Stage Prototypes for Transition-Aware Temporal Modeling in Single-Channel Sleep Staging

Anonymous Authors[1]

Abstract

Automatic sleep staging from polysomnography (PSG) remains challenging under subject-independent evaluation, particularly around stage transitions where EEG patterns are inherently ambiguous. We present **SleepTSP**, a temporal model that incorporates boundary-aware attention within a sliding-window framework for single-channel EEG. A learned boundary signal modulates temporal interactions, while stage-level representations provide global contextual anchoring. On the EDF-20 benchmark, we observe that errors are concentrated near stage transitions, and show that explicitly modeling transition structure improves robustness in these regions without increasing inference complexity. These findings highlight the importance of transition-aware temporal modeling for stable sleep staging.

1. Introduction

Sleep is a fundamental biological process essential for physical health, cognitive function, and emotional regulation. Extensive evidence links insufficient or disrupted sleep to increased risks of cardiovascular disease, metabolic disorders, depression, and mortality (Cappuccio et al., 2010; Harvey, 2011; Irwin, 2015; Medic et al., 2017). Sleep supports memory consolidation and synaptic homeostasis (Diekelmann & Born, 2010; Rasch & Born, 2013; Tononi & Cirelli, 2014), while sleep deprivation affects circadian and immune-related pathways (Möller-Levet et al., 2013; Irwin, 2015).

Clinical sleep assessment relies on polysomnography (PSG) with manual scoring following AASM guidelines. Despite its validity, manual scoring is labor-intensive and subject to variability, especially for transitional stages such as N1 (Rosenberg & Hout, 2013; Berry et al., 2017; Lee et al., 2022). This motivates automatic sleep staging using electroencephalography (EEG), which captures stage-specific neural dynamics at low cost (Supratak et al., 2017; Faust et al., 2019).

Early approaches relied on hand-crafted features (Aboalayon et al., 2016), while deep learning enabled end-to-end modeling from raw EEG. CNN-based models achieved strong performance by learning hierarchical representations (Tsinalis et al., 2016; Supratak et al., 2017; Yildirim et al., 2019). Subsequent work incorporated temporal modeling via CNN–RNN architectures (Phan et al., 2019; Michielli et al., 2019; Seo et al., 2020; Li & Gao, 2023) and sequence-to-sequence formulations (Phan et al., 2023; Lee et al., 2023), while attention and Transformer models capture temporal context without recurrence (Eldele et al., 2021; Phan et al., 2022; Pradeepkumar et al., 2024). However, many approaches increase complexity and remain sensitive under subject-independent evaluation.

We revisit sleep staging from the perspective of transition-aware temporal modeling under sliding-window inference. We propose *SleepTSP*, combining CNN-based epoch encoding with boundary-aware attention and stage-level memory tokens. The boundary-aware mechanism modulates temporal interactions based on estimated transition likelihoods, accounting for the structured and non-stationary nature of sleep stage transitions. Through ablation and transition-centric analyses, we show improved robustness near stage boundaries, while stage-level memory supports stability in steady segments. We evaluate on Sleep-EDF (EDF-20 and EDF-78) and ISRUC-Sleep under subject-independent protocols, with ablation studies conducted on EDF-20.

Our contributions are summarized as follows:

- A transition-aware sleep staging architecture incorporating boundary information into attention-based temporal modeling.
- Stage-level memory tokens for global contextual anchoring and temporal consistency.
- Transition-centric analyses clarifying model behavior under subject-independent evaluation.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Related Work

Deep learning has become the dominant approach for automatic sleep staging, replacing hand-crafted feature pipelines. CNN-based models learn hierarchical EEG representations (Supratak et al., 2017), while temporal modeling is commonly introduced using recurrent or sequence-based architectures (Phan et al., 2019; Seo et al., 2020). More recently, attention and Transformer-based models have been explored to capture temporal context without recurrence (Eldele et al., 2021; Phan et al., 2022).

Sleep staging follows structured transition patterns, and earlier work modeled these dynamics using probabilistic approaches such as hidden Markov models (Ghimatgar et al., 2019). However, modern deep learning methods typically treat transitions implicitly, relying on recurrent states or uniform attention.

Public benchmarks such as Sleep-EDF and ISRUC-Sleep are widely used for subject-independent evaluation (Kemp et al., 2000; Khalighi et al., 2016). Prior work shows that competitive performance can be achieved using single-channel EEG (Supratak et al., 2017; Phan et al., 2019). Recent studies further explore single-modality inference settings, including SleepSMC (Ma et al., 2025) and MixSleepNet (Ji et al., 2024), highlighting challenges in subject-independent generalization.

In this work, we focus on explicitly modeling transition structure within a deep learning framework to improve robustness near stage boundaries.

3. Method

3.1. Overview

We formulate sleep staging as sequence modeling over short temporal windows of single-channel EEG. Each recording is segmented into 30-second epochs, and fixed-length windows of $L = 20$ epochs are constructed with stride $S = 10$. The model predicts per-epoch labels using only the center positions of each window to avoid overlapping supervision.

Our model, *SleepTSP*, consists of three components: (i) an epoch-wise convolutional encoder, (ii) boundary-aware self-attention for transition-sensitive temporal modeling, and (iii) lightweight stage-level memory tokens for global contextual aggregation.

3.2. Epoch Encoding

Each epoch x_t is independently encoded using a 1D convolutional network to produce a representation $h_t \in \mathbb{R}^d$. Applying the encoder to a window yields a sequence

$$\mathbf{H} = [h_1, \ldots, h_L] \in \mathbb{R}^{L \times d}. \tag{1}$$

3.3. Boundary-aware Self-Attention

To explicitly model sleep stage transitions, we introduce a boundary-aware attention mechanism that modulates temporal interactions based on estimated transition likelihoods.

A boundary score $b_t \in [0, 1]$ is computed from adjacent feature differences:

$$b_t = \sigma\big(g(h_t - h_{t-1})\big). \tag{2}$$

These scores define an additive attention bias:

$$\Delta_{ij} = -\frac{|i - j|}{\tau} - \lambda(b_i + b_j), \tag{3}$$

which is applied to scaled dot-product attention. This bias suppresses interactions across likely stage boundaries while preserving local context within homogeneous segments.

3.4. Stage-level Memory

We introduce a small set of learnable stage tokens that aggregate global context within each window. These tokens attend to epoch representations and are then used to refine epoch features via cross-attention. This mechanism provides soft stage-level anchoring without imposing explicit transition constraints.

3.5. Training

The model is trained using cross-entropy applied only to center indices $\mathcal{C}$:

$$\mathcal{L} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{CE}(\mathbf{Z}_{:,c,:}, \mathbf{y}_{:,c}). \tag{4}$$

At inference, center predictions are tiled to reconstruct the full hypnogram.

4. Results

4.1. Experimental Protocol

We follow subject-independent evaluation using a single EEG channel. EDF-20 and ISRUC Subgroup 3 use leave-one-subject-out (LOSO), while EDF-78 uses subject-independent 10-fold cross-validation. Performance is reported using accuracy (ACC), macro-F1 (MF1), and Cohen's kappa (κ).

4.2. Overall Performance

Table 1 reports performance across EDF-20, EDF-78, and ISRUC Subgroup 3. SleepTSP achieves competitive performance across all datasets under subject-independent evaluation. Overall improvements in MF1 and κ are modest, as

Table 1. Single-channel sleep staging performance across benchmarks under subject-independent evaluation, where the best ensemble size M is selected per dataset based on validation performance.

Dataset	Method	ACC	MF1	κ	W	N1	N2	N3	REM
EDF-20	DeepSleepNet (Supratak et al., 2017)	82.0	76.9	–	84.7	46.6	85.9	84.8	82.4
	IITNet (Seo et al., 2020)	83.9	77.6	0.78	87.7	43.4	87.7	86.7	82.5
	AttnSleep (Eldele et al., 2021)	84.4	78.1	0.79	89.7	42.6	88.8	90.2	79.0
	SeriesSleepNet (s=10) (Lee et al., 2023)	84.1	78.9	0.78	89.1	47.5	87.8	88.7	82.6
	Ours (best-M=9)	83.8	79.2	0.78	87.8	50.5	86.6	85.3	85.8
EDF-78	SeqSleepNet (Phan et al., 2019)	82.6	76.4	0.76	92.2	47.8	84.9	77.2	79.9
	AttnSleep (Eldele et al., 2021)	81.3	75.1	0.74	92.0	42.0	85.0	82.1	74.2
	SleepTransformer (Phan et al., 2022) (w/o pretrained)	81.4	74.3	0.74	91.7	40.4	84.3	77.9	77.2
	SleepSMC (EEG) (Ma et al., 2025)	79.6	72.2	0.72	92.0	35.8	82.8	79.9	70.3
	Ours (best-M=7)	81.9	77.0	0.75	92.8	50.1	83.2	76.2	82.5
ISRUC-3	MixSleepNet (EEG-only) (Ji et al., 2024)	76.8	73.5	0.70	85.6	47.1	77.6	88.2	69.2
	Causal-aware (Hu et al., 2025) (F3-A2)	75.3	66.9	–	89.7	47.6	75.9	86.5	34.5
	(C3-A2)	77.2	71.6	–	92.5	47.6	76.0	88.6	53.5
	SleepSMC (EEG) (Ma et al., 2025)	76.5	74.0	0.70	88.8	50.7	74.7	86.4	69.3
	Ours (best-M=4)	78.2	76.7	0.72	87.4	53.3	77.1	87.7	78.1

evaluation is dominated by steady segments with limited temporal ambiguity.

Across datasets, N1 remains the most challenging stage, consistent with its transitional nature.

4.3. Ablation Study

Table 2 shows that removing either boundary-aware attention or stage tokens leads to minor changes in overall metrics, while removing both results in substantial degradation, indicating complementary effects. Per-class results show the largest degradation for N1 and REM, suggesting that the proposed components primarily affect ambiguous and transition-related predictions.

Table 2. Ablation results on EDF-20 under a fixed ensemble setting ($M = 9$). Overall epoch-level performance and per-class F1 scores are reported. The ensemble size is selected based on the full model and reused for all variants.

Var	ACC	MF1	κ	W	N1	N2	N3	R
Full	83.8	79.2	.780	87.8	50.5	86.6	85.3	85.8
noS	83.8	79.2	.781	88.2	50.2	86.7	85.4	85.5
noB	83.7	79.0	.779	88.4	50.5	86.5	85.1	84.7
noBS	80.8	74.9	.741	86.5	39.9	85.8	84.5	77.7

4.4. Transition-centric Analysis

4.4.1. Window-type analysis

We stratify windows into steady and transition subsets. As shown in Table 3 and Section A.2.1, performance is consistently lower in transition windows, confirming that errors are concentrated near true stage changes. SleepTSP achieves the highest macro-F1 in transition regions with reduced variance compared to ablation variants.

Table 3. Macro-F1 (%) on EDF-20 stratified by window type using the best single checkpoint ($M = 1$). Values are reported as mean (std) across subject-independent splits.

Variant	All	Steady	Transition
Full	77.8 (6.1)	77.6 (8.7)	68.2 (6.2)
noS	77.3 (7.1)	77.3 (8.4)	67.3 (7.0)
noB	77.2 (6.7)	77.8 (7.3)	66.5 (7.2)
noBS	73.2 (7.9)	73.2 (7.5)	65.0 (8.2)

Directional asymmetry. Transition difficulty is direction-dependent (e.g., N2→N1 vs. N1→N2), and this asymmetry persists across ablations, supporting the view that transitions are structured rather than symmetric. A detailed directional breakdown is included in Section A.1.

4.4.2. Distance to transition

Figure 1 shows macro-F1 as a function of distance d to the nearest transition. Performance is lowest near transitions ($d = 0, 1$) and rapidly recovers within a few epochs. SleepTSP consistently improves near-boundary performance and exhibits faster recovery than variants without boundary-aware attention.

At larger distances, performance differences diminish but do not fully vanish, particularly for the noBS variant, which remains consistently worse across d.

These results suggest that boundary-aware modulation primarily accelerates recovery from transition-induced ambiguity rather than uniformly improving stable regions.

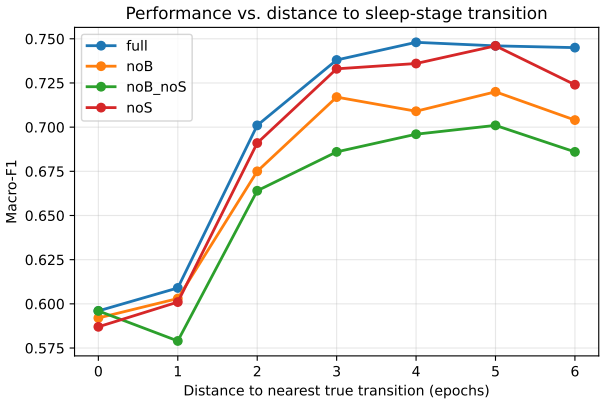


Figure 1. Macro-F1 as a function of the distance d (in epochs) to the nearest true sleep-stage transition on EDF-20. Performance is lowest near transitions and recovers rapidly within 2–3 epochs, with the full model exhibiting faster recovery and higher near-boundary performance than ablation variants.

4.4.3. Temporal Stability

Beyond per-epoch accuracy, we analyze temporal coherence of predicted hypnograms using: (i) *change rate*, the percentage of adjacent epochs with a stage change; (ii) *singleton rate*, the percentage of runs of length one; and (iii) *run length*, the mean length of continuous stage segments. All values are computed on reconstructed hypnograms and reported as mean (std) across folds ($M = 1$).

Table 4. Temporal stability metrics on EDF-20 ($M = 1$). Values are reported as mean (std) across folds. GT denotes statistics computed from the ground-truth hypnograms.

Variant	ChangeRate (%)	Singleton (%)	RunLength
GT	10.42 (2.10)	32.68 (5.33)	10.06 (2.83)
Full	12.52 (2.53)	42.32 (3.94)	8.27 (1.65)
noS	12.81 (2.92)	42.38 (4.39)	8.16 (1.82)
noB	14.76 (3.12)	46.23 (3.63)	7.05 (1.54)
noB_noS	20.96 (4.79)	51.30 (2.46)	5.00 (1.17)

As shown in Table 4, removing boundary-aware attention increases change and singleton rates while shortening stage runs. The full model produces more stable predictions closer to ground-truth statistics without post-hoc smoothing. Although overall MF1 differences are modest, transition-localized and temporal stability metrics consistently favor boundary-aware modulation, with the largest degradation when both components are removed.

5. Discussion

We examined transition-aware temporal modeling for single-channel sleep staging under a sliding-window, center-only inference setting. Across datasets, results consistently show that errors are concentrated near true sleep-stage transitions, where EEG patterns are inherently ambiguous. While overall gains in MF1 and κ are modest, transition-centric analyses reveal clearer improvements, highlighting the importance of evaluating temporal models beyond aggregate metrics.

Why are aggregate gains modest? Global metrics are dominated by steady segments, where classification is relatively easier. As a result, improvements localized near transitions are diluted when averaged over all epochs. This suggests that transition-aware evaluation provides a more sensitive measure of temporal modeling quality.

Stage prototypes as stabilizing context. Stage tokens in SleepTSP are learned end-to-end and interact with epoch representations via bidirectional attention, producing window-conditioned stage prototypes. Individually, stage tokens have limited impact on global metrics, but when combined with boundary-aware modulation they help prevent over-segmentation, as reflected by the larger degradation when both components are removed. This suggests that stage prototypes act as soft anchors that stabilize predictions within steady segments, while boundary-aware attention primarily targets mixed-stage context near transitions.

Limitations. Improvements in transition-window MF1 remain modest, and statistical significance is mainly observed when both components are removed. The boundary signal is learned without explicit supervision, and its alignment with true physiological transitions is not guaranteed. In addition, our analysis is limited to single-channel EEG and a fixed window configuration (L=20, S=10), and further validation across different settings is needed.

Future work. Future directions include incorporating transition-aware supervision (e.g., weak labels from hypnogram changes) to better calibrate boundary estimation, as well as evaluating the approach under different window configurations and datasets. Transition-specific evaluation protocols may also provide more sensitive benchmarks for temporal modeling in sleep staging.

6. Conclusion

We present SleepTSP, a temporal model for single-channel EEG staging that combines epoch-wise convolutional encoding, boundary-aware attention, and stage-level prototypes under a sliding-window framework. Across Sleep-EDF variants and ISRUC-3, we report competitive performance, while analyses on EDF-20 show that errors concentrate near stage transitions and that transition-aware modeling is best assessed through localized analysis. While overall gains are modest, the proposed components improve transition robustness and temporal stability, reducing over-segmentation without smoothing. These findings highlight the importance of explicitly modeling transition structure in sleep staging.

References

Aboalayon, K. A. I., Faezipour, M., Almuhammadi, W. S., and Moslehpour, S. Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation. *Entropy*, 18(9):272, 2016. ISSN 1099-4300. doi: 10.3390/e18090272.

Berry, R. B., R, B., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., Troester, M., and Vaughn, B. V. AASM scoring manual updates for 2017 (version 2.4), 2017.

Cappuccio, F. P., D'Elia, L., Strazzullo, P., and Miller, M. A. Sleep duration and all-cause mortality: A systematic review and meta-analysis of prospective studies. *Sleep*, 33(5):585–592, 05 2010. ISSN 0161-8105. doi: 10.1093/sleep/33.5.585.

Diekelmann, S. and Born, J. The memory function of sleep. *Nature reviews. Neuroscience*, 11(2):114–126, 2010. doi: 10.1038/nrn2762.

Eldele, E., Chen, Z., Liu, C., Wu, M., Kwoh, C.-K., Li, X., and Guan, C. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021. doi: 10.1109/TNSRE.2021.3076234.

Faust, O., Razaghi, H., Barika, R., Ciaccio, E. J., and Acharya, U. R. A review of automated sleep stage scoring based on physiological signals for the new millennia. *Computer Methods and Programs in Biomedicine*, 176:81–91, 2019. ISSN 0169-2607. doi: 10.1016/j.cmpb.2019.04.032.

Ghimatgar, H., Kazemi, K., Helfroush, M. S., and Aarabi, A. An automatic single-channel EEG-based sleep stage scoring method based on hidden markov model. *Journal of Neuroscience Methods*, 324:108320, 2019. ISSN 0165–0270. doi: 10.1016/j.jneumeth.2019.108320.

Harvey, A. G. Sleep and circadian functioning: Critical mechanisms in the mood disorders? *Annual Review of Clinical Psychology*, 7(2011):297–319, 2011. ISSN 1548-5951. doi: 10.1146/annurev-clinpsy-032210-104550.

Hu, Y., Yang, X., Xu, Y., and Sun, J. Causal-aware reliability assessment of single-channel EEG for transformer-based sleep staging. *Frontiers in Neuroscience*, 19, 2025. ISSN 1662-453X. doi: 10.3389/fnins.2025.1670124.

Irwin, M. R. Why sleep is important for health: A psychoneuroimmunology perspective. *Annual Review of Psychology*, 66(2015):143–172, 2015. ISSN 1545-2085. doi: 10.1146/annurev-psych-010213-115205.

Ji, X., Li, Y., Wen, P., Barua, P., and Acharya, U. R. MixSleepNet: A multi-type convolution combined sleep stage classification model. *Computer Methods and Programs in Biomedicine*, 244:107992, 2024. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2023.107992.

Kemp, B., Zwinderman, A., Tuk, B., Kamphuisen, H., and Oberye, J. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Khalighi, S., Sousa, T., Santos, J. M., and Nunes, U. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016. ISSN 0169–2607. doi: 10.1016/j.cmpb.2015.10.013.

Lee, M., Kwak, H. G., Kim, H. J.and Won, D. O., and Lee, S. W. SeriesSleepNet: an EEG time series model with partial data augmentation for automatic sleep stage scoring. *Frontiers in Physiology*, 14:1188678, 2023. doi: 10.3389/fphys.2023.1188678.

Lee, Y. J., Lee, J. Y., Cho, J. H., and Choi, J. H. Interrater reliability of sleep stage scoring: a meta-analysis. *Journal of Clinical Sleep Medicine*, 18(1):193–202, 2022. doi: 10.5664/jcsm.9538.

Li, W. and Gao, J. Automatic sleep staging by a hybrid model based on deep 1D-ResNet-SE and LSTM with single-channel raw EEG signals. *PeerJ Computer Science*, 9:e1561, 2023. doi: 10.7717/peerj-cs.1561.

Ma, S., Zhang, Y., Chen, Y., Wang, H., Jin, Y., Zhang, W., and Jia, Z. SleepSMC: Ubiquitous sleep staging via supervised multimodal coordination. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=B5VEi5d3p2.

Medic, G., Wille, M., and Hemels, M. E. Short- and long-term health consequences of sleep disruption. *Nature and Science of Sleep*, 9:151–161, 2017. doi: 10.2147/NSS.S134864.

Michielli, N., Acharya, U. R., and Molinari, F. Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. *Computers in Biology and Medicine*, 106:71–81, 2019. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2019.01.013.

Möller-Levet, C. S., Archer, S. N., Bucca, G., Laing, E. E., Slak, A., Kabiljo, R., Lo, J. C. Y., Santhi, N., von Schantz, M., Smith, C. P., and Dijk, D.-J. Effects of insufficient sleep on circadian rhythmicity and expression amplitude of the human blood transcriptome. *Proceedings of the*

National Academy of Sciences, 110(12):E1132–E1141, 2013. doi: 10.1073/pnas.1217154110.

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and De Vos, M. 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):400–410, 2019. doi: 10.1109/TNSRE.2019.2896659.

Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A., and De Vos, M. SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8): 2456–2467, 2022. doi: 10.1109/TBME.2022.3147187.

Phan, H., Lorenzen, K. P., Heremans, E., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., Baumert, M., Mikkelsen, K. B., and De Vos, M. L-SeqSleepNet: Whole-cycle long sequence modeling for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 27(10): 4748–4757, 2023. doi: 10.1109/JBHI.2023.3303197.

Pradeepkumar, J., Anandakumar, M., Kugathasan, V., Suntharalingham, D., Kappel, S. L., De Silva, A. C., and Edussooriya, C. U. S. Toward interpretable sleep stage classification using cross-modal transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:2893–2904, 2024. doi: 10.1109/TNSRE.2024.3438610.

Rasch, B. and Born, J. About sleep's role in memory. *Physiological Reviews*, 93(2):681–766, 2013. doi: 10.1152/physrev.00032.2012. PMID: 23589831.

Rosenberg, R. S. and Hout, S. V. The American Academy of Sleep Medicine inter-scorer reliability program: Sleep stage scoring. *Journal of Clinical Sleep Medicine*, 09(01): 81–87, 2013. doi: 10.5664/jcsm.2350.

Seo, H., Back, S., Lee, S., Park, D., Kim, T., and Lee, K. Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG. *Biomedical Signal Processing and Control*, 61:102037, 2020. ISSN 1746-8094. doi: 10.1016/j.bspc.2020.102037.

Supratak, A., Dong, H., Wu, C., and Guo, Y. 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. doi: 10.1109/TNSRE.2017.2721116.

Tononi, G. and Cirelli, C. Sleep and the price of plasticity: From synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, 81(1):12–34, 2014. doi: 10.1016/j.neuron.2013.12.025.

Tsinalis, O., Matthews, P. M., and Guo, Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Annals of Biomedical Engineering*, 44:1587–1597, 2016. doi: 10.1007/s10439-015-1444-y.

Yildirim, O., Baloglu, U. B., and Acharya, U. R. A deep learning model for automated sleep stages classification using PSG signals. *International Journal of Environmental Research and Public Health*, 16(4):599, 2019. doi: 10.3390/ijerph16040599.

A. Additional Analyses

This appendix provides additional analyses that complement the main experimental results. These analyses focus on fine-grained transition behavior, variance characteristics, and statistical properties of model predictions, and are intended to provide deeper insight into how boundary-aware temporal modeling and stage-level memory influence performance. All results are reported on EDF-20 under subject-independent evaluation unless otherwise stated.

A.1. Transition Taxonomy Analysis

Sleep stage transitions are not uniformly distributed, and certain transition types (e.g., N2↔N3 or N1↔N2) occur substantially more frequently than others. To better understand model behavior at a fine-grained level, we analyze classification performance for the most common transition pairs.

Table 5 reports window-level accuracy for the top-10 most frequent transition types, together with the number of transition windows observed in EDF-20. Results are shown for the full model and ablation variants.

Table 5. Top-10 most frequent sleep stage transition pairs on EDF-20 and corresponding window-level accuracy. Results are reported as mean (std) using single-checkpoint inference ($M = 1$).

Transition	#Windows	Full	noS	noB	noB_noS
N2→N3	349	70.3 (23.0)	70.3 (23.2)	71.8 (21.2)	70.8 (22.8)
N3→N2	220	73.8 (20.2)	75.0 (19.5)	74.1 (19.6)	75.6 (18.6)
N2→N1	193	80.8 (17.2)	81.0 (17.1)	79.6 (18.0)	78.2 (18.0)
N1→N2	181	73.8 (19.6)	72.9 (20.0)	72.1 (21.4)	67.1 (21.4)
W→N1	176	68.9 (23.2)	65.7 (23.6)	67.3 (23.2)	63.8 (23.0)
N2→W	138	75.0 (21.2)	73.8 (20.5)	73.4 (22.2)	72.8 (20.2)
N2→REM	136	71.3 (23.4)	69.0 (24.2)	69.2 (22.1)	70.3 (21.9)
REM→N1	132	73.2 (19.4)	71.1 (20.1)	71.5 (21.1)	66.7 (21.1)
REM→N2	96	63.1 (23.3)	61.6 (25.7)	61.5 (24.4)	61.9 (22.6)
REM→W	69	63.6 (24.4)	57.7 (28.5)	57.0 (28.4)	63.5 (21.6)

A.2. Ablation Statistics

A.2.1. Window-Type Metrics with Sample Counts

Table 6 reports accuracy (ACC), macro-F1 (MF1), and Cohen's κ stratified by window type, evaluated using the best single checkpoint ($M = 1$). Values are reported as mean (std) across folds, along with the number of evaluated center epochs.

Table 6. Window-type performance on EDF-20 ($M = 1$). Values are reported as mean (std) across folds. N_{epochs} denotes the number of evaluated center epochs aggregated across folds.

Variant	Subset	ACC (%)	MF1 (%)	κ (%)	N_{folds}	N_{epochs}
full	all_windows	83.10 (7.17)	77.79 (6.07)	76.93 (9.15)	20	42140
full	steady_windows	91.59 (8.07)	77.62 (8.68)	88.17 (10.76)	20	24150
full	transition_windows	71.82 (6.13)	68.25 (6.25)	60.93 (7.95)	20	17990
noS	all_windows	82.67 (7.56)	77.25 (7.08)	76.32 (9.84)	20	42140
noS	steady_windows	91.31 (8.75)	77.29 (8.46)	87.81 (11.73)	20	24150
noS	transition_windows	71.21 (6.01)	67.27 (7.03)	59.88 (8.35)	20	17990
noB	all_windows	83.04 (6.24)	77.18 (6.69)	76.63 (8.49)	20	42140
noB	steady_windows	91.92 (6.91)	77.76 (7.31)	88.52 (9.55)	20	24150
noB	transition_windows	71.21 (5.64)	66.55 (7.20)	59.56 (8.38)	20	17990
noB_noS	all_windows	80.02 (7.71)	73.19 (7.89)	72.69 (9.85)	20	42140
noB_noS	steady_windows	87.46 (8.99)	73.25 (7.51)	82.51 (11.83)	20	24150
noB_noS	transition_windows	69.95 (6.27)	64.97 (8.18)	58.11 (8.45)	20	17990

A.2.2. Paired Statistical Tests for Transition-Window MF1

Table 7 reports paired comparisons on transition-window MF1 between the full model and each ablation variant. We report mean difference (full minus ablation), 95% confidence intervals, paired effect size (Cohen's d_z), and p-values from both paired t-test and Wilcoxon signed-rank test.

Table 7. Paired statistical tests on *transition-window* MF1 (full minus ablation; $M = 1$). Values are computed across $N_{\text{folds}} = 20$ folds.

Comparison	Mean diff	95% CI (lo)	95% CI (hi)	Cohen's d_z	p (paired t)	p (Wilcoxon)
full − noS	0.00982	−0.00189	0.02129	0.361	0.123	0.189
full − noB	0.01703	0.00165	0.03329	0.466	0.0509	0.0759
full − noB_noS	0.03289	0.01641	0.04898	0.866	0.00103	0.00169

A.2.3. Paired Statistical Tests for Predicted Stage-Change Rate

Table 8 reports paired comparisons on the predicted stage-change rate. Mean differences are reported in *fraction* units (e.g., 0.02 corresponds to 2 percentage points).

Table 8. Paired statistical tests on predicted stage-change rate (full minus ablation; $M = 1$). Values are computed across $N_{\text{folds}} = 20$ folds.

Comparison	Mean diff	95% CI (lo)	95% CI (hi)	Cohen's d_z	p (paired t)	p (Wilcoxon)
full − noS	−0.00289	−0.01041	0.00401	−0.172	0.451	0.674
full − noB	−0.02236	−0.03652	−0.01291	−0.787	0.00228	0.000004
full − noB_noS	−0.08433	−0.09854	−0.07048	−2.57	0.0000000005	0.000002

B. Extended Performance Analysis

This appendix provides supplementary diagnostic analyses that complement the main quantitative results. We focus on aggregated confusion matrices and representative hypnogram reconstructions to illustrate class-wise error modes, temporal behavior, and the effect of checkpoint ensembling across datasets.

B.1. Why ISRUC-1 Is Included in the Appendix

ISRUC-Sleep is a well-established benchmark that is widely used for reporting subject-independent sleep staging performance. In our experimental environment, however, we encountered substantial data integrity issues during dataset preparation. For a large fraction of subjects, preprocessed recordings yielded degenerate statistics (e.g., mean $= 0$ and std $= 0$ across windows), indicating corrupted or non-informative signals after loading or preprocessing. As a result, only a limited subset of subjects could be reliably used in our pipeline (e.g., 16 uncorrupted subjects out of 100 attempted).

To avoid over-interpreting results derived from a reduced and potentially non-representative subset, we report ISRUC-1 results as supplementary evidence in the appendix and focus the main results on Sleep-EDF benchmarks.

B.2. Example Hypnogram Reconstructions

We provide representative hypnogram reconstructions to illustrate temporal behavior and typical failure modes beyond summary metrics at the epoch-level. All examples are reconstructed from center-only predictions, ensuring consistency with the training and evaluation protocol (Figure 2, Figure 3, Figure 4).

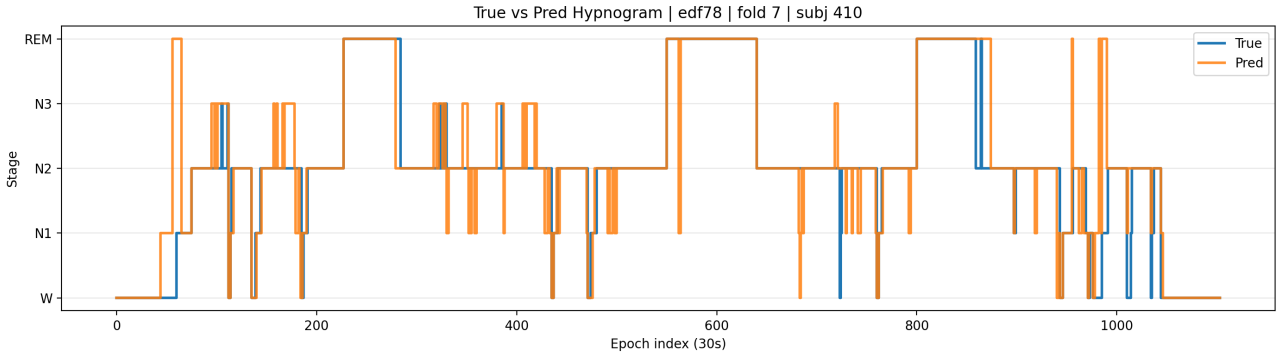


Figure 2. **EDF-78.** Example hypnogram reconstruction on EDF-78. Ground truth and predicted stage sequences are shown for a full-night recording.

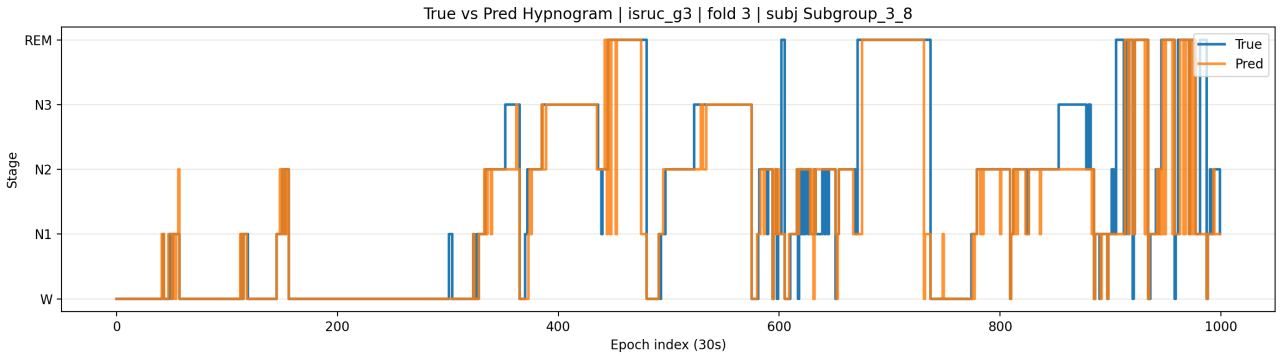
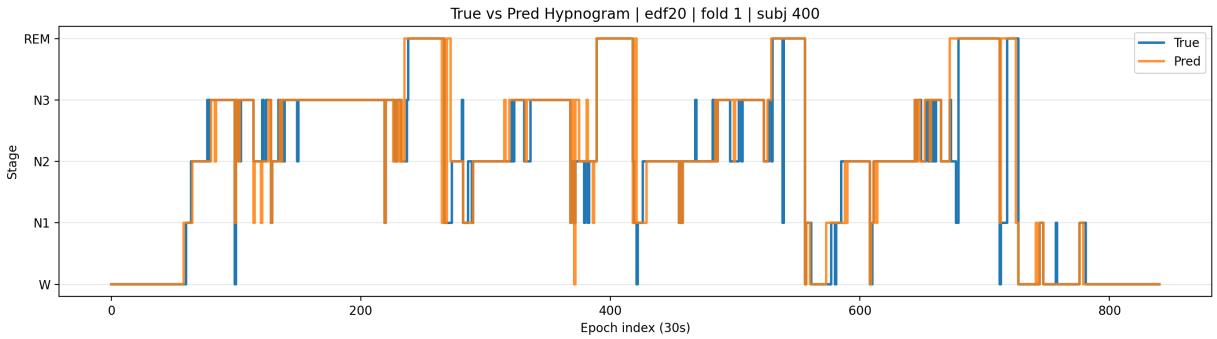
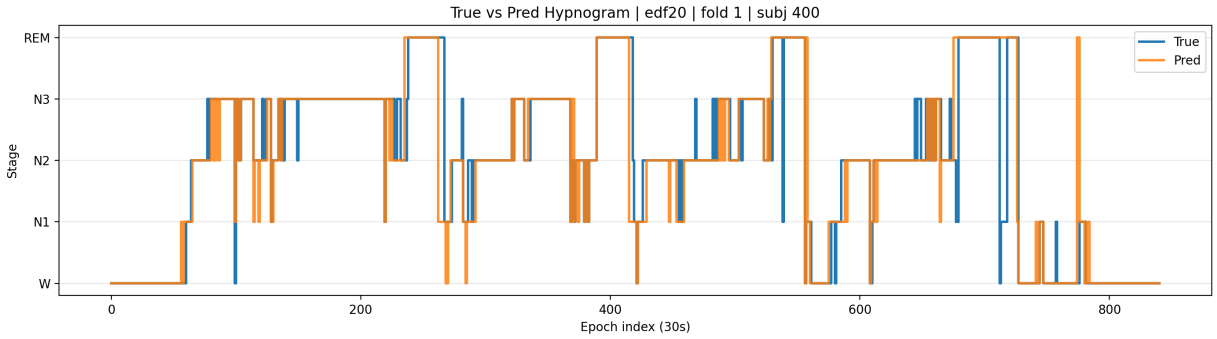


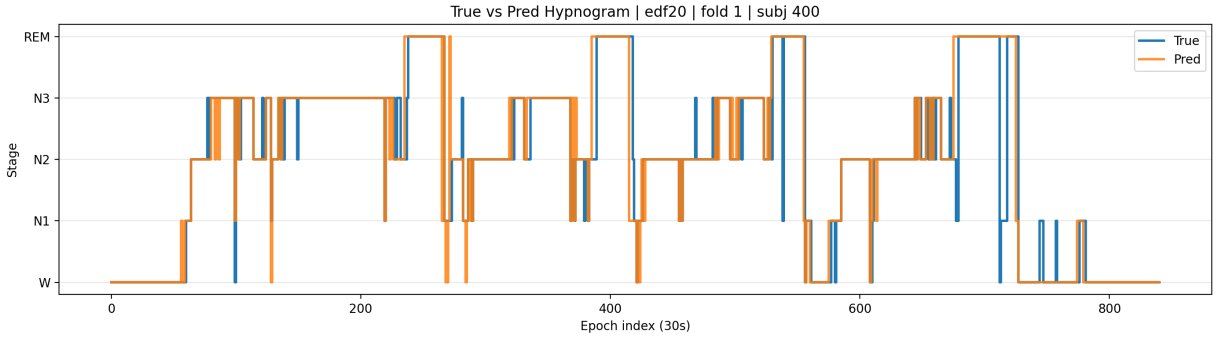
Figure 3. **ISRUC Subgroup 3.** Example hypnogram reconstruction on ISRUC Subgroup 3.



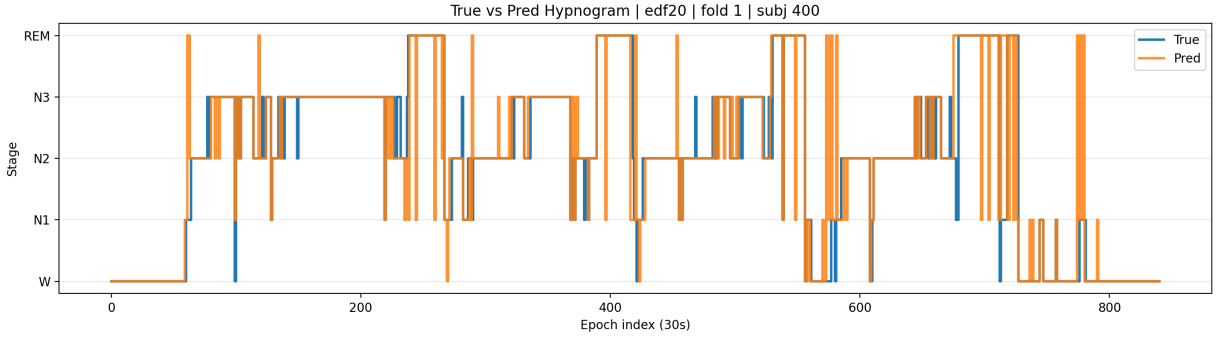
(a) Full model



(b) noS (without stage tokens)



(c) noB (without boundary-aware attention)



(d) noB_noS (without both components)

Figure 4. **EDF-20 (Ablation Comparison).** Example hypnogram reconstructions on EDF-20 for the full model and ablation variants using single-checkpoint inference ($M{=}1$). The full model produces more temporally coherent stage sequences with fewer isolated flips. Removing either component increases local instability, while removing both leads to pronounced over-segmentation near stage transitions.