CTXSLEEPNET: CONTEXT-AWARE MULTIMODAL SLEEP STAGING WITH LABEL-EFFICIENT DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalization across recording channels, acquisition devices, and clinical centers remains a major challenge in multimodal sleep staging. We propose CtxSleepNet, a context-aware framework that integrates multimodal representation learning with adaptation tailored to the available target-domain label budget. A hierarchical CNN–GRU–Transformer encoder captures local morphology, short-term dynamics, and long-range dependencies, while an LSTM-based cross-epoch module incorporates neighboring context to refine individual epoch predictions. For label-free transfer, we introduce recording-wise domain statistics alignment, combining stage reweighting with batch normalization correction relative to source references. This strategy exploits unlabeled target data without additional pretraining or gradient-based parameter updates. For adaptation with limited labels, we develop *typicality–certainty selection* (TYPCERT), which combines target-subject selection based on stage-distribution typicality and predictive certainty with budget-dependent parameter updating. By restricting updates at low label budgets and enabling full-network fine-tuning at higher budgets, this strategy balances estimation variance and model flexibility. Experiments on four public datasets demonstrate that CtxSleepNet consistently outperforms competing methods in-domain, achieving accuracies of 79.6%–87.3%. Across six transfer scenarios spanning channel, device, and center differences, domain statistics alignment improves mean zero-shot accuracy from 68.2% to 74.4%, a gain of 6.2 percentage points. With 10% labeled target data, mean accuracy reaches 79.6%, only 2.1 percentage points below the in-domain reference of 81.7%. These results support the effectiveness of contextual multimodal modeling and complementary label-free and label-efficient adaptation for sleep staging across heterogeneous recording environments.

1 INTRODUCTION

Automatic sleep staging is an important component of polysomnography analysis because it enables objective characterization of sleep architecture and facilitates the assessment of sleep quality and the screening of sleep disorders, while reducing the time and labor required for manual expert scoring (Supratak et al., 2017; Phan et al., 2021). Nevertheless, a substantial body of existing work still formulates sleep staging primarily as an epoch-wise classification problem, thereby underutilizing the temporal continuity between neighboring 30-s epochs. Since sleep stages evolve according to structured transition dynamics, contextual information from adjacent epochs can help resolve ambiguous stage boundaries and improve classification accuracy (Phan et al., 2019; 2022). Beyond temporal modeling, the generalization protocol is equally critical. Most existing studies evaluate their models under an intra-dataset setting, in which the training and test data are collected under similar acquisition conditions. Such evaluations may overestimate real-world robustness because variations in electrode montages, acquisition devices, clinical centers, and subject populations can induce substantial domain shifts. Therefore, cross-channel, cross-device, and cross-center generalization, collectively referred to as *three-cross generalization*, provides a more clinically relevant assessment of whether a sleep-staging model can be reliably deployed on unseen signal configurations, hardware platforms, and clinical environments (Wang et al., 2024b).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Unlike standard static classification tasks, sleep staging is performed on continuous physiological recordings, for which the 30-s epoch serves as a clinically defined annotation unit rather than an independent observation. Consequently, neighboring epochs provide essential information about sleep-stage transitions and should be jointly considered when assigning the label of a target epoch. Early deep learning models, such as DeepSleepNet (Supratak et al., 2017), incorporated recurrent architectures to encode transition patterns, while SeqSleepNet (Phan et al., 2019) further formulated sleep staging as a sequence-to-sequence prediction problem. Subsequent studies explored fully convolutional sequence segmentation through U-Time (Perslev et al., 2019), multi-view sequential representations through XSleepNet (Phan et al., 2021), and Transformer-based temporal modeling through SleepTransformer (Phan et al., 2022). Despite these advances, existing sleep-staging models do not consistently distinguish within-epoch feature extraction from inter-epoch sequence modeling. The respective contributions of short-term dynamics, long-range dependencies, and continuity across consecutive epochs therefore remain insufficiently characterized within a unified framework. Moreover, temporal modeling is often evaluated under a single recording configuration, leaving its robustness across electrode montages, acquisition devices, and recording centers unclear.

Designing a model with three-cross generalization is essential because widely used benchmarks exhibit substantial acquisition heterogeneity: SleepEDF (Goldberger et al., 2000; Kemp et al., 2000) typically provides two EEG channels (Fpz-Cz and Pz-Oz), whereas ISRUC-Sleep (Khalighi et al., 2016) and Physio2018 (Ghassemi et al., 2018) provide six EEG channels with different reference electrodes, sampling protocols, recording equipment, and clinical origins. Consequently, models trained on a single dataset may capture dataset-specific artifacts rather than sleep-related physiological patterns. Recent studies have explored this issue through transfer learning for arbitrary PSG montages, subject-invariant spatial-temporal representations, unsupervised domain adaptation, and multi-source domain generalization (Guillot & Thorey, 2021; Jia et al., 2021a; Fan et al., 2022; Wang et al., 2024b). However, the joint optimization of subject selection and adaptation capacity under concurrently varying electrode montages, acquisition devices, and clinical centers remains underexplored in existing methods.

Accordingly, the main contributions of this study are summarized as follows:

- We develop a context-aware framework for multimodal sleep staging that combines hierarchical CNN–GRU–Transformer feature extraction with an LSTM sequence module. By modeling interactions among neighboring epochs, the framework integrates temporal context to refine individual epoch predictions.
- Given the available label budget, we propose TYPCERT, an adaptive fine tuning strategy that combines zero shot subject selection with budget dependent parameter updating. By selecting target subjects according to stage distributional typicality and predictive certainty, while adjusting the adaptation scope to the label budget, TYPCERT balances estimation variance and model flexibility.
- We introduce recording-wise domain statistics alignment that combines stage reweighting with batch normalization correction relative to source references, exploiting unlabeled target data without additional pretraining or parameter updates. Empirically, this strategy substantially improves zero-shot accuracy, bringing it close to in-domain performance.

2 PROPOSED METHOD

This section first introduces the hierarchical architecture used to pretrain the source model and then describes the fine tuning protocol for adapting it to the target domain under different label budgets.

2.1 PRETRAIN ARCHITECTURE

Although this study primarily focuses on sequence modeling of sleep signals and evaluates generalization across channels, acquisition devices, and centers, the underlying network architecture is equally important for effectively extracting complementary information from multimodal recordings. Accordingly, we develop a hierarchical framework that transforms raw signals into discriminative epoch-level representations, models temporal dependencies across consecutive epochs, and generates final sleep-stage predictions. The proposed architecture is introduced from three complementary components: the *Epoch Encoder*, which extracts within-epoch features from EEG and

EOG signals; the *Sequence Encoder*, which captures inter-epoch continuity and context-aware dependencies; and the *Output Head*, which maps the resulting contextual representations to sleep-stage probabilities. For each component, we describe its structure and explain the corresponding design rationale.

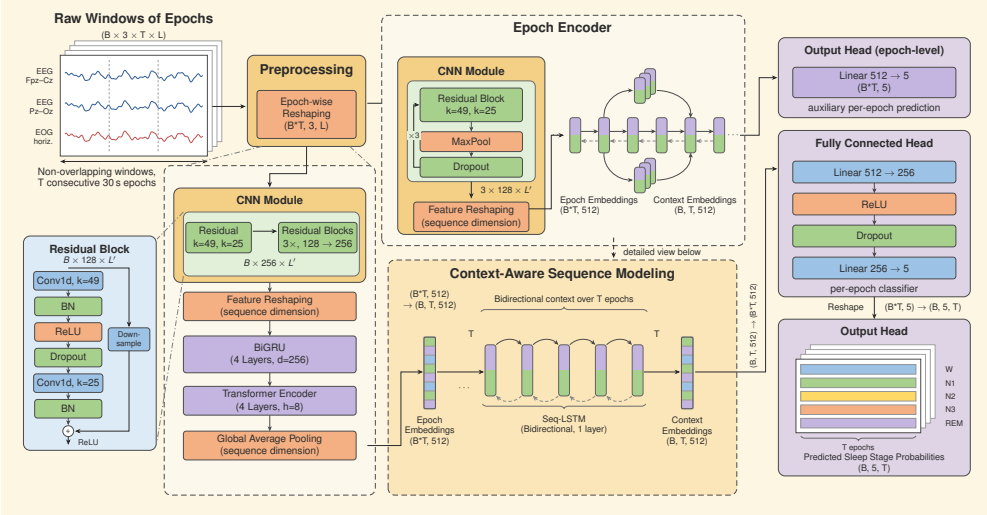

Figure 1: The pretrain architecture of our proposed approach.

First, the Epoch Encoder transforms each three-channel raw epoch into a 512-dimensional representation. Two residual CNN blocks with pooling and dropout extract local morphological patterns at multiple temporal scales, while the four-layer bidirectional GRU and Transformer encoder capture complementary short- and long-range dependencies. Global average pooling then produces a compact multimodal embedding (Chambon et al., 2018). Second, the resulting epoch embeddings are organized into an ordered window of T consecutive epochs and processed by a bidirectional LSTM. This hierarchical design separates within-epoch signal representation from inter-epoch temporal modeling, enabling bidirectional information exchange and explicitly exploiting sleep continuity. Such context-aware sequence modeling is particularly suitable for ambiguous stage transitions (Phan et al., 2019). Third, a shared fully connected head maps each contextual embedding from 512 to 5 dimensions, yielding epoch-level probabilities for the five sleep stages while maintaining parameter efficiency.

2.2 Fine-tuning

We tailored adaptation to the label budget to balance estimation variance and model flexibility. At the 1% and 2% budgets, limited subject diversity motivated *typicality–certainty selection* (TypCert) and parameter-efficient adaptation to limit subject-specific overfitting. Feature extractors were frozen, while normalization affine parameters and the final classifier were updated, and batch normalization statistics were recalibrated using target-domain data. At the 5% and 10% budgets, greater subject diversity supported full-network fine-tuning. We therefore retained the original randomly sampled nested cohorts and updated all parameters to enable target-specific representation learning.

TypCert selects target subjects using only zero shot ensemble predictions. Let $\mathbf{q}_i \in \Delta^{C-1}$ denote the predicted class probability vector for epoch i, and let $\widehat{\boldsymbol{\pi}}_{\mathcal{T}}$ denote the predicted stage marginal computed over all unlabeled target epochs. For a subject u containing N_u epochs, we define its predicted stage marginal and mean predictive entropy as

$$\widehat{\pi}_{u,c} = \frac{1}{N_u} \sum_{i \in u} \mathbf{1}\left\{ \arg\max_{1 \le j \le C} q_{i,j} = c \right\}, \quad \bar{H}_u = \frac{1}{N_u} \sum_{i \in u} \left(-\sum_{c=1}^{C} q_{i,c} \log q_{i,c} \right).$$

For any nonempty subject set $\mathcal{A}$, let $N_{\mathcal{A}} = \sum_{v \in \mathcal{A}} N_v$, and define the corresponding length weighted stage marginal by $\widehat{\pi}_{\mathcal{A},c} = N_{\mathcal{A}}^{-1} \sum_{v \in \mathcal{A}} N_v \widehat{\pi}_{v,c}$ and the mean entropy by $\bar{H}(\mathcal{A}) = N_{\mathcal{A}}^{-1} \sum_{v \in \mathcal{A}} N_v \bar{H}_v$. Starting from $\mathcal{S}_0 = \emptyset$, let $\mathcal{R}_t = \mathcal{U} \setminus \mathcal{S}_t$ denote the remaining candidates.

At iteration t, each candidate $u \in \mathcal{R}_t$ is evaluated at the set level according to

$$d_{\mathrm{TV}}(\mathcal{A}) = \frac{1}{2} \left\| \widehat{\boldsymbol{\pi}}_{\mathcal{A}} - \widehat{\boldsymbol{\pi}}_{\mathcal{T}} \right\|_1, \quad u_t = \arg\min_{u \in \mathcal{R}_t} \left\{ r_t^{\mathrm{TV}}(u) + r_t^{H}(u) \right\}, \quad \mathcal{S}_{t+1} = \mathcal{S}_t \cup \{u_t\},$$

where $r_t^{\mathrm{TV}}(u)$ and $r_t^{H}(u)$ are the ascending within iteration ranks of $d_{\mathrm{TV}}(\mathcal{S}_t \cup \{u\})$ and $\bar{H}(\mathcal{S}_t \cup \{u\})$, respectively. Rank aggregation makes the selection criterion insensitive to the numerical scales of the two quantities and eliminates the need for a data set dependent mixing coefficient. Moreover, evaluating each candidate after its inclusion in the current set promotes complementary subject selection rather than the repeated selection of individually similar subjects. The 1% and 2% cohorts are defined as prefixes of the resulting subject ordering and are therefore nested.

TYPCERT limits cohort-induced gradient bias by favoring low total variation distance between selected-cohort and target predicted stage distributions, while prioritizing low-entropy predictions to reduce uncertain updates. Complementarily, updating only normalization layers and the classification head limits adaptation variance, retains pretrained feature-extractor weights, and accommodates target covariate shift. Together, these mechanisms help mitigate accuracy degradation under severe annotation scarcity. Rank-based selection further avoids dataset-specific mixing coefficients, providing a reproducible framework that combines label-free cohort construction with parameter-efficient adaptation for cross-dataset sleep staging.

Although TYPCERT uses unlabeled target predictions for subject selection, it does not directly improve zero-shot inference. We therefore introduce recording-wise statistics alignment to exploit unlabeled target data while keeping all model parameters fixed. Inspired by domain statistics alignment (DSA) (Fan et al., 2022), we adjust batch normalization (BN) statistics to accommodate acquisition-related shifts. However, pooled feature moments also depend on sleep-stage composition, potentially confounding acquisition differences with label shift. For recording r of subject u, let $\hat{\pi}_{u,r,c}$ denote the predicted stage marginal, defined analogously to $\hat{\pi}_{u,c}$ within that recording. Using the previously defined predictions $q_{i,c}$, we assign epoch weights

$$\hat{y}_i = \arg\max_{1 \le c \le C} q_{i,c}, \qquad \omega_{u,r,i} = \frac{\pi_{\mathrm{src},\hat{y}_i}}{\hat{\pi}_{u,r,\hat{y}_i}}, \tag{1}$$

where $\pi_{\mathrm{src},c}$ is the source stage prior, and the importance weights $\omega_{u,r,i}$ are clipped to $[1/5, 5]$. This weighting approximately standardizes stage composition toward the source distribution, following the rationale of importance weighting under label shift (Tachet des Combes et al., 2020). Pseudo-labels determine the weights but are never used as training targets.

For each BN layer ℓ, we estimate the weighted feature mean $\hat{\mu}_{\ell,u,r}$ and weighted geometric mean of per-epoch variances $\hat{v}_{\ell,u,r}$ over the recording. Rather than replacing the stored statistics, we apply corrections relative to reference moments $(\bar{\mu}_\ell, \bar{v}_\ell)$, precomputed from labeled source recordings using the same estimator:

$$\mathrm{BN}_\ell(h) = \gamma_\ell \frac{h - \left(\mu_\ell^{\mathrm{src}} + \hat{\mu}_{\ell,u,r} - \bar{\mu}_\ell\right)}{\sqrt{(\sigma_\ell^{\mathrm{src}})^2 \hat{v}_{\ell,u,r} / \bar{v}_\ell + \epsilon}} + \beta_\ell,$$

where $(\mu_\ell^{\mathrm{src}}, (\sigma_\ell^{\mathrm{src}})^2)$ are the stored running statistics and $(\gamma_\ell, \beta_\ell)$ remain fixed during zero-shot inference. Geometric averaging reduces sensitivity to extreme per-epoch variances, while relative correction recovers the original BN mapping when the recording moments match the source reference. Inference uses two forward passes per recording: one for pseudo-label estimation and one for layerwise alignment, without source-data access or parameter optimization. For subsequent few-shot fine-tuning, the same normalization uses observed stage labels in equation (1) while retaining the source running statistics, limiting their distortion by a small target cohort.

3 THEORETICAL ANALYSIS

We characterize cross-epoch encoding and label-efficient adaptation under the class-weighted cross-entropy used in our experiments. Definitions, assumptions, and proofs are provided in Appendix B.

Proposition 1 (Value of cross-epoch context)**.** *Let Z_i be the epoch embedding and Z_{-i} the other embeddings in its window of T epochs. The optimal weighted risks over all measurable predictors satisfy*

$$R_w^*(Z_i) - R_w^*(Z_i, Z_{-i}) = \mathbb{E}[w(Y_i)]\, I_w(Y_i; Z_{-i} \mid Z_i) \ge 0,$$

where I_w is conditional mutual information under the law reweighted by $w(Y_i)$.

This proposition provides a theoretical basis for cross-epoch modeling by quantifying how neighboring embeddings can reduce the optimal weighted risk for the target epoch. For sequence encoder, this potential gain translates into lower risk when it exceeds the combined representation and prediction-head errors; see equation (B.3).

Theorem 1 (Cohort bias and adaptation capacity)**.** *Fix a parameter point before querying target labels. Let g_U be the conditional target-pool risk gradient, $\widehat{g}_A$ the labeled-cohort gradient, and P the orthogonal projector onto trainable parameters. Under the bounded-gradient and subject-independence conditions in Appendix B,*

$$\mathbb{E}\big[\|P\widehat{g}_A - g_U\|^2 \mid \mathcal{F}\big] \le \|(I-P)g_U\|^2 + B_A^2 + \frac{\sigma^2 \operatorname{rank}(P)}{m_{\mathrm{eff}}(A)},$$

where $m_{\mathrm{eff}}(A) = N_A^2 / \sum_{u\in A} N_u^2$ and $B_A = 2G[d_{\mathrm{TV}}(A) + \bar{H}(A) + \varepsilon_A] + \rho_A$. Here G bounds class-conditional gradients, ρ_A measures their cohort mismatch, ε_A collects posterior-error and target-entropy terms, and σ^2 bounds each subject's gradient covariance.

This theorem links adaptation-gradient error to cohort bias, trainable capacity, and effective subject coverage, providing a theoretical rationale for stable fine-tuning under limited annotation. TYPCERT targets $d_{\mathrm{TV}}(A)$ and $\bar{H}(A)$ in B_A, with the entropy criterion additionally controlling label-induced variance when posterior error is bounded. At 1%–2% budgets, restricting updates to normalization and classifier parameters lowers the variance bound at the cost of omitted gradient signal. Greater effective subject coverage reduces the variance penalty for full adaptation at 5%–10% budgets. The corresponding local risk-decrease condition is given in equation (B.8).

4 EXPERIMENT

4.1 DATASETS

We evaluated the proposed method on four public polysomnographic datasets. SleepEDF-20 and SleepEDF-78 are two subsets of the SleepEDF Expanded database, containing 20 and 78 subjects, respectively, with EEG recordings from Fpz-Cz and Pz-Oz (Kemp et al., 2000). ISRUC includes 117 subjects and 125 recordings acquired using F3-A2 and O1-A2, whereas Physio2018 contains 994 recordings with F3-M2 and O1-M2 EEG channels (Khalighi et al., 2016; Ghassemi et al., 2018). To ensure consistent multimodal inputs across datasets, we used two modalities, EEG and EOG, represented by three channels comprising two EEG derivations and one EOG channel. This configuration captures complementary neural and ocular dynamics while accommodating inter-dataset differences in electrode montage and acquisition settings.

Table 1: Summary of the four public sleep datasets.

Dataset	Subjects	Recordings	EEG Channels	Evaluation Scheme	W	N1	N2	N3	REM	Total
SleepEDF-20	20	39	Fpz-Cz & Pz-Oz	20-fold CV	10197	2804	17799	5703	7717	44220
SleepEDF-78	78	153	Fpz-Cz & Pz-Oz	10-fold CV	69824	21522	69132	13039	25835	199352
ISRUC	117	125	F3-A2 & O1-A2	10-fold CV	24857	14846	35726	21910	15040	112379
Physio2018	994	994	F3-M2 & O1-M2	10-fold CV	157945	136978	377870	102592	116877	892262

4.2 EXPERIMENTAL SETUP

During pretraining, the development windows were randomly split into training and validation subsets at a ratio of 9:1. To enhance robustness, 10% of the training windows were augmented using a channel-wise strategy. For a three-channel signal $X = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \in \mathbb{R}^{T\times 3}$, one channel $\mathbf{x}_c$ was randomly selected, while the other channels remained unchanged. Two distinct operators were sampled from $\mathcal{A} = \{\text{noise}, \text{shift}, \text{scale}, \text{filter}, \text{warp}, \text{frequency}, \text{crop}\}$ and sequentially applied:

$$\tilde{\mathbf{x}}_c = (a_2 \circ a_1)(\mathbf{x}_c), \qquad a_1, a_2 \in \mathcal{A},\ a_1 \neq a_2.$$

The augmented sample was obtained by replacing $\mathbf{x}_c$ with $\tilde{\mathbf{x}}_c$, thereby increasing sample diversity while preserving cross-channel dependencies. The validation set was used to select the ten models with the highest accuracy. Each model was trained using Adam for 50 epochs, with a learning rate of 10^{-4}, a weight decay of 10^{-3}, and a batch size of 32. Weighted cross-entropy with class weights 1:1.5:1:1:1 was employed. During testing, predictions from the selected models were aggregated by majority voting, with all test epochs retained for evaluation.

Fine-tuning was evaluated for each ordered source-to-target dataset pair using subject-wise repeated splits over 10 repetitions. Under label budgets of 1%, 2%, 5%, and 10%, a pretrained ensemble of ten models was adapted to the target domain. For the two lower budgets, TYPCERT selected target subjects based on zero-shot predictions, and only the normalization parameters and classifier were updated. For the two higher budgets, randomly sampled nested cohorts were used and the entire network was fine-tuned. In all settings, 10% of the adaptation windows were augmented. Fine-tuning was performed for 10 epochs using Adam with a learning rate of 10^{-4}, weight decay of 10^{-4}, and batch size 32. Performance was evaluated on held-out subjects under full coverage using ensemble predictions.

4.3 IN-DOMAIN EVALUATION

Table 2: In-domain performance against competing methods on four datasets.

Dataset	Method	Overall Metrics (%)			Per-class F1-Score (%)				
		Acc	MF1	κ	W	N1	N2	N3	REM
SleepEDF-20	DeepSleepNet (Supratak et al., 2017)	81.4	75.6	75.0	83.9	43.5	85.4	84.1	81.1
	DeepSleepNet-lite (Fiorillo et al., 2021)	84.0	78.0	78.0	87.1	44.4	87.9	88.2	82.4
	IITNet (Seo et al., 2020)	84.0	77.7	78.0	87.9	44.7	88.0	84.1	81.1
	SleepEEGNet (Mousavi et al., 2019)	84.3	79.7	79.0	89.2	52.2	86.8	85.1	85.0
	TinySleepNet (Supratak & Guo, 2020)	85.4	80.5	80.0	90.1	51.4	88.5	88.3	84.3
	XSleepNet (Phan et al., 2021)	86.3	80.6	81.3	**92.2**	51.8	88.0	86.8	83.9
	CtxSleepNet (Our)	**87.3**	**82.3**	**82.7**	92.1	**54.5**	**89.4**	**88.3**	**87.1**
SleepEDF-78	DeepSleepNet (Supratak et al., 2017)	76.9	70.7	69.0	90.8	44.8	78.5	67.9	71.3
	SleepEEGNet (Mousavi et al., 2019)	80.0	73.6	73.0	91.7	44.1	82.5	73.5	76.1
	DeepSleepNet-lite (Fiorillo et al., 2021)	80.3	75.2	73.0	91.5	46.0	82.9	79.2	76.4
	U-Time (Perslev et al., 2019)	81.3	76.3	74.5	92.0	51.0	83.5	74.6	80.2
	SleepTransformer (Phan et al., 2022)	81.4	74.3	74.3	91.7	40.4	84.3	77.9	77.2
	TinySleepNet (Supratak & Guo, 2020)	83.1	78.1	77.0	92.8	51.0	85.3	**81.1**	80.3
	XSleepNet (Phan et al., 2021)	84.0	77.9	77.8	**93.3**	49.9	86.0	78.7	81.8
	CtxSleepNet (Our)	**84.6**	**79.6**	**78.7**	92.7	**52.3**	**88.1**	76.7	**84.3**
ISRUC	U-Time (Perslev et al., 2019)	73.5	74.7	66.0	84.2	55.2	72.2	85.2	76.4
	TinySleepNet (Supratak & Guo, 2020)	75.8	73.5	68.9	81.2	49.9	77.0	85.7	73.4
	DeepSleepNet (Supratak et al., 2017)	76.5	76.1	69.8	**86.6**	51.9	76.8	86.7	78.5
	SalientSleepNet (Jia et al., 2021b)	76.9	76.5	70.3	84.6	57.3	75.8	82.3	77.8
	SleepWaveNet (Wang et al., 2024a)	79.2	78.2	73.3	84.0	57.4	78.9	**88.1**	82.7
	CtxSleepNet (Our)	**79.6**	**78.3**	**73.8**	83.6	**58.9**	**81.0**	83.3	**84.9**
Physio2018	DeepSleepNet (Supratak et al., 2017)	76.4	74.1	67.6	80.4	51.7	81.7	76.5	80.1
	U-Time (Perslev et al., 2019)	78.8	77.4	71.4	82.5	59.0	83.1	79.0	83.5
	TinySleepNet (Supratak & Guo, 2020)	79.0	77.0	71.3	82.9	56.4	83.6	**79.2**	83.0
	XSleepNet (Phan et al., 2021)	80.3	**79.6**	73.2	-	-	-	-	-
	CtxSleepNet (Our)	**80.8**	79.5	**73.9**	**85.2**	**64.8**	**85.1**	72.9	**87.0**

Table 2 compares CtxSleepNet with competing methods across four datasets. CtxSleepNet achieves the highest accuracy and Cohen's κ on all datasets and the highest Macro-F1 on three, while remaining competitive on Physio2018. These results support the effectiveness of incorporating information from neighboring epochs into sleep-stage representations. Notably, the improvements on all datasets are achieved with only 50 training epochs, indicating effective learning under a limited training budget. The per-class results further demonstrate its advantage in N1 recognition: CtxSleepNet achieves the highest N1 F1-score across all four datasets, outperforming the strongest reported competitor on each dataset by an average of approximately 2.7 percentage points. This advantage is consistent with the transitional nature of N1, whose ambiguous within-epoch characteristics make temporal context particularly informative for distinguishing it from other stages. These in-domain results support the model's representation capability and provide a foundation for subsequent out-of-domain evaluation and adaptation.

4.4 OUT-OF-DOMAIN EVALUATION

In this subsection, we evaluate three-cross generalization using a cross-dataset protocol. Although Subsection 4.1 introduces four datasets, namely SleepEDF-20, SleepEDF-78, ISRUC, and Physio2018, only the latter three are included in the out-of-domain evaluation because SleepEDF-78

extends SleepEDF-20 and shares 20 subjects with it. To avoid subject overlap and potential information leakage, each of the three retained datasets is alternately treated as the source domain, while the other two serve as unseen target domains. Predictive performance is evaluated under both zero-shot transfer and few-shot adaptation using 1%, 2%, 5%, and 10% of the labeled target-domain data.

Table 3: Out-of-domain performance.

Fine-tuning	ISRUC→EDF	ISRUC→Phys	EDF→ISRUC	EDF→Phys	Phys→ISRUC	Phys→EDF
Zero	67.1±0.4	73.5±0.1	62.9±0.6	69.5±0.1	64.7±0.5	71.5±0.4
1%	69.9±5.7	74.6±1.3	53.0±9.8	73.4±2.9	59.3±7.8	73.8±5.5
1%*	71.5±4.8	75.4±1.2	63.7±2.2	72.0±1.7	65.5±1.1	77.7±2.6
2%	74.6±3.9	75.8±1.6	55.0±10.8	74.3±2.1	60.1±10.8	76.1±4.2
2%*	72.8±3.1	76.1±0.7	64.1±2.3	73.0±1.2	66.9±1.0	78.7±1.2
5%	79.4±1.3	76.8±1.4	69.0±4.4	76.5±0.9	75.0±3.9	80.8±1.2
10%	80.5±1.3	77.8±0.8	73.1±3.1	77.3±0.9	77.3±2.4	81.8±1.2

Note: * indicates that TYPCERT was used to select fine-tuning samples under limited labeling budgets.

Table 3 summarizes the out-of-domain performance across channels, devices, and centers. The pretrained model achieves an average zero-shot accuracy of 68.2% across six transfer settings, suggesting that it captures transferable sleep-related patterns. With labeling budgets of 1%, 2%, 5%, and 10%, few-shot fine-tuning yields average test accuracies of 67.3%, 69.3%, 76.2%, and 78.0%, respectively. However, fine-tuning with 1% and 2% of the target-domain labels reduces accuracy when ISRUC serves as the target domain. This degradation is plausibly attributable to limited subject coverage: ISRUC contains only 125 recordings, leaving just one and three recordings available for fine-tuning at these budgets, respectively. Given intersubject heterogeneity, such small subsets may inadequately represent the target distribution. To address this limitation, TYPCERT selects fine-tuning samples based on typicality and prediction stability. For transfers to ISRUC, it improves accuracy over fine-tuning without sample selection by an average of 8.4 and 8.0 percentage points at the 1% and 2% budgets, respectively, with all corresponding accuracies exceeding their zero-shot baselines. These results support the effectiveness of TYPCERT in mitigating performance degradation during adaptation with limited target-domain labels.

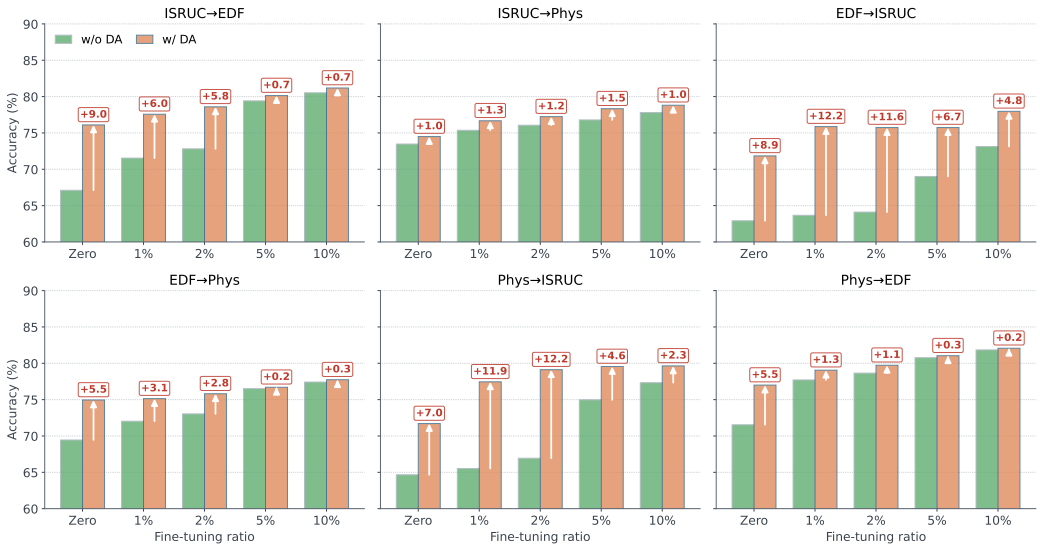

Figure 2: Out-of-domain performance with and without domain adaptation.

To improve zero-shot performance without additional training, we introduce a domain adaptation strategy combining domain statistics alignment (DSA) with batch normalization (BN) adaptation. As shown in Fig. 2, this strategy increases the mean zero-shot accuracy from 68.2% to 74.4%, a gain of 6.2 percentage points, substantially narrowing the gap to the in-domain accuracy of 81.7%. Consistent improvements are also observed across all fine-tuning budgets and transfer settings. With 10% labeled target-domain data, the mean accuracy reaches 79.6%, only 2.1 percentage points below the in-domain result. These findings highlight the importance of domain adaptation in improv-

ing cross-domain generalization, both without target-domain labels and under limited supervision. The cross-dataset performance comparison between CtxSleepNet and other methods is presented in Table 6 of Appendix A.

4.5 SENSITIVITY ANALYSIS

Table 4: Performance under different sequence lengths.

Sequence	Overall Metrics (%)			Per-class F1-Score (%)				
Length	Acc	MF1	κ	W	N1	N2	N3	REM
5	87.0	81.9	82.3	92.1	53.9	88.9	87.2	**87.5**
10	86.7	81.7	81.9	91.7	54.6	88.9	86.4	86.9
15	86.9	82.1	82.2	**92.6**	54.7	88.6	88.6	86.0
20	**87.3**	**82.3**	**82.7**	92.1	54.9	**89.4**	88.3	87.1
25	86.7	82.1	82.0	91.5	**55.9**	87.4	**90.9**	87.0
30	86.6	81.9	81.8	91.7	55.1	88.9	88.0	85.7

Table 4 investigates the effect of the sequence length T on prediction performance. The proposed model achieves the best overall performance when $T = 20$, yielding an overall accuracy of 87.3%, a Macro-F1 of 82.3%, and a Cohen's kappa of 82.7%. Since each epoch corresponds to 30 s, $T = 20$ provides a 600-s (10-min) cross-epoch context, which is sufficient for capturing the short-term continuity and gradual transitions of human sleep stages. This setting is also consistent with prior sleep-staging studies that model 20–30 consecutive 30-s epochs to capture inter-epoch temporal dependencies (Phan et al., 2022). Although the best F1-score for individual sleep stages may occur at different sequence lengths, the overall accuracy varies by no more than 0.7 percentage points for $T \in \{5, 10, 15, 20, 25, 30\}$, indicating that the proposed architecture is relatively insensitive to the choice of T. Therefore, we set $T = 20$ throughout the remaining experiments as a suitable balance between contextual modeling and overall predictive performance.

4.6 ABLATION STUDY

The ablation study evaluates the proposed framework from three complementary perspectives: data, model, and prediction. For data utilization, Case 1 compares a single EEG channel (Fpz–Cz) with a three-channel multimodal input to quantify the benefit of complementary physiological information, while Case 2 examines the contribution of data augmentation to cross-subject generalization. For architectural design, the framework combines CNN, GRU, and Transformer modules to capture local signal patterns, short-term temporal dynamics, and long-range dependencies, respectively. An additional LSTM-based Sequence mechanism models contextual interactions within a window of T consecutive epochs to exploit temporal continuity. Cases 3–5 remove the GRU, Transformer, and Sequence modules, respectively, to assess their individual contributions. Finally, Case 6 evaluates a computationally efficient ensemble strategy that aggregates multiple predictions to reduce prediction variance and improve stability. Together, these experiments characterize how multimodal inputs, feature extraction, cross-epoch context, and prediction aggregation contribute to sleep staging performance.

Table 5: Ablation results (%) on SleepEDF-20. Values in parentheses denote changes relative to Base in percentage points; red indicates degradation and blue indicates improvement.

Model	Overall Metrics			Per-class F1-Score (F1)				
	Acc	MF1	κ	W	N1	N2	N3	REM
Base	87.3	82.3	82.7	92.1	54.5	89.4	88.3	87.1
Case 1 (Single channel)	85.6 (↓ 1.7)	80.1 (↓ 2.2)	80.5 (↓ 2.2)	91.0 (↓ 1.1)	51.8 (↓ 2.7)	88.3 (↓ 1.1)	84.0 (↓ 4.3)	85.5 (↓ 1.6)
Case 2 (No augmentation)	86.7 (↓ 0.6)	82.0 (↓ 0.3)	82.0 (↓ 0.7)	90.9 (↓ 1.2)	54.8 (↑ 0.3)	88.9 (↓ 0.5)	89.0 (↑ 0.7)	86.4 (↓ 0.7)
Case 3 (Remove GRU)	86.7 (↓ 0.6)	82.0 (↓ 0.3)	82.0 (↓ 0.7)	90.7 (↓ 1.4)	54.7 (↑ 0.2)	88.6 (↓ 0.8)	89.0 (↑ 0.7)	87.0 (↓ 0.1)
Case 4 (Remove Transformer)	85.9 (↓ 1.4)	81.1 (↓ 1.2)	80.8 (↓ 1.9)	90.8 (↓ 1.3)	54.0 (↓ 0.5)	88.1 (↓ 1.3)	88.2 (↓ 0.1)	84.5 (↓ 2.6)
Case 5 (Remove Sequence)	86.2 (↓ 1.1)	80.5 (↓ 1.8)	81.2 (↓ 1.5)	91.6 (↓ 0.5)	48.5 (↓ 6.0)	88.6 (↓ 0.8)	89.3 (↑ 1.0)	84.4 (↓ 2.7)
Case 6 (Single classifier)	85.9 (↓ 1.4)	80.8 (↓ 1.5)	80.9 (↓ 1.8)	90.6 (↓ 1.5)	54.1 (↓ 0.4)	88.4 (↓ 1.0)	86.4 (↓ 1.9)	84.7 (↓ 2.4)

Table 5 provides controlled evidence for the contribution of each component in the proposed framework. In Case 1, replacing the three-channel multimodal input with a single EEG channel decreases the overall accuracy, Macro F1, and Kappa by 1.7, 2.2, and 2.2 percentage points, respectively. The class-wise F1-scores decrease across all stages, with the largest reductions observed for N3 and N1 at 4.3 and 2.7 percentage points, confirming the importance of complementary EEG–EOG information

for learning stage-discriminative representations. Removing data augmentation in Case 2 produces modest declines in the overall metrics, although the N1 and N3 F1-scores increase slightly, indicating that its regularization effect varies across sleep stages. In Case 3, excluding the GRU primarily reduces the F1-scores for W and N2, supporting its role in capturing short-term temporal dynamics.

Removing the Transformer in Case 4 decreases Kappa by 1.9 percentage points and causes notable F1-score reductions for W, N2, and REM, with the largest decline of 2.6 percentage points occurring for REM. These results demonstrate the importance of long-range dependency modeling. Case 5 yields the most pronounced degradation in N1 recognition, reducing its F1-score from 54.5% to 48.5%, together with a 1.8-point decrease in Macro F1. This finding supports the use of the Sequence mechanism to exploit continuity across consecutive epochs and distinguish ambiguous transitional stages. Finally, removing the ensemble strategy in Case 6 decreases the overall accuracy, Macro F1, and Kappa by 1.4, 1.5, and 1.8 percentage points, respectively, with particularly evident reductions for N3 and REM. Collectively, the results show that multimodal inputs, data augmentation, multi-scale temporal modeling, and ensemble aggregation address complementary sources of uncertainty in automatic sleep staging.

4.7 Visualization of Learned Representations

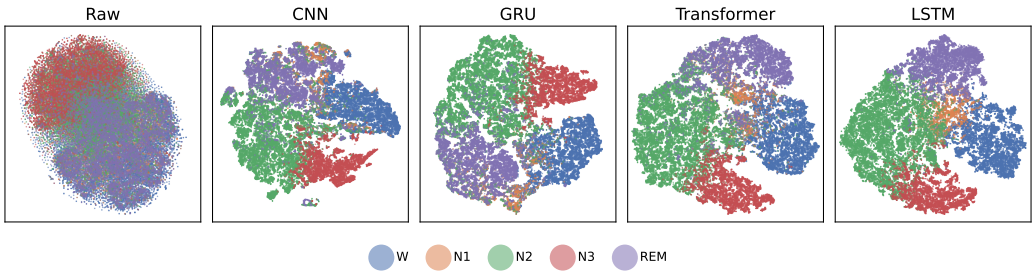

Figure 3: The feature clustering maps of CtxSleepNet on the SleepEDF-20.

Figure 3 presents t-SNE visualizations of the raw features and representations extracted by the CNN, GRU, Transformer, and LSTM modules for all 44,220 epochs from SleepEDF-20, colored according to the five sleep stages. The raw features exhibit substantial overlap across stages, whereas the CNN representations reveal more distinct stage-specific structures. However, N1 remains largely intermingled with W and REM, indicating limited separability at this level. Following processing by the GRU, Transformer, and LSTM modules, the N1 distribution becomes progressively more discernible, although its boundaries remain poorly defined. These visualizations provide qualitative evidence that the model learns increasingly discriminative representations through successive feature extraction and temporal modeling. Nevertheless, the persistent overlap between N1 and neighboring stages is consistent with its transitional nature, highlighting the greater difficulty of distinguishing N1 compared with stages exhibiting clearer separation.

5 Conclusion

This work proposed CtxSleepNet, with theoretical and empirical evidence that cross-epoch context improves sleep-stage prediction. To enhance three-cross generalization, TypCert combines subject typicality and predictive certainty to mitigate accuracy degradation under extremely limited fine-tuning budgets. Our analysis further characterizes cohort bias and the trade-off between adaptation capacity and estimation variance across annotation budgets. Complementing this strategy, DSA-inspired recording-wise BN alignment exploits unlabeled target data without additional pretraining, substantially improving zero-shot accuracy and approaching in-domain performance in the evaluated settings. Future work will extend this framework to continual adaptation for longitudinal sleep monitoring under evolving acquisition conditions (Wang et al., 2022). We will also investigate reliability-aware updates to limit model collapse under noisy and imbalanced target streams (Niu et al., 2023).

AI USE STATEMENT

We used ChatGPT to assist with language polishing, and the interpretation and presentation of experimental results. The authors reviewed and revised all AI-assisted content for consistency with the methods and findings and take full responsibility for the accuracy, originality, and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

The source code is anonymously available at `https://anonymous.4open.science/r/CtxSleepNet-047E/`, covering data preprocessing, source-model training, cross-dataset fine-tuning, TYPCERT selection, domain statistics alignment (DSA), and recording-wise normalization (BN). The repository includes fixed random seeds and subject splits, training logs, and per-split results, with a README linking reported results to the corresponding scripts and outputs. The four evaluation datasets are described in Section 4.1; the Experimental Setup subsection specifies the training configurations and fine-tuning protocol with 10 repeated subject-wise splits. The Proposed Method section details the architecture and adaptation procedures, while Section 3 and Appendix B provide the theoretical assumptions and proofs. Scripts are provided to regenerate the pretrained source ensembles, which are not hosted directly owing to their size.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4): 758–769, 2018.

Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. 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.

Jiahao Fan, Hangyu Zhu, Xinyu Jiang, Long Meng, Chen Chen, Cong Fu, Huan Yu, Chenyun Dai, and Wei Chen. Unsupervised domain adaptation by statistics alignment for deep sleep staging networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:205–216, 2022.

Luigi Fiorillo, Paolo Favaro, and Francesca Dalia Faraci. Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates. *IEEE transactions on neural systems and rehabilitation engineering*, 29:2076–2085, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

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

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

Antoine Guillot and Valentin Thorey. Robustsleepnet: Transfer learning for automated sleep staging at scale. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1441–1451, 2021.

Ziyu Jia, Youfang Lin, Jing Wang, Xiaojun Ning, Yuanlai He, Ronghao Zhou, Yuhan Zhou, and Li-wei H Lehman. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1977–1986, 2021a.

Ziyu Jia, Youfang Lin, Jing Wang, Xuehui Wang, Peiyi Xie, and Yingbin Zhang. Salientsleepnet: Multimodal salient wave detection network for sleep staging. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2614–2620. International Joint Conferences on Artificial Intelligence Organization, 2021b.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. 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.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pp. 97–105. PMLR, 2015.

Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PloS one*, 14(5):e0216456, 2019.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *Advances in neural information processing systems*, 32, 2019.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. 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.

Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2021.

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

Hogeon Seo, Seunghyeok Back, Seongju Lee, Deokhwan Park, Tae Kim, and Kyoobin Lee. 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.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops*, pp. 443–450. Springer, 2016.

Akara Supratak and Yike Guo. Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 641–644. IEEE, 2020.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE transactions on neural systems and rehabilitation engineering*, 25(11):1998–2008, 2017.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33:19276–19289, 2020.

Jing Wang, Xuehui Wang, Xiaojun Ning, Youfang Lin, Huy Phan, and Ziyu Jia. Subject-adaptation salient wave detection network for multimodal sleep stage classification. *IEEE Journal of Biomedical and Health Informatics*, 29(3):2172–2184, 2024a.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 265–273, 2024b.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201. IEEE, 2022.

Siyuan Zhao, Chenyu Liu, Yi Ding, and Xinliang Zhou. Selectivefinetuning: Enhancing transfer learning in sleep staging through selective domain alignment. *arXiv preprint arXiv:2501.03764*, 2025.

A ADDITIONAL OUT-OF-DOMAIN RESULTS

Table 6 compares the proposed domain adaptation strategy with existing methods for cross-dataset sleep staging without target labels. CtxSleepNet achieves the highest reported accuracy among the compared methods on all three target datasets, reaching 77.0%, 71.8%, and 75.0% on SleepEDF-78, ISRUC, and Physio2018, respectively. These results exceed the strongest competing baseline on each target by 1.8, 1.6, and 3.4 percentage points, respectively. Each CtxSleepNet model is trained on a single source dataset and applies DSA and BN adaptation using unlabeled target data. The consistent accuracy advantages across the three targets support the effectiveness of this strategy in transferring learned sleep-stage representations and mitigating domain shifts, enabling accurate cross-dataset prediction without target annotations.

Table 6: Out-of-domain sleep staging performance without target labels. Acc: accuracy; MF1: macro-F1; UDA: unsupervised domain adaptation; DG: domain generalization; DSA: domain statistics alignment; BN: batch normalization. Multi-source denotes the leave-one-dataset-out protocol of Wang et al. (2024b), in which the four remaining cohorts are used as sources.

Method	Strategy	Source	Acc (%)	MF1 (%)
Target: SleepEDF-78				
Selective Finetuning (Zhao et al., 2025)	UDA	SHHS	75.2	**73.7**
CtxSleepNet (Our)	DSA + BN adaptation	ISRUC	76.1	70.5
CtxSleepNet (Our)	DSA + BN adaptation	Physio2018	**77.0**	73.2
Target: ISRUC				
DeepSleepNet (Supratak et al., 2017)	No adaptation	Multi-source	64.0	62.8
U-Time (Perslev et al., 2019)	No adaptation	Multi-source	65.1	63.7
AttnSleep (Eldele et al., 2021)	No adaptation	Multi-source	64.5	62.2
TinySleepNet (Supratak & Guo, 2020)	No adaptation	Multi-source	65.9	64.0
SalientSleepNet (Jia et al., 2021b)	No adaptation	Multi-source	65.4	63.2
MMD (Long et al., 2015)	Multi-source DG	Multi-source	69.4	66.8
CORAL (Sun & Saenko, 2016)	Multi-source DG	Multi-source	70.1	68.6
IRM (Arjovsky et al., 2019)	Multi-source DG	Multi-source	70.2	68.4
DANN (Ganin et al., 2016)	Multi-source DG	Multi-source	69.3	67.7
CtxSleepNet (Our)	DSA + BN adaptation	SleepEDF-78	**71.8**	69.4
CtxSleepNet (Our)	DSA + BN adaptation	Physio2018	71.7	**70.4**
Target: Physio2018				
DeepSleepNet (Supratak et al., 2017)	No adaptation	Multi-source	66.6	60.5
U-Time (Perslev et al., 2019)	No adaptation	Multi-source	67.4	61.1
AttnSleep (Eldele et al., 2021)	No adaptation	Multi-source	66.2	60.8
TinySleepNet (Supratak & Guo, 2020)	No adaptation	Multi-source	68.3	61.4
SalientSleepNet (Jia et al., 2021b)	No adaptation	Multi-source	68.9	63.1
MMD (Long et al., 2015)	Multi-source DG	Multi-source	70.7	65.0
CORAL (Sun & Saenko, 2016)	Multi-source DG	Multi-source	71.6	67.8
IRM (Arjovsky et al., 2019)	Multi-source DG	Multi-source	71.6	67.9
DANN (Ganin et al., 2016)	Multi-source DG	Multi-source	71.0	67.4
CtxSleepNet (Our)	DSA + BN adaptation	SleepEDF-78	**75.0**	71.6
CtxSleepNet (Our)	DSA + BN adaptation	ISRUC	74.5	**73.7**

Three-cross generalization across electrode montages, acquisition devices, and recording centers introduces shifts in signal statistics and feature distributions that source-domain training alone may not adequately accommodate. Domain adaptation is therefore motivated by the need to correct recording-dependent variation while retaining transferable sleep-stage information. DSA explicitly aligns domain statistics, whereas BN adaptation recalibrates layer-wise activation means and variances using unlabeled target recordings. Together, these operations can reduce statistical mismatch in the learned representations, particularly when domain shifts contain substantial location and scale components. Compared with the source-only DG baselines in Table 6, this strategy directly exploits

the observed target distribution to calibrate the model for deployment. Such target-specific correction provides a plausible explanation for the reported accuracy advantages across the three target datasets and supports statistical adaptation as a practical complement to transferable representation learning.

B Theoretical Results and Proofs

B.1 Cross-epoch information and prediction risk

We retain the method's notation, with $C = 5$ and $w = (1, 1.5, 1, 1, 1)$ in stage order W, N1, N2, N3, REM. All logarithms are natural; $0 \log 0 = 0$ and $\|\cdot\|$ is the Euclidean norm. For the epoch encoder, sequence encoder, and output head, respectively, write

$$Z_i \in \mathbb{R}^{512}, \qquad V_i = \psi_i(Z_i, Z_{-i}) \in \mathbb{R}^{512}, \qquad f(V_i) \in \Delta^{C-1}, \tag{B.1}$$

where Z_{-i} contains the other embeddings in the window of T epochs. For any representation V and measurable predictor f, define

$$R_w(f; V) = \mathbb{E}[-w(Y_i) \log f_{Y_i}(V)], \qquad R_w^*(V) = \inf_f R_w(f; V),$$
$$\kappa_w = \mathbb{E}[w(Y_i)] \in (0, \infty), \qquad \frac{d\mathbb{P}_w}{d\mathbb{P}} = \frac{w(Y_i)}{\kappa_w}.$$

Under $\mathbb{P}_w$, the conditional posterior is

$$p_w(c \mid V) = \frac{w(c)\mathbb{P}(Y_i = c \mid V)}{\sum_{j=1}^{C} w(j)\mathbb{P}(Y_i = j \mid V)}.$$

Proof of Proposition 1. For probability vectors p, r, with the usual infinite-divergence convention,

$$D_{\mathrm{KL}}(p\|r) = -\sum_{c: p_c > 0} p_c \log \frac{r_c}{p_c} \geq \sum_{c: p_c > 0} (p_c - r_c) \geq 0.$$

Consequently,

$$\begin{aligned} R_w(f; V) &= \kappa_w \mathbb{E}_w \left[-\sum_c p_w(c \mid V) \log f_c(V) \right] \\ &= \kappa_w H_w(Y_i \mid V) + \kappa_w \mathbb{E}_w D_{\mathrm{KL}}\big(p_w(\cdot \mid V) \| f(V)\big), \\ R_w^*(V) &= R_w(p_w; V) = \kappa_w H_w(Y_i \mid V). \end{aligned} \tag{B.2}$$

Applying (B.2) to Z_i and (Z_i, Z_{-i}) yields

$$\begin{aligned} R_w^*(Z_i) - R_w^*(Z_i, Z_{-i}) &= \kappa_w \big[H_w(Y_i \mid Z_i) - H_w(Y_i \mid Z_i, Z_{-i})\big] \\ &= \kappa_w \mathbb{E}_w D_{\mathrm{KL}}\big(p_w(\cdot \mid Z_i, Z_{-i}) \| p_w(\cdot \mid Z_i)\big) \\ &= \kappa_w I_w(Y_i; Z_{-i} \mid Z_i) \geq 0. \end{aligned}$$

Equality holds iff $p_w(\cdot \mid Z_i, Z_{-i}) = p_w(\cdot \mid Z_i)$ almost surely. □

For the learned sequence encoder and head in (B.1), let

$$\begin{aligned} a_\psi &= \kappa_w \mathbb{E}_w D_{\mathrm{KL}}\big(p_w(\cdot \mid Z_i, Z_{-i}) \| p_w(\cdot \mid V_i)\big) \\ &= \kappa_w \big[H_w(Y_i \mid V_i) - H_w(Y_i \mid Z_i, Z_{-i})\big] \geq 0, \\ e_f &= \kappa_w \mathbb{E}_w D_{\mathrm{KL}}\big(p_w(\cdot \mid V_i) \| f(V_i)\big) \geq 0. \end{aligned}$$

Since V_i is a function of (Z_i, Z_{-i}), telescoping (B.2) gives

$$\begin{aligned} &R_w(f; V_i) - R_w^*(Z_i) = -\kappa_w I_w(Y_i; Z_{-i} \mid Z_i) + a_\psi + e_f, \\ &R_w(f; V_i) < R_w^*(Z_i) \iff \kappa_w I_w(Y_i; Z_{-i} \mid Z_i) > a_\psi + e_f. \end{aligned} \tag{B.3}$$

Thus, neighboring epochs benefit the implemented sequence model when their additional label information exceeds sequence encoder compression and head-fitting errors. No within-window independence assumption is required.

B.2 TypCert selection and cohort gradient bias

Condition on $\mathcal{F}$, generated by the source model, unlabeled target recordings, their lengths, zero-shot predictions, recalibrated normalization statistics, and label-free selection randomness. The nonempty cohort $A \subseteq U$ and parameter point θ are $\mathcal{F}$-measurable. For any nonempty $D \subseteq U$, define

$$p_{i,c} = \mathbb{P}(Y_i = c \mid \mathcal{F}), \qquad \widehat{y}_i = \arg\max_c q_{i,c},$$
$$\pi_D = \frac{1}{N_D}\sum_{i\in D} p_i, \qquad \widehat{\pi}_D = \frac{1}{N_D}\sum_{i\in D} e_{\widehat{y}_i},$$
$$\delta_D = \frac{1}{N_D}\sum_{i\in D} \mathrm{TV}(p_i, q_i), \qquad \varepsilon_A = \delta_A + \bar{H}(U) + \delta_U.$$

Here $\mathrm{TV}(a,b) = \frac{1}{2}\|a-b\|_1$, e_c is the c-th one-hot vector, and class ties use a fixed rule. The sum over $i \in D$ includes all epochs of subjects in D; $\widehat{\pi}_U = \widehat{\pi}_T$.

Lemma B.1 (Cohort marginal discrepancy)**.** *For every such A,*

$$\mathrm{TV}(\pi_A, \pi_U) \le d_{\mathrm{TV}}(A) + \bar{H}(A) + \varepsilon_A.$$

Proof. For each epoch and each nonempty $D \subseteq U$,

$$\mathrm{TV}(q_i, e_{\widehat{y}_i}) = 1 - q_{i,\widehat{y}_i} \le -\log q_{i,\widehat{y}_i} \le -\sum_c q_{i,c}\log q_{i,c} = H(q_i), \tag{B.4}$$

$$\begin{aligned}\mathrm{TV}(\pi_D, \widehat{\pi}_D) &\le \frac{1}{N_D}\sum_{i\in D}\mathrm{TV}(p_i, e_{\widehat{y}_i}) \\ &\le \frac{1}{N_D}\sum_{i\in D}\big[\,\mathrm{TV}(p_i,q_i) + H(q_i)\big] = \delta_D + \bar{H}(D).\end{aligned}$$

Two applications of the triangle inequality give

$$\begin{aligned}\mathrm{TV}(\pi_A,\pi_U) &\le \mathrm{TV}(\pi_A,\widehat{\pi}_A) + \mathrm{TV}(\widehat{\pi}_A,\widehat{\pi}_U) + \mathrm{TV}(\widehat{\pi}_U,\pi_U) \\ &\le \delta_A + \bar{H}(A) + d_{\mathrm{TV}}(A) + \bar{H}(U) + \delta_U.\end{aligned}$$

□

For the contextual classifier $f_\theta(i)$, define

$$\ell_i(\theta,c) = -w(c)\log f_{\theta,c}(i), \qquad R_D(\theta) = \frac{1}{N_D}\sum_{i\in D}\sum_c p_{i,c}\ell_i(\theta,c),$$
$$g_D = \nabla R_D(\theta), \qquad g_{D,c} = \frac{\sum_{i\in D} p_{i,c}\nabla\ell_i(\theta,c)}{N_D\pi_{D,c}},$$
$$\rho_A = \sum_c \pi_{A,c}\|g_{A,c} - g_{U,c}\|, \qquad G \ge \max_{c:\pi_{U,c}>0}\|g_{U,c}\|.$$

Assume these derivatives exist; zero-mass class gradients are set to zero. Since $\pi_{U,c} = 0 \Rightarrow \pi_{A,c} = 0$, the decomposition is valid for all classes:

$$\begin{aligned}g_A - g_U &= \sum_c(\pi_{A,c} - \pi_{U,c})g_{U,c} + \sum_c \pi_{A,c}(g_{A,c} - g_{U,c}), \\ \|g_A - g_U\| &\le G\|\pi_A - \pi_U\|_1 + \rho_A \\ &\le 2G\big[d_{\mathrm{TV}}(A) + \bar{H}(A) + \varepsilon_A\big] + \rho_A =: B_A.\end{aligned} \tag{B.5}$$

The two observed TypCert scores enter B_A directly. The residuals δ_D and ρ_A retain posterior error and within-stage gradient mismatch; neither is assumed to vanish.

B.3 SUBJECT-LEVEL VARIANCE AND RESTRICTED ADAPTATION

Conditionally on $\mathcal{F}$, subject label vectors are independent; within-subject dependence is unrestricted. Let $P = P^\top = P^2$ be $\mathcal{F}$-measurable and assume

$$\widehat{g}_u = \frac{1}{N_u} \sum_{i \in u} \nabla \ell_i(\theta, Y_i), \qquad \Sigma_u = \mathrm{Cov}(\widehat{g}_u \mid \mathcal{F}) \preceq \sigma^2 I.$$

Set

$$\alpha_u = \frac{N_u}{N_A}, \qquad \widehat{g}_A = \sum_{u \in A} \alpha_u \widehat{g}_u, \qquad m_{\mathrm{eff}}(A) = \left(\sum_{u \in A} \alpha_u^2 \right)^{-1}.$$

Proof of Theorem 1. Conditional measurability of A, θ implies

$$\mathbb{E}[\widehat{g}_A \mid \mathcal{F}] = \sum_{u \in A} \frac{\alpha_u}{N_u} \sum_{i \in u} \sum_c p_{i,c} \nabla \ell_i(\theta, c) = g_A,$$

$$\Sigma_A := \mathrm{Cov}(\widehat{g}_A \mid \mathcal{F}) = \sum_{u \in A} \alpha_u^2 \Sigma_u \preceq \frac{\sigma^2}{m_{\mathrm{eff}}(A)} I.$$

Let $b_A = g_A - g_U$ and $\xi_A = \widehat{g}_A - g_A$. Then

$$P\widehat{g}_A - g_U = Pb_A + P\xi_A - (I - P)g_U, \quad \mathbb{E}[\xi_A \mid \mathcal{F}] = 0, \quad P(I - P) = 0.$$

Orthogonality and centering yield

$$\begin{aligned} \mathcal{E}_A(P) &:= \mathbb{E}[\|P\widehat{g}_A - g_U\|^2 \mid \mathcal{F}] \\ &= \|(I - P)g_U\|^2 + \|Pb_A\|^2 + \mathbb{E}[\|P\xi_A\|^2 \mid \mathcal{F}] \\ &= \|(I - P)g_U\|^2 + \|Pb_A\|^2 + \mathrm{tr}(P\Sigma_A P). \end{aligned} \tag{B.6}$$

For an orthonormal basis $(v_j)_{j=1}^r$ of $\mathrm{range}(P)$,

$$V_A(P) := \mathrm{tr}(P\Sigma_A P) = \sum_{j=1}^r v_j^\top \Sigma_A v_j \le \frac{\sigma^2 r}{m_{\mathrm{eff}}(A)}, \qquad r = \mathrm{rank}(P),$$

$$\mathcal{E}_A(P) \le \|(I - P)g_U\|^2 + B_A^2 + \frac{\sigma^2 \,\mathrm{rank}(P)}{m_{\mathrm{eff}}(A)},$$

using $\|Pb_A\| \le \|b_A\| \le B_A$ from (B.5). □

For the entropy criterion, define

$$d_P = \max_{i,c,c'} \|P[\nabla \ell_i(\theta, c) - \nabla \ell_i(\theta, c')]\|, \qquad a_u = \frac{1}{N_u} \sum_{i \in u} \nabla \ell_i(\theta, \widehat{y}_i).$$

Centering at the conditional mean and applying Jensen's inequality give

$$\begin{aligned} \mathrm{tr}(P\Sigma_u P) &\le \mathbb{E}[\|P(\widehat{g}_u - a_u)\|^2 \mid \mathcal{F}] \\ &\le \frac{1}{N_u} \sum_{i \in u} \mathbb{E}[\|P(\nabla \ell_i(\theta, Y_i) - \nabla \ell_i(\theta, \widehat{y}_i))\|^2 \mid \mathcal{F}] \\ &\le \frac{d_P^2}{N_u} \sum_{i \in u} (1 - p_{i,\widehat{y}_i}) \le d_P^2 (\bar{H}_u + \delta_{\{u\}}). \end{aligned}$$

The last inequality follows from $1 - p_{i,\widehat{y}_i} \le 1 - q_{i,\widehat{y}_i} + \mathrm{TV}(p_i, q_i)$ and (B.4). Hence

$$V_A(P) \le d_P^2 \sum_{u \in A} \alpha_u^2 (\bar{H}_u + \delta_{\{u\}}) \le d_P^2 \max_{u \in A} \alpha_u \, (\bar{H}(A) + \delta_A),$$

$$1 \le m_{\mathrm{eff}}(A) = \frac{N_A^2}{\sum_{u \in A} N_u^2} \le |A| \qquad \text{(Cauchy–Schwarz)}.$$

Thus, certainty controls a variance bound through $\bar{H}(A)$, while effective subject coverage controls its scale. Repeated recordings of one person are grouped; independent epochs are not assumed.

B.4 Risk decrease and budget-dependent adaptation

Corollary B.1 (Local risk decrease)**.** *Under Theorem 1, let $\theta^+ = \theta - \eta P \widehat{g}_A$, with $0 < \eta \le L^{-1}$. If R_U is L-smooth on every possible update segment, then*

$$\mathbb{E}[R_U(\theta^+) \mid \mathcal{F}] - R_U(\theta) \le -\frac{\eta}{2}\|Pg_U\|^2 + \frac{\eta}{2}B_A^2 + \frac{L\eta^2}{2}V_A(P). \tag{B.7}$$

Proof. Smoothness, conditional expectation, and polarization imply

$$\begin{aligned} R_U(\theta^+) - R_U(\theta) &\le -\eta\langle Pg_U, P\widehat{g}_A\rangle + \frac{L\eta^2}{2}\|P\widehat{g}_A\|^2, \\ \mathbb{E}[R_U(\theta^+) \mid \mathcal{F}] - R_U(\theta) &\le -\eta\langle Pg_U, Pg_A\rangle + \frac{L\eta^2}{2}\big[\|Pg_A\|^2 + V_A(P)\big] \\ &= -\frac{\eta}{2}\|Pg_U\|^2 + \frac{\eta}{2}\|Pb_A\|^2 - \frac{\eta}{2}(1 - L\eta)\|Pg_A\|^2 + \frac{L\eta^2}{2}V_A(P) \\ &\le -\frac{\eta}{2}\|Pg_U\|^2 + \frac{\eta}{2}B_A^2 + \frac{L\eta^2}{2}V_A(P). \end{aligned}$$

□

Combining (B.7) with (B.5),

$$\begin{aligned} \|Pg_U\|^2 > \big\{2G[d_{\mathrm{TV}}(A) + \bar{H}(A) + \varepsilon_A] + \rho_A\big\}^2 + L\eta V_A(P) \\ \implies \mathbb{E}[R_U(\theta^+) \mid \mathcal{F}] < R_U(\theta). \end{aligned} \tag{B.8}$$

This condition links the selected cohort and trainable parameters to target-pool risk.

For nested orthogonal projectors $P_1P_2 = P_1$, let $Q = P_2 - P_1$. Then $Q^2 = Q$, $P_1Q = 0$, and (B.6) gives

$$\begin{aligned} \mathcal{E}_A(P_1) - \mathcal{E}_A(P_2) &= \|Qg_U\|^2 - \|Qb_A\|^2 - \mathrm{tr}(Q\Sigma_A Q), \\ \mathcal{E}_A(P_2) < \mathcal{E}_A(P_1) &\iff \|Qg_U\|^2 > \|Qb_A\|^2 + \mathrm{tr}(Q\Sigma_A Q). \end{aligned}$$

In CtxSleepNet, $P_1 = P_{\mathrm{NH}}$ selects normalization affine parameters and the classifier (1%–2%); $P_2 = I$ permits full adaptation (5%–10%). The additional variance satisfies

$$\mathrm{tr}(Q\Sigma_A Q) \le \frac{\sigma^2 \,\mathrm{rank}(Q)}{m_{\mathrm{eff}}(A)}.$$

Thus, greater effective subject coverage relaxes the noise penalty for updating the encoder; selection bias and omitted target signal remain explicit.

For the rank rule, set $d_t(u) = d_{\mathrm{TV}}(S_t \cup \{u\})$, $h_t(u) = \bar{H}(S_t \cup \{u\})$, and $s_t(u) = r_t^{\mathrm{TV}}(u) + r_t^{H}(u)$. With equal values assigned equal ranks, suppose $v \in R_t$ dominates a minimizer u_t:

$$\begin{aligned} d_t(v) \le d_t(u_t), \quad h_t(v) \le h_t(u_t), \quad (d_t(v), h_t(v)) \ne (d_t(u_t), h_t(u_t)) \\ \implies s_t(v) < s_t(u_t) = \min_{u \in R_t} s_t(u), \end{aligned}$$

a contradiction. Strictly increasing transforms preserve both ranks. Therefore the selected union is Pareto-undominated among current candidates; global minimization of B_A is not implied.

All adaptation statements are conditional, local results with fixed recalibrated running statistics. They do not assert multi-step Adam convergence or risk guarantees for unobserved subjects.