled detection and assessment of parkinson’s disease using nocturnal breathing signals. *Nature Medicine*, 28(10):2207–2215, 2022.
- [42] Terry Young, Mari Palta, Jerome Dempsey, Paul E Peppard, F Javier Nieto, and K Mae Hla. Burden of sleep apnea: rationale, design, and major findings of the wisconsin sleep cohort study. *WMJ: official publication of the State Medical Society of Wisconsin*, 108(5):246, 2009.
- [43] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [44] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [45] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.
- [46] 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.
- [47] Hongjun Zhang, Jing Wang, Jiahong Xiong, Yuxuan Ding, Zhenliang Gan, and Youfang Lin. Expert knowledge inspired contrastive learning for sleep staging. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2022.
- [48] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed A Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, et al. Sensorlm: Learning the language of wearable sensors. *arXiv preprint arXiv:2506.09108*, 2025.

A. Training Details

Training ReCoCa: We train ReCoCa using the AdamW optimizer with a learning rate of $1e-4$ and a cosine annealing schedule. The training runs for 15 epochs with a 5,000-step linear warmup. We utilize a global batch size of 384 distributed across 4 NVIDIA H100 GPUs. The loss components are weighted as follows: $\lambda_{con} = 1.0$, $\lambda_{cap} = 2.0$, and $\lambda_{rec} = 0.1$. Gradients are clipped at a norm of 1.0. Training typically converges within approximately 48 hours.

Baseline Finetuning Strategy: For the multimodal LLM baselines (Qwen3-VL-8B-Instruct [6] and LLaVA-Next [24]), we adopt the two-stage modality adaptation protocol proposed in LLaVA-Next.

- **Stage 1 (Alignment):** We freeze both the pretrained Sleep Encoder (initialized from ReCoCa) and the LLM backbone, training only the projector and token pooler layers. This stage aligns the sleep feature space with the LLM’s embedding space.
- **Stage 2 (Finetuning):** We unfreeze the projector, pooler, and Sleep Encoder, and apply Low-Rank Adaptation (LoRA) to the LLM backbone.

Due to computational constraints, we use LoRA with rank $r = 16$, $\alpha = 32$, and dropout $p = 0.05$. We use a batch size of 8 with gradient accumulation every 16 steps. The models are warmed up for 500 steps followed by one full epoch of finetuning. These baseline experiments require significantly higher compute, taking 96–144 hours on the same 4×H100 hardware setup.

B. Dataset Details

B.1. Data Split & Preprocessing

To evaluate generalization and zero-shot performance, we implement a strict splitting strategy. We partition SHHS and MrOS into pretraining and internal evaluation sets on a subject level. We utilize CCSHS exclusively for training. Crucially, we hold out CFS and WSC entirely to serve as external validation datasets. For evaluation, we sample a fixed subset of 2,000 epochs from the validation partition of each dataset. This sampling strategy is necessary to accommodate the prohibitive computational and financial costs associated with performing inference on the full validation set using proprietary LLMs. To preserve statistical validity, this subset is strictly stratified at the subject level, ensuring maximal biological variance despite the reduced sample size. Additionally, to ensure high representativeness, we strictly stratified these validation samples across 930 distinct patients (SHHS: 409, MrOS: 193, CFS: 328), capturing a truly diverse population. The complete split information is available in Table 5, and we provide the demographic distribution over different cohorts used in Table 6.

Table 5 | Distribution of number of epochs across different datasets and splits

Split	Internal Datasets			External Datasets		Total
	SHHS	MROS	CCSHS	CFS	WSC	
Train	6,985,633	3,569,317	552,693	0	0	11,107,643
Valid	2,000	2,000	0	2,000	2,000	8,000
Total	6,987,633	3,571,317	552,693	2,000	2,000	11,115,643

Missing channels are zero-padded to maintain consistency. Given the heterogeneity of the source data (varying sampling rates, device ranges), we unify all signals to a fixed sampling rate of 64Hz. We apply manual quality control over all night’s data to trim excessive wakefulness or non-wear

period at the start and end of recordings to reduce extreme sensor noise. Finally, we apply z-score normalization on a per-night basis. For respiratory channels specifically, we apply area-dependent z-score normalization to ensure consistent amplitude scaling over time.

Table 6 | **Age distribution across cohorts in train and validation splits.** Values are percentages. This demographic skew is expected and aligns with real world clinical baselines, as the prevalence of sleep disorders dramatically increases with advanced age.

Train					Validation				
Age	SHHS	MROS	CCSHS	All	Age	SHHS	MROS	CFS	All
0–18	0.0	0.0	78.1	3.3	0–18	0.0	0.0	20.7	7.3
18–35	0.0	0.0	21.9	0.9	18–35	0.0	0.0	21.0	7.4
35–50	12.9	0.0	0.0	8.4	35–50	11.0	0.0	26.5	14.2
50+	87.1	100.0	0.0	87.3	50+	89.0	100.0	31.7	71.1

B.2. Statistics Details

For our [Channel Captions](#), we defined a variable set of statistics for each channel in order to provide diverse, fine-grained, clinically relevant information. The specific statistics and their relevant channels are provided in [Table 5](#).

C. Tasks Setup

C.1. Zero-Shot Task Definitions

We assess the model’s zero-shot capabilities across the following four categories.

1. Sleep Staging Classification: We perform standard 5-class classification (Wake, N1, N2, N3, REM). To classify an epoch, we calculate the cosine similarity between the signal embedding z_{cls} and the average text embedding of a diverse set of template captions (e.g., “The patient is in N2 sleep”) for each stage. We report the AUC and balanced accuracy. The full list of template prompts is provided in [Appendix E.3](#).

2. Sleep Event Localization & Classification: We evaluate the ability to identify and localize specific clinical events: arousal, central apnea, hypopnea, and oxygen desaturation. Note that obstructive apnea and mixed apnea are explicitly held out from this evaluation to test generalization on unseen concept.

- **Classification:** We treat this as a binary classification task per event type and report balanced accuracy.
- **Localization:** We parse the start and end timestamps from the generated caption and compare them against ground truth annotations using IoU.

3. Implicit Physiological Inference: As described in the method section, Heart Rate and SpO2 signals are excluded from the encoder input. We evaluate the model’s ability to infer these vitals purely from cross-channel correlations in the remaining sensors.

- **Statistics:** We parse the predicted statistics (mean, min, max) from the generated caption and calculate the MAE against the ground truth.

Table 7 | **Channel Statistics Configuration:** a detailed list of what statistics we compute for each channel and included into the channel caption

Channel	Statistic	Description
ECG	hr_mean rmssd_30s	estimated beat rate ultra-short HRV (RMSSD)
HR	min max mean	minimum value maximum value mean
SPO2	min mean	minimum value mean
ABD	rr_bpm rr_iqr	respiratory rate breath interval variability
THX	rr_bpm rr_iqr	respiratory rate breath interval variability
AF	rr_bpm_af flow_flatness	airflow rate inspiratory flow flatness
EOG_E1_A2	sem_power rem_power	slow eye movement relative power REM saccadic relative power
EOG_E2_A1	sem_power rem_power	slow eye movement relative power REM saccadic relative power
EMG_Chin	mdf_hz_10_30 tail_ratio env_centroid_hz	median frequency burst intensity ratio burst modulation rate
EEG_C3_A2	delta_power theta_power alpha_power beta_power	delta relative power theta relative power alpha relative power beta relative power
EEG_C4_A1	delta_power theta_power alpha_power beta_power	delta relative power theta relative power alpha relative power beta relative power
POS	-	-

- **Trends:** We extract predicted trend windows. Due to the subjective definition of “trends” in physiological signals, we report Recall rather than IoU to avoid penalizing valid detections that slightly differ from heuristic ground truth boundaries.

4. Explicit Signal Grounding (Channel Statistics): To assess the model’s physical understanding of the visible signals provided to the encoder, we evaluate its ability to estimate clinically relevant signal statistics (e.g., EEG variance, EMG power). We parse the numeric values from the generated channel-specific captions and report the Symmetric Mean Absolute Percentage Error (sMAPE).

C.2. Unseen Concept Classification

To rigorously test generalization, we curate a balanced binary classification dataset from the external WSC dataset, which was held out entirely during pretraining. The dataset consists of 500 positive and 500 negative samples for each of the two unseen events (mixed apnea and obstructive apnea).

Zero-Shot Prototype Construction: Since the specific labels for these apneas were not present in the pretraining vocabulary, we utilize a zero-shot prototype approach by computing the cosine similarity between the signal embedding and two text anchors:

- **Positive Anchor:** The average embedding of a diverse set of “event presence” templates (e.g., “An obstructive apnea event”).
- **Negative Anchor:** Constructing a negative anchor is non-trivial, as our pretraining corpus lacks explicit negation concepts (e.g., “No obstructive apnea”). We therefore utilize a global “No event at all” anchor as a substitute.

Note on Anchor Noise: It is important to note that the negative anchor described above is inherently noisy. A “negative” sample in this binary task is defined as “not obstructive Aapnea,” but the epoch may still contain other events (e.g., a hypopnea or arousal). By using “No event at all” as the negative anchor, we inadvertently penalize the model if it detects these other valid events. The high performance reported in the main text is achieved despite this structural disadvantage, underscoring the model’s discriminative precision.

C.3. Few-Shot Evaluation

To assess representation quality in data-scarce regimes, we conduct a linear probing evaluation on the external WSC dataset using the following setup.

Baselines: We compare ReCoCa against four distinct baselines, all controlled to have approximately the same parameter count:

- **SSL Baselines:** MAE (Masked Autoencoder) and SimCLR (Contrastive Learning), trained on the same internal corpus as ReCoCa to ensure fair comparison of pretraining objectives.
- **Supervised Baselines:** WideResNet [44] and ViT [17], trained from scratch.

Protocol: We simulate data scarcity by providing strictly K labeled samples per class, where $K \in \{1, 5, 10, 20, 50\}$. **Frozen Encoder:** The backbone weights of all models are frozen to evaluate the quality of the static pretrained representations. **Linear Probe:** A simple linear classifier is trained on top of the fixed embeddings using the limited K -shot training set. **Tasks:** Evaluation is performed on two downstream tasks: 5-class sleep stage classification and binary oxygen desaturation detection.

D. Additional Results

This section presents supplementary experimental results that substantiate the claims in the main manuscript and include the full, unaggregated results underlying the summarized evaluations reported in the main text.

D.1. Scaling on Dataset

Table 8 | **Effect of pretraining data scale on downstream performance** (internal: SHHS; external: CFS). Best is **bold**; second best is underlined.

Internal					External				
<i>Pretraining data</i>	Stage Acc [↑]	Event IoU [↑]	TtS R@1 [↑]	SMAPE [↓]	<i>Pretraining data</i>	Stage Acc [↑]	Event IoU [↑]	TtS R@1 [↑]	SMAPE [↓]
SHHS only	77.0	31.2	96.0	3.80	SHHS only	70.5	27.2	73.5	5.38
Multi-source	79.6	31.9	96.1	2.73	Multi-source	74.2	28.2	78.7	3.99

In Table 8, we present the results of training on single vs multisource datasets. Our results demonstrate that multisource data pretraining not only helps the performance on the external datasets, which is expected, but also brings a significant performance on the internal dataset. This shows that our captioning pipeline is robust and extends beyond datasets’ boundary.

D.2. Scaling on Model size

Table 9 | **Effect of ReCoCa parameter size on downstream performance** (internal: (SHHS+MROS), external: CFS). Best is **bold**; second best is underlined.

Internal					External				
<i>Pretraining data</i>	Stage Acc [↑]	Event IoU [↑]	TtS R@1 [↑]	SMAPE [↓]	<i>Pretraining data</i>	Stage Acc [↑]	Event IoU [↑]	TtS R@1 [↑]	SMAPE [↓]
ReCoCa-Tiny	75.2	25.7	63.1	8.23	SleepLM-T	71.9	27.2	52.5	8.98
ReCoCa-Small	77.1	29.4	76.9	5.24	SleepLM-S	72.8	27.2	65.4	6.38
ReCoCa-Base	78.3	31.3	84.4	3.13	SleepLM-B	74.2	28.2	78.7	3.99

Table 10 | **Architecture specs for ReCoCa variants.**

Model	Sleep encoder			Text encoder			Sleep decoder			#Params
	Head	Layer	Dim	Head	Layer	Dim	Head	Layer	Dim	
SleepLM-T	8	2	256	8	3	256	8	3	256	38M
SleepLM-S	12	2	768	12	4	768	12	4	768	180M
SleepLM-B	12	6	768	12	12	768	12	12	768	410M

To assess the scalability of the framework, we evaluate three variants of SleepLM with increasing capacity: SleepLM-T (38M), SleepLM-S (180M), and SleepLM-B (410M). Detailed architectural specifications for each configuration are provided in Table 10.

As presented in Table 9, we observe consistent monotonic improvements in performance as parameter count increases. This trend is particularly pronounced in tasks requiring fine-grained semantic understanding, such as cross-modal retrieval and channel statistics regression. Notably, performance does not saturate at the Base scale (410M), suggesting that the SleepLM architecture is capable of effectively leveraging larger parameter budgets for further gains.

D.3. Ablation Study

Table 11 | **Ablation results on sleep reconstruction:** We show results of SleepLM trained with and without sleep reconstruction. **Ablation results on multilevel captions:** We compare the results between SleepLM trained with and without low level channel caption. All results are shown on the high-level sleep stage classification and sleep event identification tasks and averaged across all datasets.

Setting	Sleep Stage		Sleep Event	
	AUC $\uparrow$	Acc $\uparrow$	IoU $\uparrow$	Acc $\uparrow$
<i>Sleep reconstruction</i>				
w/o recon. loss	83.9	74.1	29.6	73.2
with recon. loss	85.4	76.9	30.4	74.3
<i>Channel caption</i>				
w/o channel caption	84.6	75.6	30.2	73.7
with channel caption	85.4	76.9	30.4	74.3

We conduct two ablation studies to validate our design choices, with full numerical comparisons presented in the Table 11.

Sleep Reconstruction as Regularization: We posit that reliance on sparse text supervision alone creates an “information density gap” when paired with dense physiological signals, potentially leading the encoder to discard nuanced waveform features which leads to feature collapse. By enforcing a reconstruction objective ($\mathcal{L}_{rec}$), we compel the model to retain a complete representation of the input signal, ensuring that fine-grained morphological details are preserved alongside semantic abstractions. This acts as a critical regularizer, grounding the latent space in the physical reality of the signal rather than just the linguistic approximation. As shown in the results, removing this objective consistently degrades performance across all discriminative tasks.

Multilevel Supervision of Captions: Our data pipeline integrates low-level channel captions alongside high-level global summaries to foster a “bottom-up” understanding of sleep physiology. The intuition is that robust high-level inference (e.g., sleep staging) relies on the accurate detection of fundamental waveform statistics and local morphologies. By explicitly supervising the model on these granular details, we prevent it from overfitting to abstract labels and instead encourage a hierarchical learning process. The results confirm that this low-level grounding provides an improvement in performance even on global classification tasks compared to a high-level-only baseline.

D.4. Fewshot Results

In this section, we provide a comprehensive breakdown of the few-shot transfer learning experiments introduced in the main text. To assess the quality of the learned representations in data-scarce regimes, we isolate the sleep encoder of SleepLM and compare it against state-of-the-art SSL baselines (MAE [22], SimCLR [12]) and supervised architectures (ViT [17]) on the held-out WSC cohort. We freeze the encoder bodies of all models and train a linear probe using strictly K labeled samples per class, where $K \in \{1, 5, 10, 20, 50\}$. This setup explicitly tests the discriminative power of the static, pretrained features without the benefit of fine-tuning. While the main text highlights performance on sleep staging, we present here the extended evaluation covering additional oxygen desaturation detection task. As detailed in the Table 18, SleepLM consistently outperforms specialized SSL and supervised baselines across these diverse physiological targets, confirming that the semantic structure imposed by our captioning objective yields features that are significantly more robust and transferable

Table 12 | Few-shot classification results on test set (AUROC, %). Best is **bold**; second best is underlined.

Model	# shots				
	1	5	10	20	50
<i>Oxygen Desaturation</i>					
Wide-ResNet-50	53.9	51.6	55.1	54.3	61.5
ViT-Base	55.5	57.6	51.4	58.3	56.6
MAE	54.0	60.8	<u>58.9</u>	57.4	59.0
SimCLR	52.2	57.7	<u>56.1</u>	<u>61.7</u>	<u>63.4</u>
ReCoCa	56.4	<u>59.3</u>	59.6	65.0	65.2
<i>Sleep Stage (5-class, macro-OvR)</i>					
Wide-ResNet-50	46.4	52.1	64.8	59.1	<u>79.1</u>
ViT-Base	54.1	58.5	62.7	70.4	<u>74.1</u>
MAE	59.8	69.0	<u>71.6</u>	<u>75.4</u>	70.0
SimCLR	<u>24.9</u>	<u>34.6</u>	<u>38.0</u>	<u>44.0</u>	47.2
ReCoCa	62.3	77.6	80.7	83.3	88.3

than those learned via standard reconstruction or invariance-based objectives.

Table 13 | Full finetuning AUC for SleepLM against domain specific models.

Model	Sleep Stage	Arousal	Hypopnea	Oxy. Desat.
Chronos-2	87.6	<u>87.9</u>	85.0	<u>83.7</u>
SleepFM	85.4	<u>86.7</u>	81.7	<u>80.3</u>
RobustSleepNet	85.8	–	–	–
U-Sleep	83.8	–	–	–
SleepLM	<u>87.2</u>	89.2	<u>83.0</u>	84.1

Table 14 | Few shot linear probing AUC under 10-shot and 50-shot settings for SleepLM against domain specific models.

Model	Sleep Stage		Arousal		Hypopnea		Oxygen Desaturation	
	10	50	10	50	10	50	10	50
Chronos-2	59.6	60.4	64.6	68.3	50.1	65.1	60.0	66.8
SleepFM	64.9	74.7	70.2	76.8	60.4	74.0	61.8	72.8
RobustSleepNet	<u>69.1</u>	<u>76.4</u>	–	–	–	–	–	–
U-Sleep	61.8	68.9	–	–	–	–	–	–
SleepLM	71.1	79.9	<u>66.5</u>	79.5	<u>59.3</u>	75.1	62.3	76.8

In addition to general pretraining recipe, we further conduct few-shot linear probing runs and full finetuning against diverse types of domain specific models: (1) Supervised sleep stage models: RobustSleepNet [20], U-Sleep [30], (2) Sleep FMs: SleepFM [37], and (3) General time series FMs: Chronos2 [3]. When necessary, models are retrained using the exact same settings in Appendix A and B for direct comparisons on the external CFS dataset.

To compare against domain-specific models on their native tasks, while these specialized models are designed for a ‘pretrain-then-finetune’ paradigm, we evaluated SleepLM under the exact conditions for the fairest comparison. Specifically, we pretrained these models using the same splits as SleepLM. We then isolated SleepLM’s sleep encoder and perform few-shot linear probing (Table 14) and fully finetuning (Table 13) over all models on the training split of the unseen CFS dataset. Final AUCs are reported on the full test split. For Chronos-2, we finetuned its public pretrained checkpoint to appropriately assess it as a general time-series FM.

Overall, SleepLM remains competitive across tasks and settings. Also, generally FMs perform better than smaller specialized models. We attribute this in part to the substantially larger parameter scale of FMs relative to these specialized baselines. This likely stems from two factors: (1) the established scaling behavior with model size that we also observe for SleepLM (Table 9, Appendix D.2), and (2) the ability of larger models to learn representations that transfer better across tasks and datasets. This is further supported by the strong performance of Chronos-2 in our full fine-tuning experiments.

D.5. Classification Raw Results

Table 15 | **Zero-shot sleep stage and event detection results across datasets.** Best is **bold**; second best is underlined.

Model	Sleep Stage			Central Apnea			Hypopnea			Oxygen Desaturation			Arousal		
	F1 [↑]	AUC [↑]	BAcc [↑]	IoU [↑]	F1 [↑]	BAcc [↑]	IoU [↑]	F1 [↑]	BAcc [↑]	IoU [↑]	F1 [↑]	BAcc [↑]	IoU [↑]	F1 [↑]	BAcc [↑]
CFS:															
Qwen3-VL-8B-Instruct	46.9	67.2	46.3	-	-	-	4.4	15.5	52.4	15.7	39.2	61.0	17.9	46.7	65.9
LLaVA-Next	32.4	58.3	32.7	-	-	-	2.9	14.6	50.4	12.4	32.7	54.5	7.6	36.2	59.3
DeepSeek R1	13.9	50.0	19.4	-	-	-	4.1	44.2	49.0	0.0	0.0	49.3	2.3	40.9	57.2
Gemini 2.5 Pro	16.0	51.1	21.4	-	-	-	2.4	20.0	53.8	5.1	25.0	48.1	4.2	32.7	45.2
SleepLM (Cap)	71.2	82.2	70.2	-	-	-	<u>26.0</u>	<u>58.9</u>	<u>74.2</u>	<u>18.9</u>	<u>55.1</u>	71.0	38.5	<u>73.3</u>	<u>84.6</u>
SleepLM (CLIP)	68.7	83.0	72.9	-	-	-	-	-	-	-	-	-	-	-	-
SleepLM (CoCa)	<u>70.3</u>	84.0	74.6	-	-	-	23.8	58.0	73.5	19.1	55.6	<u>70.9</u>	40.1	72.8	83.3
SleepLM (ReCoCa)	69.5	<u>83.7</u>	<u>74.2</u>	-	-	-	26.9	61.5	78.1	17.6	52.2	69.1	<u>40.0</u>	75.0	86.9
MROS:															
Qwen3-VL-8B-Instruct	48.0	68.6	49.1	4.8	9.5	52.6	7.2	22.4	54.7	29.1	61.7	62.7	19.0	51.1	67.7
LLaVA-Next	27.6	56.6	29.8	1.6	7.7	52.4	1.9	12.7	50.7	21.0	49.3	51.3	9.0	39.2	59.9
DeepSeek R1	12.5	50.8	22.3	0.0	0.0	45.4	2.8	25.3	53.7	1.7	8.3	52.2	1.1	32.0	57.6
Gemini 2.5 Pro	17.0	52.1	21.9	1.2	10.5	82.7	0.0	11.8	52.4	7.7	39.5	52.4	3.2	31.0	55.7
SleepLM (Cap)	70.1	81.8	69.5	<u>23.0</u>	50.0	<u>68.8</u>	13.2	<u>34.5</u>	<u>60.5</u>	<u>36.1</u>	<u>74.3</u>	<u>66.9</u>	<u>43.0</u>	<u>72.0</u>	<u>81.3</u>
SleepLM (CLIP)	69.5	83.5	73.7	-	-	-	-	-	-	-	-	-	-	-	-
SleepLM (CoCa)	70.8	84.6	<u>75.6</u>	22.6	<u>43.1</u>	64.8	<u>14.0</u>	33.3	59.9	34.4	72.4	66.3	42.2	70.5	80.6
SleepLM (ReCoCa)	72.3	85.5	77.0	23.6	42.9	66.0	18.3	39.5	63.0	36.3	74.3	67.0	44.6	74.5	83.6
SHHS:															
Qwen3-VL-8B-Instruct	55.7	74.7	58.5	0.0	0.0	50.0	15.9	43.4	58.7	12.0	43.6	60.4	22.7	57.3	71.4
LLaVA-Next	39.3	61.9	38.5	0.0	0.0	49.5	6.0	30.3	52.5	10.3	36.0	52.6	8.5	40.7	61.3
DeepSeek R1	14.9	51.9	20.7	0.0	0.0	42.0	4.6	57.6	58.6	0.7	8.0	51.5	0.9	26.8	42.5
Gemini 2.5 Pro	22.4	53.3	30.1	0.0	0.0	38.0	3.2	13.3	50.1	2.0	16.7	45.1	2.1	27.6	45.6
SleepLM (Cap)	76.8	85.8	76.4	25.2	<u>48.3</u>	72.3	<u>35.7</u>	<u>66.3</u>	<u>74.1</u>	<u>18.5</u>	<u>51.6</u>	<u>66.3</u>	43.7	<u>75.7</u>	<u>84.6</u>
SleepLM (CLIP)	72.3	85.6	77.2	-	-	-	-	-	-	-	-	-	-	-	-
SleepLM (CoCa)	73.0	<u>85.9</u>	<u>77.8</u>	13.6	37.5	64.3	33.0	66.0	74.0	17.0	50.7	65.6	42.1	74.7	84.4
SleepLM (ReCoCa)	<u>74.6</u>	87.1	79.6	<u>24.7</u>	51.0	<u>70.8</u>	37.6	69.2	76.4	22.4	57.8	70.3	<u>42.8</u>	76.8	85.7

In this section, we present the granular, unaggregated results for sleep stage classification and sleep event localization & classification. We benchmark ReCoCa against a comprehensive suite of baselines, categorized into three groups:

- **Proprietary LLMs:** Leading general-purpose models (Gemini 2.5 Pro [14], DeepSeek-R1 [21]).
- **Finetuned VLMs:** Open-source vision-language models adapted for sleep (Qwen3-VL-8B-Instruct [6], LLaVA-Next [24]).
- **Standard Architectures:** SleepLM variants utilizing standard multimodal objectives (CLIP [32], Cap [38], CoCa [43]).

As detailed in Table 15, ReCoCa demonstrates superior zero-shot capabilities, achieving the top performance across the majority of tasks and securing the second-best position in the remaining metrics. This consistent dominance highlights the efficacy of our proposed architecture over both generic multimodal baselines and larger proprietary models.

D.6. Retrieval Raw Results

Table 16 | **Zero-shot retrieval results across datasets.** † indicates retrieval evaluated on 100 samples due to context length constraints. Best is **bold**; second best is underlined.

Model	Text→Signal					Signal→Text				
	R@1	R@5	R@10	Median Rank	Mean Rank	R@1	R@5	R@10	Median Rank	Mean Rank
CFS:										
DeepSeek R1†	-	-	-	-	-	2.0	5.0	13.0	-	-
Gemini 2.5 Pro†	-	-	-	-	-	5.0	8.0	11.0	-	-
SleepLM (CLIP)	66.0	87.8	92.5	1.00	5.97	57.6	83.0	89.3	1.00	8.12
SleepLM (CoCa)	<u>71.3</u>	<u>90.2</u>	<u>93.7</u>	1.00	4.69	<u>65.7</u>	87.0	91.7	1.00	5.96
SleepLM (ReCoCa)	78.7	91.5	94.5	1.00	6.14	70.0	<u>86.5</u>	<u>90.8</u>	1.00	9.31
MROS:										
DeepSeek R1†	-	-	-	-	-	2.0	6.0	13.0	-	-
Gemini 2.5 Pro†	-	-	-	-	-	6.0	17.0	28.0	-	-
SleepLM (CLIP)	67.8	83.5	87.4	1.00	8.39	68.7	85.4	90.3	1.00	5.05
SleepLM (CoCa)	<u>70.5</u>	<u>83.3</u>	<u>87.7</u>	1.00	7.54	<u>71.4</u>	<u>86.1</u>	<u>91.0</u>	1.00	<u>4.36</u>
SleepLM (ReCoCa)	72.7	84.5	88.3	1.00	<u>7.91</u>	76.2	88.2	92.5	1.00	4.03
SHHS:										
DeepSeek R1†	-	-	-	-	-	1.0	6.0	11.0	-	-
Gemini 2.5 Pro†	-	-	-	-	-	1.0	7.0	13.0	-	-
SleepLM (CLIP)	89.1	98.9	99.6	1.00	1.28	90.7	99.1	99.6	1.00	1.23
SleepLM (CoCa)	<u>90.9</u>	<u>99.0</u>	<u>99.7</u>	1.00	<u>1.23</u>	<u>92.9</u>	<u>99.5</u>	<u>99.7</u>	1.00	<u>1.19</u>
SleepLM (ReCoCa)	96.1	99.8	100.0	1.00	1.07	96.7	99.8	100.0	1.00	1.06

In this section, we present the comprehensive, unaggregated results for cross-modal retrieval. Note that due to context window limitations, proprietary LLMs are evaluated on a reduced pool size of $N = 100$, whereas all SleepLM variants are tested on a significantly more challenging pool of $N = 2000$. Despite this structural advantage, the LLM baselines yield performance near random chance, falling far behind the specialized pretraining variants. As detailed in Table 16, ReCoCa consistently achieves top-tier performance across the majority of metrics, validating its superior semantic alignment even when compared against baselines operating under easier test conditions.

D.7. Regression Raw Results

In this section, we present the comprehensive, unaggregated results for zero-shot regression, categorized into two distinct tasks:

Explicit Signal Grounding (Channel Stats): To assess the model’s understanding of visible input signals, we parse statistical descriptors from the generated channel-specific captions. We calculate the MAE and sMAPE for each statistic individually.

Implicit Physiological Inference (HR & SpO2): We isolate Heart Rate (HR) and Blood Oxygen (SpO2) results as these signals are explicitly excluded from the encoder input.

- **Scalar Estimation:** We evaluate the model’s ability to estimate (Max, Min, Mean) for HR and

Table 17 | **Zero-shot regression and trend statistics results across datasets.** Best is **bold**; second best is underlined.

Model	Heart Rate				SpO ₂			Channel Stats
	Mean MAE [↓]	Max MAE [↓]	Min MAE [↓]	Trend Recall [↑]	Mean MAE [↓]	Min MAE [↓]	Trend Recall [↑]	Avg sMAPE [↓]
CFS:								
Qwen-VL-8B-Instruct	18.07	23.56	14.69	11.8	<u>2.13</u>	<u>3.46</u>	10.0	33.56
LLaVA-Next	17.98	23.21	16.77	18.3	<u>2.56</u>	<u>4.73</u>	16.7	37.41
DeepSeek R1	8.54	18.06	11.08	28.6	1.84	3.57	9.6	-
Gemini 2.5 Pro	2.70	4.96	6.19	14.2	2.44	2.73	12.7	-
SleepLM (Cap)	3.57	6.23	9.79	18.4	2.40	3.82	35.4	<u>8.85</u>
SleepLM (CoCa)	3.25	5.67	<u>9.74</u>	19.7	2.42	3.89	42.8	9.02
SleepLM (ReCoCa)	<u>3.11</u>	<u>5.41</u>	9.97	<u>20.0</u>	2.52	4.15	<u>40.8</u>	3.99
SHHS:								
Qwen-VL-8B-Instruct	9.44	10.07	9.69	32.6	2.15	3.16	26.9	28.63
LLaVA-Next	11.37	12.20	12.16	26.0	2.60	4.44	20.9	35.81
DeepSeek R1	10.30	20.19	11.35	26.0	1.93	5.91	9.3	-
Gemini 2.5 Pro	1.32	8.12	6.06	7.2	1.96	1.78	11.5	-
SleepLM (Cap)	0.86	1.22	1.72	<u>51.1</u>	2.00	3.10	37.2	6.37
SleepLM (CoCa)	<u>0.85</u>	<u>1.19</u>	1.63	50.2	1.99	3.02	41.6	6.51
SleepLM (ReCoCa)	0.84	1.18	<u>1.64</u>	51.6	<u>1.95</u>	<u>2.94</u>	<u>37.5</u>	2.73

(Mean, Min) for SpO₂ based solely on cross-channel correlations (e.g., deriving heart rate from ECG).

- **Trend Detection:** We calculate the Recall for estimated morphological changes (trends) in these implicit signals.

Analysis: As shown in the results, proprietary LLMs demonstrate remarkably strong performance on HR and SpO₂ estimation. This suggests that these reasoning-heavy models effectively leverage well-established clinical algorithms (e.g., QRS detection logic) to derive vital signs from related inputs like ECG. However, ReCoCa maintains highly competitive performance on these implicit tasks while achieving a dominant margin on the “Channel Stats” metric, which reports the averaged sMAPE across all these signal descriptor except HR and SpO₂ trend. This confirms that while LLMs excel at specific, algorithmic derivations, ReCoCa possesses a superior, generalized grounding in the raw signal morphology across the full montage.

D.8. Statistical Results for Full Night Metrics

Arousal Index (all), Arousal Index (NREM), Central Apnea Index, CAI (4% desat), CAI (4% desat + arousal), AHI (3% desat), AHI (3% desat + arousal), AHI (4% desat), AHI (4% desat + arousal), RDI (no desat), RDI (arousal only) In this section, we present the quantitative evaluation of full-night clinical variable estimation. To assess real-world utility, we randomly sample 5 full-night recordings from the SHHS dataset. We process each recording using a sliding window approach to generate fine-grained, epoch-level predictions for sleep stages, event locations, and vital signal statistics. These temporal outputs are then aggregated to compute a comprehensive suite of clinically relevant indices.

Table 18 provides the complete comparison between ReCoCa and the fine-tuned Qwen3-VL-8B-Instruct. We observe the following key trends:

- **Sleep Architecture:** ReCoCa dominates on macro-structural metrics, including Sleep Efficiency (slpeffp), Sleep Latency (slplatp), and Wake After Sleep Onset (waso).
- **Respiratory Indices:** ReCoCa achieves significantly lower error rates on Apnea-Hypopnea Index (AHI) metrics (e.g., ahi_a0h3, ahi_a0h4) and their arousal-linked variants. Notably, it achieves

Table 18 | **Full-Night PSG Metrics Comparison: ReCoCa vs Qwen3 (5 SHHS Subjects)**. $\Delta = |\text{Qwen} - \text{Doc}| - |\text{ReCoCa} - \text{Doc}|$. Positive values (bold) indicate ReCoCa is closer to ground truth.

Metric	Doc GT	ReCoCa	Qwen3	ReCoCa-Doc	Qwen-Doc	Δ
Sleep Efficiency (%)	87.95	88.38	89.28	+0.42	+1.33	+0.90
Sleep Latency (min)	17.40	14.50	4.30	-2.90	-13.10	+10.20
WASO (min)	38.40	39.40	44.30	+1.00	+5.90	+4.90
Arousal Index (all)	23.04	28.01	20.64	+4.97	-2.41	-2.56
Arousal Index (NREM)	24.12	26.10	19.29	+1.98	-4.83	+2.84
Central Apnea Index	0.21	0.11	0.13	-0.10	-0.08	-0.02
CAI (4% desat)	0.08	0.00	0.00	-0.08	-0.08	+0.00
CAI (4% desat + arousal)	0.08	0.04	0.10	-0.03	+0.03	-0.00
AHI (3% desat)	25.02	16.01	6.69	-9.01	-18.33	+9.32
AHI (3% desat + arousal)	26.97	23.52	17.02	-3.45	-9.95	+6.50
AHI (4% desat)	21.91	11.21	3.55	-10.70	-18.36	+7.66
AHI (4% desat + arousal)	24.31	20.99	15.80	-3.33	-8.51	+5.19
RDI (no desat)	37.46	38.02	28.71	+0.56	-8.75	+8.19
RDI (arousal only)	12.93	18.94	15.05	+6.01	+2.12	-3.89

nearly perfect performance on the Respiratory Disturbance Index (RDIOP), showing a delta of +0.56 compared to Qwen3’s deviation of -8.75.

- **Arousal Metrics:** Qwen3 demonstrates marginally better performance on the aggregate Arousal Index (ai_all) and specific Central Apnea Index (CAI) sub-metrics, though it fails to maintain consistency across the broader respiratory reporting.

This evaluation confirms that ReCoCa successfully translates epoch-level understanding into accurate, longitudinal clinical summaries, outperforming larger fine-tuned VLMs in generating consistent full-night reports.

D.9. Full Generation Quality Results

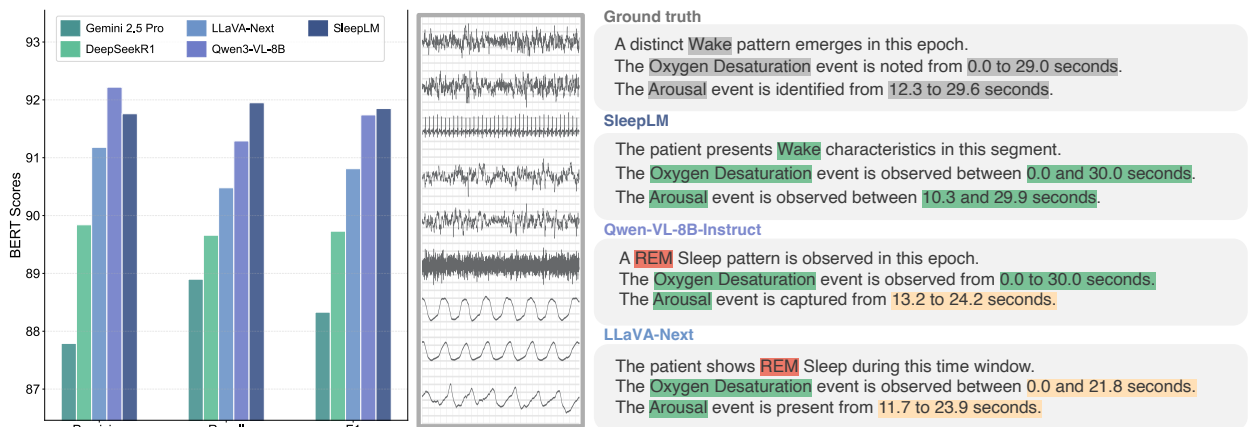


Figure 9 | Full generation from Qwen3 and DeepSeek-R1 from the paper example.

In this section, we extend our qualitative assessment to the finetuned vision-language baselines, Qwen3-VL-8B-Instruct and LLaVA-Next. As illustrated in Fig. 9, while these specialized models

Table 19 | Example templated sleep report generated from our predicted variables.

Example templated sleep report (Patient: shhs1-201545).

Full-night summary: Wake time accounts for 12.3%, sleep time accounts for 87.7%. AHI=60.91 events/hr, ODI₄=46.94 events/hr. Mean SpO₂=93.7%, minimum SpO₂=40.0%. Event composition: hypopnea 100.0%, central apnea 0.0%.

Sleep structure

Total recording time (TRT)	08:09:30
Total sleep time (TST)	07:09:30
Sleep efficiency (SE)	87.7%
Wake after sleep onset (WASO)	59.5 min

Respiratory events (from generated variables)

	Count	Index (events/hr)
Central apnea	0	0.00
Hypopnea	436	60.91
Total (all resp. events)	436	60.91
Oxygen desaturation ($\geq 4\%$)	336	46.94

Oxygen saturation (SpO₂)

Mean SpO ₂ (TRT)	93.7%
Minimum SpO ₂ (TRT)	40.0%

generate marginally more accurate descriptions than the LLMs, they fail to close the performance gap. Both models remain prone to significant misconceptions and lack the granular spatial awareness required for precise event localization.

D.10. Targeted Generation Ability

See Figure 10 as a demonstration for SleepLM’s targeted generation ability.

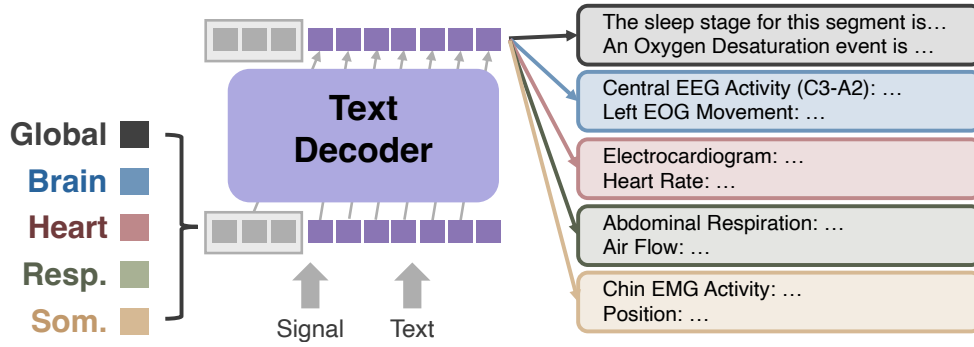


Figure 10 | **Illustration of targeted generation:** Given a pair of input signal and text, our model is able to perform precise, targeted generations on one or multiple designated modalities by prepending corresponding condition tokens during text decoding.

D.11. Caption Generation performance

We provide the raw generation performance metrics in this section to compare the caption quality generated by SleepLM and other LLM and finetuned VLM baselines. See Table 20 for the raw scores. SleepLM outperforms the baselines across all evaluated metrics, demonstrating a more accurate caption quality.

Table 20 | **Caption generation performance comparison between SleepLM and baselines.**

Model	BERTPrecision	BERTRecall	BERTF1	METEOR
DeepSeekR1	89.8	89.7	89.7	32.5
Gemini2.5pro	87.8	88.9	88.3	33.0
LLaVA-Next	91.2	90.5	90.8	42.3
SleepLM	91.8	92.0	91.8	44.2

E. Additional Examples

We provide extra qualitative examples in this section for reference purposes.

E.1. Multilevel Caption Examples

Examples Captions Generated by Our Data Pipeline (Categorized by Modalities)

===== Example 1 =====

SLEEP STAGE: During this epoch, the patient exhibits Stable Light Sleep (N2).

SLEEP EVENT: The Hypopnea event manifests from 0.0 to 3.9 seconds. The Oxygen Desaturation event is observed between 0.0 and 30.0 seconds. Analysis reveals a Arousal event from 3.5 to 8.5 seconds. A Hypopnea event is noted between 24.3 and 30.0 seconds.

BRAIN: Left EOG Movement: slow eye movement relative power measuring roughly 30%, REM saccadic relative power measuring 53.16%. Right EOG Movement: slow eye movement relative power equal to 28.50%, REM saccadic relative power valued at 55.23%. Central EEG Activity (C3-A2): delta relative power valued at 75.83%, combined with theta relative power of 9.82%, along with alpha relative power equal to approximately 8%, beta relative power at 4.07%. Central EEG Activity (C4-A1): delta relative power at 69.12%, plus theta relative power valued at 11.11%, a alpha relative power of 13.49%, and a beta relative power of roughly 4%.

HEART: Electrocardiogram: estimated beat rate valued at 78.2 bpm, ultra-short HRV (RMSSD) valued at 21.41 ms. Heart Rate: minimum value equal to 75.20, together with a maximum value at 85.20, mean valued at 78.51, an decreasing tendency from 0.9 to 5.6 seconds, along with a increasing pattern from 6.6 to 15.0 seconds, along with an decreasing behavior from 15.9 to 24.4 seconds. Blood Oxygen Saturation: a minimum value valued at 89.31, combined with a mean equal to 91.87, combined with a decreasing interval from 0.9 to 18.8 seconds, together with a increasing progression from 19.7 to 30.0 seconds.

RESPIRATORY: Abdominal Respiration: respiratory rate measuring 18.73 bpm, and lastly a breath interval variability of 0.70 s. Thoracic Respiration: respiratory rate at 18.11 bpm, and breath interval variability equal to 0.80 s. Air Flow: airflow rate measuring 18.55 bpm, a inspiratory flow flatness equal to 0.65.

SOMATIC: Chin EMG Activity: a median frequency measuring 16.69 Hz, along with burst intensity ratio of 2.11, and finally burst modulation rate of 0.98 Hz. The patient stays lying supine for the full epoch.

===== Example 2 =====

SLEEP STAGE: A distinct Wake pattern emerges in this epoch.

SLEEP EVENT: The epoch showed no notable sleep disruptions.

BRAIN: Left EOG Movement: slow eye movement relative power of 25.87%, plus REM saccadic relative power equal to 39.43%. Right EOG Movement: slow eye movement relative power of 27.01%, REM saccadic relative power of 40.66%. Central EEG Activity (C3-A2): a delta relative power of 20.64%, as well as a theta relative power valued at around 15%, along with a alpha relative power measuring roughly 32%, and beta relative power valued at around 29%. Central EEG Activity (C4-A1): delta relative power valued at 24.38%, plus a theta relative power at approximately 12%, including alpha relative power valued at 31.34%, beta relative power measuring 28.54%.

HEART: Electrocardiogram: estimated beat rate valued at 83.0 bpm, a ultra-short HRV (RMSSD) valued at 10.91 ms. Heart Rate: a minimum value equal to 82.23, plus a maximum value measuring 85.36, together with a mean equal to 83.73, including a increasing interval from 0.9 to 3.8 seconds, a decreasing period from 4.7 to 11.3 seconds, together with a increasing movement from 12.2 to 15.0 seconds, together with an increasing segment from 15.9 to 20.6 seconds, plus a decreasing movement from 19.7 to 26.3 seconds, combined with an increasing trend from 27.2 to 30.0 seconds. Blood Oxygen Saturation: minimum value valued at 96.09, accompanied by mean equal to 97.16, in addition to a decreasing period from 0.9 to 15.0 seconds, and also an increasing behavior from 10.3 to 30.0 seconds.

RESPIRATORY: Abdominal Respiration: a respiratory rate at 12.21 bpm, plus breath interval variability of 1.83 s. Thoracic Respiration: respiratory rate valued at 11.93 bpm, and a breath interval variability equal to 1.84 s. Air Flow: airflow rate measuring 11.36 bpm, inspiratory flow flatness equal to 0.55.

SOMATIC: Chin EMG Activity: median frequency measuring 17.87 Hz, burst intensity ratio of 1.68, along with burst modulation rate measuring 1.03 Hz. The patient lies prone throughout the recording.

===== Example 3 =====

SLEEP STAGE: A clear REM Sleep pattern is detected in this epoch.

SLEEP EVENT: A Arousal event is present between 0.0 and 9.2 seconds. We detect a Hypopnea event from 16.3 to 30.0 seconds.

BRAIN: Left EOG Movement: slow eye movement relative power valued at 25.54%, and a REM saccadic relative power of 57.11%. Right EOG Movement: a slow eye movement relative power measuring 31.46%, and lastly a REM saccadic relative power measuring around 47%. Central EEG Activity (C3-A2): a delta relative power valued at 41.63%, accompanied by theta relative power of 10.02%, including a alpha relative power of 11.44%, and beta relative power valued at 34.41%. Central EEG Activity (C4-A1): delta relative power measuring approximately 54%, including theta relative power equal to 12.56%, alpha relative power of 7.81%, and beta relative power of 22.95%.

HEART: Electrocardiogram: a estimated beat rate of 48.9 bpm, as well as ultra-short HRV (RMSSD) measuring 85.54 ms. Heart Rate: minimum value measuring roughly 44, as well as a maximum value at 66.21, including mean valued at 53.96, a increasing pattern from 0.9 to 7.5 seconds, and finally an decreasing tendency from 10.3 to 20.6 seconds. Blood Oxygen Saturation: a minimum value equal to 92.19, in addition to a mean of roughly 94, as well as a decreasing period from 0.9 to 9.4 seconds, and lastly a increasing pattern from 4.7 to 28.1 seconds.

RESPIRATORY: Abdominal Respiration: respiratory rate measuring 15.90 bpm, and also a breath interval variability of 0.20 s. Thoracic Respiration: respiratory rate valued at 15.61 bpm, and breath interval variability valued at 0.41 s. Air Flow: a airflow rate equal to 17.41 bpm, in addition to a inspiratory flow flatness measuring 0.70.

SOMATIC: Chin EMG Activity: a median frequency valued at 15.05 Hz, burst intensity ratio at 3.05, a burst modulation rate valued at 0.78 Hz. The patient maintains the right position for this period.

E.2. Prompts Used for LLM

Zeroshot Evaluation Prompt

You are an expert sleep medicine AI assistant analyzing polysomnography (PSG) signals.

INPUT FORMAT

You will receive raw biosignal data from a 30-second sleep epoch. The signals are sampled at 64 Hz (1920 samples per channel).

The channels provided are:

- ECG: Electrocardiogram - heart electrical activity
- ABD: Abdominal respiration belt - breathing movement
- THX: Thoracic respiration belt - chest breathing movement
- AF: Airflow - nasal/oral airflow signal
- EOG_E1_A2: Left electrooculogram - left eye movement
- EOG_E2_A1: Right electrooculogram - right eye movement
- EEG_C3_A2: Central EEG (C3-A2) - brain electrical activity
- EEG_C4_A1: Central EEG (C4-A1) - brain electrical activity
- EMG_Chin: Chin electromyogram - muscle activity
- POS: Body position - somatic state

The data is formatted as a CSV table with columns for time and each channel:

Time(s),ECG,ABD,THX,AF,...

0.0000,-0.49,-0.76,0.12,...

0.0156,-0.55,-0.78,0.15,...

Each row represents a single time point with all channel values.

YOUR TASK

Analyze the signals and provide:

1. ****Sleep Stage****: Classify the epoch into one of these stages:

```

- Wake
- Light Sleep (N1)
- Stable Light Sleep (N2)
- Deep Sleep (N3)
- REM Sleep

2. **Sleep Events**: Detect any of these events with their start and end times
(in seconds from 0.0 to 30.0):
- Central Apnea
- Hypopnea
- Oxygen Desaturation
- Arousal

Look for patterns like:
- Reduced airflow (AF channel) for apnea/hypopnea
- Oxygen desaturation patterns (infer from overall signal changes)
- Arousal patterns (sudden changes in EEG/EMG)

3. **Heart Rate Trends**: Analyze the ECG channel to detect periods of:
- "increasing" heart rate
- "decreasing" heart rate
- "stable" heart rate
Provide start and end times for each trend segment.

4. **SpO2 (Blood Oxygen) Trends**: Based on the overall physiological
patterns, infer:
- "increasing" blood oxygen
- "decreasing" blood oxygen
- "stable" blood oxygen
Provide start and end times for each trend segment.

## IMPORTANT NOTES
- All times must be between 0.0 and 30.0 seconds
- If you don't detect any events, return an empty list
- If you cannot determine trends, return an empty list
- Provide confidence scores (0.0 to 1.0) for all predictions
- For trends, focus on detecting the main patterns - you don't need to capture
every tiny fluctuation

## OUTPUT FORMAT
Respond with a JSON object matching the specified schema.

```

Zeroshot Retrieval Prompt

You are an expert sleep medicine AI assistant performing signal-to-text retrieval.

INPUT FORMAT

You will receive:

1. Raw polysomnography (PSG) biosignal data from a 30-second sleep epoch
2. A list of candidate text captions describing different sleep epochs

The signals are sampled at 64 Hz (1920 samples per channel).

The channels provided are:

- ECG: Electrocardiogram - heart electrical activity
- ABD: Abdominal respiration belt - breathing movement
- THX: Thoracic respiration belt - chest breathing movement
- AF: Airflow - nasal/oral airflow signal
- EOG_E1_A2: Left electrooculogram - left eye movement
- EOG_E2_A1: Right electrooculogram - right eye movement
- EEG_C3_A2: Central EEG (C3-A2) - brain electrical activity
- EEG_C4_A1: Central EEG (C4-A1) - brain electrical activity
- EMG_Chin: Chin electromyogram - muscle activity
- POS: Body position - somatic state

The signal data is formatted as a CSV table with columns for time and each channel:

```
Time(s),ECG,ABD,THX,AF,...
0.0000,-0.49,-0.76,0.12,...
0.0156,-0.55,-0.78,0.15,...
```

YOUR TASK

Analyze the biosignal data and rank the captions by how well they describe this specific signal epoch.

Consider:

- Sleep stage characteristics (EEG patterns, eye movements, muscle tone)
- Respiratory events (airflow, chest/abdominal movement patterns)
- Cardiac patterns (heart rate variability)
- Temporal patterns and event timing
- Overall physiological coherence

OUTPUT FORMAT

Respond with a JSON object containing:

- top_captions: List of up to 10 best matching captions, ranked from best to worst
 - Each entry has: caption_idx (0-based index) and confidence (0.0 to 1.0)
 - MUST include at least the top caption, preferably top 10
- reasoning: Brief explanation of your ranking

Be precise and analytical in your ranking.

Unseen classification Prompt

You are an expert sleep medicine AI assistant performing signal-to-text retrieval.

INPUT FORMAT

You will receive:

1. Raw polysomnography (PSG) biosignal data from a 30-second sleep epoch
2. A list of candidate text captions describing different sleep epochs

The signals are sampled at 64 Hz (1920 samples per channel).

The channels provided are:

- ECG: Electrocardiogram - heart electrical activity
- ABD: Abdominal respiration belt - breathing movement
- THX: Thoracic respiration belt - chest breathing movement
- AF: Airflow - nasal/oral airflow signal

- EOG_E1_A2: Left electrooculogram - left eye movement
- EOG_E2_A1: Right electrooculogram - right eye movement
- EEG_C3_A2: Central EEG (C3-A2) - brain electrical activity
- EEG_C4_A1: Central EEG (C4-A1) - brain electrical activity
- EMG_Chin: Chin electromyogram - muscle activity
- POS: Body position - somatic state

The signal data is formatted as a CSV table with columns for time and each channel:

```
Time(s),ECG,ABD,THX,AF,...
0.0000,-0.49,-0.76,0.12,...
0.0156,-0.55,-0.78,0.15,...
```

YOUR TASK

Determine if **Obstructive Apnea** is present (1) or absent (0) in this epoch.

OUTPUT FORMAT

Respond with a JSON object containing:

- prediction: 0 or 1
- confidence: your confidence score (0.0 to 1.0)
- reasoning: brief explanation of your decision

E.3. Templates Examples for Classification

Templates Pool

1. Templates for embedding similarity based sleep stage classification

```
SLEEP_STAGE_TEMPLATES = [
    "The sleep stage for this epoch is {sleep_stage}.",
    "This epoch is classified as {sleep_stage}.",
    "The patient is in {sleep_stage} during this period.",
    "Sleep stage analysis shows {sleep_stage} for this epoch.",
    "This 30-second segment corresponds to {sleep_stage}.",
    "The sleep stage classification indicates {sleep_stage}.",
    "During this epoch, the patient exhibits {sleep_stage}.",
    "Sleep stage assessment reveals {sleep_stage} for this period.",
    "This time window is characterized by {sleep_stage}.",
    "The sleep stage for this segment is {sleep_stage}.",
    "The patient demonstrates {sleep_stage} during this epoch.",
    "A {sleep_stage} pattern is observed in this epoch.",
    "The sleep stage analysis reveals {sleep_stage} for this period.",
    "This epoch shows evidence of {sleep_stage}.",
    "The patient presents {sleep_stage} characteristics in this segment.",
    "A clear {sleep_stage} pattern is detected in this epoch.",
    "The patient manifests {sleep_stage} during this time window.",
    "Sleep stage evaluation indicates {sleep_stage} for this epoch.",
    "The patient displays {sleep_stage} behavior in this segment.",
    "A distinct {sleep_stage} pattern emerges in this epoch.",
    "The patient exhibits {sleep_stage} during this period.",
    "Sleep stage assessment shows {sleep_stage} for this epoch.",
    "The patient reveals {sleep_stage} characteristics in this segment.",
```

```

"A {sleep_stage} state is observed during this epoch.",
"The patient demonstrates {sleep_stage} patterns in this period.",
"Sleep stage analysis indicates {sleep_stage} for this epoch.",
"The patient shows {sleep_stage} during this time window.",
"A {sleep_stage} phase is detected in this epoch.",
"The patient exhibits {sleep_stage} features in this segment.",
"Sleep stage evaluation reveals {sleep_stage} for this epoch.",
"The patient presents {sleep_stage} during this period.",
"A {sleep_stage} condition is observed in this epoch.",
]

```

2. Templates for embedding similarity based unseen events classification:

```

PRESENCE_TEMPLATES = [
    "a {event_name} event is present during this epoch.",
    "the {event_name} event occurs in this period.",
    "we observe a {event_name} event.",
    "a {event_name} event is detected.",
    "the {event_name} event is identified in this epoch.",
    "this epoch contains a {event_name} event.",
    "a {event_name} event is recorded during this period.",
    "the presence of a {event_name} event is noted.",
    "we detect a {event_name} event in this epoch.",
    "a {event_name} event is observed.",
    "the {event_name} event manifests during this period.",
    "analysis reveals a {event_name} event.",
    "a {event_name} event occurs in this recording.",
    "the {event_name} event is present.",
    "we identify a {event_name} event.",
    "this period shows a {event_name} event.",
    "a {event_name} event is documented.",
    "the {event_name} event is captured in this epoch.",
    "we observe the occurrence of a {event_name} event.",
    "a {event_name} event is evident during this period.",
    "the {event_name} event appears in this epoch.",
    "this epoch exhibits a {event_name} event.",
    "a {event_name} event is noted.",
    "the {event_name} event is found in this recording.",
    "we record a {event_name} event during this period.",
    "a {event_name} event is visible in this epoch.",
    "the {event_name} event is apparent.",
    "this period includes a {event_name} event.",
    "a {event_name} event is present in the recording.",
    "the {event_name} event is confirmed in this epoch.",
]

ABSENCE_TEMPLATES = [
    "no significant sleep events were detected during this epoch.",
    "the epoch showed no notable sleep disruptions.",
    "sleep remained undisturbed throughout the epoch.",
    "the patient experienced uninterrupted sleep during this period.",
    "sleep continuity was maintained without events.",
    "no respiratory or movement-related sleep disturbances were observed.",
]

```

"the recording showed no significant sleep-related events.",
 "no pathological sleep events were identified in this epoch.",
 "sleep proceeded without notable interruptions or events.",
 "the patient maintained stable sleep without events.",
 "no clinically significant sleep events were recorded.",
 "sleep architecture remained undisturbed during this period.",
 "the epoch was characterized by uninterrupted sleep patterns.",
 "no sleep-disordered breathing events were detected.",
 "this epoch is free of significant sleep events.",
 "the recording shows normal sleep without events.",
 "no abnormal sleep events were found in this period.",
 "sleep was uninterrupted by events.",
 "the epoch demonstrates absence of sleep events.",
 "no respiratory events or arousals were detected.",
 "the patient slept without disturbances.",
 "this period lacks any significant sleep events.",
 "no events were observed during this epoch.",
 "sleep quality was maintained without events.",
 "the recording is negative for sleep events.",
 "no clinically relevant events were identified.",
 "this epoch shows normal sleep patterns without events.",
 "the analysis found no pathological events.",
 "the absence of sleep events characterizes this period.",
 "no pathological findings were detected.",

]

E.4. More Retrieval Quality Examples

More retrieval Examples

===== Paper Example =====

Original Caption: "A REM Sleep pattern is observed in this epoch. The Oxygen Desaturation event is documented from 13.0 to 28.0 seconds. The Hypopnea event is observed from 20.0 to 30.0 seconds."

MOST SIMILAR:

1. "A REM Sleep pattern is observed in this epoch. The Oxygen Desaturation event is documented from 13.0 to 28.0 seconds. The Hypopnea event is observed from 20.0 to 30.0 seconds."
2. "The sleep stage for this segment is REM Sleep. The Hypopnea event is recorded from 0.0 to 19.2 seconds. A Oxygen Desaturation event occurs from 17.6 to 30.0 seconds. A Hypopnea event is identified between 25.8 and 30.0 seconds."
3. "During this epoch, the patient exhibits REM Sleep. We observe a Hypopnea event between 0.0 and 3.9 seconds. The Hypopnea event spans 16.1 to 28.3 seconds. The Oxygen Desaturation event is observed from 17.3 to 30.0 seconds."

INTERMEDIATE SIMILAR:

1. "Sleep stage evaluation reveals REM Sleep for this epoch. A Oxygen Desaturation event is observed between 15.1 and 27.0 seconds. The Hypopnea event is documented from 20.9 to 30.0 seconds."

2. "The patient exhibits Light Sleep (N1) features in this segment. The patient maintained stable sleep without events."
3. "The sleep stage classification indicates Deep Sleep (N3). No significant sleep events were detected during this epoch."

LEAST SIMILAR:

1. "The patient shows Deep Sleep (N3) during this time window. No sleep-disordered breathing events were detected."
2. "The patient is in Stable Light Sleep (N2) during this period. No clinically significant sleep events were recorded."
3. "Sleep stage evaluation reveals Stable Light Sleep (N2) for this epoch. No sleep-disordered breathing events were detected."

===== Extra Example 1 =====

Original Caption: "The patient exhibits Stable Light Sleep (N2) features in this segment. No sleep-disordered breathing events were detected."

MOST SIMILAR:

1. "The patient exhibits Stable Light Sleep (N2) features in this segment. No sleep-disordered breathing events were detected."
2. "This time window is characterized by Stable Light Sleep (N2). No respiratory or movement-related sleep disturbances were observed."
3. "Sleep stage analysis shows Stable Light Sleep (N2) for this epoch. The epoch showed no notable sleep disruptions."

INTERMEDIATE SIMILAR:

1. "The patient displays Stable Light Sleep (N2) behavior in this segment. Sleep proceeded without notable interruptions or events."
2. "The patient displays Stable Light Sleep (N2) behavior in this segment. No significant sleep events were detected during this epoch."
3. "The patient exhibits Stable Light Sleep (N2) during this period. A Hypopnea event is detected between 9.1 and 29.3 seconds."

LEAST SIMILAR:

1. "The patient presents REM Sleep during this period. The Hypopnea event is present between 29.3 and 30.0 seconds."
2. "The patient presents Stable Light Sleep (N2) characteristics in this segment. The Oxygen Desaturation event is documented from 0.0 to 12.0 seconds. A Oxygen Desaturation event is noted between 0.0 and 16.0 seconds. The Hypopnea event is observed between 18.0 and 30.0 seconds. A Oxygen Desaturation event occurs from 21.0 to 30.0 seconds."
3. "The patient shows REM Sleep during this time window. The Oxygen Desaturation event is documented from 0.0 to 30.0 seconds. We observe a Oxygen Desaturation event between 14.0 and 30.0 seconds."

===== Extra Example 2 =====

Original Caption: "This 30-second segment corresponds to Stable Light Sleep (N2). Sleep continuity was maintained without events."

MOST SIMILAR:

1. "This 30-second segment corresponds to Stable Light Sleep (N2). Sleep continuity was maintained without events."
2. "During this epoch, the patient exhibits Stable Light Sleep (N2). No sleep-disordered breathing events were detected."
3. "Sleep stage analysis indicates Stable Light Sleep (N2) for this epoch. The recording showed no significant sleep-related events."

INTERMEDIATE SIMILAR:

1. "The sleep stage classification indicates Stable Light Sleep (N2). The recording showed no significant sleep-related events."
2. "The sleep stage analysis reveals Stable Light Sleep (N2) for this period. A Hypopnea event is detected from 0.0 to 2.8 seconds. Analysis reveals a Oxygen Desaturation event from 4.7 to 30.0 seconds."
3. "The sleep stage for this segment is Stable Light Sleep (N2). The Arousal event is documented from 17.9 to 26.1 seconds."

LEAST SIMILAR:

1. "The patient displays Stable Light Sleep (N2) behavior in this segment. A Hypopnea event is noted between 0.0 and 13.0 seconds. The Arousal event is identified from 11.6 to 17.3 seconds. A Oxygen Desaturation event is identified between 20.0 and 30.0 seconds."
2. "The sleep stage for this segment is REM Sleep. The Hypopnea event is recorded from 0.0 to 19.2 seconds. A Oxygen Desaturation event occurs from 17.6 to 30.0 seconds. A Hypopnea event is identified between 25.8 and 30.0 seconds."
3. "Sleep stage evaluation indicates REM Sleep for this epoch. The Oxygen Desaturation event manifests from 0.0 to 17.1 seconds. The Oxygen Desaturation event is noted from 14.0 to 30.0 seconds."

===== Extra Example 3 =====

Original Caption: "Sleep stage assessment reveals Stable Light Sleep (N2) for this period. The Hypopnea event is recorded from 0.0 to 10.7 seconds. The Hypopnea event is present from 21.2 to 30.0 seconds."

MOST SIMILAR:

1. "Sleep stage assessment reveals Stable Light Sleep (N2) for this period. The Hypopnea event is recorded from 0.0 to 10.7 seconds. The Hypopnea event is present from 21.2 to 30.0 seconds."
2. "The patient displays Stable Light Sleep (N2) behavior in this segment. The Hypopnea event is documented from 4.0 to 14.0 seconds."
3. "Sleep stage assessment shows Stable Light Sleep (N2) for this epoch. We observe a Hypopnea event between 26.5 and 30.0 seconds."

INTERMEDIATE SIMILAR:

1. "The patient shows REM Sleep during this time window. The Hypopnea event lasts from 9.0 to 26.0 seconds. The Arousal event is captured from 26.9 to 30.0 seconds."
2. "The sleep stage for this segment is REM Sleep. The patient experienced uninterrupted sleep during this period."
3. "A Stable Light Sleep (N2) state is observed during this epoch. The recording showed no significant sleep-related events."

LEAST SIMILAR:

1. "The sleep stage for this segment is Wake. No clinically significant sleep events were recorded."
2. "A distinct Stable Light Sleep (N2) pattern emerges in this epoch. Sleep architecture remained undisturbed during this period."
3. "Sleep stage evaluation indicates Deep Sleep (N3) for this epoch. Sleep proceeded without notable interruptions or events."

E.5. More Perturbation Study

More Perturbation Examples

===== Paper Example =====

Similarity = 0.3114 (original)

"A REM Sleep phase is detected in this epoch. The Hypopnea event is documented from 0.0 to 5.6 seconds. The Oxygen Desaturation event lasts from 14.2 to 30.0 seconds. The Hypopnea event is present from 15.4 to 30.0 seconds."

Similarity = 0.2743

"... The Hypopnea event is documented from 0.0 to 2.0 seconds. ... The Hypopnea event is documented from 28.0 to 30.0 seconds."

Similarity = 0.3195

"... The Hypopnea event is documented from 0.0 to 4.0 seconds. ... The Hypopnea event is documented from 22.0 to 30.0 seconds."

Similarity = 0.3511

"... The Hypopnea event is documented from 0.0 to 6.0 seconds. ... The Hypopnea event is documented from 16.0 to 30.0 seconds."

Similarity = 0.3152

"... The Hypopnea event is documented from 0.0 to 8.0 seconds. ... The Hypopnea event is documented from 12.0 to 30.0 seconds."

Similarity = 0.3044

"... The Hypopnea event is documented from 0.0 to 10.0 seconds. ... The Hypopnea event is documented from 10.0 to 30.0 seconds."

===== Extra Example 1 =====

Similarity = 0.4007 (original)

"Sleep stage analysis indicates Stable Light Sleep (N2) for this epoch. The Central Apnea event is identified from 0.0 to 23.9 seconds. We observe a Oxygen Desaturation event between 29.3 and 30.0 seconds."

Similarity = 0.2440

"... The Central Apnea event is identified from 0.0 to 6.0 seconds. ..."

Similarity = 0.3053

"... The Central Apnea event is identified from 0.0 to 12.0 seconds. ..."

Similarity = 0.3525

"... The Central Apnea event is identified from 0.0 to 18.0 seconds. ..."

Similarity = 0.3568

"... The Central Apnea event is identified from 0.0 to 24.0 seconds. ..."

Similarity = 0.3326

"... The Central Apnea event is identified from 0.0 to 30.0 seconds. ..."

===== Extra Example 2 =====

Similarity = 0.3294 (original)

"Sleep stage assessment shows Stable Light Sleep (N2) for this epoch. We detect a Arousal event from 15.0 to 30.0 seconds."

Similarity = 0.1243

"... We detect a Arousal event from 0.0 to 30.0 seconds."

Similarity = 0.2723

"... We detect a Arousal event from 5.0 to 30.0 seconds."

Similarity = 0.3294

"... We detect a Arousal event from 15.0 to 30.0 seconds."

Similarity = 0.2248

"... We detect a Arousal event from 25.0 to 30.0 seconds."

Similarity = 0.2290

"... We detect a Arousal event from 29.0 to 30.0 seconds."