Subject-Relative Micro-Motion and Sleep Dynamics for Near-Infrared Video Sleep Staging

Kunmin Jang[*]
Seoul National University
jangk10@snu.ac.kr

You Rim Choi[*]
Seoul National University
yrchoi@snu.ac.kr

Hun Heo
Seoul National University
greathunh@snu.ac.kr

Heonjun Lee
Seoul National University
johnlee0321@snu.ac.kr

Suahn Bae
Seoul National University
snubbb@snu.ac.kr

Dongik Park
Seoul National University
breadsora@snu.ac.kr

Hyun-Woo Shin[†]
Seoul National University
Seoul National University Hospital
charlie@snu.ac.kr

Hyung-Sin Kim[†]
Seoul National University
hyungkim@snu.ac.kr

Abstract

Near-infrared (NIR) video is a promising modality for contactless sleep monitoring, but recent video-based sleep staging methods often use it as a route to reconstructed respiratory/cardiac proxies or cross-modal physiological representations. We study video-only sleep staging under labels defined by polysomnography (PSG), where the model infers sleep stages from NIR video alone without explicit physiological proxy reconstruction or auxiliary physiological signal supervision. This tests whether NIR video itself can provide informative sleep-stage evidence, rather than only serving as an input for recovering physiological proxies. We propose *ViNUSS* (Video-Native Unmediated Sleep Staging), a framework that combines subject-relative micro-motion learning with full-night sleep dynamics modeling. Spatially anchored pre-spatial micro-motion encoding preserves localized temporal variation with its spatial context. Within-subject stage contrast learns stage cues with respect to each subject's night-specific baseline. Two-scale sleep dynamics modeling captures within-epoch motion evolution and organizes epoch-level evidence into a coherent full-night sleep-stage trajectory. On 475 overnight NIR recordings ($\sim$3,250 hours), *ViNUSS* achieves 0.80 accuracy and 0.78 macro-F1 for four-class sleep staging, and adding reconstructed or contact-sensor waveform proxies to its representation yields no measurable improvement. Interpretability analysis suggests attention to thoraco-abdominal periodic motion and gross body movements associated with arousals and position changes. These results support NIR video as an independently informative modality for PSG-defined sleep-stage estimation.

1 Introduction

Sleep staging is a central procedure in sleep medicine, assigning consecutive 30-second epochs of an overnight recording to Wake, Light (N1/N2), Deep (N3), or REM [2, 36]. The clinical reference is polysomnography (PSG), where trained technicians score multimodal physiological signals such

[*]These authors contributed equally.
[†]Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

as EEG, EOG, and EMG [35, 41]. Although PSG provides the standard definition of sleep stages, it requires contact sensors, overnight instrumentation, and expert scoring, limiting scalability for longitudinal and home-based monitoring [30, 52]. These limitations have motivated contactless approaches that estimate sleep stages under PSG-defined labels while reducing sensing burden.

Near-infrared (NIR) video is a promising modality for contactless sleep monitoring because it operates in darkness and captures spatially localized body motion, including respiration-related motion, limb movement, position changes, and brief arousals. Recent video-based sleep staging methods often use video as a route to another representation, such as reconstructed respiratory/cardiac signals, optical-flow-derived motion features, or EEG-guided cross-modal features [6, 17, 33, 45, 50, 53]. Proxy-first approaches are valuable and interpretable, but blankets, position changes, and face anonymization can make respiratory or remote photoplethysmography (rPPG) extraction unreliable, while one-dimensional proxies may discard regional motion patterns and transient movements. Cross-modal physiological approaches are complementary, but they answer a different question: how video can align with auxiliary physiological representations.

In this work, we study video-only sleep staging under labels defined by PSG. The model infers sleep stages from NIR video alone, without explicit physiological proxy reconstruction or auxiliary physiological representation learning. This formulation does not replace PSG as the clinical reference. Instead, it tests whether NIR video itself can provide informative sleep-stage evidence, rather than only serving as an input for recovering physiological proxies or aligning to representations derived from other physiological modalities.

This setting differs from generic video understanding. Standard video models are typically designed for actions, objects, and scene changes, where discriminative cues often appear as large motion or semantic interaction [3, 11, 22, 27, 43]. Action-oriented temporal modules further enhance motion representation through temporal shifts, motion excitation, or temporal differences [25, 26, 49]. In NIR sleep video, however, the scene is mostly static, useful cues may appear as weak local micro-motion, and subject-specific factors such as body shape, bedding, camera geometry, and baseline respiration-related motion can dominate the representation. The interpretation of motion also depends on its location, since weak thoraco-abdominal periodicity, fast limb or body movements, and broad position changes provide different sleep-stage evidence. Effective video-only sleep staging therefore requires preserving localized temporal variation with its spatial context, learning stage evidence relative to each subject's night-specific baseline, and modeling sleep dynamics across the full night.

We propose *ViNUSS*, a framework for subject-relative micro-motion learning and full-night sleep dynamics modeling. First, spatially anchored pre-spatial micro-motion encoding densely scans the full 30-second epoch with short temporal filters before spatial aggregation, retaining the spatial context of each temporal change. This design targets weak, spatially localized traces of sleep-relevant motion over the clinically defined epoch before they are diluted by static appearance cues. Second, within-subject stage contrast learns stage evidence relative to the subject's own visual and motion baseline rather than as an absolute appearance template. Third, two-scale sleep dynamics modeling captures motion evolution within each 30-second epoch and organizes epoch-level evidence into a coherent full-night sleep-stage trajectory.

Evaluation is conducted on the Hospital A subset of the Korea Video Sleep Study (KVSS) [9], which contains 475 overnight NIR recordings, approximately 3,250 hours of video, synchronized with PSG labels. The recordings include far-field viewpoints, bedding occlusion, weak motion, and face anonymization, making them challenging for both proxy extraction and raw-video modeling. *ViNUSS* achieves 0.80 accuracy and 0.78 macro-F1 for four-class staging. Without any contact sensor, it performs on par with a proxy-first pipeline that receives contact-sensor breathing and cardiac waveforms, and supplying such waveforms to *ViNUSS* adds no measurable benefit, suggesting that the physiological evidence they carry is already captured from video. Ablations show complementary benefits from micro-motion representation, within-subject stage contrast, and full-night sleep dynamics. Sleep Video EigenCAM (SVECAM) further suggests that *ViNUSS* uses physiologically meaningful motion, including thoraco-abdominal periodicity and larger movements associated with arousals and position changes.

Our contributions are summarized as follows:

- **Video-only formulation under PSG-defined labels.** We study NIR video sleep staging where PSG provides the clinical labels but the model uses NIR video alone, without explicit physiological proxy reconstruction or EEG-guided representation learning.
- **Subject-relative micro-motion learning.** We introduce pre-spatial micro-motion encoding and within-subject stage contrast to preserve sleep-relevant temporal traces over the 30-second epoch while learning stage evidence relative to each subject's night-specific baseline.
- **Two-scale sleep dynamics.** We model within-epoch motion evolution and full-night temporal context, organizing video-derived epoch evidence into coherent overnight sleep-stage trajectories.
- **Large-scale evaluation and visual evidence analysis.** We evaluate *ViNUSS* on 475 overnight NIR recordings, with proxy-fusion, failure-mode, cross-site, and visual evidence analyses.

2 Related Work

2.1 Sleep Staging and Contactless Monitoring

The clinical reference for sleep staging is polysomnography (PSG), where trained technicians score multi-channel physiological signals such as EEG, EOG, and EMG according to the American Academy of Sleep Medicine (AASM) guidelines [2]. Although PSG provides high-fidelity labels, it requires contact sensors, overnight instrumentation, and expert scoring, which limits scalability for longitudinal monitoring [52]. Wearable devices based on actigraphy, photoplethysmography (PPG), or simplified EEG reduce this burden [7, 13, 34], but still require physical contact and can be affected by comfort, adherence, and motion artifacts. Fully contactless sensing, including audio [20], under-mattress pressure [47], and radio-frequency methods [19], avoids body-worn sensors but often captures only limited aspects of sleep behavior. Among contactless modalities, near-infrared (NIR) video is particularly relevant because it captures spatially localized body motion in darkness.

2.2 Video-Based and Cross-Modal Sleep Staging

Video-based sleep staging has been studied through handcrafted motion features and proxy-first pipelines. Early camera-based methods used frame-difference features with SVMs [33], while recent methods estimate physiological proxies before classification. For example, van Meulen *et al.* [45] used camera-based PPG and heart-rate variability, Video-PSG [50] extracted respiratory and cardiac signals, and SleepVST [6] modeled respiratory/cardiac waveforms and optical-flow features with pre-trained transformers. These proxy-first methods are valuable because they provide physiologically interpretable inputs. However, explicit proxy reconstruction can be unreliable in realistic NIR recordings: bedding and position changes can corrupt camera-based respiratory motion [51], while far-field views, occlusion, and face anonymization can weaken rPPG-based cardiac estimation [4, 16, 29]. Moreover, proxy compression can discard regional motion patterns and transient movements that are not captured by predefined physiological signals.

Recent work has also connected infrared video with EEG-derived knowledge through EEG-teacher distillation [53] or cross-modal retrieval between infrared videos and EEG representations [17]. These studies show that IR video contains sleep-relevant visual evidence when guided by physiological supervision. Our work is complementary: PSG defines the labels, but our model does not use EEG-teacher features, EEG retrieval databases, or explicit respiratory/cardiac proxy reconstruction. We instead study NIR video as an independently informative modality through subject-relative micro-motion learning and full-night video-conditioned dynamics.

2.3 Video Understanding and Full-Night Temporal Modeling

Modern video understanding has developed strong spatiotemporal representations for action recognition and semantic video learning, including efficient 3D CNNs, spatiotemporal transformers, and self-supervised video encoders [3, 11, 22, 27, 43]. Temporal modules further improve action recognition by exchanging information across frames, exciting motion-sensitive channels, or modeling temporal differences [25, 26, 49]. These methods motivate motion-aware video processing, but they primarily target salient action cues, whereas NIR sleep video requires preserving dense, low-amplitude temporal traces over a clinically defined 30-second epoch before spatial abstraction.

Sleep staging also requires temporal context beyond isolated epochs, and prior PSG/EEG-based neural models commonly exploit neighboring-epoch structure [24, 37, 38, 42]. For NIR video, the

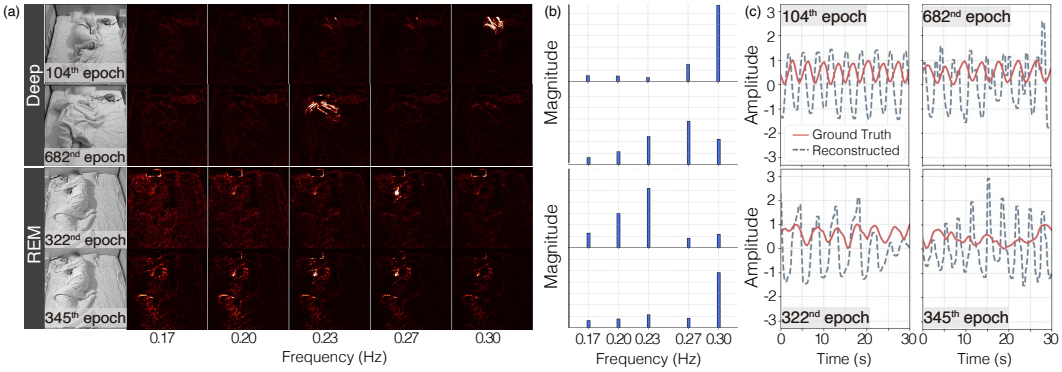


Figure 1: Frequency analysis and respiratory reconstruction from NIR sleep video. (a) *Video frequency maps* from pixel-wise FFT over 0.17–0.30 Hz; brighter regions indicate stronger periodic motion. (b) *Contact thoracic spectra* from chest-belt signals. (c) *Video-reconstructed respiratory waveforms*.

evidence is indirect and subject-dependent, making longer temporal context important for interpreting local motion. *ViNUSS* therefore summarizes motion within each 30-second epoch and then models the resulting representations across the subject's full-night trajectory using state-space sequence modeling [15]. This couples full-night temporal context with pre-spatial micro-motion encoding and within-subject stage contrast.

3 Method

Each overnight recording is divided into 30-second *epochs*, following clinical sleep-staging practice. Given a full-night NIR video sequence $\mathcal{V} = \{X_1, \ldots, X_N\}$ with N epochs, our goal is to predict a stage sequence $\hat{\mathcal{Y}} = \{\hat{y}_1, \ldots, \hat{y}_N\}$ for ground-truth labels $\mathcal{Y} = \{y_1, \ldots, y_N\}$, where $y_t \in \{\text{Wake}, \text{Light}, \text{Deep}, \text{REM}\}$. The labels are defined by PSG scoring, but the model input is NIR video only. We do not reconstruct respiratory or cardiac waveforms, nor use auxiliary physiological signals such as EEG as model inputs or for cross-modal supervision.

3.1 Design Motivation and *ViNUSS* Overview

Figure 1 illustrates why explicit proxy reconstruction is informative but incomplete in realistic NIR sleep videos. We reconstruct respiratory waveforms using video motion magnification and camera-based respiratory extraction methods [48, 51]. Figure 1(a) shows localized periodic responses around the thoraco-abdominal region, with different dominant frequencies across epochs, indicating respiration-related video evidence. However, panel (b) shows that video-dominant frequencies can deviate from contact thoracic spectra under position changes, occlusion, weak motion, or non-respiratory body movement, and panel (c) shows attenuation or frequency mismatch in waveforms. For example, the Deep 682nd and REM 322nd epochs show frequency mismatch, while the REM 345th epoch shows non-respiratory movement contamination. Thus, representing video with a single proxy may discard spatially localized evidence or introduce proxy-specific errors.

Naively training a generic video encoder on raw NIR clips is also insufficient. As shown in Figure 2(a), a baseline 3D CNN (MoViNet-A2 [22]) attends to static high-contrast structures such as bedding folds, sensor cables, or limb boundaries, even when frame differences indicate motion around the thoraco-abdominal region. Figure 2(b) shows that epoch-level embeddings cluster primarily by subject rather than by sleep stage, with substantial dispersion across subjects even within the same stage. This subject-dominant structure reflects differences in body shape, bedding, camera geometry, obstructive sleep apnea (OSA) severity, and baseline respiration-related motion. Moreover, motion patterns can vary across the night even within the same subject and stage, reinforcing the need to interpret local video evidence within a full-night trajectory.

Taken together, these analyses motivate learning directly from spatially localized video evidence rather than reducing each epoch to a reconstructed proxy or a generic raw-video embedding. Since single-epoch visual evidence remains indirect and sleep stages evolve across the night, *ViNUSS* preserves weak local micro-motion before spatial aggregation, learns stage evidence relative to each

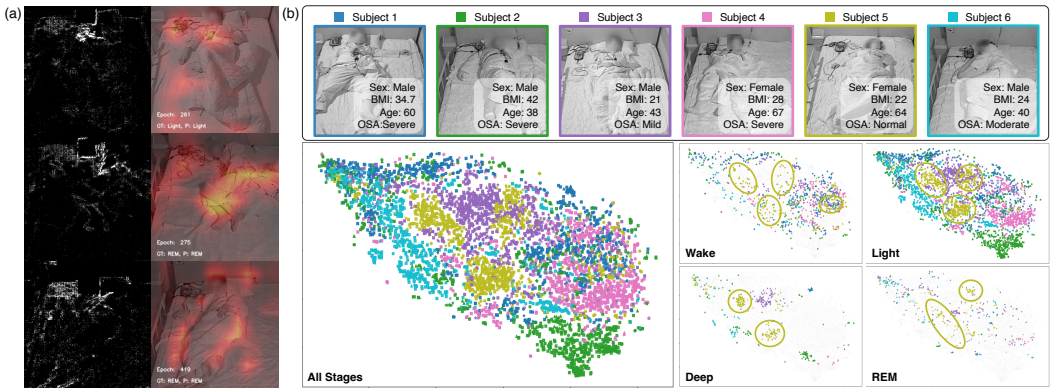


Figure 2: Baseline 3D CNN behavior on raw NIR sleep videos. (a) Frame-difference and spatial attention maps highlight static high-contrast regions rather than temporally coherent thoraco-abdominal motion. (b) t-SNE [44] projection of epoch-level embeddings from six subjects shows subject-dominant clustering and within-stage subject dispersion.

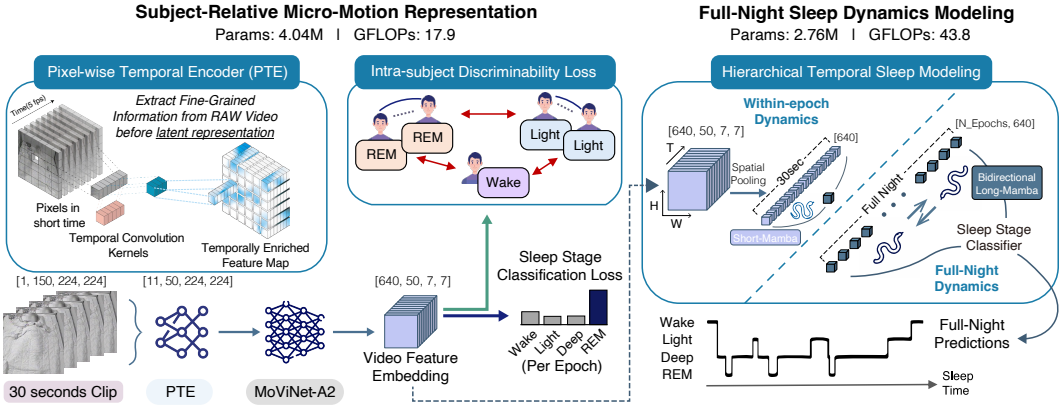


Figure 3: Overview of *ViNUSS*.

subject's overnight baseline, and models sleep dynamics over the full night. Figure 3 summarizes how *ViNUSS* implements these design goals.

3.2 Subject-Relative Micro-Motion Representation

We first describe how *ViNUSS* forms epoch-level video representations before full-night modeling. The representation module addresses the first two design goals by enriching each 30-second clip with pixel-wise temporal responses and using within-subject stage contrast to reduce subject-dominant structure during training.

Spatially anchored pre-spatial micro-motion encoding. Let $X_t \in \mathbb{R}^{C\times T\times H\times W}$ denote the video clip for epoch t, with channel, frame, height, and width dimensions. Standard video backbones may mix and downsample local regions before weak temporal evidence becomes separable. This is problematic for NIR sleep video because sleep-relevant motion can be highly localized and may be diluted unless temporal variation is first encoded at each pixel. To preserve dense pixel-level temporal traces over the epoch before spatial reduction, we introduce a Pixel-Wise Temporal Encoder (PTE):

$$\tilde{X}_t = \mathrm{Proj}\left(\mathrm{Concat}\left[X_t, M_3(X_t), M_5(X_t)\right]\right), \tag{1}$$

where $M_3(\cdot)$ and $M_5(\cdot)$ are learnable temporal encoding branches with kernel sizes 3 and 5, applied along the frame axis independently at each spatial location. $\mathrm{Concat}(\cdot)$ denotes channel-wise concatenation with the original clip, and $\mathrm{Proj}(\cdot)$ is a shallow temporal projection layer that reduces the frame dimension via temporal striding before the video backbone. PTE does not introduce a separate spatial stream; spatial mixing is deferred to the backbone. The detailed architecture is provided in Appendix B.

PTE applies short temporal filters over the full 30-second epoch while preserving the spatial location of each temporal response. Although each convolution uses a small kernel, stacking expands the effective receptive field to local sub-epoch intervals of roughly one to three seconds at 5 fps. Thus, PTE targets local motion traces rather than full-epoch or longer-range respiratory dynamics. It exposes displacement-like variation, short oscillatory changes, and transient movement onsets before they are mixed with static appearance. Unlike hand-crafted frame differences or late-fusion motion branches, PTE produces learnable, spatially localized micro-motion traces before backbone-level spatial mixing. The enriched clip is then encoded by a compact MoViNet-A2 backbone, producing a spatiotemporal feature map F_t.

Within-subject stage contrast. Even after motion-sensitive encoding, epoch features may reflect subject- and acquisition-specific factors rather than stage-related motion. We therefore contrast sleep stages within the same overnight recording. For subject s, an anchor and a positive epoch are sampled from the same stage, while a negative epoch is sampled from a different stage of the same subject. This within-subject sampling makes the comparison subject-relative, encouraging the representation to emphasize stage-related motion differences with respect to the subject's night-specific baseline.

We implement this objective as an Intra-Subject Discriminability Loss (ISDL). Let $(x^a_{s,i}, x^p_{s,i}, x^n_{s,i})$ denote the i-th anchor-positive-negative triplet of epoch features for subject s, and let $f(\cdot)$ be a projection head. Given N_{tri} sampled triplets and margin α, ISDL is

$$\mathcal{L}_{\text{ISDL}} = \frac{1}{N_{\text{tri}}} \sum_i \left[\|f(x^a_{s,i}) - f(x^p_{s,i})\|_2^2 - \|f(x^a_{s,i}) - f(x^n_{s,i})\|_2^2 + \alpha \right]_+ . \tag{2}$$

This follows a triplet-style margin objective [18], but with within-subject sampling. The representation objective combines the sleep-stage classification cross-entropy loss $\mathcal{L}_{\text{CE}}$ with ISDL:

$$\mathcal{L}_{\text{rep}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{ISDL}}, \tag{3}$$

where λ controls the weight of the contrastive regularizer.

The key role of ISDL is not the margin loss itself, but the sampling semantics. Unlike dataset-level contrastive objectives, ISDL forms comparisons within the same overnight recording, reducing variation from cross-subject and acquisition differences.

3.3 Full-Night Sleep Dynamics Modeling

Although the representation module produces motion-sensitive features for each epoch, sleep staging is inherently sequential. We introduce Sleep Dynamics Modeling (SDM) as a two-scale sequence module that first summarizes within-epoch motion evolution and then models the resulting epoch representations across the full night. This design allows temporal modeling to operate on compact video-derived motion summaries rather than frame-level features or post-hoc label sequences. We instantiate both temporal scales with Mamba [15]; implementation details are provided in Appendix B.

Within-epoch motion dynamics. The backbone output F_t retains spatial and temporal dimensions. We first pool the spatial dimensions while preserving the temporal axis:

$$R_t = \text{SpatialPool}(F_t). \tag{4}$$

ShortMamba then summarizes the temporal evolution within the 30-second epoch into an epoch-level motion representation:

$$z_t = \text{ShortMamba}(R_t). \tag{5}$$

This step condenses within-epoch temporal dynamics, such as respiration-related periodicity and transient body movement, before full-night sequence modeling.

Full-night sleep dynamics. The epoch-level motion representations are arranged chronologically as $\mathcal{Z} = \{z_1, \ldots, z_N\}$ and processed by a bidirectional LongMamba module:

$$(s_1, \ldots, s_N) = \text{LongMamba}(z_1, \ldots, z_N). \tag{6}$$

The sleep-stage probability for epoch t is then

$$p(y_t \mid \mathcal{V}) = \text{softmax}(\text{Classifier}(s_t)). \tag{7}$$

Table 1: Performance comparison of video-based sleep staging methods. Results for non-KVSS datasets are taken from their respective papers, while KVSS results are from our own implementation. *Oracle* indicates experiments using ground-truth physiological signals from contact sensors. SleepVST variants with the waveform encoder fine-tuned on KVSS are reported in Appendix C.4.

Method	Input / Proxy	Dataset	# of Data	Acc.	κ	MF1
van Meulen et al. [45]	HRV	HealthBed	46	0.68	0.49	0.63
Video-PSG [50]	HR+HRV+BR+M	Shenzhen	20	0.73	0.62	0.74
SleepVST [6]	BW+CW+M	Oxford	50	0.79	0.71	0.78
SleepVST [6]	BW*+CW*+M	KVSS	475	0.78	0.63	0.73
	BW+CW*+M			0.61	0.24	0.37
MoViNet-A2 [22]	Video only	KVSS	475	0.60	0.35	0.50
MoViNet-A2 [22] + PTE	Video only	KVSS	475	0.62	0.40	0.55
MoViNet-A5 [22] + PTE	Video only	KVSS	475	0.60	0.38	0.53
Video Swin-Tiny [27] + PTE	Video only	KVSS	475	0.51	0.27	0.46
ViNUSS	**Video only**	KVSS	475	**0.80**	**0.68**	**0.78**

Acc: accuracy; κ: Cohen's kappa; MF1: macro F1-score. BR: breathing rate; HR: heart rate; HRV: heart-rate variability; BW: breathing waveform; CW: cardiac waveform; M: motion feature. Asterisks (*) indicate waveforms from PSG sensors (oracle setting). Rows marked "+PTE" include only the Pixel-Wise Temporal Encoder, excluding ISDL and SDM.

SDM operates on video-derived epoch representations before classification, rather than smoothing predicted labels after classification. By relating local motion summaries to the subject's full-night trajectory, SDM helps calibrate subject-specific motion baselines, maintain coherent sleep progression, and reduce abrupt epoch-wise fluctuations.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on the Korea Video Sleep Study (KVSS) dataset [1, 9], using 475 overnight NIR sleep recordings ($\sim$3,250 hours) from Hospital A paired with synchronized PSG. Sleep stages were annotated by certified technicians according to AASM guidelines [2]. The cohort includes healthy and sleep-disordered individuals with diverse age, BMI, sex, and OSA severity. Videos were captured at 5 fps using a wall-mounted NIR camera under routine overnight sleep-study conditions, including far-field viewpoints, bedding occlusion, weak motion, and face anonymization. We follow the fixed subject-level 335/70/70 train/validation/test split of the dataset paper [9]. Since each subject contributes one night, the splits are subject-independent and temporally disjoint. Additional acquisition details, site characteristics, and the split protocol are provided in Appendix A.

Implementation. We implement *ViNUSS* in PyTorch and train on four NVIDIA RTX A6000 GPUs. The model is trained with five AASM sleep-stage labels (Wake, N1, N2, N3, REM), and N1/N2 are merged into Light sleep for the reported four-class evaluation. We report accuracy, Cohen's kappa (κ), and macro-F1 (MF1). MF1 is emphasized because sleep stages are imbalanced and minority stages such as Deep and REM are clinically important. Uncertainty is estimated by a subject-level bootstrap over test subjects (5,000 resamples); main-text tables report point estimates, and 95% confidence intervals (CIs) are given in the appendix (Appendix B.3). Architecture, preprocessing, optimization, and hyperparameter details are provided in Appendix B. Code, the preprocessing pipeline, and trained checkpoints are available at `https://github.com/alex-kunmin-jang/ViNUSS`.

4.2 Overall Performance

Table 1 compares *ViNUSS* with prior video-based sleep staging methods. Results from non-KVSS datasets are included for reference, as datasets differ in acquisition protocols, cohort size, camera setup, and label distribution. For controlled comparisons on KVSS, we include SleepVST [6], a representative proxy-first method, and video-only backbone variants.

Reproducing the fully video-based SleepVST pipeline on KVSS is not feasible because its video-to-waveform extraction module is not publicly released, and face anonymization and bedding occlusion make rPPG and respiratory reconstruction unreliable. We therefore evaluate two oracle-input variants: a partial-oracle setting with contact-sensor cardiac waveforms and a full-oracle setting with both contact-sensor respiratory and cardiac waveforms. These variants are oracle-input versions of one

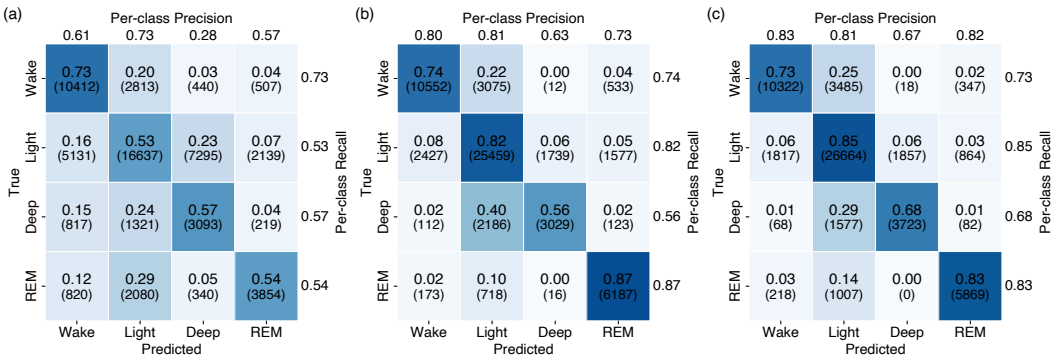


Figure 4: Confusion matrices on the KVSS test set for four-class sleep staging. (a) Without SDM. (b) SDM with LongMamba only. (c) Full SDM with ShortMamba and LongMamba.

representative proxy-first architecture, not universal upper bounds for physiological-signal-based sleep staging; details are provided in Appendix C.

Video-only backbone variants perform substantially below *ViNUSS*. PTE improves the vanilla MoViNet-A2 baseline, but neither the larger MoViNet-A5 nor Video Swin-Tiny closes the gap to the full *ViNUSS*. This indicates that pre-spatial micro-motion encoding or backbone scaling alone is insufficient without within-subject stage contrast and full-night dynamics, and MoViNet-A2 provides the best accuracy-efficiency tradeoff among the evaluated backbones (Appendix D.2).

On KVSS, *ViNUSS* achieves 0.80 accuracy, 0.68 κ, and 0.78 MF1 from face-anonymized video alone. This is on par with SleepVST with contact-sensor breathing and cardiac waveforms and a waveform encoder fine-tuned on KVSS (0.75 MF1), and well above it when breathing is reconstructed from video, the setting closest to contactless use (Appendix C.4). Adding these waveform proxies to the *ViNUSS* video representation provides no measurable incremental benefit (Appendix D.6), supporting NIR video as an independently informative modality for PSG-defined sleep-stage estimation. Stratified analyses across recording conditions and subject characteristics are in Appendix D.3.

4.3 Ablation Study

Effect of PTE and ISDL. Table 2 shows the contribution of each component, starting with the representation modules without SDM. The baseline (Exp. 1) achieves 0.60 accuracy and 0.50 MF1. Adding PTE (Exp. 2) improves accuracy to 0.62 and MF1 to 0.55, indicating that pre-spatial micro-motion encoding helps preserve weak local temporal cues before spatial processing. Adding ISDL (Exp. 3) further improves MF1 to 0.56, although accuracy remains comparable, suggesting that within-subject stage contrast improves stage-wise separability under subject-specific variation. Overall, PTE and ISDL improve representation quality from complementary perspectives: PTE enhances motion sensitivity, while ISDL regularizes the embedding space to reduce subject-dependent bias.

Table 2: Ablation study of key components on KVSS. Exp. 1 corresponds to the vanilla MoViNet-A2. ✓ indicates that each component is applied.

Exp.#	Method			Acc.	MF1
	PTE	ISDL	SDM		
1	–	–	–	0.60	0.50
2	✓	–	–	0.62	0.55
3	✓	✓	–	0.60	0.56
4	–	–	✓	0.75	0.67
5	✓	–	✓	0.77	0.72
6	✓	✓	✓	**0.80**	**0.78**

Effect of SDM. Adding SDM to the vanilla backbone (Exp. 4) improves MF1 from 0.50 to 0.67, showing the importance of full-night temporal context. However, the benefit of SDM depends strongly on the quality of the epoch-level representation. With the same SDM module, adding PTE (Exp. 5) increases MF1 from 0.67 to 0.72, and adding both PTE and ISDL (Exp. 6) further increases MF1 to 0.78. Thus, the representation and dynamics modules are complementary: PTE and ISDL make the epoch-level video evidence more informative and subject-relative, while SDM organizes this evidence into a coherent full-night sleep-stage trajectory.

Figure 4 shows how SDM changes class-wise prediction behavior beyond the aggregate scores in Table 2. Without SDM (a), predictions are broadly less balanced, with substantial confusion

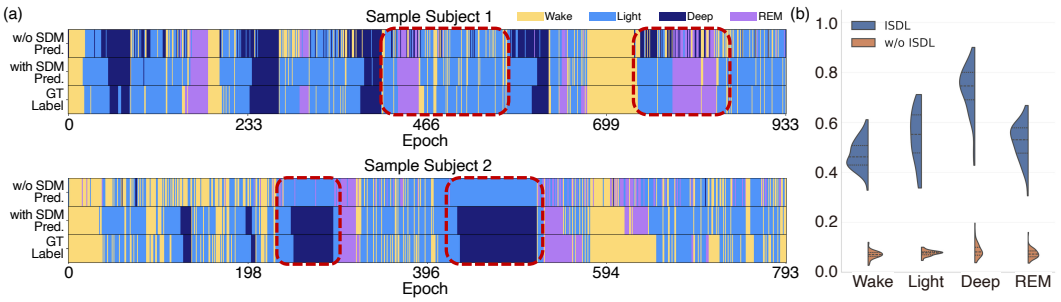


Figure 5: Full-night prediction behavior and temporal coherence. (a) Full-night predictions. Red boxes highlight representative segments. (b) Temporal coherence.

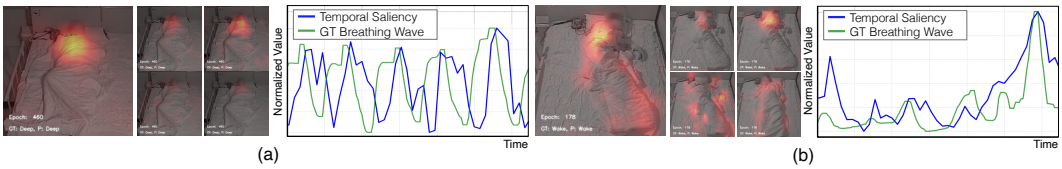


Figure 6: SVECAM visualizations for representative NIR sleep-video epochs. (a) Stable Deep sleep, showing thoraco-abdominal saliency and temporal saliency aligned with breathing. (b) Brief arousals and position change, showing both respiration-related motion and larger body movement.

between Light and neighboring stages and low precision for Deep sleep. Adding LongMamba only (b) improves the overall prediction structure by incorporating full-night temporal context, increasing recall for Light and REM and improving precision for Wake and Deep. This indicates that full-night context helps organize epoch-level predictions into a more consistent sleep-stage trajectory. The full SDM (c) further incorporates within-epoch motion summarization. Compared with LongMamba only, it improves Deep precision and recall from 0.63/0.56 to 0.67/0.68 and increases REM precision from 0.73 to 0.82, although REM recall slightly decreases from 0.87 to 0.83. Together, these results support the two-scale design, where LongMamba provides trajectory-level context and ShortMamba refines the epoch-level motion evidence used by the full-night model. Additional experiments on SDM design choices are provided in Appendix D.4, including temporal context-length sweeps and comparisons with alternative sequence models.

Full-night prediction behavior. Figure 5(a) illustrates how SDM affects complete overnight predictions. For Subject 1, predictions without SDM frequently oscillate between adjacent stages in the highlighted regions, whereas SDM suppresses abrupt fluctuations and preserves smoother stage trajectories. For Subject 2, predictions without SDM fail to recover sustained Deep sleep segments, while SDM accumulates temporally consistent evidence and restores continuous Deep predictions. These examples show that full-night dynamics modeling not only smooths noisy local predictions, but also helps recover sustained sleep structure when single-epoch visual evidence is weak.

Figure 5(b) further shows how ISDL interacts with full-night dynamics by improving the temporal coherence of epoch-level representations. We measure coherence as the cosine similarity between consecutive epoch embeddings, $s_t = \cos(\mathbf{f}_t, \mathbf{f}_{t+1})$. ISDL increases this coherence across stages, indicating that within-subject stage contrast produces smoother feature trajectories within each subject. These more stable epoch representations provide better inputs for SDM, explaining why combining PTE, ISDL, and SDM yields the strongest performance. Together with Table 2 and Figure 4, these results show that micro-motion representation, subject-relative contrast, and full-night temporal modeling provide complementary benefits for video-only sleep staging.

4.4 Interpretation and Failure Analysis

Visual evidence. To inspect the visual evidence used by *ViNUSS*, we apply Sleep Video EigenCAM (SVECAM), which adapts EigenCAM [31] to 3D sleep-video features. SVECAM produces class-agnostic spatial saliency maps and temporal saliency curves, showing where and when the model attends within each epoch (Appendix E). Figure 6 shows two cases. During stable Deep sleep (a), saliency concentrates around the thoraco-abdominal region, and temporal saliency is aligned with

respiratory rhythm despite the absence of explicit respiratory reconstruction. During arousals and position changes (b), the model still captures respiration-related motion but also attends to larger body movements. These visualizations suggest that *ViNUSS* uses physiologically meaningful motion evidence encoded in NIR video, without explicitly reconstructing predefined physiological proxies.

These patterns also help explain the difficulty of single-epoch video-only staging. Stable Deep sleep often depends on subtle periodic motion with little gross movement, whereas arousal-related epochs can produce broader and more transient saliency. Light sleep and REM can remain visually heterogeneous, often sharing low-motion or transition-heavy patterns with neighboring stages. This ambiguity supports the need for both subject-relative micro-motion representation and full-night sleep dynamics modeling.

Failure modes. About half of all errors lie within a few epochs of a stage transition, indicating that the model largely identifies the correct sleep stage but places the transition a few epochs early or late. Most of the remaining errors are Deep epochs misclassified as Light, concentrated during respiratory events and in subjects with irregular breathing during Deep sleep, where the periodic motion noted above is disrupted. REM, in contrast, is largely unaffected by respiratory events, as its assignment rests substantially on full-night context from SDM. Sleep position, limb movement, and OSA severity show no detectable effect on overall staging, and full blanket coverage is associated with only a small accuracy drop (Appendix D.8).

4.5 Cross-Site Transfer

We evaluate the Hospital A model on two held-out KVSS hospitals with different acquisition setups and patient populations (Appendix A). We compare zero-shot transfer, fine-tuning SDM (Section 3.3), and additionally fine-tuning the subject-relative micro-motion representation (Section 3.2), using site-specific subjects disjoint from the test subjects. Zero-shot performance drops at both sites (Table 3), but the errors reflect stage boundaries shifted toward a different stage at each site rather than absent evidence. Recalibrating SDM with only 30 subjects recovers much of this loss, and adapting the micro-motion representation brings accuracy at Hospital B to the Hospital A level (0.80), with a larger gain at Hospital C, whose camera angle differs. MF1 nevertheless remains below Hospital A in every setting, mainly in Deep sleep, which is about a third as prevalent at the external sites and leaves few Deep epochs for calibration. Detailed results are provided in Appendix D.7.

Table 3: Cross-site transfer of *ViNUSS* trained at Hospital A. SDM fine-tunes only SDM, and +Rep. also fine-tunes the micro-motion representation. n is the number of fine-tuning subjects at Hospital B/C.

Fine-tuned	**Hospital B**		**Hospital C**	
	Acc.	**MF1**	**Acc.**	**MF1**
None (zero-shot)	0.61	0.46	0.63	0.48
SDM (n=30)	0.76	0.62	0.71	0.61
SDM (n=75/76)	0.77	0.62	0.70	0.60
+Rep. (n=30)	0.79	0.63	0.71	0.60
+Rep. (n=75/76)	**0.80**	**0.64**	**0.75**	**0.65**
In-site (Hosp. A)	*0.80 / 0.78 (Acc. / MF1)*			

5 Conclusion

We presented *ViNUSS*, a framework for PSG-defined sleep staging from near-infrared (NIR) video that preserves localized temporal motion before spatial reduction, learns stage evidence relative to each subject's overnight baseline, and models full-night sleep dynamics. On 475 overnight NIR recordings, *ViNUSS* achieves strong four-class staging performance, with complementary benefits from its micro-motion representation, within-subject stage contrast, and sleep dynamics. Supplying contact-sensor or video-reconstructed waveform proxies provides no measurable incremental benefit, supporting NIR video as an independently informative modality for contactless sleep-stage estimation under PSG-defined labels. Sleep Video EigenCAM further suggests that the model uses physiologically meaningful visual motion, including thoraco-abdominal periodicity and broader body movements associated with arousals and position changes, and residual errors concentrate where this motion is disrupted, such as Deep sleep during respiratory events. Although modest site-specific fine-tuning recovers much of the zero-shot transfer loss at two held-out hospitals, MF1 remains below in-site performance, particularly for Deep sleep, reflecting single-site training and the sensitivity of PTE to camera geometry. Multi-site training, validation in home environments, and recordings above the 5 fps of KVSS remain important, and future work may explore integration with wearable modalities for more comprehensive sleep monitoring.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Research Foundation (NRF) of Korea grant funded by the Korean government (MSIT) (RS-2023-00212780, RS-2023-00222663, and RS-2025-25407378); the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2026-25507282 and RS-2025-25463302); grant No. 3720252500 from the SNUH Research Fund; and the data construction project for artificial intelligence through the National Information Agency of Korea (NIA), funded by the Ministry of Science and ICT. Hyun-Woo Shin is listed as an inventor on a patent application related to respiratory state monitoring and sleep disorder assessment (PCT/KR2021/005672; US application 17/930,569, filed by Seoul National University R&DB Foundation), which has been examined and is no longer being pursued. Hyun-Woo Shin is also a founder of OUaR LaB, Inc., serves on its Board of Directors and as Chief Executive Officer, and owns stock in OUaR LaB, Inc., which is subject to certain restrictions under university policy. The remaining authors declare no competing interests.

References

[1] AI Hub. Korea Video Sleep Study (KVSS) dataset. `https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=&aihubDataSe=data&dataSetSn=638`, 2021.

[2] Richard B. Berry, Rita Brooks, Charlene E. Gamaldo, Susan M. Harding, Carole L. Marcus, Bradley V. Vaughn, et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, Darien, IL, version 2.0 edition, 2012.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 813–824, 2021.

[4] Saksham Bhutani, Mohamed Elgendi, and Carlo Menon. Preserving privacy and video quality through remote physiological signal removal. *Communications Engineering*, 4(1):66, 2025.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[6] Jonathan F Carter, João Jorge, Oliver Gibson, and Lionel Tarassenko. SleepVST: Sleep staging from near-infrared video signals using pre-trained transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12479–12489, 2024.

[7] Amelia A. Casciola, Sebastiano K. Carlucci, Brianne A. Kent, Amanda M. Punch, Michael A. Muszynski, Daniel Zhou, Alireza Kazemi, Maryam S. Mirian, Jason Valerio, Martin J. McKeown, and Haakon B. Nygaard. A deep learning strategy for automatic sleep staging based on two-channel EEG headband data. *Sensors*, 21(10):3316, 2021.

[8] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, 06 2015.

[9] You Rim Choi, Dongik Park, Hyun Kyung Lee, Suahn Bae, Sehyun Park, Dong-Kyu Kim, Yong Min Kim, Chan-Soon Park, Hyung-Sin Kim, and Hyun-Woo Shin. Non-contact on-device detection of obstructive sleep apnea from infrared video. *Nature Communications*, 17(1):7163, 2026.

[10] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[11] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] Christopher A Field and Alan H Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(3):369–390, 2007.

[13] Pedro Fonseca, Marco Ross, Andreas Cerny, Peter Anderer, Fokke van Meulen, Hennie Janssen, Angelique Pijpers, Sylvie Dujardin, Pauline van Hirtum, Merel van Gilst, and Sebastiaan Overeem. A computationally efficient algorithm for wearable sleep staging in clinical populations. *Scientific Reports*, 13(1):9182, 2023.

[14] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[15] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[16] Divij Gupta and Ali Etemad. Privacy-preserving remote heart rate estimation from facial videos. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 706–712, 2023.

[17] Jianan Han, Aidong Men, Yang Liu, Ziming Yao, Shaoxing Zhang, Yan Yan, and Qingchao Chen. IoT-V2E: An uncertainty-aware cross-modal hashing retrieval between infrared-videos and EEGs for automated sleep state analysis. *IEEE Internet of Things Journal*, 11(3):4551–4569, 2024.

[18] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[19] Hong Hong, Li Zhang, Chen Gu, Yusheng Li, Guangxin Zhou, and Xiaohua Zhu. Noncontact sleep stage estimation using a CW Doppler radar. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8(2):260–270, 2018.

[20] Joonki Hong, Hai Hong Tran, Jinhwan Jung, Hyeryung Jang, Dongheon Lee, In-Young Yoon, Jung Kyung Hong, and Jeong-Whun Kim. End-to-end sleep staging using nocturnal sounds from microphone chips for mobile devices. *Nature and Science of Sleep*, 14:1187–1201, 2022.

[21] Yoo-Min Jung and Leekyung Kim. MambaSL: Exploring single-layer Mamba for time series classification. In *International Conference on Learning Representations (ICLR)*, 2026.

[22] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16020–16030, 2021.

[23] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. Fast optical flow using dense inverse search. In *European conference on computer vision*, pages 471–488. Springer, 2016.

[24] Hyojin Lee, You Rim Choi, Hyun Kyung Lee, Jaemin Jeong, Joopyo Hong, Hyun-Woo Shin, and Hyung-Sin Kim. Explainable vision transformer for automatic visual sleep staging on multimodal PSG signals. *npj Digital Medicine*, 8(1):55, 2025.

[25] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. TEA: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 909–918, 2020.

[26] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022.

[28] Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2024.

[29] Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55(9):1–40, 2023.

[30] Tomi Miettinen, Katja Myllymaa, Taina Hukkanen, Juha Töyräs, Kirsi Sipilä, and Sami Myllymaa. Home polysomnography reveals a first-night effect in patients with low sleep bruxism activity. *Journal of Clinical Sleep Medicine*, 14(8):1377–1386, 2018.

[31] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

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

[33] Teruaki Nochino, Yuko Ohno, Takafumi Kato, Masako Taniike, and Shima Okada. Sleep stage estimation method using a camera for home use. *Biomedical Engineering Letters*, 9(2):257–265, 2019.

[34] Keondo Park, Joopyo Hong, Wooseok Lee, Hyun-Woo Shin, and Hyung-Sin Kim. DistillSleep: real-time, on-device, interpretable sleep staging from single-channel electroencephalogram. *Sleep*, 48(12):zsaf240, 2025.

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

[36] Aakash K Patel, Vamsi Reddy, Karlie R Shumway, and John F Araujo. Physiology, sleep stages. In *StatPearls [Internet]*. StatPearls Publishing, 2024.

[37] Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2022.

[38] Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, 2022.

[39] 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, and Patricia W. Wahl. The Sleep Heart Health Study: Design, rationale, and methods. *Sleep*, 20(12):1077–1085, 12 1997.

[40] Rajeev Ratnavadivel, Nuy Chau, Daniel Stadler, Aeneas Yeo, R Doug McEvoy, and Peter G Catcheside. Marked reduction in obstructive sleep apnea severity in slow wave sleep. *Journal of Clinical Sleep Medicine*, 5(6):519–524, 2009.

[41] Jessica Vensel Rundo and Ralph Downey III. Polysomnography. *Handbook of Clinical Neurology*, 160:381–392, 2019.

[42] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

[43] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022.

[44] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[45] Fokke B van Meulen, Angela Grassi, Leonie van den Heuvel, Sebastiaan Overeem, Merel M van Gilst, Johannes P van Dijk, Henning Maass, Mark JH Van Gastel, and Pedro Fonseca. Contactless camera-based sleep staging: The healthbed study. *Bioengineering*, 10(1):109, 2023.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[47] Nikhil Vyas, Kelly Ryoo, Hosanna Tesfaye, Ruhan Yi, and Marjorie Skubic. Sleep stage classification using non-invasive bed sensing and deep learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3232–3238. IEEE, 2021.

[48] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics*, 32(4):1–10, 2013.

[49] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. TDN: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1895–1904, 2021.

[50] Qiongyan Wang, Hanrong Cheng, and Wenjin Wang. Video-psg: An intelligent contactless monitoring system for sleep staging. *IEEE Transactions on Biomedical Engineering*, 72(3):965–977, 2025.

[51] Wenjin Wang and Albertus C. den Brinker. Algorithmic insights of camera-based respiratory motion extraction. *Physiological Measurement*, 43(7):075004, 2022.

[52] Magdy Younes, Jill Raneri, and Patrick Hanly. Staging sleep in polysomnograms: analysis of inter-scorer variability. *Journal of Clinical Sleep Medicine*, 12(6):885–894, 2016.

[53] Zhen Yu, Shaoxing Zhang, Yang Liu, and Qingchao Chen. IR-based sleep monitoring: Movement, respiration, and sleep staging with a structure-aware cross-modal EEG knowledge distillation. *IEEE Internet of Things Journal*, 12(18):37482–37499, 2025.

Appendix

A KVSS Dataset

A.1 Dataset Composition

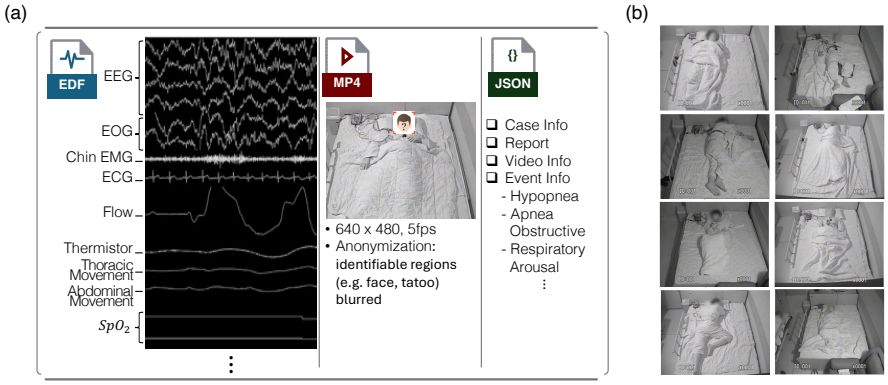


Figure A.1: Composition of the KVSS dataset. (a) Overnight PSG acquisition with synchronized multimodal signals, anonymized NIR video, and expert sleep-stage annotations. (b) Representative NIR frames (Hospital A) showing variation in subjects, body positions, bedding conditions, and room layouts.

Figure A.1 summarizes the data composition and acquisition pipeline of the Korea Video Sleep Study (KVSS) [1, 9]. As shown in Figure A.1(a), the Hospital A subset contains 475 overnight NIR sleep recordings (∼3,250 hours) paired with synchronized PSG and structured clinical metadata. PSG signals include EEG, EOG, EMG, ECG, airflow, thoracic and abdominal respiratory effort, and blood oxygen saturation (SpO_2), from which sleep stages were annotated by certified technicians following AASM guidelines [2]. NIR videos were captured at 640×480 resolution and 5 fps using a wall-mounted camera positioned approximately 3.5 m from the subject with an oblique view. Identifiable regions, including faces and tattoos, were blurred to protect privacy while preserving motion cues relevant to sleep analysis. The cameras are infrared surveillance-class devices already installed as clinical infrastructure, and Hospital A spans multiple rooms with cameras from two manufacturers [9]. The illumination wavelength is not documented in the installation records, but such cameras typically operate in the 850–940 nm near-infrared band.

The recordings were collected during routine overnight sleep studies with a fixed far-field viewpoint, where subjects were typically covered by bedding. Figure A.1(b) illustrates substantial variation in body position, bedding, occlusion, and room layout. Compared with closer-range or more controlled camera setups, these conditions attenuate visible motion, increase occlusion, and introduce inter-subject variability, making KVSS a challenging testbed for video-based sleep staging.

Table A.1: Acquisition setup at the three KVSS sites. Hospitals B and C are held out from model development.

Site	Role	Patients	Camera angle	Distance	Referring department
A	Training	475	45°	3.5 m	Otolaryngology
B	Held out	205	45°	3.0 m	Neurology†
C	Held out	256	30°	3.0 m	Otolaryngology

†The bed is oriented diagonally relative to the camera.

Beyond Hospital A, KVSS contains recordings from two further hospitals, which the dataset paper designates as external test sets [9]. All model development uses the Hospital A subset, and Hospitals B and C are held out for cross-site evaluation. As summarized in Table A.1, the three sites were not recorded under a shared protocol: viewing angle, camera distance, and bed orientation differ, which changes the apparent scale, direction, and image location of the same thoraco-abdominal displacement. The viewing angles in Table A.1 are nominal site values; mattress geometry in the held-out test videos

suggests that the effective angle also varies between rooms within Hospital A, by up to about 10°. The sites also draw on different referring departments and were scored by their own technicians. The resolution and frame rate, however, were standardized across all sites during dataset construction [9]. Representative frames from Hospitals B and C are shown in the dataset paper [9].

For all experiments, we use the fixed 335/70/70 training/validation/test split of Hospital A defined in the dataset paper [9] (stratified by sex and OSA severity), rather than K-fold cross-validation, so that results remain comparable to other work on the same cohort. The split is made at the subject level with no subject overlap across sets, and both training stages use the same partition. Since each subject contributes a single night, the sets are also temporally disjoint. Model selection uses only the validation subjects, and test subjects are used only for final evaluation.

A.2 Demographic and Clinical Characteristics

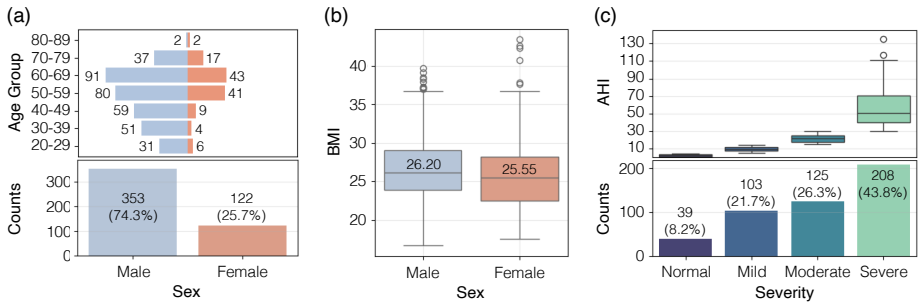


Figure A.2: Demographic and clinical characteristics of the Hospital A cohort. (a) Sex and age. (b) BMI by sex. (c) OSA severity and AHI.

Figure A.2 summarizes the demographic and clinical characteristics of the 475 subjects from Hospital A. The cohort spans a broad age range and includes 353 males (74.3%) and 122 females (25.7%), reflecting the composition of patients undergoing clinical sleep testing. BMI distributions show substantial inter-individual variability, with median BMI values of 26.20 for males and 25.55 for females. The cohort also covers a wide OSA severity spectrum, with 39 normal (8.2%), 103 mild (21.7%), 125 moderate (26.3%), and 208 severe cases (43.8%) according to the apnea–hypopnea index (AHI).

Together, Figures A.1 and A.2 show that the KVSS Hospital A cohort is diverse in demographics, body type, clinical severity, and video acquisition conditions. This diversity supports evaluation under heterogeneous sleep populations and realistic clinical recording conditions. The demographic and clinical characteristics of the held-out Hospitals B and C cohorts are reported in the dataset paper [9].

B *ViNUSS* Architecture and Experimental Details

B.1 *ViNUSS* Architecture

B.1.1 Subject-Relative Micro-Motion Representation

Figure B.1 illustrates the detailed architecture of the subject-relative micro-motion representation module. Each 30-second epoch, comprising 150 frames at 5 fps, is first processed by the Pixel-Wise Temporal Encoder (PTE). PTE consists of temporal-only convolution branches applied independently at each spatial location, without spatial mixing. The branches use stacked $k \times 1 \times 1$ temporal convolutions with short kernels, producing local temporal responses that are concatenated with the original video along the channel dimension. A strided $3 \times 1 \times 1$ temporal projection then reduces the temporal resolution before MoViNet-A2 encoding.

Although the individual kernels are short, stacking expands the effective temporal receptive field. Including the temporal projection, the two PTE branches cover local neighborhoods of up to 9 and 13 frames, corresponding to about 1.8 s and 2.6 s at 5 fps. Thus, PTE is designed to expose local sub-epoch motion traces before spatial mixing, while longer-range within-epoch dynamics

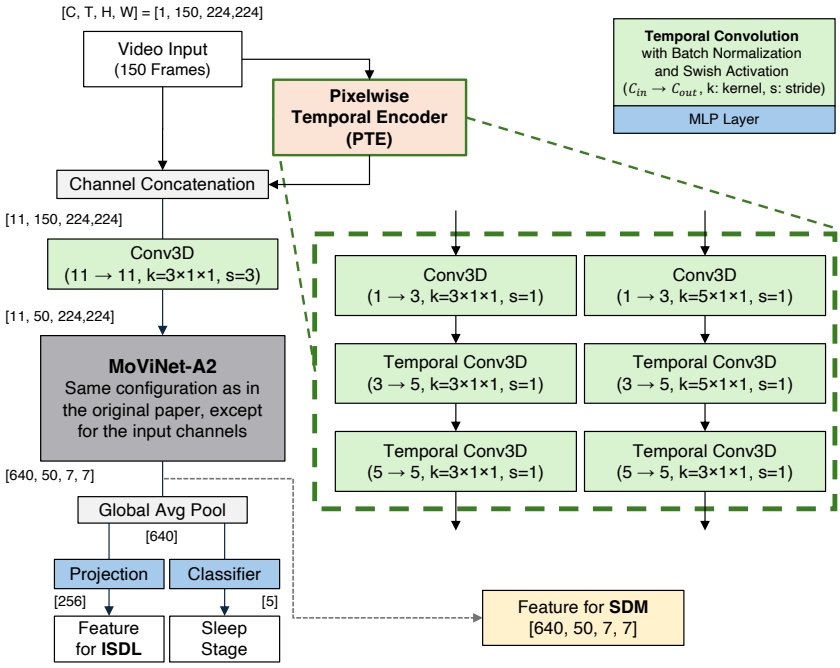


Figure B.1: Architecture of the subject-relative micro-motion representation module. PTE enriches each 30-second epoch with pixel-wise temporal responses, which are concatenated with the original video before strided temporal projection and MoViNet-A2 encoding. GAP features are used for epoch-level classification and ISDL, while pre-GAP feature maps are passed to SDM.

are handled by the sleep dynamics modeling module. All MoViNet-A2 settings follow the original specification [22], except that the first convolutional layer is modified to accommodate the increased number of input channels after concatenation.

The MoViNet-A2 backbone produces spatiotemporal feature maps. Global average pooling (GAP) yields epoch-level features used for sleep-stage classification and the Intra-Subject Discriminability Loss (ISDL). The pre-GAP feature maps are retained as input to the sleep dynamics modeling module.

B.1.2 Full-Night Sleep Dynamics Modeling

Figure B.2 illustrates the architecture of Sleep Dynamics Modeling (SDM), which operates on the pre-GAP spatiotemporal feature maps from the representation module. SDM has two temporal scales: ShortMamba summarizes motion evolution within each 30-second epoch, and bidirectional LongMamba models the resulting epoch tokens across the full night.

For each epoch, pre-GAP feature maps are spatially pooled into a temporal sequence with 50 temporal positions. After projection, a learnable summary token is appended to the sequence. ShortMamba models within-epoch temporal dynamics, and the output corresponding to the summary token is used as the compact epoch-level motion representation.

The epoch-level tokens are stacked chronologically across the full night. Each token is augmented with a learnable positional embedding, and the resulting sequence is processed by bidirectional LongMamba. The updated tokens are fed into a shared classifier to produce sleep-stage predictions for all epochs. This two-scale design captures within-epoch motion evolution and full-night sleep progression while avoiding post-hoc smoothing of predicted labels. The Mamba implementation enables linear-time modeling of the full-night epoch sequence.

B.2 *ViNUSS* Training Details

Video Preprocessing. All NIR sleep videos in KVSS were stored as MP4 files at 640×480 resolution and 5 fps. For each night, frames were cropped to a 400×400 region centered on the

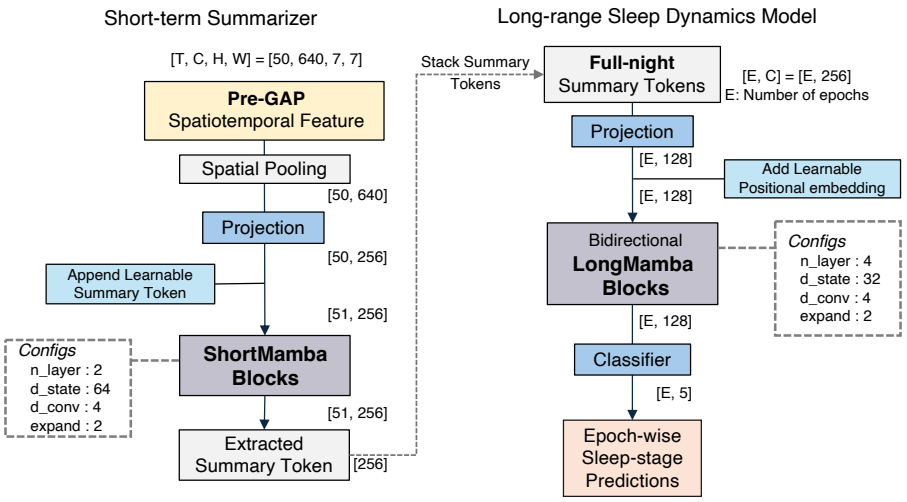


Figure B.2: Architecture of the sleep dynamics modeling module. ShortMamba summarizes each epoch by spatially pooling pre-GAP feature maps, appending a learnable summary token, and extracting the summary-token output. Bidirectional LongMamba models the stacked epoch tokens over the full night with positional embeddings and outputs stage predictions for all epochs.

bed to retain body motion while removing irrelevant background, and then resized to 224×224 to reduce computational cost. The continuous video stream was segmented into 30-second epochs aligned with sleep-stage annotations, so that each epoch contained 150 frames. For epochs with fewer than 150 frames, such as those at recording boundaries, we applied last-frame padding to enforce a fixed temporal length. For efficient I/O during training, all epochs from each subject were stored as fixed-size tensors in a single HDF5 file and accessed directly by the data loader.

Subject-Relative Micro-Motion Representation Training. The MoViNet-A2 backbone in the subject-relative micro-motion representation module is initialized from a publicly available Kinetics-600-pre-trained checkpoint [5, 22] and fine-tuned on KVSS. Training is performed on four NVIDIA RTX A6000 GPUs using weighted random sampling. Each optimizer epoch samples approximately 9% of the full training set. The effective batch size is 8 triplet samples, corresponding to 24 sleep epochs per optimization step because each sample contains an anchor, positive, and negative epoch for within-subject contrast. We use AdamW with a learning rate of 10^{-3}. The module is optimized with the sum of the epoch-level cross-entropy loss and ISDL. For ISDL triplet construction, we use class-balanced anchor sampling to reduce the effect of sleep-stage imbalance. For each sampled subject, an anchor stage is first sampled uniformly from the five AASM stages with valid candidate epochs for that subject. An anchor epoch is then sampled from the selected subject and stage. The positive epoch is sampled from the same subject and stage, and the negative epoch is sampled from the same subject but a different stage. Positive and negative epochs are selected uniformly from their respective candidate pools. We apply the following on-the-fly augmentations: normalization with dataset mean and standard deviation, random brightness jitter (± 0.2), random contrast jitter (± 0.2), horizontal flip with probability 0.5, random rotation up to $\pm 30^{\circ}$, and random scaling uniformly sampled from $[0.8, 1.1]$. The model is trained with five AASM stage labels, Wake, N1, N2, N3, and REM. For four-class evaluation, N1 and N2 are merged into Light sleep.

Full-Night Sleep Dynamics Modeling Training. After training the subject-relative micro-motion representation module, we extract and store the pre-GAP spatiotemporal feature maps for all epochs. These saved features are used as input to the sleep dynamics modeling module. SDM is trained on four NVIDIA RTX A6000 GPUs using subject-level sequences, where each data point corresponds to one full-night recording. To handle variable-length nights without padding, sequences are processed individually and gradient accumulation is used to obtain an effective batch size of 64 subject-level sequences. We use AdamW with a learning rate of 10^{-4}.

B.3 Evaluation Metrics

We evaluate four-class sleep staging performance using accuracy, macro-F1, and Cohen's κ. For completeness, we summarize the definitions used in the main paper.

Accuracy. Accuracy measures the proportion of correctly classified epochs:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i),$$

where N is the total number of evaluated epochs.

Macro-F1 (MF1). For each class c, precision, recall, and F1-score are computed as

$$\mathrm{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \qquad \mathrm{Recall}_c = \frac{TP_c}{TP_c + FN_c},$$
$$\mathrm{F1}_c = 2 \cdot \frac{\mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}.$$

Macro-F1 is the unweighted average across classes:

$$\mathrm{MF1} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{F1}_c,$$

where C is the number of sleep-stage classes. This metric assigns equal weight to each stage and is therefore useful under class imbalance.

Cohen's Kappa (κ). κ measures agreement between predictions and ground truth while accounting for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed agreement and p_e is the expected agreement under random labeling:

$$p_e = \sum_{c=1}^{C} \frac{n_{c,\mathrm{pred}} \cdot n_{c,\mathrm{true}}}{N^2}.$$

Here, $n_{c,\mathrm{pred}}$ and $n_{c,\mathrm{true}}$ denote the number of predicted and true labels for class c.

Per-subject accuracy for stratified analysis. For robustness analyses across recording conditions and subject characteristics, we compute accuracy separately for each subject:

$$\mathrm{Acc}_s = \frac{1}{N_s}\sum_{i=1}^{N_s} \mathbb{I}(\hat{y}_{s,i} = y_{s,i}),$$

where N_s is the number of epochs for subject s. For a subgroup $\mathcal{G}$, we report the mean per-subject accuracy:

$$\mathrm{Acc}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|}\sum_{s\in\mathcal{G}} \mathrm{Acc}_s.$$

This assigns equal weight to each subject within a subgroup, regardless of recording length.

Subject-level bootstrap confidence intervals. Because evaluation uses a single fixed split, we estimate uncertainty with a nonparametric bootstrap [10] in which subjects, rather than epochs, are the resampling unit. Epochs within one night are strongly correlated, so resampling at the subject level preserves within-subject dependence [12]. In each of 5,000 iterations, we draw test subjects with replacement, pool all epochs of the drawn subjects, and recompute each metric. We report 95% confidence intervals from the 2.5th and 97.5th percentiles of the resulting distribution. For a difference between two conditions, such as event versus event-free epochs or supine versus non-supine epochs, both quantities are computed on the same resample in each iteration, and the interval is taken over the resulting differences; this pairs the comparison within subjects whenever both conditions occur in the same recording. Correlations between subject-level quantities are bootstrapped in the same way. For cross-site evaluation, resampling is performed over the test subjects of each external site. We refer to a difference as reliable when its interval excludes zero. Main-text tables report point estimates, and intervals are given in the appendix where reported.

C SleepVST Re-implementation

Following the original SleepVST setup [6], we pre-train the model on the SHHS [39] and MESA [8] contact-sensor datasets for four-class sleep staging (Wake, N1/N2, N3, REM).

C.1 Datasets for SleepVST Training

Sleep Heart Health Study (SHHS) [39]. SHHS is a large multicenter cohort designed to study the cardiovascular and respiratory consequences of sleep-disordered breathing in adults. It includes overnight home polysomnography from more than 6,400 participants, together with demographic, anthropometric, and clinical questionnaires. PSG recordings include EEG, EOG, EMG, airflow, respiratory effort, and SpO_2, with sleep stages and respiratory events manually scored. SHHS is widely used for sleep staging and sleep-disordered breathing research, and access requires approval through the National Sleep Research Resource (NSRR).

Multi-Ethnic Study of Atherosclerosis (MESA) [8]. MESA is a longitudinal NHLBI cohort originally consisting of 6,814 Black, White, Hispanic, and Chinese-American adults aged 45–84. Between 2010 and 2012, 2,237 participants completed the MESA Sleep Exam, which included unattended overnight PSG, 7-day wrist actigraphy, and sleep questionnaires. The dataset provides PSG recordings with rich multi-ethnic demographic and cardiovascular phenotyping, and is available through approved data access requests.

C.2 Architecture and Pre-training

Electrocardiogram (ECG) and thoracic respiratory effort signals are extracted from each dataset and converted into cardiac waveforms (CW) and breathing waveforms (BW), respectively. After filtering, normalization, and 30-second patchification, the waveforms are encoded through convolutional layers and a Transformer encoder to produce per-epoch representations. The model is pre-trained on 240-epoch, 2-hour windows using four-class cross-entropy loss with the AdamW optimizer.

C.3 Transfer to KVSS

The KVSS dataset provides NIR video recordings with time-aligned contact-sensor signals. SleepVST requires reliable cardiac and respiratory waveforms, but direct recovery of these waveforms from KVSS videos is challenging. Facial regions are blurred for privacy, which limits recovery of subtle facial color changes needed for rPPG-based cardiac waveform estimation. We also attempted to extract respiratory waveforms from NIR video using two representative video-based respiratory estimation methods, phase-based motion magnification [48] and a camera-based respiratory extraction method [51]. On KVSS, the reconstructed waveforms showed substantial deviations in dominant frequency and amplitude compared with contact thoracic-belt signals, especially under bedding, position changes, weak motion, and non-respiratory body movement. Consequently, a faithful video-only implementation of the original SleepVST pipeline could not be realized under KVSS conditions.

To enable a controlled comparison on KVSS, we evaluated two oracle-input variants of SleepVST. In the partial-oracle variant, the cardiac waveform was replaced by the corresponding contact-sensor signal from PSG, while the breathing waveform was video-extracted using phase-based motion magnification [48]. In the full-oracle variant, both cardiac and breathing waveforms were obtained directly from contact sensors. These variants should be interpreted as oracle-input versions of one representative proxy-first architecture, not as universal upper bounds for physiological-signal-based sleep staging or practical video-only baselines.

In the main comparison (Table 1), the preprocessed waveforms are passed through the frozen SleepVST encoder and Transformer backbone to obtain per-epoch signal features. To incorporate motion information from video, optical flow is estimated using the DIS (Dense Inverse Search) algorithm [23], and the resulting motion feature vector is concatenated with the SleepVST signal feature. A Random Forest classifier is then trained on the combined features for four-class sleep staging on KVSS.

C.4 Fine-Tuning the Waveform Encoder on KVSS

In the variants above, the optical-flow motion features and the classifier are fitted on KVSS, as in the original SleepVST pipeline, but the waveform encoder is pre-trained on SHHS and MESA and never exposed to KVSS. To adapt every component of the baseline to the target cohort, we additionally fine-tune the waveform encoder on the same Hospital A training subjects as ViNUSS.

Here, the waveform encoder comprises the cardiac and breathing convolutional encoders and the Transformer described in Appendix C.2. Fine-tuning updates these modules and the classification head, starting from the pre-trained weights with no layers frozen. Training uses 240-epoch windows with a stride of 10 epochs, unweighted four-class cross-entropy, and AdamW (learning rate 10^{-6}, weight decay 0.05, batch size 8) for up to 30 epochs, with early stopping on validation loss (patience 3). We keep the checkpoint with the lowest validation loss. Each variant is fine-tuned on the waveforms it is evaluated with, since an encoder fine-tuned on contact-sensor waveforms does not transfer to the video-reconstructed breathing waveform.

After fine-tuning, we freeze the waveform encoder, extract its per-epoch features, concatenate them with the motion features, and fit a new Random Forest (300 trees, balanced subsample class weights) for each variant on the training subjects. The same subjects are used for fine-tuning and for fitting the Random Forest, without cross-fitting.

Table C.1: SleepVST on KVSS before and after fine-tuning the waveform encoder. All SleepVST rows include the optical-flow motion feature (M). BW* and CW* are contact-sensor waveforms, and BW is reconstructed from video. "Pre-trained" denotes the encoder pre-trained on SHHS and MESA (as in Table 1), and "Fine-tuned" denotes the same encoder further fine-tuned on KVSS. Intervals are subject-level bootstrap 95% CIs (Appendix B.3).

Input	Waveform encoder	Acc.	κ	MF1
BW*+CW*+M	Pre-trained	0.784 [0.763, 0.802]	0.635 [0.599, 0.667]	0.726 [0.698, 0.751]
BW*+CW*+M	Fine-tuned	0.799 [0.779, 0.817]	0.663 [0.628, 0.694]	0.748 [0.721, 0.772]
BW+CW*+M	Pre-trained	0.612 [0.592, 0.631]	0.236 [0.208, 0.266]	0.371 [0.344, 0.400]
BW+CW*+M	Fine-tuned	0.674 [0.648, 0.700]	0.422 [0.375, 0.470]	0.571 [0.533, 0.610]
Video only	*ViNUSS*	0.804 [0.784, 0.823]	0.684 [0.650, 0.713]	0.777 [0.754, 0.797]

Table C.1 shows that fine-tuning improves the baseline in both configurations, by 0.022 MF1 with contact-sensor waveforms and by 0.200 with the video-reconstructed breathing waveform. With contact-sensor breathing and cardiac waveforms, the fine-tuned baseline reaches 0.748 MF1 against 0.777 for video-only *ViNUSS*, and the intervals overlap substantially, so we do not claim a reliable advantage over this configuration. Rather, staging from face-anonymized video alone performs on par with a proxy-first pipeline that receives clean contact measurements of both breathing and cardiac activity and is adapted to our cohort. This configuration, however, requires contact sensors. In the configuration closest to contactless operation, with the breathing waveform reconstructed from video, the fine-tuned baseline reaches 0.571 MF1, and its intervals do not overlap with those of *ViNUSS* on any metric. Even this configuration remains optimistic, since the cardiac waveform still comes from a contact sensor.

D Additional Experiments and Analyses

D.1 *ViNUSS* Results

Table D.1 reports the detailed class-wise performance of *ViNUSS*. The model is trained with five AASM sleep-stage labels (Wake, N1, N2, N3, REM). For the four-class evaluation used in the main paper, both predictions and ground-truth labels are mapped by merging N1 and N2 into Light sleep and treating N3 as Deep sleep. Thus, the 5-class and 4-class results are obtained from the same trained model, with only the evaluation label mapping changed. For the four-class evaluation, we additionally report subject-level bootstrap 95% CIs (Appendix B.3). Intervals are narrow for the overall metrics, about 0.04 wide for accuracy and MF1, and widest for Deep, the sparsest stage.

Table D.1: Detailed class-wise performance of *ViNUSS*. (a) Five-class evaluation. (b) Four-class evaluation after merging N1/N2 into Light and treating N3 as Deep, with subject-level bootstrap 95% CIs (Appendix B.3) for F1, accuracy, and κ.

(a) 5-Class Classification

Class	Prec.	Rec.	F1
Wake	0.83	0.73	0.78
N1	0.58	0.57	0.58
N2	0.69	0.75	0.72
N3	0.67	0.68	0.67
REM	0.82	0.83	0.82
Macro Avg.	0.72	0.71	**0.72**
Weighted Avg.	0.72	0.71	0.72
Acc.		**0.71**	
κ		**0.62**	

(b) 4-Class Classification

Class	Prec.	Rec.	F1	95% CI
Wake	0.83	0.73	0.78	[0.74, 0.81]
Light	0.81	0.85	0.83	[0.81, 0.85]
Deep	0.67	0.68	0.67	[0.62, 0.72]
REM	0.82	0.83	0.82	[0.78, 0.86]
Macro Avg.	0.78	0.77	**0.78**	[0.75, 0.80]
Weighted Avg.	0.81	0.80	0.80	
Acc.		**0.80**		[0.78, 0.82]
κ		**0.68**		[0.65, 0.71]

In the four-class setting, Light denotes N1+N2 and Deep denotes N3.

D.2 Backbone Selection for Subject-Relative Micro-Motion Representation

Table D.2: Backbone selection for the subject-relative micro-motion representation module. All models are trained with pre-spatial micro-motion encoding (PTE) only, without within-subject stage contrast (ISDL) or full-night dynamics modeling (SDM), as in Table 1.

Model	**Param.**	**Acc.**	**MF1**
Video Swin-Tiny [27]	28.3M	0.51	0.46
MoViNet-A5 [22]	17.4M	0.60	0.53
MoViNet-A2 [22]	4.0M	**0.62**	**0.55**

To select the video backbone for the subject-relative micro-motion representation module, we evaluated three candidate architectures: Video Swin-Tiny [27], MoViNet-A5, and MoViNet-A2 [22]. All models in this comparison are trained with the same pre-spatial micro-motion encoding and without within-subject stage contrast or full-night dynamics modeling, so the results isolate the effect of the backbone architecture within the representation module. Table D.2 summarizes model size and classification performance.

Video Swin-Tiny performs worse than the MoViNet variants, suggesting that generic attention-based video representations do not directly solve NIR sleep staging. This may reflect the mismatch between action-oriented video representations and NIR sleep video, where discriminative evidence is weak, localized, and easily dominated by static subject or bedding cues. The convolutional inductive bias of MoViNet provides a more suitable local spatiotemporal prior for this setting.

Among the MoViNet variants, MoViNet-A2 provides the best accuracy-efficiency tradeoff. Although MoViNet-A5 has more than four times the parameters of MoViNet-A2, it yields lower accuracy and MF1. This suggests that increasing generic backbone capacity alone is insufficient for NIR sleep video, where the key challenge is preserving weak, localized motion evidence under strong subject- and acquisition-specific variability. We therefore select MoViNet-A2 as the backbone of *ViNUSS*, balancing performance, model size, and practical deployment constraints.

D.3 Stratified Performance across Recording Conditions and Subject Characteristics

To assess performance consistency under deployment-relevant variation, we conduct a stratified analysis on the held-out test set. Subjects are grouped by recording condition, including blanket occlusion, camera angle, room re-entry, and light intensity, and by subject characteristic, including sex, age group, BMI class, and AHI group. Recording-condition groups are assigned by the authors through visual review of each recording. For blanket occlusion, a recording is labeled *full* when the blanket covers the body throughout the night, *none* when it does not, and *partial* otherwise, a category that includes both partial body coverage and coverage that changes during the night. Figure D.1 reports per-subject accuracy for each subgroup. *ViNUSS* maintains comparable performance across the observed subgroups without clear systematic degradation, suggesting stable behavior across the

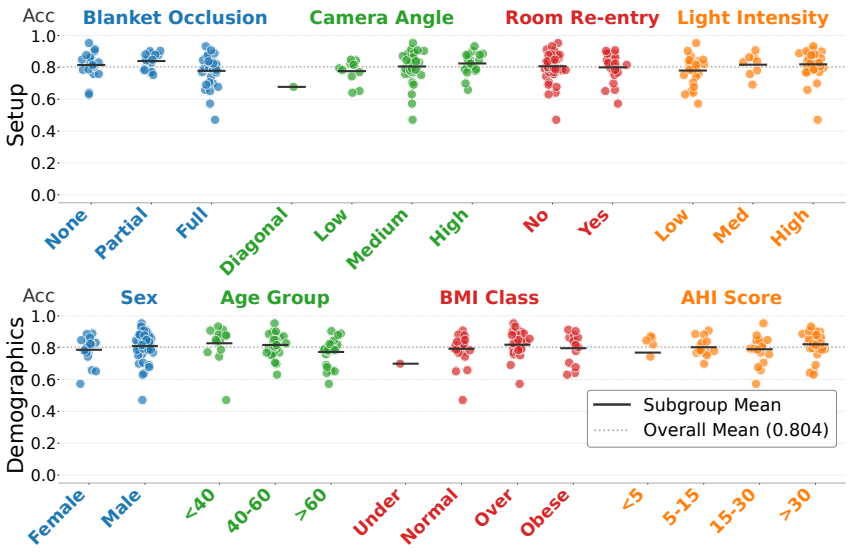


Figure D.1: Per-subject accuracy stratified by recording conditions and subject characteristics on the held-out test set. Each dot represents one subject. The upper row shows recording-condition groups, including blanket occlusion, camera angle, room re-entry, and light intensity. The lower row shows subject-characteristic groups, including sex, age, BMI class, and AHI group.

heterogeneous recording environments and patient characteristics represented in KVSS. Appendix D.8 complements this subject-level view with epoch-level analyses of where errors concentrate, including Acc, κ, and MF1 by blanket coverage, sleep position, and respiratory events.

D.4 Additional Experiments on Full-Night Sleep Dynamics Modeling

D.4.1 Comparison with Alternative Sequence Models

We compare the full-night sequence modeling component of *ViNUSS* with temporal backbones from four families: recurrent neural networks (RNNs), Transformer-based models, convolutional neural networks (CNNs), and state-space models (SSMs). In this experiment, each 30-second epoch is treated as one timestamp in the full-night sequence. All inter-epoch-only sequence models receive the same epoch-level representations, so their comparison isolates the effect of the inter-epoch temporal model. The "Mamba (Long only)" row corresponds to the bidirectional LongMamba path without the within-epoch ShortMamba summarizer, whereas "SDM (Short+Long)" reports our full two-scale design.

Table D.3 reports overall and per-class results. Mamba (Long only) achieves the strongest performance among inter-epoch-only sequence models, with 0.78 accuracy, 0.74 MF1, and 0.65 κ. BiLSTM is the closest alternative, while vanilla Transformer, PatchTST, and ModernTCN perform worse under the same epoch-level input setting. This suggests that full-night sleep staging benefits from sequence models that maintain and update compact temporal states over the overnight trajectory, and supports Mamba as an efficient state-space backbone for long-sequence modeling.

Per-class results show that Mamba (Long only) outperforms the other inter-epoch-only sequence models across all classes, with the clearest advantage on Deep sleep. The full SDM further improves every class over LongMamba alone, especially Deep sleep, where F1 increases from 0.59 to 0.67. These results support our design choice: LongMamba is effective for inter-epoch full-night modeling, while ShortMamba improves the epoch-level motion summaries passed to the full-night model.

Implementation details of compared sequence models. All compared models are trained to produce per-epoch five-class logits (Wake, N1, N2, N3, REM); reported four-class results are obtained by merging N1 and N2 into Light and treating N3 as Deep.

Table D.3: Comparison of sequence models for full-night sleep dynamics modeling. All models operate on epoch-level representations, where each 30-second epoch is treated as one timestamp. Best results are shown in **bold**, and second-best results are shown in *italics*.

Model	Family	Overall			Per-class F1			
		Acc.	MF1	κ	Wake	Light	Deep	REM
BiLSTM [14]	RNN	0.77	0.72	0.62	0.76	0.80	0.53	*0.80*
PatchTST [32]	Transformer	0.74	0.70	0.58	0.71	0.78	0.53	0.76
ModernTCN [28]	CNN	0.73	0.68	0.56	0.70	0.77	0.50	0.75
Vanilla Transformer [46]	Transformer	0.71	0.62	0.52	0.72	0.77	0.33	0.66
Mamba (Long only) [15]	SSM	*0.78*	*0.74*	*0.65*	*0.77*	*0.81*	*0.59*	*0.80*
SDM (Short+Long) [15]	SSM	**0.80**	**0.78**	**0.68**	**0.78**	**0.83**	**0.67**	**0.82**

RNN: recurrent neural network; CNN: convolutional temporal model; SSM: state-space model. Light denotes N1+N2 and Deep denotes N3.

BiLSTM. We use a two-layer bidirectional LSTM [14] with hidden size 256. Epoch representations are projected to 256 dimensions using LayerNorm and a linear layer, processed over the full-night sequence, and decoded by a two-layer MLP head at each epoch.

Vanilla Transformer. We use a vanilla Transformer [46] with three encoder layers, $d_{\text{model}} = 256$, 8 attention heads, $d_{\text{ff}} = 1024$, pre-normalization, GELU activations, and dropout 0.1. This baseline models the full-night epoch sequence with bidirectional self-attention and a per-epoch MLP classification head.

PatchTST. We adapt PatchTST [32] to epoch-level sequence classification. Patches are formed over consecutive epoch-level timestamps, not raw video frames, using patch_len $= 16$ and stride $= 8$. We use the channel-independent TSTiEncoder with $d_{\text{model}} = 128$, 16 heads, 3 layers, and $d_{\text{ff}} = 256$. Because PatchTST is originally designed for forecasting, we replace the forecasting head with a classification head that maps patch-level outputs back to epoch-level predictions by overlap averaging.

ModernTCN. We use a ModernTCN [28] encoder with ETTh1-style hyperparameters and attach a per-epoch classification head. The encoder uses RevIN, a stem layer, and one ModernTCN stage with large-kernel depthwise temporal convolution and grouped ConvFFN blocks. The output is compressed and decoded by a two-layer MLP head at each epoch.

Mamba (Long only). The Long-only Mamba baseline uses the full-night LongMamba path without the within-epoch ShortMamba summarizer. It consists of four bidirectional LongMamba blocks with $d_{\text{model}} = 128$ and $d_{\text{state}} = 32$, using epoch-index embeddings and a shared per-epoch classifier.

SDM (Short+Long). The full SDM uses both ShortMamba and LongMamba. ShortMamba first summarizes within-epoch motion evolution into compact epoch tokens, and LongMamba then models the stacked epoch tokens across the full night.

D.4.2 Effect of Temporal Context Length

To examine how temporal context length affects full-night sleep dynamics modeling, we train the Mamba (Long only) backbone with context windows of $W \in \{10, 30, 50, 100, 200, 500, \text{full}\}$, while keeping all other hyperparameters fixed. Each 30-second epoch is treated as one timestamp, and each training iteration samples $K = 5$ random length-W crops per subject. When a recording contains fewer than W epochs, zero padding is used with `ignore_index` masking. At inference, we use overlapping sliding windows with stride $W/10$ and aggregate predictions by averaging logits for epochs covered by multiple windows. This setup isolates the effect of temporal context length while keeping the sequence model and training recipe fixed.

Table D.4 shows that longer context consistently improves sleep dynamics modeling. Accuracy increases from 0.72 at $W = 10$ to 0.78 with full-night context, while MF1 improves from 0.64 to 0.74 and κ from 0.53 to 0.65. The largest gains occur as the window grows up to $W = 100$, after which accuracy begins to plateau. However, MF1 and κ continue to improve with longer context, suggesting that full-night information is especially helpful for less frequent or temporally structured stages such as Deep and REM.

Table D.4: Effect of temporal context length on full-night sleep dynamics modeling. Each 30-second epoch is treated as one timestamp.

Window size	Acc.	MF1	κ
10	0.72	0.64	0.53
30	0.74	0.67	0.58
50	0.76	0.71	0.61
100	0.77	0.72	0.63
200	0.77	0.73	0.63
500	0.78	0.74	0.64
full	**0.78**	**0.74**	**0.65**

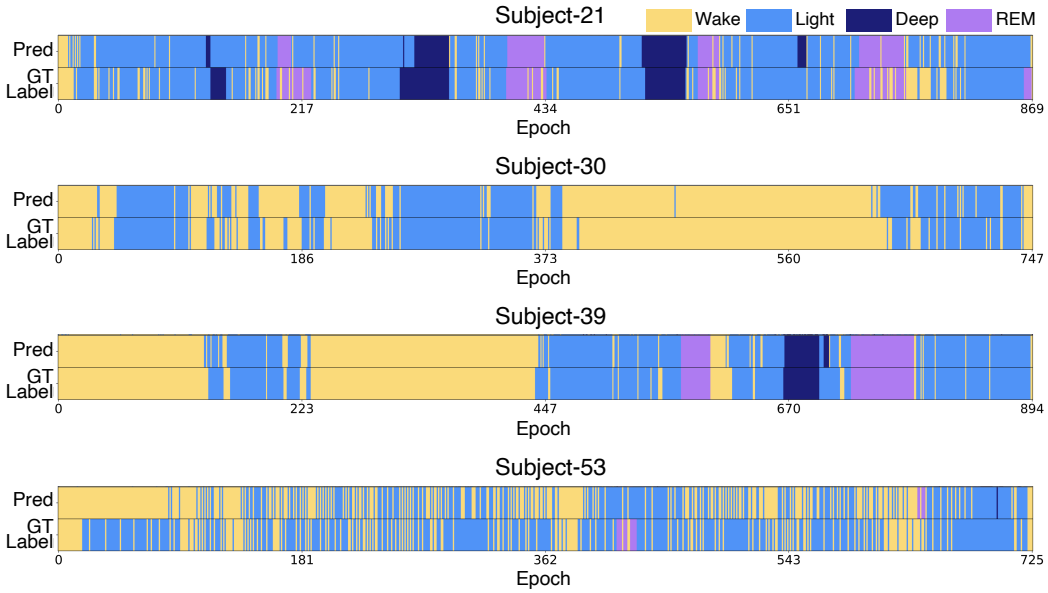


Figure D.2: Example full-night predictions on heterogeneous test subjects. Ground-truth (GT) labels and model predictions are shown for four representative subjects with different sleep-stage trajectories.

D.5 Example Full-Night Test Predictions

Figure D.2 provides additional full-night prediction examples complementing Figure 5. The selected subjects exhibit heterogeneous sleep architectures, ranging from multi-cycle sleep patterns to irregular profiles dominated by Wake and Light sleep. Subject 21 shows multiple stage transitions across the night, and the model follows the changing proportions of Light, Deep, and REM stages without collapsing to a fixed template. Subjects 30, 39, and 53 illustrate more atypical patterns, including recordings with little or no Deep/REM sleep, prolonged wakefulness, and frequent Wake–Light alternations. Across these examples, most prediction errors occur near stage boundaries or short transient segments, while longer sleep-stage structures are largely preserved, a pattern quantified in Appendix D.8. These cases further illustrate that full-night modeling helps *ViNUSS* track subject-specific overnight trajectories under diverse sleep patterns.

D.6 Physiological Proxies as Inputs

This section tests whether NIR video itself carries sleep-stage evidence, by comparing *ViNUSS* with models that stage from directly measured respiratory effort and by fusing waveform proxies into the *ViNUSS* video representation. Throughout, BW* and CW* denote contact-sensor breathing and cardiac waveforms, and BW denotes the breathing waveform reconstructed from video (Appendix C).

The split, labels, and evaluation protocol are identical to those of *ViNUSS*, and intervals are subject-level bootstrap 95% CIs (Appendix B.3).

D.6.1 Staging from Contact Respiratory Effort

Effort belts measure the thoraco-abdominal motion that *ViNUSS* attends to in video (Figure 6), without attenuation from camera distance, viewing angle, or bedding. We train two recent multivariate time-series models, ModernTCN [28] and MambaSL [21], on the thoracic and abdominal effort channels (BW*) of the KVSS training subjects, and pass their epoch representations to the same SDM used in *ViNUSS*. Unlike the inter-epoch comparison in Appendix D.4, ModernTCN here serves as the epoch-level signal encoder rather than the full-night sequence model.

Table D.5: Four-class staging from contact respiratory effort (thoracic and abdominal BW*) with the *ViNUSS* SDM, compared with video-only *ViNUSS*.

Method	Input	Acc.	κ	MF1
MambaSL + SDM	BW*	0.773 [0.747, 0.797]	0.628 [0.586, 0.666]	0.721 [0.688, 0.752]
ModernTCN + SDM	BW*	0.800 [0.778, 0.822]	0.673 [0.636, 0.708]	0.761 [0.734, 0.787]
ViNUSS	Video only	0.804 [0.784, 0.823]	0.684 [0.650, 0.713]	0.777 [0.754, 0.797]

As shown in Table D.5, video-only *ViNUSS* is at least as strong as both signal models on every metric, with overlapping intervals. This parity does not imply that the two modalities carry identical information. The signal models recover only 45–53% of true Deep epochs, with most of the remainder labeled Light, against 68% for *ViNUSS*. SVECAM also shows that *ViNUSS* attends to larger body movements as well as respiration-related motion (Figure 6(b)), which belt waveforms do not represent.

D.6.2 Proxy Fusion

We next feed waveform proxies to *ViNUSS* as additional inputs. A waveform encoder following SleepVST, fine-tuned on KVSS (Appendix C.4), produces an epoch-level waveform representation, which is concatenated with the *ViNUSS* video representation. The combined embedding then feeds the epoch classifier and SDM exactly as the video representation does in *ViNUSS*, so the pipeline is unchanged apart from the added waveform features, and all results are measured after SDM. Because face anonymization rules out rPPG-based cardiac recovery, the cardiac waveform CW* comes from the contact sensor in all conditions, while the breathing waveform is evaluated both as reconstructed from video (BW) and as a contact-sensor oracle (BW*).

Table D.6: Proxy fusion. Waveform features are concatenated with the *ViNUSS* video representation, with the rest of the pipeline unchanged.

Input	Acc.	κ	MF1
Video + BW + CW*	0.762 [0.739, 0.782]	0.614 [0.581, 0.644]	0.721 [0.697, 0.744]
Video + BW* + CW*	0.802 [0.781, 0.821]	0.683 [0.651, 0.711]	0.769 [0.743, 0.792]
Video only (*ViNUSS*)	0.804 [0.784, 0.823]	0.684 [0.650, 0.713]	0.777 [0.754, 0.797]

Table D.7: Per-stage F1 for the proxy-fusion configurations in Table D.6.

Input	Wake	Light	Deep	REM
Video + BW + CW*	0.738 [0.700, 0.773]	0.802 [0.781, 0.821]	0.594 [0.539, 0.645]	0.751 [0.703, 0.794]
Video + BW* + CW*	0.787 [0.756, 0.815]	0.831 [0.812, 0.850]	0.631 [0.573, 0.685]	0.829 [0.775, 0.869]
Video only (*ViNUSS*)	0.776 [0.739, 0.810]	0.834 [0.814, 0.853]	0.674 [0.622, 0.720]	0.823 [0.784, 0.856]

Neither fusion configuration improves on video-only *ViNUSS* (Table D.6). With the reconstructed breathing waveform, fusion is worse than video alone by 0.042 accuracy, 0.070 κ, and 0.056 MF1, with clearly separated intervals on κ and MF1 and marginally separated intervals on accuracy. This is consistent with Figure 1, where reconstructed respiration departs from the contact signal in dominant frequency and amplitude under bedding, position changes, and non-respiratory movement, so an input channel carrying such errors can hurt rather than help. Under the reconstructed waveform, all

four stages decline together (Table D.7), as expected from a degraded input channel rather than a stage-specific interaction.

Replacing BW with the contact-sensor waveform removes this error, and fusion then becomes indistinguishable from video alone rather than better than it. The differences of -0.002 accuracy, -0.001 κ, and -0.008 MF1 are an order of magnitude smaller than the interval widths, and per-stage intervals overlap throughout, with the largest difference on Deep, the sparsest stage with the widest interval. Even with clean contact measurements of both breathing and cardiac activity and an encoder fine-tuned on KVSS, the evaluated waveform proxies therefore provide no measurable incremental benefit to the video representation.

D.7 Cross-Site Transfer

Protocol. We evaluate the Hospital A model on Hospitals B and C (Table A.1). For Hospital C we follow the adaptation/test split defined in the dataset paper [9], and for Hospital B we use a matched design so that both sites receive comparable amounts of site-specific data. This yields 60 test subjects at Hospital B and 110 at Hospital C, with adaptation pools of 75 and 76 subjects, and we also evaluate a reduced setting with 30 adaptation subjects per site. Adaptation subjects are disjoint from test subjects, and all settings for a site are evaluated on the same test subjects. We compare three settings: (i) *zero-shot*, which applies the Hospital A model unchanged; (ii) *SDM only*, which fine-tunes the sleep dynamics modeling module with the representation module frozen, recalibrating full-night stage structure to the new site; and (iii) *both modules*, which additionally fine-tunes the subject-relative micro-motion representation (Section 3.2) together with SDM, so that the feature extractor adapts to the new camera geometry (denoted +Rep. in Table 3).

Table D.8: Cross-site performance of the Hospital A model. n is the number of test subjects at each site, and "Adapt." is the number of adaptation subjects, disjoint from test subjects. Intervals are subject-level bootstrap 95% CIs.

Site	Setting	Adapt.	Acc.	κ	MF1
B (n=60)	Zero-shot	0	0.606 [0.560, 0.650]	0.291 [0.238, 0.347]	0.463 [0.418, 0.506]
	SDM only	30	0.763 [0.737, 0.786]	0.545 [0.503, 0.583]	0.615 [0.576, 0.650]
	SDM only	75	0.774 [0.747, 0.801]	0.535 [0.489, 0.579]	0.624 [0.579, 0.661]
	Both modules	30	0.786 [0.759, 0.813]	0.549 [0.499, 0.598]	0.627 [0.574, 0.674]
	Both modules	75	0.800 [0.774, 0.824]	0.587 [0.542, 0.630]	0.637 [0.588, 0.679]
C (n=110)	Zero-shot	0	0.634 [0.616, 0.651]	0.319 [0.285, 0.352]	0.478 [0.442, 0.510]
	SDM only	30	0.712 [0.694, 0.731]	0.498 [0.466, 0.530]	0.612 [0.578, 0.644]
	SDM only	76	0.700 [0.683, 0.716]	0.494 [0.466, 0.521]	0.598 [0.569, 0.624]
	Both modules	30	0.714 [0.697, 0.729]	0.502 [0.476, 0.528]	0.598 [0.570, 0.623]
	Both modules	76	0.745 [0.730, 0.759]	0.566 [0.541, 0.590]	0.648 [0.616, 0.675]
A (n=70)	In-site	–	0.804 [0.784, 0.823]	0.684 [0.650, 0.713]	0.777 [0.754, 0.797]

Table D.9: Per-stage F1 across adaptation settings, using 75 (B) and 76 (C) adaptation subjects, with Hospital A for reference.

	Hospital B			Hospital C			Hospital A
Stage	Zero-shot	SDM only	Both	Zero-shot	SDM only	Both	In-site
Wake	0.385	0.629	0.663	0.498	0.688	0.712	0.776
Light	0.723	0.844	0.861	0.736	0.762	0.803	0.834
Deep	0.249	0.357	0.300	0.254	0.388	0.438	0.674
REM	0.494	0.665	0.724	0.426	0.556	0.638	0.823

Zero-shot transfer. Without adaptation, performance drops clearly at both sites (Table D.8), and each site collapses toward a different stage. At Hospital B, Light and REM epochs are absorbed into Wake, which is predicted about twice as often as it occurs, whereas at Hospital C about half of the true Wake epochs are labeled Light. This pattern indicates that the stage boundaries sit at the wrong operating point for the new site, rather than that epoch-level evidence is absent from the recordings.

Adaptation. Recalibrating SDM alone, without changing the representation module, raises MF1 to 0.624 at Hospital B and 0.612 at Hospital C, and reverses the zero-shot errors: the Light-to-Wake rate at Hospital B falls from 23.7% to 5.0%, and Wake recall at Hospital C rises from 42.5% to 70.4%. Adapting the representation module as well raises MF1 to 0.637 and 0.648. The additional gain from adapting the extractor is larger at Hospital C (+0.036 MF1 with 76 subjects) than at Hospital B (+0.013), consistent with Hospital C being the site whose camera angle differs from Hospital A. With 30 adaptation subjects, adapting both modules already reaches 0.786 accuracy at Hospital B, close to the 75-subject result, while at Hospital C additional data continue to help. At Hospital C, SDM-only adaptation with 30 subjects slightly exceeds that with 76 subjects, within overlapping intervals.

Remaining gap. MF1 and κ remain below Hospital A in every setting. Per-stage F1 (Table D.9) shows that Wake, Light, and REM move substantially toward Hospital A after adaptation, with Light exceeding it at Hospital B and approaching it at Hospital C, whereas Deep remains far below Hospital A at both sites (0.300 and 0.438 against 0.674). Because MF1 weights all stages equally, Deep accounts for most of the remaining MF1 gap. Deep is also the stage whose prevalence differs most across sites, covering 9.4% of epochs at Hospital A against 3.0% at Hospital B and 3.6% at Hospital C, which leaves the adaptation sets with roughly a third as many Deep epochs per subject. The remaining Deep errors go predominantly to Light at both sites, the same Deep-to-Light confusion that dominates residual errors at Hospital A (Appendix D.8). These results suggest that much of the zero-shot loss reflects calibration to the acquisition setup and local stage distribution, which responds to a modest amount of site-specific data. A clear gap nevertheless remains, particularly for Deep sleep. Training used a single site, and the external cohorts differ from it in both patient population and recording setup, so training on more sites and a wider range of populations and acquisition conditions, as well as stronger domain adaptation than the deliberately simple settings used here, are important directions.

D.8 Failure-Mode Analysis

We characterize the errors of *ViNUSS* on the Hospital A test set from two perspectives: what kinds of errors the model makes, and under which recording and physiological conditions they concentrate. All analyses use the frozen final checkpoint, and intervals are subject-level bootstrap 95% CIs over the 70 test subjects (Appendix B.3). Unless otherwise stated, metrics are computed by pooling epochs within each group.

D.8.1 Error Types

About half of the errors are boundary-timing offsets rather than stage confusions. When a prediction is counted as correct if it matches the label within ± 1 epoch, accuracy rises from 0.804 to 0.859 and MF1 from 0.777 to 0.828, and with a tolerance of ± 3 epochs to 0.898 and 0.869, recovering 48% of all errors. The model therefore identifies the sleep stage of each region but often places stage transitions a few epochs early or late. Among the remaining errors, confusions between adjacent sleep stages dominate, with Light↔Deep the most frequent pair.

D.8.2 Recording Conditions

Table D.10: Performance by blanket coverage (subject-level labels, Appendix D.3) and by sleep position (epoch-level PSG annotation). Fifteen epochs with missing position annotation are excluded.

Condition	Group	n	Acc.	κ	MF1
Blanket coverage	None	18 subj.	0.817 [0.778, 0.854]	0.692 [0.624, 0.754]	0.770 [0.717, 0.811]
	Partial	19 subj.	0.840 [0.820, 0.860]	0.735 [0.706, 0.765]	0.801 [0.771, 0.828]
	Full	33 subj.	0.777 [0.743, 0.810]	0.649 [0.595, 0.698]	0.762 [0.727, 0.795]
Position	Supine	37,807 ep.	0.802 [0.777, 0.825]	0.666 [0.626, 0.703]	0.752 [0.725, 0.777]
	Non-supine	20,096 ep.	0.809 [0.786, 0.832]	0.709 [0.674, 0.742]	0.800 [0.774, 0.823]

Blanket coverage. Performance is not ordered by the extent of coverage, since partial coverage is highest on all three metrics (Table D.10). Contrasting fully covered recordings with the rest gives Δ accuracy -0.052 [-0.091, -0.014], $\Delta\kappa$ -0.067 [-0.129, -0.009], and ΔMF1 -0.026 [-0.069,

+0.016]. Heavy occlusion is thus associated with a small drop in accuracy and κ. Coverage is a coarse subject-level label rather than a per-epoch annotation of visible body area, so this compares 33 subjects against 37, and subject-level variation within each coverage group is larger than the difference between groups.

Sleep position. Position is annotated as supine or non-supine, with lateral and prone positions grouped together. Accuracy differs by only 0.007 between positions with heavily overlapping intervals, whereas MF1 differs by 0.048, indicating that the difference lies in the minority stages. Since position is annotated per epoch, we compute the paired within-subject change per stage (Table D.11), reporting recall alongside F1 because the two positions have different stage compositions, to which F1 is sensitive and recall is not. Deep and REM are the only stages with a change reliably above zero, Deep on F1 and REM on recall. Respiratory events are more frequent in the supine position and themselves degrade staging (below), but restricting to event-free epochs, REM recall remains 0.764 supine against 0.897 non-supine, so the difference is not driven by events. Supine is also the majority position at 65% of epochs, so the overall results already come mostly from the less favorable position.

Table D.11: Per-stage change from supine to non-supine epochs, with paired subject-cluster bootstrap 95% CIs. Deep is present in both positions for 19 of 70 subjects, which widens its recall interval.

Stage	Δ Recall	Δ F1
Wake	-0.018 [-0.074, +0.034]	+0.010 [-0.037, +0.056]
Light	+0.011 [-0.029, +0.054]	-0.009 [-0.040, +0.022]
Deep	+0.046 [-0.082, +0.173]	+0.147 [+0.049, +0.246]
REM	+0.099 [+0.020, +0.182]	+0.043 [-0.025, +0.114]

D.8.3 Breathing Regularity

The contact breathing signal in KVSS is normalized per subject, with RMS close to 1 for every subject, so baseline respiratory amplitude cannot be compared across subjects. A shallow signal scaled to unit RMS, however, may carry relatively more noise and less well-defined periodicity than a deep one, so an amplitude difference can reappear as a difference in waveform regularity. We therefore measure regularity per epoch by peak-to-peak swing (P2P), spectral energy in the 0.10–0.50 Hz band (band-power), and the share of total power in that band (peak-sharpness); an irregular signal shows high P2P with low band-power and peak-sharpness.

Averaged over each subject's whole night, none of these features predicts overall MF1 (Spearman ρ: P2P -0.21 [-0.44, +0.05], band-power +0.17 [-0.09, +0.41], peak-sharpness +0.16 [-0.10, +0.39]). Because a whole-night average would dilute an effect confined to one stage, we recompute the features over each subject's Deep epochs and relate them to how often those epochs are labeled Light, using the 60 subjects with at least one true Deep epoch. All three features then point the same way: the Deep-to-Light rate correlates positively with P2P (ρ = +0.39 [+0.11, +0.62]) and negatively with band-power (-0.53 [-0.72, -0.28]) and peak-sharpness (-0.45 [-0.66, -0.19]). The mean Deep-to-Light rate is 45.7% in the lowest quartile of band-power against 11.3% in the highest, and the correlations hold when restricted to the 53 subjects with at least 30 Deep epochs. The association is specific to this confusion: it does not appear for Deep-to-Wake or Deep-to-REM, and the same features do not predict Wake, Light, or REM performance. The model thus misreads Deep as Light more often in subjects whose breathing during Deep sleep departs from the regular slow pattern typical of the stage.

D.8.4 Respiratory Events and OSA Severity

We compare epochs containing a scored apnea or hypopnea (event epochs) with those that do not. Event epochs account for 20,121 of the 57,918 test epochs, and their prevalence scales with severity, from 3.9% in Normal to 58.8% in Severe subjects.

During events, accuracy rises while MF1 falls (Table D.12), and both follow from the difference in stage composition, since event epochs are dominated by Light. We therefore decompose by stage using recall, which is computed over the true epochs of each stage and is unaffected by that difference (Table D.13). The effect is directional rather than a uniform loss. Deep recall falls while Light recall rises, the signature of Deep epochs being relabeled as Light. A respiratory event alters

Table D.12: Performance on event-free epochs and epochs containing a scored apnea or hypopnea.

Epochs	n	Acc.	κ	MF1
Event-free	37,797	0.784 [0.756, 0.808]	0.676 [0.634, 0.713]	0.778 [0.750, 0.802]
Event	20,121	0.843 [0.822, 0.861]	0.651 [0.609, 0.686]	0.645 [0.604, 0.684]

Table D.13: Per-stage recall on event-free and event epochs. Δ is event minus event-free. The first column gives the share of each stage's epochs that contain an event.

Stage	With event	Event-free	Event	Δ
Wake	14.6%	0.755	0.574	-0.181 [-0.236, -0.124]
Light	46.2%	0.818	0.897	+0.079 [+0.042, +0.124]
Deep	5.1%	0.703	0.314	-0.389 [-0.553, -0.186]
REM	47.4%	0.836	0.818	-0.018 [-0.088, +0.054]

thoraco-abdominal motion within the epoch, by reducing effort or producing paradoxical effort followed by a recovery breath, and the errors move along the same Deep-to-Light axis identified above. Its reach is bounded by physiology: respiratory events are markedly reduced in slow-wave sleep [40], and only 5.1% of Deep epochs contain one.

Wake recall also falls during events, but this does not indicate that arousal-related motion is mistaken for wakefulness. If it were, the model would predict Wake more liberally during events, yet Wake is predicted for 4.4% of event-free sleep epochs and 5.4% of event sleep epochs. The drop is instead a boundary effect: respiratory events are scored during sleep, while an epoch is scored Wake when wake occupies most of its 30 seconds, so an epoch that is both spans a transition, where errors concentrate.

REM shows no detectable effect despite the highest event exposure of any stage. This is consistent with how REM is identified: REM recall is 0.54 from the epoch-level classifier and 0.83 after SDM (Figure 4), so REM assignment rests substantially on full-night context, much as AASM scoring continues REM across epochs lacking rapid eye movements [2]. The SDM lift for REM is essentially unchanged within event epochs (0.54 to 0.82) and event-free epochs (0.55 to 0.84), whereas Deep gains far less from context (0.57 to 0.68), consistent with Deep being where events hurt most.

Table D.14: Four-class MF1 by AHI severity and epoch type. The Normal event subset contains no true Deep epochs, so four-class MF1 is not defined there.

Severity	Subjects	Event-free	Event	Δ
Normal	6	0.746	–	–
Mild	15	0.797	0.723	-0.074 [-0.219, -0.004]
Moderate	19	0.762	0.666	-0.095 [-0.168, -0.032]
Severe	30	0.775	0.610	-0.165 [-0.208, -0.115]
Overall	70	0.778	0.645	-0.133 [-0.177, -0.085]

Because these are pooled comparisons, subjects with more events contribute more to the event subset, so we also stratify by AHI severity (Table D.14). Event-free MF1 spans only 0.051 across the four severity groups and is not ordered by severity, so OSA severity itself does not shift average staging performance, consistent with the subject-level view in Appendix D.3. The event-related drop is present in Mild, Moderate, and Severe subjects with intervals excluding zero, so it is not confined to severe disease. The point estimates grow with severity, but the intervals overlap and we do not treat this ordering as established.

D.8.5 Limb Movement

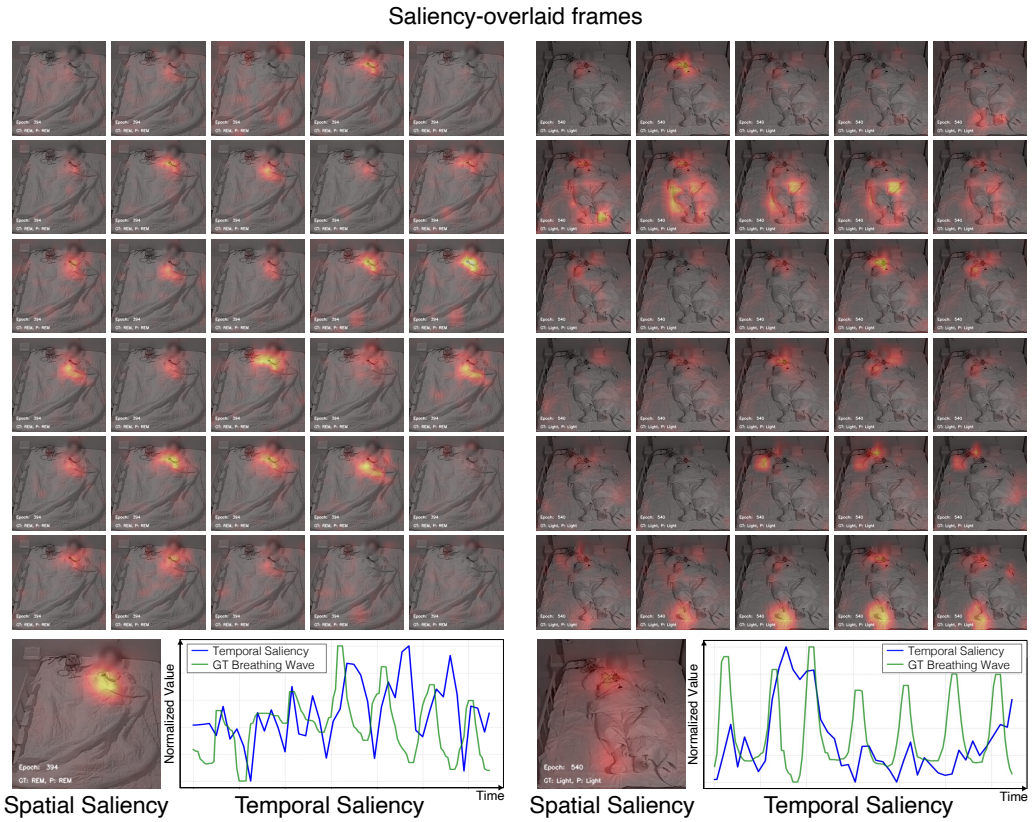
Limb movement is a different class of abnormal motion that could plausibly be confused with arousal-related motion. The 21 of 70 test subjects with a periodic limb movement index of 5 or higher reach 0.766 MF1 [0.731, 0.798] against 0.781 [0.752, 0.806] for the remaining subjects, a difference of -0.015 [-0.059, +0.028], with accuracy and κ likewise indistinguishable. These subjects are on

average 11 years older, so the point estimate cannot be attributed cleanly to limb movement, but we find no detectable degradation.

D.8.6 Summary

Beyond boundary-timing offsets, residual errors concentrate on a single axis, Deep read as Light, and occur where Deep is least distinctive: in subjects whose breathing during Deep sleep lacks the regular slow pattern typical of the stage, and in epochs containing a respiratory event. Conditions commonly expected to be difficult either leave overall staging intact or act through this one boundary. Sleep position and limb movement do not change overall staging, respiratory events act through Deep rather than across all stages, OSA severity does not lower event-free performance, and heavy occlusion is associated with a drop smaller than the subject-level spread within any coverage group.

E Sleep Video EigenCAM (SVECAM)



(a) Subject-64 Epoch-394 (b) Subject-53 Epoch-540

Figure E.1: SVECAM visualizations for two representative NIR sleep-video epochs. For each epoch, we show spatial saliency maps, temporal saliency curves, and saliency-overlaid frames. Frames are sampled at 1 Hz over the 30-second epoch and arranged sequentially from left to right, wrapping every five seconds.

NIR sleep videos provide weak and indirect visual evidence for sleep-stage discrimination, making it important to inspect which spatiotemporal regions the model uses. We therefore extend EigenCAM [31] to 3D activation tensors from sleep-video models.

Given an activation tensor $\mathbf{A} \in \mathbb{R}^{C\times T\times H\times W}$, a direct EigenCAM application would flatten temporal and spatial dimensions before singular value decomposition (SVD), which can mix motion-related and appearance-related structure. We instead use Sleep Video EigenCAM (SVECAM), which derives temporal and spatial channel weights separately and then combines them.

- **Temporal weighting.** For each spatial location (h, w), we apply SVD to $\mathbf{A}_{h,w} \in \mathbb{R}^{C\times T}$ and use the leading left singular vector $\mathbf{u}_T^{(h,w)} \in \mathbb{R}^C$ as a channel-weight vector capturing temporal co-variation at that location.
- **Spatial weighting.** For each frame t, we apply SVD to $\mathbf{A}_t \in \mathbb{R}^{C\times(H\cdot W)}$ and use the leading left singular vector $\mathbf{u}_S^{(t)} \in \mathbb{R}^C$ as a channel-weight vector capturing spatially coherent activation at that time.

We form a spatiotemporal weight tensor $\mathbf{W} \in \mathbb{R}^{C\times T\times H\times W}$ by broadcasting and summing the normalized temporal and spatial channel weights:

$$\mathbf{W}(c, t, h, w) = \tilde{\mathbf{u}}_T^{(h,w)}(c) + \tilde{\mathbf{u}}_S^{(t)}(c),$$

where $\tilde{\mathbf{u}}$ denotes an ℓ_2-normalized vector. The final SVECAM saliency is computed as a channel-weighted activation map followed by ReLU:

$$\mathrm{SVECAM}(t, h, w) = \mathrm{ReLU}\left(\sum_c \mathbf{A}_{c,t,h,w}\mathbf{W}_{c,t,h,w}\right).$$

This formulation highlights activation channels that are temporally informative at each spatial location and spatially coherent within each frame. Because NIR sleep videos are captured from a fixed camera and the body layout is relatively stable within many 30-second epochs, preserving this spatiotemporal structure is useful for interpreting localized sleep-related motion.

SVECAM is applied to the subject-relative micro-motion representation module. Figure E.1 presents visualizations for representative 30-second epochs from different subjects. In example (a), saliency is concentrated around the thoraco-abdominal region, and the temporal saliency curve follows respiration-related periodic motion. This suggests that the model uses breathing-related visual motion without explicitly reconstructing a respiratory waveform.

In example (b), the epoch contains transient body movements in addition to respiration-related motion. SVECAM highlights larger movements when they occur, while saliency returns to the thoraco-abdominal region during quieter intervals. These examples suggest that *ViNUSS* uses both localized respiration-related motion and transient behavioral motion as visual evidence for sleep staging.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include theoretical results, theorems, or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide full architectural and training details in Section 3 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code, the preprocessing pipeline, and trained checkpoints are publicly released at `https://github.com/alex-kunmin-jang/ViNUSS`. We do not redistribute the KVSS data due to patient privacy, and the dataset is available through the controlled-access procedures of the AI Hub portal, after which our released configurations reproduce the reported results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide details of the experimental setup in Section 4.1 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report subject-level bootstrap 95% confidence intervals (5,000 resamples over test subjects) for the experiments supporting the main claims, including the overall performance of *ViNUSS*, the SleepVST comparison, the proxy-fusion ablation, cross-site transfer, and the failure-mode analysis (Appendices B.3, C.4, and D.1, D.6–D.8). These intervals capture variability over the test cohort rather than over training runs, since each configuration is trained once on 335 full-night recordings. Component ablations and architecture comparisons (Table 2, Appendices D.2 and D.4) report point estimates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Training hardware is reported in Section 4.1 and Appendix B. The model parameter counts are stated in Figure 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms in every respect to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impacts (scalable contactless sleep monitoring) are discussed in Section 1 and Section 5. Privacy considerations from nighttime video capture are addressed through the dataset's face anonymization design described in Appendix A.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All data, models, and code used in the paper are properly cited and respect their respective licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the training and evaluation code, the preprocessing pipeline, and trained checkpoints at `https://github.com/alex-kunmin-jang/ViNUSS`, with documentation of the environment, data preparation, training, and evaluation, together with the license and usage terms. No new dataset is released, and the checkpoints are trained on KVSS, which remains subject to the AI Hub access terms.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects beyond the use of a pre-existing clinical dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve new human subject recruitment. The KVSS dataset was collected under institutional oversight at the participating hospitals, with identifiable regions blurred to protect participant privacy, as described in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This paper does not involve the use of large language models in research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.