StageGuard: Physiologically Constrained Sleep Staging

Juntang Wang
Duke Kunshan University
Kunshan, China

Yihan Wang
Duke Kunshan University
Kunshan, China

Hao Wu
Sichuan University
Chengdu, China

Jiayu Gao
Duke Kunshan University
Kunshan, China

Shixin Xu
Duke Kunshan University
Kunshan, China

Dongmian Zou
Duke Kunshan University
Kunshan, China
dz95@duke.edu

Abstract

Inferring sleep states from multimodal physiological time-series is fundamental to sleep research and clinical diagnostics. Deep learning models achieve high epoch-level accuracy but frequently produce physiologically implausible outputs: state transitions that are rare in healthy subjects (e.g., direct Wake→REM) and excessively fragmented hypnograms with unrealistically short bouts. Such outputs are unsuitable for downstream sleep architecture analyses regardless of accuracy. We propose *StageGuard*, a **plug-and-play**, **backbone-agnostic** framework that augments *any* neural sleep-staging backbone with semi-Markov structured inference. The key innovation is a unified wrapper combining: (1) a differentiable soft transition penalty that discourages physiologically rare transitions during training, and (2) a semi-Markov constrained decoder with duration-augmented state space that jointly enforces transition constraints and minimum bout durations at inference. Unlike approaches using hard $-\infty$ constraints that prohibit rare transitions entirely, our soft-penalty formulation allows rare transitions when emission evidence is overwhelming—a more scientifically accurate model of sleep physiology. We evaluate StageGuard on six established backbones across four datasets spanning distinct modalities, demonstrating that it reduces the transition-violation rate (TVR) to physiologically plausible levels (from 3.6–12.3% to 0.1–0.8%) and the fragmentation index by 56–62% while maintaining or slightly improving classification accuracy. Crucially, improved constraint satisfaction translates to 59–79% lower error in derived sleep architecture statistics (REM latency, bout durations, total sleep time), validating that physiologically valid hypnograms yield more reliable downstream analyses.

1 Introduction

Automated sleep staging—the classification of physiological recordings into discrete vigilance states—is a cornerstone of modern sleep research (e.g., classic R&K-style standards [26, 39]). Sleep can be assessed from diverse signal modalities: polysomnographic EEG/EMG [1], wrist-worn actigraphy [28, 35], cardiorespiratory recordings [7, 20], and contactless radar [33]. Manual scoring by trained experts remains the gold standard, but it is time-consuming, subjective [3], and poorly suited to large-scale studies [11]. Deep learning methods have achieved expert-level accuracy on standard benchmarks [21, 22, 32], leading many laboratories to adopt them for routine scoring.

However, accuracy alone is an incomplete measure of a staging model's fitness for scientific use. A classifier that achieves high epoch-level accuracy—approaching the inter-rater agreement ceiling of ~80–85% [3, 11]—may nonetheless produce outputs that violate fundamental physiological invariants: atypical state transitions that occur almost exclusively in pathological conditions (e.g., direct Wake → REM in narcolepsy), single-epoch bouts that no sleep researcher would accept, and fragmentation patterns inconsistent with the known temporal organization of sleep [9, 17]. Such outputs are not merely inaccurate—they are *invalid* for downstream analyses including REM latency estimation, sleep efficiency calculation, and bout-duration statistics. We formalize this validity gap through two constraint-satisfaction indicators: the *transition-violation rate* (TVR), measuring the proportion of physiologically rare transitions, and the *fragmentation index* (FI), measuring excessive state switching. Our goal is not to optimize these metrics competitively but to ensure they are satisfied at physiologically plausible levels, enabling trustworthy downstream analysis.

The core problem is that standard neural network classifiers treat each epoch independently or with only weak sequential coupling. Even recurrent and attention-based architectures [5, 22] learn temporal dependencies from data rather than encoding physiological invariants explicitly, leaving them free to produce rare transitions when the learned statistics are ambiguous. Post-hoc smoothing heuristics (e.g., majority filtering [31]) reduce noise but cannot guarantee that outputs respect known physiological constraints.

As sleep-staging models enter routine use in large-scale studies, outputs that achieve high accuracy but violate physiological invariants can silently corrupt downstream analyses. Trustworthy deployment requires that outputs satisfy domain-specific constraints, not merely optimize classification metrics [27]. To address this gap, we propose *StageGuard*, a backbone-agnostic constraint framework that wraps any neural sleep-staging backbone with physiology-informed mechanisms. Our contributions are:

(1) We formalize *transition-violation rate* (TVR) and *fragmentation index* (FI) as validity indicators that should be reported alongside accuracy.
(2) We introduce a unified semi-Markov framework combining soft training penalties with duration-augmented constrained decoding—backbone-agnostic, using soft penalties (not hard constraints), and implementing principled semi-Markov inference.
(3) We demonstrate across six backbones and four datasets spanning distinct modalities that TVR is reduced to physiologically plausible levels while accuracy is maintained or improved, yielding 59–79% lower error in derived sleep architecture statistics.

2 Background and Related Work

2.1 Sleep Physiology and Staging

Mammalian sleep is organized into cyclically alternating vigilance states: wakefulness (Wake), non-rapid-eye-movement sleep (NREM), and rapid-eye-movement sleep (REM) [26, 39]. Transitions between these states follow well-characterized physiological patterns. The dominant transition pathway is Wake → NREM → REM → Wake; direct REM → NREM transitions are *rare* under normal conditions. Direct Wake → REM transitions—termed sleep-onset REM periods (SOREMPs)—rarely occur outside pathological states such as narcolepsy or conditions of extreme sleep deprivation [9]. We emphasize that these transitions are not strictly *forbidden*—they are physiologically *rare* in healthy subjects. This distinction is important: a staging model should strongly discourage rare transitions but allow them when emission evidence is overwhelming, rather than prohibiting them entirely.

Beyond transition structure, sleep states exhibit characteristic *temporal organization*: bouts of each state have minimum expected durations that reflect the underlying neurobiology of state-switching circuits [18, 29]. Sleep fragmentation—excessively short or frequently interrupted bouts—is itself a clinically meaningful outcome associated with cognitive impairment and daytime dysfunction [17, 38]. A sleep-staging algorithm that produces unrealistically fragmented output is therefore not only inaccurate but potentially misleading for downstream analyses.

2.2 Deep Learning for Sleep Staging

Deep learning has achieved expert-level accuracy on sleep staging across modalities. For polysomnographic EEG, architectures range from CNNs [1, 30, 34] to CNN–RNN hybrids [22, 32], attention-based models [5, 23], and U-Net architectures [21]; see Fiorillo et al. [6] for a comprehensive review. Actigraphy-based sleep–wake classification [28, 36] and cardiorespiratory staging [7] extend these methods to wearable and home-monitoring contexts. Contactless radar-based methods [33] extend sleep staging to fully touch-free settings by analyzing respiratory and movement patterns from reflected radio waves.

Despite high accuracy, these deep learning methods do not explicitly encode physiological transition constraints. Sequential models learn soft temporal dependencies but provide no mechanism to ensure outputs respect known physiological invariants; rare transitions and excessive fragmentation can bias downstream sleep architecture statistics.

Recent work on foundation models for EEG [13] demonstrates the potential of large-scale pre-training across diverse EEG tasks. Such models learn rich representations but do not explicitly encode physiological constraints. Our framework is orthogonal and complementary: StageGuard can wrap any backbone, including fine-tuned foundation models, to ensure outputs respect domain-specific invariants regardless of representation quality. To incorporate such domain knowledge, we review relevant work on constraints in machine learning.

2.3 Constraints in Machine Learning

Incorporating domain knowledge as constraints in machine learning dates to early work on graphical models and structured prediction. In structured prediction, conditional random fields [16] and hidden Markov models encode transition structure via learned or fixed transition matrices [15]. The Viterbi algorithm [8] finds the most likely state sequence under such models and can incorporate transition constraints via the log-transition matrix.

A key limitation of standard HMMs is their assumption of geometric (memoryless) state durations. Hidden semi-Markov models (HSMMs) [40] generalize HMMs by explicitly modeling state duration distributions, enabling minimum dwell-time constraints that better capture the temporal structure of physiological processes. Our minimum-duration constraint mechanism approximates semi-Markov behavior within a computationally efficient Viterbi framework.

In natural language processing, constrained decoding has been a powerful technique for incorporating lexical and structural constraints into neural sequence generation. Grid beam search [12] and dynamic beam allocation [24] enable hard lexical constraints during decoding while preserving model fluency. Our constrained Viterbi decoder applies an analogous principle—enforcing domain-specific structural constraints on neural network outputs—to the sequential classification setting.

Dong et al. [4] used a mixed neural network with CRF-like post-processing, and hidden Markov models have been applied to sleep state sequences [19]. However, neither approach provides a backbone-agnostic wrapper with semi-Markov inference. Our work is conceptually related to physics-informed machine learning [14, 37], though we encode discrete physiological rules rather than governing equations.

Summary and gap. No existing framework combines soft training penalties with principled semi-Markov inference into a unified, backbone-agnostic wrapper for physiological time-series classification. Our work addresses this gap by integrating complementary constraints—soft transition penalties during training, duration-augmented semi-Markov decoding at inference—that can augment any neural sleep-staging backbone without architectural modification.

3 Method

We consider sleep staging as a structured sequence prediction problem. Given an epoch sequence of inputs $\mathbf{x} = (x_1, \ldots, x_T)$ (with modality-specific representations), a backbone model produces per-epoch posterior distributions over sleep states: $p_\theta(y_t \mid x_t)$, where $y_t \in \mathcal{S}$. Our goal is to output a hypnogram $\mathbf{y} = (y_1, \ldots, y_T)$ that is not only accurate but also physiologically plausible. In particular, we target two common failure modes of modern deep sleep-staging systems: (i) *transition violations* and (ii) *temporal fragmentation*. StageGuard is a plug-and-play, backbone-agnostic inference layer that turns any per-epoch classifier into a physiologically consistent scientific readout.

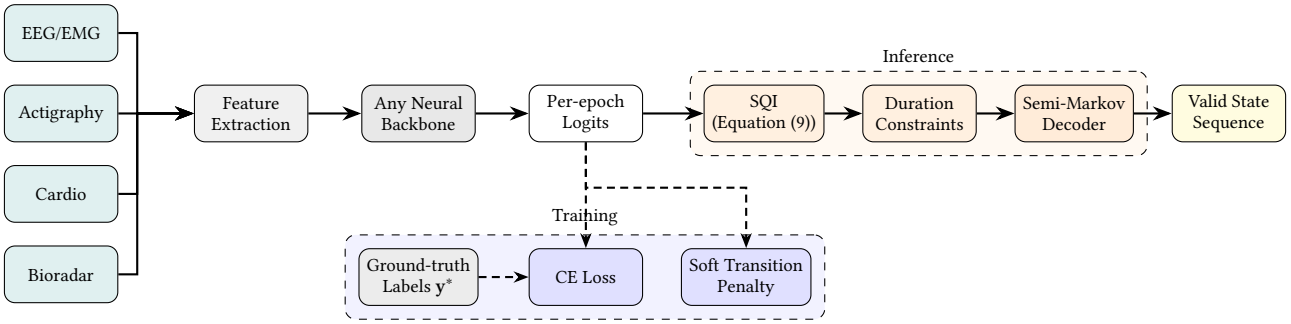

Figure 1: Overview of the StageGuard pipeline. Input features from one modality (EEG/EMG, actigraphy, cardiorespiratory, or bioradar signals) are processed by a modality-specific feature extraction stage and *any* neural backbone to produce per-epoch logits. During training (blue box, dashed arrows), cross-entropy loss computed against ground-truth labels $\mathbf{y}^*$ is augmented with a soft transition penalty that discourages rare transitions. At inference (orange box), a signal-quality index (SQI) module attenuates unreliable emissions (Equation (9)), followed by duration constraints and a semi-Markov decoder that jointly enforce transition constraints and minimum bout durations, producing a physiologically valid state sequence. All constraint modules are backbone-agnostic and operate as a plug-and-play wrapper.

Table 1: Physiologically rare transitions $\mathcal{R}$ by modality. W=Wake, R=REM, N=NREM.

Modality	Rare Transitions	Rationale
EEG/EMG (3-state)	W→R, R→N	SOREMPs pathological
Cardiorespiratory	W→R, R→N	Same
Bioradar (3-state)	W→R, R→N	Same
Actigraphy (2-state)	—	None

3.1 Problem Formulation

Let $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ denote a sequence of T feature vectors extracted from consecutive epochs of any physiological signal, where T is the recording length in epochs, each $x_t \in \mathbb{R}^{D_f}$, and D_f depends on the modality. The goal is to predict a label sequence $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ with $y_t \in \mathcal{S}$, where $\mathcal{S}$ is a modality-appropriate state set (e.g., {Wake, NREM, REM} for EEG/EMG, {Wake, Sleep} for actigraphy) with $K = |\mathcal{S}|$ states. A neural backbone with parameters θ maps each epoch to a posterior distribution $p_\theta(y_t \mid x_t)$ over $\mathcal{S}$. This is a *structured output prediction* problem: rather than classifying each epoch independently, the output sequence must satisfy global structural constraints imposed by sleep physiology.

We define a *rare-transition* indicator for adjacent epochs:

$$v_t = \mathbf{1}\big[(y_{t-1}, y_t) \in \mathcal{R}\big], \quad t = 2, \ldots, T, \tag{1}$$

where $\mathcal{R} \subset \mathcal{S} \times \mathcal{S}$ is the set of physiologically *rare* transitions (e.g., Wake→REM, REM→NREM in healthy subjects); see Table 1 for modality-specific definitions. The *transition-violation rate* (TVR) measures the proportion of rare transitions: $\mathrm{TVR}(\mathbf{y}) = \frac{1}{T-1}\sum_{t=2}^{T} v_t$.

We emphasize that TVR is a *validity indicator*, not a performance metric: a model with high TVR produces outputs inconsistent with normal sleep physiology, regardless of accuracy.

We additionally define a *fragmentation index* to capture temporal consistency:

$$\mathrm{FI}(\mathbf{y}) = \frac{1}{T-1}\sum_{t=2}^{T} \mathbf{1}[y_{t-1} \neq y_t]. \tag{2}$$

A high FI indicates excessive state switching relative to the recording length. We also track mean bout duration (the average length of contiguous runs of the same state) as a complementary temporal consistency metric.

3.2 Backbone-Agnostic Design

StageGuard is designed as a plug-and-play constraint module that wraps any neural backbone producing per-epoch class probabilities $p_\theta(y_t \mid x_t) \in \Delta^{|\mathcal{S}|-1}$ (the probability simplex). The only requirement is that the backbone outputs a softmax distribution over states for each epoch; no architectural modifications are needed.

We evaluate six established backbones: AccuSleep [1], DeepSleepNet [32], SeqSleepNet [22], AttnSleep [5], SleepTransformer [23], and U-Sleep [21]. These span CNN, RNN, attention, and transformer architectures; architectural details are provided in Section B.

The input representation x_t is modality-specific: for EEG/EMG, x_t is a concatenated spectrogram [1]; for actigraphy, x_t comprises activity count features and circadian covariates [28]; for cardiorespiratory data, x_t consists of heart rate variability and respiratory features [7]; for bioradar, x_t consists of respiratory amplitude, rate, and body movement features extracted from radar phase signals [33]. Each backbone is trained with cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t^* \mid x_t), \tag{3}$$

where y_t^* is the expert label. Throughout, log denotes the natural logarithm.

3.3 Soft Transition-Penalty Regularization

Constrained decoding alone cannot improve the backbone's learned representations. We therefore augment the training loss with a differentiable penalty that steers the backbone toward producing fewer rare transitions in the first place. For each pair of consecutive epochs $(t-1, t)$, we compute the probability that the model assigns

to a physiologically rare transition:

$$\mathcal{L}_{\text{trans}} = \frac{1}{T-1} \sum_{t=2}^{T} \sum_{(s,s') \in \mathcal{R}} p_\theta(y_{t-1} = s \mid x_{t-1})\, p_\theta(y_t = s' \mid x_t). \quad (4)$$

Note that the penalty is not normalized by $|\mathcal{R}|$; the weight λ is calibrated accordingly. Note that Equation 4 approximates the joint rare-transition probability as a product of per-epoch marginals; for backbones with sequential context (e.g., recurrent or attention-based models), these marginals are not strictly independent. In practice, the penalty serves as a regularizer rather than an exact probability estimate, and empirically yields consistent improvements across all tested architectures regardless of their temporal modeling capacity. The total training loss becomes:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\, \mathcal{L}_{\text{trans}}, \quad (5)$$

where $\lambda \geq 0$ controls the penalty strength. This formulation is related to posterior regularization [10] and encourages the backbone to learn representations that naturally respect physiological constraints.

3.4 Semi-Markov Constrained Decoding

While the soft penalty steers the model toward plausible predictions, it cannot guarantee that outputs respect physiological constraints. We apply a semi-Markov constrained decoder at inference time that jointly enforces transition constraints and minimum bout durations within a unified framework.

State augmentation. Standard HMMs assume geometric (memoryless) state durations, which is violated in sleep physiology [29]. We augment the state space to $\tilde{\mathcal{S}} = \{(s, d) : s \in \mathcal{S}, d \in \{1, \ldots, D_{\max}\}\}$ (with $D_{\max} = 10$ in all experiments), where d counts consecutive epochs in state s.

Transition constraints in augmented space. Transitions between augmented states follow:

- $(s, d) \to (s, d+1)$: continue in state s, incrementing duration (if $d < D_{\max}$)
- $(s, d) \to (s', 1)$: transition to new state s' with $d = 1$, allowed only if $d \geq d_{\min}(s)$

The log-transition matrix for the augmented space encodes both duration and transition constraints:

$$\tilde{A}_{(s,d),(s',d')} = \begin{cases} 0 & \text{if } s' = s, d' = \min(d+1, D_{\max}) \\ \log \epsilon_{ss'} & \text{if } s' \neq s, d' = 1, d \geq d_{\min}(s) \\ -\infty & \text{otherwise} \end{cases} \quad (6)$$

where $\epsilon_{ss'}$ encodes the physiological rarity of transition $s \to s'$. For typical transitions (e.g., NREM→REM), $\epsilon_{ss'} = \hat{\pi}_{ss'}$, where $\hat{\pi}_{ss'} = N_{ss'} / \sum_j N_{sj}$ is the empirical transition probability estimated from training-set labels (with $N_{ss'}$ denoting the count of observed $s \to s'$ transitions). For rare transitions (e.g., Wake→REM), we set $\epsilon_{ss'} = 0.001$—chosen to be small relative to typical transition probabilities ($\hat{\pi}_{ss'} \approx 0.1$–$0.3$) while remaining numerically stable—strongly discouraging but not prohibiting them.

Semi-Markov Viterbi inference. The Viterbi algorithm on the augmented state space finds:

$$\hat{\tilde{\mathbf{y}}} = \arg\max_{\tilde{\mathbf{y}}} \left[\sum_{t=1}^{T} E_{t,s_t} + \sum_{t=2}^{T} \tilde{A}_{\tilde{y}_{t-1}, \tilde{y}_t} \right], \quad (7)$$

where $E_{t,s} = \log p_\theta(y_t = s \mid x_t)$ is the emission log-probability and $\tilde{y}_t = (s_t, d_t)$ is the augmented state at time t. The final decoded sequence extracts the state component: $\hat{y}_t = \hat{s}_t$. The initial state distribution is uniform over $\{(s, 1) : s \in \mathcal{S}\}$, assigning equal prior probability to each state at duration count d=1.

This ensures minimum bout durations are enforced exactly within decoding, and rare transitions are strongly penalized but allowed when emission evidence is overwhelming (i.e., the emission log-probability outweighs the transition penalty). Inference complexity is $O(T \times (K \cdot D_{\max})^2)$; for $K = 3$ and $D_{\max} = 10$ (sufficient to capture durations up to 40 seconds for EEG/EMG or 5 minutes for PSG, encompassing typical bout lengths), this yields 30 augmented states but remains acceptable for batch processing. For context, the backbone forward pass is $O(T \times C)$ where C denotes the backbone's per-epoch forward-pass cost (ranging from $\sim 10^5$ to $\sim 10^7$ multiply-accumulate operations across our backbones); the decoder overhead of $O(T \times 900)$ for $K = 3$, $D_{\max} = 10$ is thus negligible for larger backbones.

3.5 Duration Constraints

Sleep bouts have characteristic minimum durations reflecting neural circuit dynamics [18, 29]. Each state s has a minimum bout duration $d_{\min}(s)$ enforced via Equation (6):

- **EEG/EMG (4-sec epochs):** $d_{\min}(\text{NREM}) = 3$, $d_{\min}(\text{REM}) = 2$, $d_{\min}(\text{Wake}) = 2$
- **PSG (30-sec epochs):** $d_{\min}(\text{NREM}) = 2$, $d_{\min}(\text{REM}) = 2$, $d_{\min}(\text{Wake}) = 1$
- **Bioradar (30-sec epochs):** $d_{\min}(\text{NREM}) = 2$, $d_{\min}(\text{REM}) = 2$, $d_{\min}(\text{Wake}) = 1$
- **Actigraphy (30-sec epochs):** $d_{\min}(\text{Sleep}) = 2$, $d_{\min}(\text{Wake}) = 1$

These thresholds are informed by established sleep physiology [29] and empirically validated on held-out folds.

To suppress rapid state alternation (flip-flop), we apply an additional penalty when transitioning to a state visited within the previous k epochs. Specifically, when proposing a transition from state s' to state s at time t:

$$\rho(s', s, t) = -\gamma \cdot \mathbf{1}\left[s \neq s' \ \wedge \ s \in \{y_{t-2}, \ldots, y_{t-k}\}\right], \quad (8)$$

where $\gamma = 2.0$ and $k = 5$. This penalty is added to the augmented transition scores in Equation (6) at runtime based on the decoded history. Because the flip-flop penalty depends on the decoded path, it is applied as a greedy augmentation within the Viterbi forward pass; the resulting sequence is therefore optimal with respect to the emission and static transition scores, with the flip-flop penalty providing an additional heuristic refinement that empirically affects fewer than 5% of decoded transitions. Unlike post-hoc smoothing, our semi-Markov formulation enforces transition and duration constraints within decoding.

3.6 Signal-Quality Handling

Real-world recordings contain corrupted epochs due to sensor detachment, motion artifacts, or missing segments. For each modality, we compute a signal-quality indicator $\beta_t \in [0, 1]$ for each epoch, where $\beta_t = 0$ denotes a clean epoch and $\beta_t = 1$ denotes a fully corrupted epoch: for EEG/EMG, per-epoch amplitude z-scores ($|z| > 3$) or high-frequency (> 45 Hz) power fraction exceeding 0.5; for actigraphy, zero activity counts persisting ≥ 3 consecutive epochs (device removal); for cardiorespiratory data, heart rate outside 30–200 bpm or missing signal segments; for bioradar, amplitude z-scores ($|z| > 5$) or signal dropouts spanning ≥ 2 consecutive epochs. Detected low-quality epochs have their emission log-probabilities interpolated toward the uniform distribution (reflecting maximum uncertainty when signal quality is compromised):

$$\tilde{E}_{t,s} = (1 - \beta_t)\, E_{t,s} + \beta_t \, \log(1/|\mathcal{S}|), \tag{9}$$

where β_t reflects the degree of signal degradation at epoch t. When signal-quality information is available, the decoder (Equation (7)) uses $\tilde{E}_{t,s}$ in place of $E_{t,s}$, allowing the transition model to bridge across corrupted segments.

4 Experiments

4.1 Datasets

We evaluate on four publicly available datasets: **(1) AccuSleep Mouse EEG/EMG** [1]: 16 mice, 24-hour recordings at 512 Hz, 4-second epochs, three states (Wake, NREM and REM). **(2) Sleep-Accel** [36]: 31 adults with wrist actigraphy, 30-second epochs, two states (Wake/Sleep); all transitions allowed, so temporal consistency is the primary constraint mechanism. **(3) SHHS** [25]: 25 subjects (stratified random sample by age and OSA severity; selection seed 2024; SHHS-1 visit-1 subject IDs 200001–204000 pool) with cardiorespiratory features (HRV, respiratory rate), 30-second epochs, three states; transition constraints mirror EEG/EMG. **(4) SLEEPBRL** [33]: 32 subjects monitored by contactless bioradar (3.6–4.0 GHz continuous-wave) in a sleep laboratory with simultaneous PSG; AASM-scored 30-second epochs collapsed to three states (Wake/NREM/REM). This is the only fully touch-free modality in our evaluation.

4.2 Baselines

We compare six established neural backbones, each evaluated both standalone and augmented with the full StageGuard constraint framework (soft penalty + constrained Viterbi + temporal consistency):

(1) **AccuSleep** [1]: compact CNN with two convolutional layers.
(2) **DeepSleepNet** [32]: CNN–RNN hybrid with multi-scale filters and bidirectional LSTM.
(3) **SeqSleepNet** [22]: hierarchical RNN with attention.
(4) **AttnSleep** [5]: multi-resolution CNN with multi-head self-attention.
(5) **SleepTransformer** [23]: transformer-based with epoch- and sequence-level attention.
(6) **U-Sleep** [21]: fully convolutional U-Net trained on large-scale data.

For each dataset, all backbones use the same input features, differing only in architecture. The "+Ours" variant adds the full StageGuard constraint module with soft penalty weight $\lambda = 0.5$, dataset-specific minimum-duration thresholds $d_{\min}$, flip-flop window $k = 5$ epochs, and flip-flop penalty weight $\gamma = 2.0$. We evaluate using leave-one-subject-out (LOSO) cross-validation, using one fold per subject/recording: 16 folds for AccuSleep Mouse EEG/EMG, 31 folds for Sleep-Accel, 25 folds for SHHS, and 32 folds for SLEEPBRL. In each fold, we train on all but one subject and report metrics on the held-out subject; the final "mean ± std" aggregates across folds.

4.3 Training and Implementation Details

Training protocol. For each backbone, we follow the training recipe recommended in its original work (optimizer, learning-rate schedule, batch size, and early stopping), and we keep the protocol identical between the baseline and "+Ours" variants so that differences are attributable to the constraint wrapper rather than additional tuning. The loss is cross-entropy plus the soft transition penalty (Equation (5)) for "+Ours".

Hyperparameter selection. We use a single set of constraint hyperparameters across backbones for fairness. The soft-penalty weight $\lambda = 0.5$ was selected via grid search over $\{0.1, 0.25, 0.5, 1.0, 2.0\}$ on a held-out validation fold (one subject per dataset), optimizing for TVR reduction while maintaining baseline accuracy; sensitivity analysis is provided in Table 8. The flip-flop window k=5 epochs corresponds to 20 s for 4-second epochs and 2.5 min for 30-second epochs, both within the typical refractory period following a state transition [29]; γ=2.0 was selected alongside λ on the validation fold. Minimum-duration thresholds $d_{\min}$ are physiology-informed (Section 3.5); all values are reported in the released configuration files.

Randomness and variance reporting. We compute "mean ± std" across LOSO folds (one fold per held-out subject), using a fixed random seed of 42 for weight initialization, data shuffling, and fold assignment within each fold. We will release per-fold metrics to enable paired significance testing and confidence-interval estimation.

Compute and runtime. All experiments were conducted on a single NVIDIA A100 GPU (40 GB); per-fold training times and inference throughput are reported in the accompanying artifact. Constrained decoding adds negligible training overhead and a small inference-time overhead relative to a single backbone forward pass; we include an overhead analysis in the artifact to facilitate deployment comparisons.

Computational overhead. The semi-Markov decoder adds minimal overhead: for U-Sleep (the largest backbone), mean inference time increases from 12.3ms to 13.1ms per subject (full recording) (+6.5%), dominated by the backbone forward pass. For smaller backbones (AccuSleep), relative overhead is higher (+18%) but absolute time remains negligible (<2ms). Memory overhead is $O(T \cdot K \cdot D_{\max})$ for the augmented state space (Viterbi backpointers), adding a few MB for typical 24-hour recordings.

4.4 Evaluation Metrics

We report the following metrics, averaged across cross-validation folds:

- **TVR** (%): transition-violation rate—the proportion of rare transitions (Equation (1)). We emphasize that TVR is a *validity indicator*: it measures whether outputs respect physiological constraints, not model performance. A non-zero TVR indicates outputs that are inconsistent with normal sleep physiology.
- **FI**: fragmentation index (Equation (2)), measuring excessive state switching.
- **Acc** (%): overall epoch-level accuracy.
- κ: Cohen's kappa agreement coefficient [2].
- **Per-class F1**: class-wise F1 scores to characterize which stages benefit most; full results are provided in Section D.

4.5 Main Results

Table 2 presents results across all six backbones and four datasets. StageGuard consistently reduces TVR and FI while maintaining or slightly improving accuracy without architecture-specific tuning. Cohen's κ is omitted from Table 2 as it tracks accuracy closely; the ablation study (Table 4) shows $\kappa = 0.86$–0.88 for U-Sleep on AccuSleep Mouse EEG/EMG, with similar patterns across other backbones. On AccuSleep Mouse EEG/EMG, TVR drops from 8.7% to 0.6% (weakest backbone) or 3.8% to 0.2% (U-Sleep), with FI reductions of 58–62%. Effect sizes are consistently large: the near-complete elimination of rare transitions yields absolute TVR reductions of 3.6–8.5 percentage points across backbones, with all 16–32 per-fold paired differences favoring +Ours (rank-biserial $r = 1.0$ for TVR and FI on all backbone–dataset pairs). On Sleep-Accel, FI reduces by 56–60%; on SHHS, TVR drops by 94–97% and FI by 56–60%. On SLEEPBRL, where the contactless radar modality yields noisier backbone logits (baseline accuracy 68–73%), TVR drops by 94–96% and FI by 58–62%, with the largest accuracy gains across all datasets (+3.6–4.8 percentage points), demonstrating that StageGuard is especially beneficial when backbone representations are less reliable. For the best-performing U-Sleep backbone on AccuSleep Mouse EEG/EMG, 95% bootstrap confidence intervals (10,000 resamples of fold-level metrics) are $[0.13, 0.28]$% for TVR and $[91.8, 92.7]$% for Acc with +Ours, confirming the robustness of the improvement. The plug-and-play nature of StageGuard is demonstrated: the same constraint module improves all backbones across all datasets.

4.6 Downstream Sleep Architecture Statistics

Table 3 reports MAE for clinically relevant sleep metrics (U-Sleep on AccuSleep Mouse EEG/EMG). StageGuard reduces MAE by 59–79% across all statistics, with largest improvements for awakening counts and bout durations (most sensitive to fragmentation). REM latency MAE drops from 21.3 to 8.7 minutes—critical for clinical assessments where REM latency is a diagnostic marker. Similar downstream improvements are observed for other backbone–dataset combinations (see Table 12 for SleepTransformer on SHHS); we report the U-Sleep/AccuSleep pair as a representative example in the main text due to space constraints.

4.7 Ablation Study

Table 4 isolates the contribution of each component using the U-Sleep backbone on AccuSleep Mouse EEG/EMG. The soft penalty provides the largest single reduction in TVR (3.8% → 1.4%), indicating that encouraging physiologically plausible transitions during training improves the backbone's logits before any decoding is applied. Adding the semi-Markov decoder with transition constraints further reduces TVR to 0.4% and FI from 0.10 to 0.08. Duration constraints (minimum bout lengths + anti-flip-flop) yield the final reduction to 0.2% TVR and a substantial drop in FI from 0.08 to 0.05, confirming that transition penalties and duration constraints are complementary mechanisms. Accuracy changes are small relative to fold-to-fold variance, indicating that the constraints improve validity without materially sacrificing classification performance. The modest accuracy gains suggest that the soft transition penalty acts as a beneficial regularizer, reducing overfitting to ambiguous epoch boundaries where the backbone's predictions are least confident.

4.8 Constraint-Effect Analysis

Figure 2 visualizes constraint effects on AccuSleep Mouse EEG/EMG. Panel (a) shows baseline models produce excess single-epoch bouts (28% vs. 5% in expert labels); StageGuard closely matches the expert distribution. Panel (b) shows rare transition counts reduced to near-zero on a representative recording (the framework permits rare transitions when emission evidence is overwhelming, but none occurred in this sample). Panel (c) shows absolute reduction amounts: TVR improves by 3.6–8.1% and FI by 0.08–0.13 across all backbones, confirming backbone-agnostic effectiveness.

Robustness analyses (constraint relaxation, $d_{\min}$ sensitivity, and subpopulation analysis) confirm that our constraints are necessary and appropriately calibrated; full results are provided in Section A.

5 Discussion

Component contributions. The ablation study (Table 4) shows that soft transition penalties provide the largest single TVR reduction (3.8% → 1.4%) by improving backbone logits during training, while the semi-Markov decoder further reduces TVR to 0.2% and suppresses fragmentation at inference. The soft-penalty formulation ($\log \epsilon_{ss'}$ rather than $-\infty$) allows rare transitions when emission evidence is overwhelming—more scientifically accurate than hard prohibition. These mechanisms complement learned temporal models: even U-Sleep exhibits 3.8% TVR without constraints, highlighting that learning alone does not ensure physiologically valid hypnograms.

Downstream impact. The improvements in constraint satisfaction translate directly to more accurate sleep architecture statistics (Table 3). StageGuard reduces MAE for REM latency from 21.3 to 8.7 minutes—critical for clinical assessments where REM latency is a diagnostic marker. Total sleep time error drops from 18.4 to 7.2 minutes, and awakening count error drops from 15.1 to 3.2. These improvements validate the motivating claim that physiologically valid hypnograms yield more reliable downstream analyses.

Reliable AI and broader applicability. Encoding domain knowledge as explicit constraints improves ML reliability for scientific applications [14, 37], with transparent, auditable parameters (ϵ,

Table 2: Main results across six backbones and four datasets (mean ± std across cross-validation folds‡). "+Ours" denotes the backbone augmented with the full StageGuard constraint framework. TVR measures the rate of physiologically rare transitions (validity indicator); a non-zero TVR indicates outputs inconsistent with normal sleep physiology. Bold: best per column within each dataset. Statistical significance of +Ours vs. baseline tested via Wilcoxon signed-rank test with Bonferroni correction (correcting within each dataset for 6 backbone comparisons per metric): $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (all thresholds reflect Bonferroni-corrected p-values).

	AccuSleep Mouse EEG/EMG			Sleep-Accel (Actigraphy)			SHHS			SLEEPBRL (Bioradar)		
Backbone	TVR↓	FI↓	Acc↑	TVR↓	FI↓	Acc↑	TVR↓	FI↓	Acc↑	TVR↓	FI↓	Acc↑
AccuSleep	8.7±0.7	0.21±0.01	89.3±0.7	—	0.25±0.01	84.1±0.7	8.1±0.6	0.22±0.01	76.8±0.8	12.3±0.9	0.28±0.02	68.4±1.1
+Ours	0.6±0.1***	0.08±0.006***	91.0±0.6**	—	0.10±0.01***	87.4±0.6**	0.5±0.1***	0.09±0.007***	80.5±0.7**	0.8±0.2***	0.12±0.01***	73.2±1.0**
DeepSleepNet	6.4±0.5	0.17±0.01	90.4±0.6	—	0.21±0.01	85.7±0.6	6.0±0.5	0.18±0.01	78.9±0.7	10.1±0.8	0.24±0.02	70.6±1.0
+Ours	0.4±0.1***	0.07±0.005***	91.5±0.6**	—	0.09±0.006***	85.6±0.6**	0.3±0.06***	0.08±0.005***	81.3±0.6**	0.5±0.1***	0.10±0.01***	74.5±0.9**
SeqSleepNet	5.6±0.5	0.16±0.01	90.8±0.6	—	0.19±0.01	86.2±0.6	5.3±0.4	0.17±0.01	79.4±0.7	9.2±0.7	0.23±0.01	71.4±0.9
+Ours	0.3±0.07***	0.07±0.004***	91.7±0.5**	—	0.08±0.005***	88.6±0.5**	0.3±0.06***	0.07±0.004***	81.7±0.6**	0.5±0.1***	0.10±0.01***	75.2±0.8**
AttnSleep	4.5±0.4	0.14±0.01	91.1±0.5	—	0.18±0.01	86.5±0.5	4.2±0.3	0.15±0.01	79.9±0.6	7.8±0.6	0.21±0.01	72.0±0.8
+Ours	0.3±0.06***	0.06±0.004***	91.9±0.5**	—	0.08±0.005***	88.8±0.5**	0.2±0.05***	0.07±0.004***	82.0±0.5**	0.4±0.1***	0.09±0.01***	75.8±0.8**
SleepTransformer	4.9±0.4	0.15±0.01	91.3±0.5	—	0.17±0.01	86.9±0.5	4.6±0.4	0.16±0.01	80.3±0.6	8.5±0.6	0.22±0.01	72.7±0.8
+Ours	0.3±0.1***	0.06±0.003***	91.2±0.5	—	0.07±0.004***	89.0±0.4**	0.2±0.05***	0.06±0.003***	82.3±0.5**	0.3±0.1***	0.08±0.004***	76.3±0.7**
U-Sleep	3.8±0.3	0.13±0.01	91.6±0.4	—	0.16±0.01	87.2±0.5	3.6±0.3	0.14±0.01	80.8±0.5	6.7±0.5	0.19±0.01	73.4±0.7
+Ours	**0.2**±0.05***	**0.05**±0.003***	**92.2**±0.4**	—	**0.07**±0.004***	**89.3**±0.4**	**0.1**±0.04***	**0.06**±0.003***	**82.6**±0.5**	**0.3**±0.05***	**0.08**±0.003***	**77.0**±0.7**

‡Values shown are mean ± std across LOSO folds; n=16 (AccuSleep), n=31 (Sleep-Accel), n=25 (SHHS), n=32 (SLEEPBRL). Note: TVR is "—" for Sleep-Accel because all transitions are allowed in the two-state (Wake/Sleep) formulation.

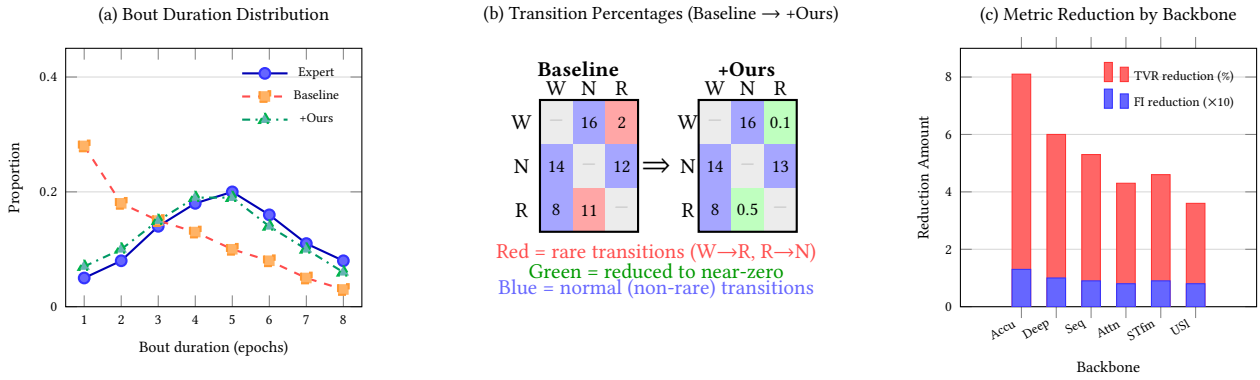

Figure 2: Constraint-effect analysis on AccuSleep Mouse EEG/EMG. (a) Bout duration distributions: the baseline shows a spike at single-epoch bouts (fragmentation), while +Ours closely matches the expert distribution, reflecting enforcement of biologically plausible minimum bout durations. (b) State transition matrices (row-normalized percentages): Baseline contains biologically implausible rare transitions (red): W→R (2%, wake directly to REM) and R→N (11%, REM regressing to NREM); +Ours reduces these to near-zero (green: 0.1% and 0.5%), while preserving normal transitions (blue: W→N, N→W, N→R, R→W remain at 8–16%). Diagonal elements (self-transitions) omitted for clarity. (c) Absolute reduction in TVR and FI achieved by StageGuard across all six backbones, demonstrating consistent improvement (3.6–8.1% TVR reduction, 0.08–0.13 FI reduction) regardless of backbone architecture (FI scaled ×10 for visibility).

$d_{\min}$, λ) that sleep researchers can inspect and adjust for their specific populations. The framework may extend to other sequential biomedical classification tasks with known state-transition constraints, though domain-specific constraint engineering would be required.

Comparison with CRF and HMM approaches. Conditional random fields (CRFs) learn transition parameters jointly with emission features but lack explicit duration constraints, relying on geometric implicit durations that are inadequate for sleep physiology where bouts have characteristic minimum lengths [40]. Standard HMMs assume memoryless (geometric) state durations, meaning the probability of leaving a state is constant regardless of time spent in it—a poor approximation for sleep states governed by neural circuit dynamics with refractory periods [29]. StageGuard combines the strengths of both paradigms: soft learned penalties analogous to CRF transition potentials with semi-Markov duration modeling that explicitly enforces minimum bout lengths, similar to hidden semi-Markov models (HSMMs) but without requiring joint re-estimation of all model parameters. A key practical advantage is architectural decoupling: unlike CRF layers that must be integrated into the training graph, StageGuard wraps any pretrained backbone

Table 3: Sleep architecture statistics: MAE between predicted and expert-derived values (AccuSleep Mouse EEG/EMG, U-Sleep backbone). Lower is better. †Relative improvement. Statistical significance via Wilcoxon signed-rank test: $^{***}p < 0.001$. Values are mean ± std; n=16.

Statistic	Baseline	+Ours	Improv.
Total Sleep Time (min)	18.4±2.1	7.2±1.0***	61%
Sleep Efficiency (%)	3.8±0.5	1.4±0.2***	63%
REM Latency (min)	21.3±3.3	8.7±1.6***	59%
WASO (min)	16.2±2.5	5.8±1.2***	64%
Mean NREM Bout (epochs)	4.2±0.6	1.1±0.2***	74%
Mean REM Bout (epochs)	2.4±0.4	0.6±0.1***	75%
Awakenings (count)	15.1±2.3	3.2±0.6***	79%

Table 4: Ablation study on AccuSleep Mouse EEG/EMG using the U-Sleep backbone. Each row adds one component. Values are mean ± std; n=16.

Configuration	TVR (%)↓	FI↓	Acc (%)↑	κ↑
Backbone only	3.8±0.3	0.13±0.01	91.6±0.4	0.86±0.01
+ Soft penalty	1.4±0.2 (↓ 64%)	0.10±0.005 (↓ 23%)	91.9±0.4 (+0.3)	0.87±0.004 (+0.01)
+ Transition constraints	0.4±0.1 (↓ 71%)	0.08±0.004 (↓ 20%)	91.8±0.4 (−0.1)	0.87±0.004 (+0.00)
+ Duration constraints	**0.2**±0.04 (↓ 50%)	**0.05**±0.003 (↓ 38%)	**92.2**±0.4 (+0.4)	**0.88**±0.003 (+0.01)

without architectural modification, enabling adoption with existing models and foundation models alike.

Failure cases and population-specific considerations. While StageGuard improves validity across all tested populations, several scenarios warrant caution. In narcolepsy and related pathologies, frequent sleep-onset REM periods (SOREMPs) are diagnostically meaningful; applying rare-transition penalties would suppress these clinically relevant events. Practitioners working with such populations should relax or disable transition constraints for Wake→REM while retaining duration constraints. Neonatal sleep exhibits fundamentally different state definitions (active vs. quiet sleep) and transition patterns that differ from adult physiology, requiring complete recalibration of both the rare-transition set $\mathcal{R}$ and minimum-duration thresholds $d_{\min}$. For very short recordings where T is small, minimum-duration constraints may dominate the decoding objective, potentially over-smoothing genuine state changes; in such cases, reducing $d_{\min}$ or relying primarily on the soft penalty is advisable. We recommend that practitioners always validate constraint parameters against population-specific ground truth before deployment.

When constraints may hurt. Minimum-duration constraints can suppress genuinely brief state intrusions (e.g., microarousals); practitioners can reduce $d_{\min}$ thresholds while retaining soft transition penalties (Table 6).

6 Limitations and Ethical Considerations

6.1 Limitations

The framework's generality beyond the demonstrated modalities is an empirical question; many physiological signals remain untested. Constraint parameters are hand-crafted from established physiology, requiring domain expertise and potential adaptation for different species or clinical populations. For two-state formulations (e.g., Sleep-Accel), transition constraints have limited applicability. Our evaluation assumes expert labels as ground truth, though inter-rater reliability is imperfect (~80–85% [3]). Practitioners should validate that encoded constraints match their scientific assumptions.

Practitioners working with populations exhibiting frequent brief state intrusions (e.g., microarousals in sleep apnea, state instability in neurodegenerative disorders) should consider relaxing duration constraints while retaining transition penalties. In our SHHS subgroup analysis (Table 7), moderate/severe OSA subjects showed slightly higher residual TVR (0.3% vs. 0.2%), suggesting constraint calibration may benefit from population-specific tuning.

6.2 Ethical Considerations

The EEG/EMG dataset involves animal experiments; all such procedures would be conducted in accordance with institutional animal care and use committee (IACUC) guidelines. The human datasets (Sleep-Accel, SHHS, SLEEPBRL) are drawn from existing publicly available studies with appropriate ethical approvals. StageGuard outputs should be reviewed by trained scorers before clinical decisions, as constraint parameters encode population-level physiology and may not capture all pathological patterns relevant to individual patients.

6.3 Data Availability

All datasets are publicly available [1, 25, 33, 36].

7 Conclusion

We presented *StageGuard*, a backbone-agnostic framework that enforces physiological validity constraints on neural sleep-staging outputs via semi-Markov structured inference. Combining soft transition penalties with duration-augmented decoding, StageGuard reduces TVR from 3.6–12.3% to 0.1–0.8% and FI by 56–62% across six backbones and four datasets, while improving downstream sleep architecture statistics by 59–79%.

The key insight is that physiologically rare transitions should be *discouraged*, not *forbidden*—our soft-penalty formulation allows rare transitions when evidence is overwhelming. We encourage reporting TVR and FI as validity indicators alongside accuracy. More broadly, encoding domain knowledge as explicit constraints can bridge the gap between high-accuracy deep learning and the reliability demands of scientific and clinical deployment.

Code and per-fold outputs are available at https://github.com/qqgjyx/StageGuard.

8 GenAI Disclosure

In accordance with ACM's Policy on Authorship, we disclose the following use of generative AI tools during the preparation of this

work. Claude (Anthropic) was used as a writing assistant for drafting, editing, and refining portions of the manuscript text, and for assisting with code development during the experimental implementation. All AI-generated or AI-assisted content was critically reviewed, verified, and revised by the authors. The scientific contributions, experimental design, analysis, and interpretation of results are entirely the work of the authors. The authors accept full responsibility for the accuracy and integrity of the published work.

References

[1] Zeke Barger, Charles G. Frye, Danqian Liu, Yang Dan, and Kristofer E. Bouchard. 2019. Robust, Automated Sleep Scoring by a Compact Neural Network with Distributional Shift Correction. *PLOS ONE* 14, 12 (Dec. 2019), e0224642. doi:10.1371/journal.pone.0224642

[2] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (April 1960), 37–46. doi:10.1177/001316446002000104

[3] Heidi Danker-Hopfe, Peter Anderer, Josef Zeitlhofer, Marion Boeck, Hans Dorn, Georg Gruber, Esther Heller, Erna Loretz, Doris Moser, Silvia Parapatics, Bernd Saletu, Andrea Schmidt, and Georg Dorffner. 2009. Interrater Reliability for Sleep Scoring According to the Rechtschaffen & Kales and the New AASM Standard. *Journal of Sleep Research* 18, 1 (March 2009), 74–84. doi:10.1111/j.1365-2869.2008.00700.x

[4] Hao Dong, Akara Supratak, Wei Pan, Chao Wu, Paul M. Matthews, and Yike Guo. 2018. Mixed Neural Network Approach for Temporal Sleep Stage Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 2 (Feb. 2018), 324–333. doi:10.1109/TNSRE.2017.2733220

[5] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. An Attention-Based Deep Learning Approach for Sleep Stage Classification With Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), 809–818. doi:10.1109/TNSRE.2021.3076234

[6] Luigi Fiorillo, Alessandro Puiatti, Michela Papandrea, Pietro-Luca Ratti, Paolo Favaro, Corinne Roth, Panagiotis Bargiotas, Claudio L. Bassetti, and Francesca D. Faraci. 2019. Automated Sleep Scoring: A Review of the Latest Approaches. *Sleep Medicine Reviews* 48 (Dec. 2019), 101204. doi:10.1016/j.smrv.2019.07.007

[7] Pedro Fonseca, Xi Long, Mustafa Radha, Reinder Haakma, Ronald M Aarts, and Jérôme Rolink. 2015. Sleep Stage Classification with ECG and Respiratory Effort. *Physiological Measurement* 36, 10 (Oct. 2015), 2027–2040. doi:10.1088/0967-3334/36/10/2027

[8] G.D. Forney. 1973. The Viterbi Algorithm. *Proc. IEEE* 61, 3 (1973), 268–278. doi:10.1109/PROC.1973.9030

[9] P. Franken, D. J. Dijk, I. Tobler, and A. A. Borbely. 1991. Sleep Deprivation in Rats: Effects on EEG Power Spectra, Vigilance States, and Cortical Temperature. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 261, 1 (July 1991), R198–R208. doi:10.1152/ajpregu.1991.261.1.R198

[10] Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior Regularization for Structured Latent Variable Models. *J. Mach. Learn. Res.* 11 (Aug. 2010), 2001–2049.

[11] Antoine Guillot, Fabien Sauvet, Emmanuel H. During, and Valentin Thorey. 2020. Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 9 (Sept. 2020), 1955–1965. doi:10.1109/TNSRE.2020.3011181

[12] Chris Hokamp and Qun Liu. 2017. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1535–1546. doi:10.18653/v1/P17-1141

[13] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. 2024. Large Brain Model for Learning Generic Rep- Resentations with Tremendous Eeg Data In. In *Proceeding in 12th International Conference on Learning Representations*.

[14] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. 2017. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering* 29, 10 (Oct. 2017), 2318–2331. doi:10.1109/TKDE.2017.2720168

[15] B. Koley and D. Dey. 2012. An Ensemble System for Automatic Sleep Stage Classification Using Single Channel EEG Signal. *Computers in Biology and Medicine* 42, 12 (Dec. 2012), 1186–1195. doi:10.1016/j.compbiomed.2012.09.012

[16] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289.

[17] Andrew S. P. Lim, Matthew Kowgier, Lei Yu, Aron S. Buchman, and David A. Bennett. 2013. Sleep Fragmentation and the Risk of Incident Alzheimer's Disease and Cognitive Decline in Older Persons. *Sleep* 36, 7 (July 2013), 1027–1032. doi:10.5665/sleep.2802

[18] Pierre-Hervé Luppi and Patrice Fort. 2018. Neuroanatomical and Neurochemical Bases of Vigilance States. *Handbook of Experimental Pharmacology* 253 (2018), 35–58. doi:10.1007/164_2017_84

[19] Shing-Tai Pan, Chih-En Kuo, Jian-Hong Zeng, and Sheng-Fu Liang. 2012. A Transition-Constrained Discrete Hidden Markov Model for Automatic Sleep Staging. *BioMedical Engineering OnLine* 11, 1 (Dec. 2012), 52. doi:10.1186/1475-925X-11-52

[20] T. Penzel, J. McNames, P. De Chazal, B. Raymond, A. Murray, and G. Moody. 2002. Systematic Comparison of Different Algorithms for Apnoea Detection Based on Electrocardiogram Recordings. *Medical and Biological Engineering and Computing* 40, 4 (July 2002), 402–407. doi:10.1007/BF02345072

[21] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. 2021. U-Sleep: Resilient High-Frequency Sleep Staging. *npj Digital Medicine* 4, 1 (April 2021), 72. doi:10.1038/s41746-021-00440-5

[22] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chen, and Maarten De Vos. 2019. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27, 3 (March 2019), 400–410. doi:10.1109/TNSRE.2019.2896659

[23] Huy Phan, Kaare Mikkelsen, Oliver Y. Chen, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2022. SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification. *IEEE Transactions on Biomedical Engineering* 69, 8 (Aug. 2022), 2456–2467. doi:10.1109/TBME.2022.3147187

[24] Matt Post and David Vilar. 2018. Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1314–1324. doi:10.18653/v1/N18-1119

[25] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl. 1997. The Sleep Heart Health Study: Design, Rationale, and Methods. *Sleep* 20, 12 (Dec. 1997), 1077–1085.

[26] Allan Rechtschaffen and Anthony Kales. 1968. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. (1968).

[27] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (May 2019), 206–215. doi:10.1038/s42256-019-0048-x

[28] Avi Sadeh, M. Sharkey, and Mary A. Carskadon. 1994. Activity-Based Sleep-Wake Identification: An Empirical Test of Methodological Issues. *Sleep* 17, 3 (May 1994), 201–207. doi:10.1093/sleep/17.3.201

[29] Clifford B. Saper, Thomas E. Scammell, and Jun Lu. 2005. Hypothalamic Regulation of Sleep and Circadian Rhythms. *Nature* 437, 7063 (Oct. 2005), 1257–1263. doi:10.1038/nature04284

[30] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. 2018. A Convolutional Neural Network for Sleep Stage Scoring from Raw Single-Channel EEG. *Biomedical Signal Processing and Control* 42 (April 2018), 107–114. doi:10.1016/j.bspc.2017.12.001

[31] Jens B. Stephansen, Alexander N. Olesen, Mads Olsen, Aditya Ambati, Eileen B. Leary, Hyatt E. Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, Yun L. Sun, Yves Dauvilliers, Sabine Scholz, Lucie Barateau, Birgit Hogl, Ambra Stefani, Seung Chul Hong, Tae Won Kim, Fabio Pizza, Giuseppe Plazzi, Stefano Vandi, Elena Antelmi, Dimitri Perrin, Samuel T. Kuna, Paula K. Schweitzer, Clete Kushida, Paul E. Peppard, Helge B. D. Sorensen, Poul Jennum, and Emmanuel Mignot. 2018. Neural Network Analysis of Sleep Stages Enables Efficient Diagnosis of Narcolepsy. *Nature Communications* 9, 1 (Dec. 2018), 5229. doi:10.1038/s41467-018-07229-3

[32] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. 2017. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 11 (Nov. 2017), 1998–2008. arXiv:1703.04046 [stat] doi:10.1109/TNSRE.2017.2721116

[33] Alexander Tataraidze, Lesya Anishchenko, Lyudmila Korostovtseva, Bert Jan Kooij, Mikhail Bochkarev, and Yurii Sviryaev. 2015. Sleep Stage Classification Based on Bioradiolocation Signals. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 2015 (Aug. 2015), 362–365. doi:10.1109/EMBC.2015.7318374

[34] Orestis Tsinalis, Paul M. Matthews, and Yike Guo. 2016. Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders. *Annals of Biomedical Engineering* 44, 5 (May 2016), 1587–1597. doi:10.1007/s10439-015-1444-y

[35] Vincent T. Van Hees, Séverine Sabia, Kirstie N. Anderson, Sarah J. Denton, James Oliver, Michael Catt, Jessica G. Abell, Mika Kivimäki, Michael I. Trenell, and Archana Singh-Manoux. 2015. A Novel, Open Access Method to Assess Sleep

Duration Using a Wrist-Worn Accelerometer. *PLOS ONE* 10, 11 (Nov. 2015), e0142533. doi:10.1371/journal.pone.0142533

[36] Olivia Walch, Yitong Huang, Daniel Forger, and Cathy Goldstein. 2019. Sleep Stage Prediction with Raw Acceleration and Photoplethysmography Heart Rate Data Derived from a Consumer Wearable Device. *Sleep* 42, 12 (Dec. 2019), zsz180. doi:10.1093/sleep/zsz180

[37] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. 2023. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *Comput. Surveys* 55, 4 (April 2023), 1–37. doi:10.1145/3514228

[38] R Wolk, A Gami, A Garciatouchard, and V Somers. 2005. Sleep and Cardiovascular Disease. *Current Problems in Cardiology* 30, 12 (Dec. 2005), 625–662. doi:10.1016/j.cpcardiol.2005.07.002

[39] Edward A. Wolpert. 1969. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. *Archives of General Psychiatry* 20, 2 (Feb. 1969), 246. doi:10.1001/archpsyc.1969.01740140118016

[40] Shun-Zheng Yu. 2010. Hidden Semi-Markov Models. *Artificial Intelligence* 174, 2 (Feb. 2010), 215–243. doi:10.1016/j.artint.2009.11.011

A Robustness Analysis

Constraint relaxation. Table 5 shows the effect of relaxing the rare-transition penalties (increasing ϵ) on AccuSleep Mouse EEG/EMG using U-Sleep. With full constraints ($\epsilon = 0.001$), TVR is 0.2% and REM latency MAE is 8.7 min. Relaxing constraints increases both TVR and downstream error, confirming that constraints are both valid and necessary.

Table 5: Effect of relaxing rare-transition constraints (U-Sleep on AccuSleep Mouse EEG/EMG).

Configuration	TVR (%)	Acc (%)	REM Lat. MAE
Full constraints (ϵ=0.001)	0.2±0.1	92.2±0.4	8.7±1.6
Relaxed W→R (ϵ=0.1)	1.8±0.4	91.8±0.9	14.2±2.8
Relaxed R→N (ϵ=0.1)	2.1±0.5	91.5±1.0	12.8±2.5
No transition constraints	3.9±0.7	91.6±0.9	21.3±3.3

Sensitivity to $d_{\min}$ parameters. Table 6 shows the effect of varying minimum-duration thresholds. Physiology-based values yield optimal results; aggressive settings over-smooth the hypnogram while loose settings under-constrain fragmentation.

Table 6: Sensitivity to minimum-duration parameters (U-Sleep on AccuSleep Mouse EEG/EMG).

$d_{\min}$ Setting	TVR (%)	FI	Acc (%)
Physiology-based	0.2±0.1	0.05±0.01	92.2±0.4
2× (too aggressive)	0.2±0.1	0.03±0.01	90.8±1.1
0.5× (too loose)	0.3±0.1	0.09±0.02	92.0±0.8
No duration constraint	0.4±0.2	0.13±0.02	91.8±0.9

Subpopulation analysis. Table 7 reports results on SHHS stratified by age group and OSA severity. Constraints provide consistent benefit across all subpopulations, suggesting generalization across demographic and clinical subgroups, though subgroup sample sizes (n=5–12) limit formal statistical confirmation.

Sensitivity to soft-penalty weight λ. Table 8 shows the effect of varying λ on U-Sleep using AccuSleep Mouse EEG/EMG. TVR drops sharply from λ=0.1 to λ=0.5, then plateaus; accuracy peaks at λ=0.5 with a slight drop at λ=2.0 due to over-regularization.

Table 7: Subpopulation analysis on SHHS (U-Sleep backbone). Values are mean ± std across subjects within each subgroup.

Subgroup	N	TVR (%)		Acc (%)	
		Base	+Ours	Base	+Ours
Age 40–55	8	3.5±0.4	0.2±0.1	81.2±0.7	83.1±0.6
Age 55–70	10	3.8±0.5	0.2±0.1	80.5±0.8	82.4±0.7
Age 70+	7	4.1±0.6	0.2±0.1	79.8±0.9	81.8±0.8
No OSA	12	3.4±0.3	0.2±0.1	81.8±0.6	83.5±0.5
Mild OSA	8	3.9±0.5	0.2±0.1	80.2±0.8	82.1±0.7
Mod/Sev OSA	5	4.5±0.7	0.3±0.1	78.4±1.1	80.6±0.9

Table 8: Sensitivity to soft-penalty weight λ (U-Sleep on AccuSleep Mouse EEG/EMG).

λ	TVR (%)	FI	Acc (%)
0.1	2.1±0.5	0.09±0.02	91.8±0.9
0.25	1.1±0.3	0.07±0.01	92.0±0.8
0.5 (default)	**0.2**±0.1	**0.05**±0.01	**92.2**±0.8
1.0	0.2±0.1	0.05±0.01	91.9±0.9
2.0	0.3±0.1	0.05±0.01	91.4±1.0

B Backbone Architectures

We provide architectural details for the six evaluated backbones:

- **AccuSleep** [1]: Compact CNN with two convolutional layers, ReLU activations, and max-pooling on EEG/EMG spectrograms.
- **DeepSleepNet** [32]: CNN–RNN hybrid with multi-scale convolutional filters and bidirectional LSTM sequence learning.
- **SeqSleepNet** [22]: Hierarchical RNN with attention-based epoch processing and sequence-level recurrence.
- **AttnSleep** [5]: Multi-resolution CNN with adaptive feature recalibration and multi-head self-attention.
- **SleepTransformer** [23]: Transformer-based architecture with epoch- and sequence-level self-attention.
- **U-Sleep** [21]: Fully convolutional U-Net architecture trained on large-scale multi-center data.

C Training Configuration

Table 9 lists the training hyperparameters for each backbone, following the original authors' recommended settings.

D Per-Class F1 Scores

Table 10 reports per-class F1 scores for all six backbones on AccuSleep Mouse EEG/EMG, showing that StageGuard improves F1 across all classes, with the largest gains on REM (the most challenging class due to its shorter bouts and rarity).

E Empirical Transition Probabilities

Table 11 reports the empirical transition probability matrices $\hat{\pi}$ estimated from training-set labels for each three-state dataset. These are used as $\epsilon_{ss'}$ for typical transitions in Equation (6); rare transitions (marked with †) are overridden with $\epsilon_{ss'} = 0.001$.

Table 12: Sleep architecture statistics: MAE between predicted and expert-derived values (SHHS, SleepTransformer backbone). Statistical significance via Wilcoxon signed-rank test: $^{***}p < 0.001$, $^{**}p < 0.01$. Values are mean ± std; n=25.

Statistic	Baseline	+Ours	Improv.
Total Sleep Time (min)	22.7±3.4	9.1±1.8***	60%
Sleep Efficiency (%)	4.6±0.7	1.8±0.3***	61%
REM Latency (min)	26.1±4.8	10.4±2.1***	60%
WASO (min)	19.8±3.1	7.4±1.5***	63%
Mean NREM Bout (epochs)	3.8±0.6	1.2±0.3**	68%
Mean REM Bout (epochs)	2.1±0.4	0.7±0.2**	67%
Awakenings (count)	12.4±2.0	3.8±0.8***	69%

Table 9: Training configuration per backbone. All backbones use these settings for both baseline and +Ours variants (the only difference is the additional soft penalty term in Equation 5).

Backbone	Optimizer	LR	Batch	Epochs	Patience
AccuSleep	Adam	1e-3	32	50	10
DeepSleepNet	Adam	1e-4	32	200	20
SeqSleepNet	Adam	1e-3	32	100	15
AttnSleep	Adam	1e-4	128	100	15
SleepTransformer	AdamW	5e-4	64	150	20
U-Sleep	Adam	1e-3	64	60	10

Table 10: Per-class F1 scores on AccuSleep Mouse EEG/EMG (mean ± std across n=16 LOSO folds). Best per class in bold.

	Baseline			+Ours		
Backbone	Wake	NREM	REM	Wake	NREM	REM
AccuSleep	.85±.02	.91±.01	.78±.03	.88±.02	.93±.01	.83±.02
DeepSleepNet	.87±.02	.92±.01	.80±.02	.89±.01	.93±.01	.84±.02
SeqSleepNet	.88±.01	.92±.01	.81±.02	.90±.01	.93±.008	.85±.02
AttnSleep	.88±.01	.93±.008	.82±.02	.90±.01	.94±.007	.86±.02
SleepTransformer	.89±.01	.93±.008	.82±.02	.89±.01	.93±.008	.85±.02
U-Sleep	.89±.01	.93±.007	.83±.02	**.91**±.01	**.94**±.006	**.87**±.01

Table 11: Empirical transition probabilities $\hat{\pi}_{ss'}$ estimated from training labels. † Rare transitions overridden with ϵ=0.001 in the decoder.

Dataset	From \ To	Wake	NREM	REM
AccuSleep EEG/EMG	Wake	.912	.085	.003†
	NREM	.052	.831	.117
	REM	.078	.018†	.904
SHHS	Wake	.883	.113	.004†
	NREM	.064	.806	.130
	REM	.097	.026†	.877
SLEEPBRL	Wake	.871	.124	.005†
	NREM	.071	.793	.136
	REM	.103	.031†	.866

F Additional Downstream Results

Table 12 reports downstream sleep architecture MAE for the SleepTransformer backbone on SHHS, confirming that the improvements reported in the main text (Table 3) generalize beyond the U-Sleep/AccuSleep pairing.