sDREAMER: Self-distilled Mixture-of-Modality-Experts Transformer for Automatic Sleep Staging

Jingyuan Chen[1]   Yuan Yao[1]   Mie Anderson[2]   Natalie Hauglund[2]   Celia Kjaerby[2]
Verena Untiet[2]   Maiken Nedergaard[2]   Jiebo Luo[1]

[1]University of Rochester   [2]University of Copenhagen

jchen157@u.rochester.edu   yyao39@ur.rochester.edu   jluo@cs.rochester.edu
{mieandersen,natalie.hauglund,celia.kjaerby,verena,nedergaard}@sund.ku.edu

Abstract—Automatic sleep staging based on electroencephalography (EEG) and electromyography (EMG) signals is an important aspect of sleep-related research. Current sleep staging methods suffer from two major drawbacks. First, there are limited information interactions between modalities in the existing methods. Second, current methods do not develop unified models that can handle different sources of input. To address these issues, we propose a novel sleep stage scoring model sDREAMER, which emphasizes cross-modality interaction and per-channel performance. Specifically, we develop a mixture-of-modality-expert (MoME) model with three pathways for EEG, EMG, and mixed signals with partially shared weights. We further propose a self-distillation training scheme for further information interaction across modalities. Our model is trained with multi-channel inputs and can make classifications on either single-channel or multi-channel inputs. Experiments demonstrate that our model outperforms the existing transformer-based sleep scoring methods for multi-channel inference. For single-channel inference, our model also outperforms the transformer-based models trained with single-channel signals.

Index Terms—sleep scoring, distillation, transformer, mixture-of-modality experts

I. INTRODUCTION

Sleep is one of the most basic animal behaviours with wide biological implications. Classifying sleep into stages (typically Wakefulness, REM and non-REM) is an important aspect of sleep-related research [37]. Currently, one of the most common automated sleep staging techniques is based on electrophysiological time-series data. Polysomnography (PSG) by experts requires specialized knowledge of sleep architecture, as well as considerable time (*i.e.*experts have to observe the whole sleep process, which lasts about several hours), thus validating the necessity of automatic sleep staging.

In recent years, deep-learning-based methods have shown potential in sleep staging. Most of these works involve the electroencephalography (EEG) time series signals signals [6], [12], [29]. However, it has been shown that sleep staging with only EEG signals is insufficient for sleep classification (typically in distinguishing REM sleep and non-REM sleep) [4], [17], thus necessitating multi-modality sleep staging (*i.e.* staging with multiple types of input signals) [1], [15], [30], [32]. Although
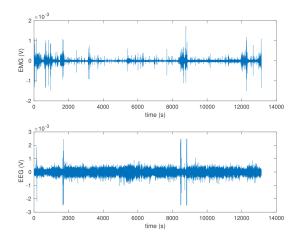

Fig. 1: The raw EMG and EEG time-series signals.

these multi-modality methods have improved the performance of sleep staging, there are two major drawbacks.

First, previous methods suffer from limited information interactions between modalities. While previous studies have attempted to develop multi-modal models that integrate information across different modalities, they ignore the possible interrelationship of different electrophysiological signals. Specifically, these models embed different signal inputs into separate feature spaces and use a late fusion scheme to generate classification results. Since the electrophysiological signals are quite noisy (see Fig. 1), the cross-reference from multiple input signals can be essential in improving the robustness of automatic sleep staging models.

Second, previous methods have not developed unified models that can handle different sources of input, such as multi-channel and single-channel signals. For instance, a model trained on EEG and EMG signals cannot perform reasonable inferences on new data with only EEG signals. This limitation is brought naturally since such models focus on multi-channel joint embedding, but neglect the quality of per-channel embed-

ding. As a result, the performance of these models is restricted, and their potential applications are hindered. Therefore, there is a need for a unified model that can handle different sources of input and improve the performance of automatic sleep staging models.

To address these problems, this study proposes a novel sleep stage scoring model that emphasizes the cross-modality alignment and per-channel performance. To achieve this, our model introduces the mixture-of-modality-experts (MoME) transformer. More specifically, we develop a model consisting of three pathways for EEG, EMG, and mixed signals with partially shared weights. Specifically, the attention weights are shared across the three paths, and the weights of the feed-forward networks are not fully shared (see Section IV for details). The partially shared weights implicitly instruct our model to align across different modalities. The three pathways ensure that our model can promote high-quality per-channel performance. Furthermore, we propose a self-distillation method to ensure better information interaction across modalities. We build an epoch MoME transformer for one-to-one sleep staging and a sequence MoME transformer for many-to-many sleep staging. Although our model is trained with multi-channel time-series signals (*i.e.* EEG and EMG), our model can accept single-channel (*i.e.* EEG or EMG) as well as multi-channel time-series signal as input in the inference phase.

We demonstrate the effectiveness of our proposed self-distilled MoME transformer, referred to as sDREAMER (self-distilled Mixture-of-Modality-Experts Transformer), through experiments on a public mice sleep dataset [18] labeled with sleep stages by our expert. For multi-channel sleep staging, our epoch MoME model and sequence MoME model outperform the transformer-based automatic sleep scoring method by significant margins, respectively. For single-channel sleep staging, our sequence MoME model also produces better inference results than the transformer-based models trained with single-channel signals.

In summary, the main contributions of our work can be summarized as follows:

- We propose the unified sleep staging framework sDREAMER that can handle either single-channel or multi-channel input based on the MoME transformer.
- We propose a self-distillation method for our proposed MoME transformer to ensure the multi-modality information interaction across different pathways (*i.e.*EEG, EMG, and mixed).
- We demonstrate through extensive experiments that our proposed model is effective for both single-channel and multi-channel sleep staging.

II. RELATED WORK

A. Deep-learning-based Sleep Staging

The success of deep learning has inspired several deep-learning-based sleep staging algorithms based on electrophysiological signals. Basically, there are two ways to deal with the raw data. One way is to directly learn from the unprocessed signals [25], [34], [38]. Another way is to transfer them to spectrograms and design models to learn from the spectrograms using computer vision-based techniques [22], [42]. Although introducing spectrograms may produce more interpretable visual understanding for humans, Phan *et al.* have shown that introducing spectrogram inputs in addition to raw signal inputs may actually lead to a performance drop [30]. Therefore, we focus on automatic sleep staging based on raw electrophysiological signals in our work.

Most automatic sleep staging works involve the time-series EEG signals [6], [10], [12], [26], [27], [29], [31], [36]. However, previous works have shown that sleep staging with only EEG signals is insufficient for sleep classification (typically in distinguishing REM sleep and non-REM sleep) [4], [17], thus validating the necessity of multi-modality sleep staging (*i.e.*staging with multiple types of input signals) [1], [15], [30], [32]. One common way of modeling the different types of input signals is the late fusion strategy [14], [21]. Despite their success in introducing different time-series signals into sleep staging, one major problem of these methods is the lack of information interaction across modalities. Although Jathurshan *et al.* have proposed cross-modal attention [32], the shallow design is insufficient for information interactions.

B. Transformer-based Cross-modality Learning

Transformer [41] is a deep-learning framework that can handle sequence input. The recent success in vision-language has demonstrated the power of transformer in cross-modality learning [19], [33], [43]. Typically, there are two types of structures for cross-modality learning: 1) fusion encoder, where different input modalities share the same encoder structure [39], [48] and 2) dual encoder, where different input modalities have different encoder structure but are mixed together in the feature space for cross-modal interactions [13], [44], [46]. Bao *et al.* suggest that mixture-of-modality-experts transformer can facilitate cross-modality learning efficiently and competitively [5]. Despite the effectiveness of cross-modality learning, most works on sleep staging ignore such techniques for modeling different types of input signals.

C. Self Distillation

Self distillation is a type of knowledge distillation where the teacher network and the student network share the same model. In one type of self-distillation methods, knowledge can be transferred from the earlier epochs to the later epochs [35], [45], [51]. Knowledge can also be transferred from deeper parts of a neural network to the more shallow parts [11], [47]. Other self-distillation methods focus on the data and label augmentation [8], [50] or weak labels [3], [49]. The effectiveness

of self-distillation has been analyzed and validated by some works [2], [23]. Recently, Wang *et al.* have proposed to distill across vision and language domains [44]. Inspired by their work, we propose a novel knowledge distillation scheme in the MoME module to distill across signal modalities for sleep staging.

III. DATA COLLECTION AND PREPROCESSING

A. Data Collection

This section presents the data collection procedures of the mouse sleep staging dataset. The dataset is a publicly available mice sleep dataset [18], while our experts manually labeled the dataset with sleep stages. All the sleep monitoring data are recorded from wildtype C57BL/6 mice subjects including both male and female ones. In particular, the EEG and EMG signals are collected when the subjects are placed in the chamber room. To denoise the collected EEG signals, experts apply filters including a high-pass filter at 1 Hz and a low-pass filter at 100 Hz. For the EMG signals, they apply a high-pass filter at 10 Hz and a low-pass filter at 100 Hz. In addition, a notch filter of 50 Hz is applied to eliminate the power line noise. Based on these raw data and the videos of sleeping mice, our experts annotate each collected signal epoch with its corresponding sleep stage. Specifically, wake stages show high muscle tonus and a high-frequency, low-amplitude EEG pattern. SWS sleep shows no muscle tonus and low-frequency, high-amplitude EEG pattern. REM sleep also shows no muscle tonus but with high frequency, low-amplitude EEG. A detailed description of the dataset can be found in Section V-A.

B. Data Preprocessing

The dataset mainly contains the EEG and EMG signal records collected from sleeping mice, as well as the sleep stage labels for every second. These raw data may suffer from ambiguity in labels, data noise and excessive sequence length. Data preprocessing is applied to the raw data, including irregularity removal, subject-wise normalization, and temporal slicing.

Irregularity Removal. During the whole span of mice sleep, there exists some time periods when experts are not sure which stage these periods belong to. Such periods are not labeled, which causes data irregularity. To handle these missing values, a neurally controlled differential equation approach [16] was considered. However, since missing values only occurs in a small percentage of the total data, it is unnecessary to apply this method to the large electrophysiological signal dataset. Instead, we remove missing values from the time series during loss calculation.

Subject-wise Normalization. Raw signal normalization is performed using a subject-wise normalization approach, which normalizes the signals of the same subject with their mean and standard deviation. This method facilitates model learning as electrophysiological signal features often differ greatly in amplitude among individual subjects.

Temporal Slicing. To obtain input data of a fixed time span, temporal slicing is performed. As our model aims to offer accurate prediction at the second level, the input signals are sliced into per-second signals. The sampling frequency for both signals are equal, denoted as T. Since we choose one second as the slicing time span, the sliced EEG or EMG time-series signal can be denoted as $x^{\text{EEG}} \in \mathbb{R}^{T \times 1}$ or $x^{\text{EMG}} \in \mathbb{R}^{T \times 1}$, respectively.

IV. METHODOLOGY

In this study, we propose sDREAMER, an efficient multi-modal learning framework to stage sleep using a self-distilled mixture-of-modality-experts transformer. We first establish the problem hierarchically into epoch and sequence settings. Then we present the data input under these two settings. Afterward, the MoME module is introduced as the core for the unified multi-modal learning framework. The MoME module also serves as the foundation for our proposed Epoch sDREAMER and Sequence sDREAMER. Both models support multi-modal and mono-modal input cases. To further enhance the performance of mono-modality input, we propose a self-distillation pipeline that employs a multi-modal expert to guide the mono-modality expert.

A. Problem Setup

Sleep stage scoring in mice is commonly approached as a multi-class classification task, requiring the model makes predictions according to the electrophysiological signals. Here we introduce two different settings for sleep stage scoring: epoch and sequence setup.

Epoch Setting. For sleep staging, the smallest annotation window (one second for our dataset) of a signal is referred to as an epoch. As previously defined, the sliced signals x^{EEG} or x^{EMG} belongs to $\mathbb{R}^{T \times 1}$, where T is the size of time dimension, representing number of signal data points in an epoch. Since the input signals can be single-channeled or multi-channeled, an additional modality dimension is introduced, and the epoch extends to $x \in \mathbb{R}^{M \times T \times 1}$, where M is the size of modality dimension. The epoch setting can then be formulated as producing a sleep stage prediction $\hat{y}$ for each epoch signal $x \in \mathbb{R}^{M \times T \times 1}$.

Sequence Setting. The sequence setting differs in considering multiple epochs as input and producing multiple predictions as output. A sequence is a larger window within the whole signal trace. It consists of multiple consecutive epochs. A K-size sequence is denoted as $(x_i)_{i=1}^{K}$ where x_i is the i-th epoch within the sequence. The sleep staging for a sequence is defined as the process of using a set of epoch signals, represented as $(x_i)_{i=1}^{K}$, to predict a corresponding set of predictions, denoted as $(\hat{y}_i)_{i=i}^{K}$ by the model. These two settings are also referred to as one-to-one and many-to-many predictions conventionally [28] [32].

B. Data Input

Patching Operator To process the epoch signal x_i in a K-size sequence $(x_1, x_2, \cdots, x_K)$, a patch operator is used to divide
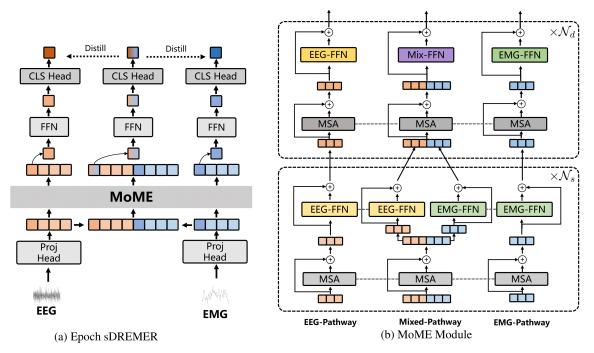
Fig. 2: The overall structure of the mixture-of-modality-experts module and the epoch sDREAMER model.

each epoch signal into a series of non-overlapping windows of equal length W. This operation generates a patched signal $x_i \in \mathbb{R}^{P \times W}$, where $P = \lfloor \frac{T}{W} \rfloor$. The patch operator reduces the number of tokens for transformer input by a factor of P, resulting in a significant decrease in the computation complexity for multi-head self-attention by a factor of P^2, thereby enabling more efficient modeling.

Initial Representation Given an EEG and EMG epoch x, we apply the aforementioned patching operator to obtain their signal patches. The resulting patches are then embedded in latent spaces to form intermediate initial tokens $\boldsymbol{I}_0^m$ as in Equation 1, where m denotes the modality, m_i denotes the i-th epoch of the modality, and $\boldsymbol{E}_{\text{trans}}^m$ is the linear transformation matrix of this modality.

$$\boldsymbol{I}_0^{m(i)} = \left[m_i^1 \boldsymbol{E}_{\text{trans}}^m; m_i^2 \boldsymbol{E}_{\text{trans}}^m; \cdots m_i^P \boldsymbol{E}_{\text{trans}}^m \right], \boldsymbol{E}_{\text{trans}}^m \in \mathbb{R}^{W \times D}$$ (1)

Following the Vision Transformer [9], we initialize two [CLS] tokens for EEG and EMG semantic representation within an epoch. To preserve the temporal and modality relationships among patch tokens, we introduced joint attribute encoding which combines learnable positional and modality encodings. The intermediate initial representations are concatenated with the [CLS] tokens, and the attribute encoding is then added elementwise to the combined representation as shown in Equation 2.

$$\boldsymbol{E}_{\text{attr}}^m = \boldsymbol{E}_{\text{pos}}^m + \boldsymbol{E}_{\text{mod}}^m, \qquad \boldsymbol{E}_{\text{pos}}^m, \boldsymbol{E}_{\text{mod}}^m \in \mathbb{R}^{P \times D}$$
$$\boldsymbol{T}_0^{m(i)} = \text{Concat}\left(\boldsymbol{T}_{\text{cls}}^m; \boldsymbol{I}_0^{m(i)} \right) + \boldsymbol{E}_{\text{attr}}^m$$ (2)

Here, $\boldsymbol{T}_{\text{cls}}^m$, $\boldsymbol{E}_{\text{pos}}^m$, $\boldsymbol{E}_{\text{mod}}^m$ and $\boldsymbol{T}_0^{m(i)}$ represents the [CLS] token, positional encoding, modality encoding and final initial representations for modality m. Afterward, the initial tokens for EEG and EMG at i-th epoch are created and denotes as $\boldsymbol{T}_0^{\text{eeg(i)}}$ and $\boldsymbol{T}_0^{\text{emg(i)}}$, respectively.

C. Mixture-of-Modality-Experts Module

The success of mixture-of-modality-experts (MoME) [5] in the visual language understanding task led us to investigate its potential strength in mice sleep scoring. In our paper, the MoME module is a module of three pathways for EEG, EMG and mix data, with partially shared weights. Our proposed epoch sDREAMER and sequence sDREAMER models are all built upon the MoME module to learn the contextual information within a given signal window. In this section, we discuss a unified token processing framework for the MoME module that applies to both models.

Specifically, a MoME module is composed of multiple MoME layers, as shown in Fig. 2. Each MoME layer is an efficient variant of the transformer layer but differs in the use of modality-agnostic shared multi-head self-attention (MSA) and modality-aware feedforward neuron networks (FFN). These modality-aware FNNs that capture modality-specific latent features are conventionally referred to as modality experts. To specify, our sleep staging model has three modality experts: EEG, EMG, and mix experts. By designing different forward propagation pathways that pass through these modality experts, the MoME module is able to generate both mono-modal and multi-modal representations.

Given the initial representations of N tokens from a modality m denoted as $\boldsymbol{T}_0^m$, a pathway of L layers is represented as
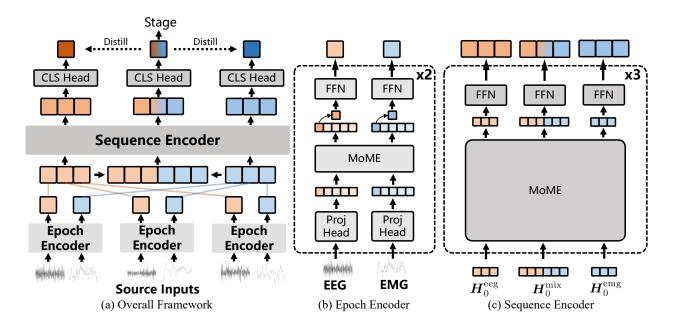
Fig. 3: Overview of the sequence sDREAMER model. The structure of MoME module has already been illustrated in Fig. 2.

follows:

$$T_0^m = \left[t_0^1; t_0^2; \cdots; t_0^N \right]^m \quad (3)$$

$$I_\ell^m = \text{MSA}\left(\text{LN}\left(T_{\ell-1}^m \right) \right) + T_{\ell-1}^m, \quad \ell = 1 \ldots L \quad (4)$$

$$T_\ell^m = \psi\left(m, \ell\right) \left(\text{LN}\left(I_\ell^m \right) \right) + T_\ell^{m'}, \quad \ell = 1 \ldots L \quad (5)$$

Here, the $\psi\left(m, \ell\right)$ is a mapping function from the layer-modality joint space D to the modality experts space F as shown in Equation 6. In other words, the activation of the modality experts is determined by the input modality and layer index. Such a mapping design allows the model to handle various inputs and produce multiple outputs with a unified model. Fig. 2 provides a graphical illustration of this mapping function.

$$\psi\left(m, \ell\right) : D \to F, D = \left\{ (m, \ell) | m \in M, \ell \in \{i\}_{i=1}^L \right\}$$
$$F = \left\{ \text{FFN}^{\text{eeg}}, \text{FFN}^{\text{emg}}, \text{FFN}^{\text{mix}} \right\}, M = \{\text{eeg}, \text{emg}, \text{mix}\} \quad (6)$$

EEG and EMG Pathways. Our MoME modules consist of two mono-modal pathways, namely the EEG-pathway and the EMG-pathway. The EEG/EMG pathway takes the input tokens corresponding to each modality and feeds them through each MoME layer to learn mono-modal contextual information. In each MoME layer, the representations first pass through the multi-head self-attention layer to learn the temporal dependencies between tokens. Next, the representations are projected into modality-specific latent space. Finally, the token representations T_L^{eeg} and T_L^{emg} at the last MoME layer L are outputted.

Mix Pathway. The mix pathway serves to capture both intra-modal and cross-modal information by concatenating tokens

from EEG and EMG and feeding them to the MoME layers. Initially, the mix tokens T_0^{mix} pass through MSA in the early MoME layers before being split and mapped back to their original latent space to maintain modality coherence. Using EEG/EMG modality experts ensures stable modality interaction, as earlier layers' low-level features are susceptible to noise. By projecting representations to their own space, the potential noise entering another space is constrained to a certain degree. However, at later layers, token representations are mapped to a multi-modal space with mixed experts to enable deeper modality interaction. The MoME module's output is the token series from the last layer, denoted as T_L^{mix}.

D. Epoch Mixture-of-Modality-Experts Transformer

Following epoch settings in previous works [28], [32], we proposed an Epoch sDREAMER model for sleep state classification with a one-to-one paradigm. The Epoch sDREAMER, depicted in Fig. 2, consists of an epoch-level MoME module with three pathways, each assigned a modality-specific classification head. The model takes the initial representations for patches of a given epoch signal as input, and the epoch-level MoME module aggregates representative information among patches to mine epoch context.

During training, all three pathways are enabled. The EEG and EMG initial representations propagate along their respective pathways to produce the final representations, and their corresponding [CLS] tokens $T_{\text{cls}}^{\text{eeg}}$ and $T_{\text{cls}}^{\text{emg}}$ are fed to their respective classifiers for prediction. However, the model still lacked cross-modal information, so we leveraged the mix pathway to learn rich cross-modal information while maintaining the information within each modality. The final $T_{\text{cls}}^{\text{mix}}$ token is utilized, which is a combination of the EEG

and EMG representations. During inference, the model can flexibly select the pathway according to the specific token type and generate predictions using only the relevant information within the input token.

E. Sequence Mixture-of-Modality-Experts Transformer

To tackle sleep staging in a sequence setting, we proposed a sequence sDREAMER model that captures contextual information hierarchically. Unlike the epoch setting, where each epoch is considered as a separate sequence, Sequence sDREAMER views each epoch in relation to its neighboring epochs. The architecture of the proposed Sequence sDREAMER model is illustrated in Fig. 3, which employs an epoch-level MoME module at a lower hierarchy to capture epoch-level contexts. Next, a sequence-level MoME transformer is used to capture sequence-level contexts at a higher hierarchy.

Epoch Encoder. Although the epoch-level MoME module in Epoch sDREAMER and Sequence sDREAMER transformer shared the same network architecture, their information flow differs for both training and inference stages. The mix pathway is not enabled in the epoch-level MoME module because multi-modality information is better modeled during the sequence stage at a higher level. This design decision ensures robustness to noise during training and speeds up inference. Thus, the epoch-level MoME module functions as a context extractor for EEG and EMG epoch signals. To compute the initial representations for an EEG/EMG sequence consisting of K epochs, we concatenate the initial tokens from each epoch to form the whole sequence's epoch-level initial representations, denoted as $\boldsymbol{H}_0^m$. Each epoch's initial representation is then passed through the EEG/EMG pathway. Finally, the resulting [CLS] tokens that have passed through the MoME module for all K epochs are concatenated to form the sequence-level initial representations of the given modality.

$$\boldsymbol{H}_0^m = \left[\boldsymbol{T}_0^{m(1)}; \boldsymbol{T}_0^{m(2)}; \cdots ; \boldsymbol{T}_0^{m(K)}\right] \tag{7}$$

$$\boldsymbol{T}_{\text{cls}}^{m(i)} = \text{Epoch-MoME}\left(\boldsymbol{T}_0^{m(i)}\right), i = 1 \ldots K \tag{8}$$

$$\boldsymbol{Z}_0^m = \text{Concat}\left(\boldsymbol{T}_{\text{cls}}^{m(1)}; \boldsymbol{T}_{\text{cls}}^{m(2)}; \cdots ; \boldsymbol{T}_{\text{cls}}^{m(K)}\right) \tag{9}$$

After passing through both pathways, two sequence-level tokens are generated: $\boldsymbol{Z}_0^{\text{eeg}}$ and $\boldsymbol{Z}_0^{\text{emg}}$.

Sequence Encoder. Considering a series of EEG/EMG tokens outputted from the epoch-level MoME module as the context features of each epoch, we now transform them into initial representations to sequence MoME module. To denote the temporal relationship between epochs, two sequence-level positional embeddings $\boldsymbol{E}_{\text{pos}}^{\text{eeg}} \in \mathbb{R}^{K \times D}$ and $\boldsymbol{E}_{\text{pos}}^{\text{emg}} \in \mathbb{R}^{K \times D}$ are added elementwise to EEG and EMG tokens, respectively.
Next, the sequence-level MoME blocks begin to learn dependencies between epochs within the same sequence. Like the Epoch sDREAMER, the EEG and EMG pathways process the input tokens separately and map their learned representations to modality-specific latent space. The output

tokens from each pathway are then projected with pathway-specific classification heads to the label space. Regarding the mix pathway, the mixed EEG-EMG tokens are first projected separately before being mapped to a multi-modal latent space. To better aggregate the multi-modal information learned in the mix pathway, the output EEG-EMG tokens from the same epoch are concatenated along the feature dimension. These concatenated tokens are then mapped to the label space with another pathway-specific head. Finally, three sets of output predictions are generated: $\left\{\boldsymbol{z}_i^{\text{eeg}}\right\}_{i=1}^K$, $\left\{\boldsymbol{z}_i^{\text{emg}}\right\}_{i=1}^K$ and $\left\{\boldsymbol{z}_i^{\text{mix}}\right\}_{i=1}^K$.

F. Self-Distillation

Provided a multi-pathways structure, we seek an effective approach that enables such pathways to co-improve with each other. To this end, a self-distillation framework for better mono-modal and multi-modal representation learning is presented. To specify, the output from the mix pathway is leveraged to distill the outputs from EEG and EMG pathways. Given the outputs of each pathway, the self-distilled loss formulates as the KL-divergence between the predictions from EEG/EMG and pseudo-targets provided by the mixed pathway. Let $\boldsymbol{p}_\tau$ denotes the softmax function with temperature factor τ defined as

$$\boldsymbol{p}_\tau(\boldsymbol{z}_i^{m(k)}) = \frac{\exp(\boldsymbol{z}_i^{m(k)}/\tau)}{\sum_j^C \exp(\boldsymbol{z}_i^{m(j)}/\tau)}, \tag{10}$$

where $z_i^{m(k)}$ denotes the k-th dimension of the logits vector for i-th prediction given by m modality pathway. The self-distillation loss for EEG and EMG formulates as shown in Equation 11 and 12, respectively.

$$\mathcal{L}_{\text{sd-eeg}} = \text{KL}\left(\boldsymbol{p}_{\tau_{\text{eeg}}}\left(\boldsymbol{z}^{\text{eeg}}\right) \| \boldsymbol{p}_{\tau_{\text{eeg}}}\left(\boldsymbol{z}^{\text{mix}}\right)\right) \tag{11}$$

$$\mathcal{L}_{\text{sd-emg}} = \text{KL}\left(\boldsymbol{p}_{\tau_{\text{emg}}}\left(\boldsymbol{z}^{\text{emg}}\right) \| \boldsymbol{p}_{\tau_{\text{emg}}}\left(\boldsymbol{z}^{\text{mix}}\right)\right) \tag{12}$$

The logits vectors of predictions from three pathways are denoted as $\boldsymbol{z}^{\text{eeg}}$, $\boldsymbol{z}^{\text{emg}}$, and $\boldsymbol{z}^{\text{mix}}$. τ_{eeg} and τ_{emg} are temperature factors for the distillation of EEG and EMG. Along with the self-distillation loss, a cross-entropy loss is also incorporated into the optimization process.

$$\mathcal{L}_{\text{ce}} = -\sum_{i=1}^L \boldsymbol{y}_i \log(\boldsymbol{p}_{\tau=1}(\boldsymbol{z}_i^{\text{mix}})) \tag{13}$$

The $\boldsymbol{y}_i$ and $\boldsymbol{z}_i^{\text{mix}}$ indicate the i-th ground truth and prediction from the mix pathway. L is the number of epoch signals. The total loss for MoME models, shown in Equation 14, is a linear combination of self-distillation and cross-entropy loss.

$$\mathcal{L}_{\text{mome}} = (1 - \alpha)\,\mathcal{L}_{\text{ce}} + \frac{\alpha}{2}\left(\mathcal{L}_{\text{sd-emg}} + \mathcal{L}_{\text{sd-emg}}\right) \tag{14}$$

Here, the α is a scaling factor used to balance the weightage of the loss terms.
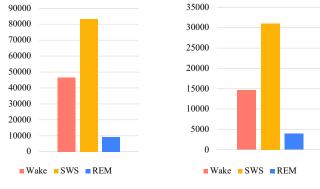
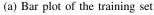TABLE I: Performance comparison between different model architectures on the mouse sleep staging dataset.

Method	Hierarchy	Shared Self-Attention	CA Level	# of Epochs	Acc(%)	F1-Score(%)
Decision Tree	Epoch	-	-	1	79.25	67.65
Random Forest	Epoch	-	-	1	85.47	74.55
AdaBoost	Epoch	-	-	1	85.85	77.78
XGBoost [7]	Epoch	-	-	1	85.29	75.79
MLP	Epoch	-	-	1	81.46	67.05
Bi-LSTM	Epoch	-	-	1	85.57	76.09
Channel-Independent Transformer [24]	Epoch	✓	-	1	86.70	78.40
Dual-Encoder Transformer	Epoch	-	-	1	87.17	78.68
Cross-Attention Transformer	Epoch	-	Epoch level	1	87.07	79.13
Cross-Modal Transformer [32]	Epoch	-	-	1	87.01	-
Epoch sDREAMER	Epoch	✓	-	1	**88.25**	**81.30**
MLP + Transformer	Sequence	-	-	16	88.88	83.33
Bi-LSTM	Sequence	-	-	16	90.07	85.21
Dual-Encoder Transformer	Sequence	-	-	16	90.65	86.37
Cross-Attention Transformer	Sequence	-	Epoch level	16	90.37	86.31
SeqCross-Attention Transformer	Sequence	-	Sequence level	16	90.93	86.48
Cross-Modal Transformer [32]	Sequence	-	Sequence level	16	87.84	-
Sequence sDREAMER	Sequence	✓	-	16	**91.72**	**87.64**

V. EXPERIMENTS

A. Experiment Setup

Dataset. We conduct an evaluation of our proposed epoch sDREAMER and sequence sDREAMER models using the mouse sleep staging dataset, which so far comprises EEG-EMG paired signals collected from mice during sleep and will include additional modalities in the future. Experts have labeled the sleep stages of the mice every second based on these electrophysiological signals, assigning each second of recorded data one of the following labels: Wake, SWS (slow-wave sleep), or REM (rapid eye movement). An illustrative example of the data is shown in Fig. 4.

(a) Bar plot of the training set   (b) Bar plot of the test set

Fig. 5: Bar plot of the class distribution, including (a) training set and (b) test set.

Given that the EEG and EMG signals are sampled at 512 Hz, there can be about 5-10 million EEG/EMG data points for each mouse record. To align with the expert labeling span, we set the epoch window to 1 second, resulting in a total of 10,000 epoch data samples. To evaluate the performance of our model, we adopted subject-wise split criteria, where the records of 12 mice subjects were used for training, while the records of 4 mice subjects were used for testing. The distribution of stages in the training set and test set can be seen in Fig. 5a and Fig. 5b, respectively.

B. Implementation Details

In this section, we explain the epoch and sequence settings of our sDREAMER model. The epoch MoME model involves a 4-layer MoME module, with the first three layers having only EEG and EMG experts, and the last layer incorporating a mixed expert. The feed-forward network dimension for each
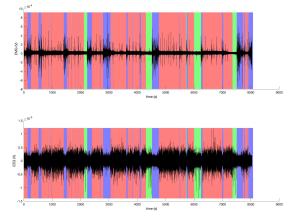

Fig. 4: An example of the collected raw data. The two graphs represent the EMG and EEG data, respectively. The background colors of blue, green, and red are the labels of sleep stages.

The dataset consists of EEG and EMG records of 16 mouse subjects, with each record spanning approximately 4 hours.

expert is 512. The Sequence sDREAMER model comprises a 2-layer MoME module for epoch modeling and a 3-layer MoME module for sequence modeling. The entire epoch-level MoME module and the first two layers of the sequence MoME module have only EEG and EMG experts, and the final layer of the sequence MoME module has an additional mix expert. All experiments are performed on an NVIDIA RTX 3090 GPU with AdamW [20] as the optimizer. The learning rate and weight decay are set to 1e-3 and 1e-4, respectively. The non-hierarchical model utilizes a batch size of 256, while the hierarchical model uses a batch size of 16. The EEG and EMG self-distillation temperature factors are set to $\tau_{\text{eeg}} = 1.0$ and $\tau_{\text{emg}} = 3.0$, respectively, and the distillation weight α is 0.33.

C. Baselines

We compare our method with both machine learning-based and deep learning-based methods. Due to the limited work for mice sleep staging, we implemented multiple baseline model architectures with different settings, from simple to complex. We also include a state-of-the-art cross-modal transformer method [32] as one of the baselines.

Machine-learning-based Methods. Decision tree, AdaBoost, XGBoost, and Random forest were chosen as the baselines for the machine learning-based methods. However, due to the high dimensionality of the EEG-EMG signal, capturing the underlying patterns within the raw signal is challenging for these methods. Therefore, we engineered several features of EEG-EMG traces. In particular, statistical features(*e.g.* mean, max, min, std, and skew) for the raw signal and its first-order derivative are leveraged. Additionally, we applied a discrete Fourier transform to both the raw signal and its derivative and incorporated the statistics in the frequency domain as well.

Deep-learning-based Methods. For deep-learning methods, we developed several baseline models and also compared others' methods. To start with, MLP is a straightforward model with multiple feedforward layers. Bi-LSTM is a sequential modeling model that captures temporal contexts from the forward and backward directions. The channel-independent transformer [24] is an architecture that shares the attention weight for multi-channel input signals. A dual-encoder transformer represents a typical late-fusion model that uses two independent encoders for each input signal and merges the information late in the architecture. The cross-attention transformer is a variant of the dual-encoder architecture that add the cross-attention layer to capture cross-modal information. Lastly, the state-of-the-art cross-modal transformer [32], another variant of the dual-encoder model, uses a self-attention layer to integrate information from two modalities.

D. Quantitative Results

Multi-modal Comparison. In order to demonstrate the efficacy of our model, we conducted a comparative analysis against the aforementioned baseline methods on the multi-modal input setting. Table I presents the experiment results for EEG-EMG paired input, where we report the accuracy

and F1-score for each model. Results show that our proposed epoch-level sDREAMER outperforms the machine-learning-based methods by a significant margin. Furthermore, our sDREAMER model outperforms all the deep-learning-based methods, including the state-of-the-art method cross-modal transformer, on both epoch level and sequence level.

Mono-modal Comparison. To demonstrate our model's generalizability on single-modal inputs, we evaluated our model's ability to produce accurate sleep stage predictions using a single EEG or EMG signal input in the inference stage. It is noteworthy that none of the baseline methods trained with multi-modal input can make inference with single-modal input. Therefore, we compare our model's performance with the baseline dual-encoder transformer model trained solely on EEG or EMG. Results are presented in Table II. The results indicate that our proposed model was able to learn a superior mono-modal representation by utilizing the multi-modality as a form of supervision during training.

E. Ablation Studies

In this section, we present the ablation studies for our model. Specifically, as the effectiveness of the MoME transformer has been validated, we hereby conduct an ablation study regarding positional and modality embedding to investigate the contribution of different attribute embedding. As shown in Table IV, the joint attribute encoding improves the performance by an increase of approximately 1.0% in accuracy. Among the two encodings, positional encoding is of greater importance than modality encoding. One possible explanation for this is that there are only two kinds of electrophysiological signals incorporated within this sleep staging task, thus weakening the impacts of modality encoding. In addition, we also perform an ablation study on the contribution of self-distillation. Table III indicates that the self-distillation not only improved the mono-modality performance significantly but also promoted the multi-modal performance by a large margin.

F. Visualizations

t-SNE Analysis. To explore the embeddings learned by our model, we utilized t-SNE [40] to visualize the high-dimensional embedding in a two-dimensional space. The resulting visualization, shown in Fig. 6, demonstrates the disentangled and aligned latent space learned by Sequence sDREAMER. In particular, the representations spaces for all three pathways are visualized, and they indicate a similar spatial pattern. This pattern alignment is non-trivial because we did not provide any supervision directly that would promote such alignment in the latent space—suggesting shared attention mechanism and self-distillation have implicitly benefited the multi-modal alignment. The t-SNE plot also reveals spatial continuity among the Wake, REM, and SWS sleep stages in the embedding space, consistent with real-world scenarios. Our model implicitly learns to maintain the relative spatial relationships in the ground truth space without injecting any prior knowledge of actual spatial relativity. These findings

(a) EEG-EMG embeddings   (b) EEG embeddings   (c) EMG embeddings
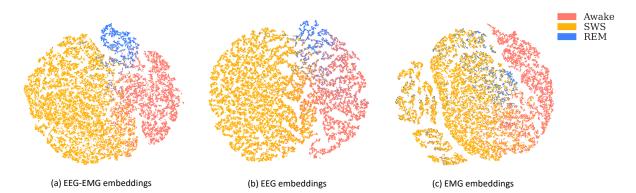
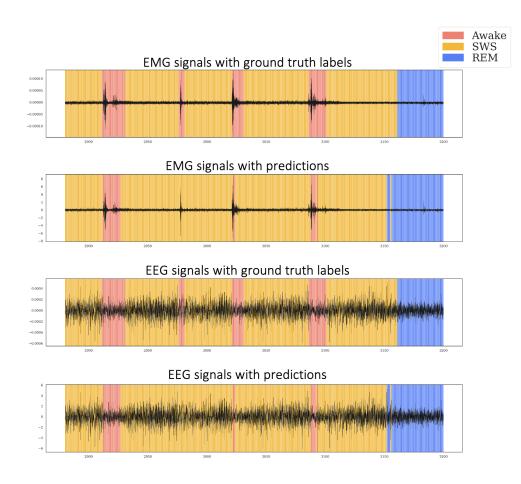Fig. 6: Visualization of learned embeddings using t-SNE.

Fig. 7: Visualization of the Predictions given by the sDREAMER.

TABLE II: Mono-modal inference performance comparison of the vanilla transformer and Sequence sDREAMER on the mouse sleep dataset. The results with the best performance are highlighted in bold. "EEG/EMG-Acc" represents the prediction accuracy when taking EEG/EMG as input. Similarly, "EEG/EMG-F1" represents the F1-score in the same context.

Method	Train Signal	EEG-Acc(%)	EEG-F1(%)	EMG-Acc(%)	EMG-F1(%)
Dual-Encoder Transformer	EEG	87.16	72.32	-	-
Dual-Encoder Transformer	EMG	-	-	83.78	57.57
Sequence sDREAMER	EEG-EMG	**88.12**	**77.83**	**83.78**	**62.82**

TABLE III: Ablation study on modality-experts self-distillation. The accuracy is reported.

Method	EEG-SD	EMG-SD	Acc	EEG-Acc	EMG-Acc
Sequence sDREAMER	✓	✓	**91.72**	**88.12**	**83.78**
Ablations	✓	-	91.45	87.57	22.67
	-	✓	90.02	60.18	83.49
	-	-	77.13	62.71	7.85

TABLE IV: Ablation study on positional encoding and modality encoding. The accuracy is reported.

Method	Pos Encoding	Mod Encoding	Acc(%)
Sequence sDREAMER	✓	✓	**91.72**
Ablations	-	✓	91.45
	✓	-	91.65
	-	-	90.99

highlight the effectiveness and interpretability of our proposed method in sleep stage scoring.

Visualizations. In Fig. 7, we give some predictions results of our sequence sDREAMER model. From the visualization results, we see that our model predicts most of the sleep stages correct.

VI. CONCLUSIONS AND DISCUSSIONS

Automatic sleep staging is a crucial first step for sleep-related research. In this paper, we propose sDREAMER, a sleep staging model that can handle both single-channel and multi-channel inputs. Our sDREAMER emphasizes the cross-modality interaction and per-channel performance, and therefore generates high-quality staging results on the sleep staging task. Extensive experiments demonstrate the effectiveness of our proposed sDREAMER.

There are some limitations in our work. The performance of the model is currently compared to one expert, who is well-trained but not perfect. Our model may show greater potential if the dataset is scored by more experts. In addition, if the dataset is scored by two newly trained students, this experiment will show whether the model can outperform newly trained students and therefore replace human scoring.

ACKNOWLEDGMENTS

REFERENCES

[1] Keishi Akada, Takuya Yagi, Yuji Miura, Carsten T Beuckmann, Noriyuki Koyama, and Ken Aoshima. A deep learning algorithm for sleep stage scoring in mice based on a multimodal network with fine-tuning technique. *Neuroscience Research*, 173:99–105, 2021.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[3] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16430–16441, 2022.

[4] Fernando Andreotti, Huy Phan, Navin Cooray, Christine Lo, Michele TM Hu, and Maarten De Vos. Multichannel sleep stage classification and transfer learning using convolutional neural networks. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 171–174. IEEE, 2018.

[5] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[6] Siddharth Biswal, Joshua Kulas, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Matt T Bianchi, and Jimeng Sun. Sleepnet: automated sleep staging system via deep learning. *arXiv preprint arXiv:1707.08262*, 2017.

[7] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[8] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] 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.

[11] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1013–1021, 2019.

[12] Jing Huang, Lifeng Ren, Xiaokang Zhou, and Ke Yan. An improved neural network based on senet for sleep stage classification. *IEEE Journal of Biomedical and Health Informatics*, 26(10):4948–4956, 2022.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[14] Ziyu Jia, Xiyang Cai, and Zehui Jiao. Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging. *IEEE Sensors Journal*, 22(4):3464–3471, 2022.

[15] Ziyu Jia, Youfang Lin, Jing Wang, Xuehui Wang, Peiyi Xie, and Yingbin Zhang. Salientsleepnet: Multimodal salient wave detection network for sleep staging. *arXiv preprint arXiv:2105.13864*, 2021.

[16] Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33:6696–6707, 2020.

[17] Hyungjik Kim and Sunwoong Choi. Automatic sleep stage classification using eeg and emg signal. In *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 207–212. IEEE, 2018.

[18] Celia Kjaerby, Mie Andersen, Natalie Hauglund, Verena Untiet, Camilla Dall, Björn Sigurdsson, Fengfei Ding, Jiesi Feng, Yulong Li, Pia Weikop, et al. Memory-enhancing properties of sleep depend on the oscillatory amplitude of norepinephrine. *Nature neuroscience*, 25(8):1059–1070, 2022.

[19] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[21] Kaare Mikkelsen and Maarten De Vos. Personalizing deep learning models for automatic sleep staging. *arXiv preprint arXiv:1801.02645*, 2018.

[22] Đorđe Miladinović, Christine Muheim, Stefan Bauer, Andrea Spinnler, Daniela Noain, Mojtaba Bandarabadi, Benjamin Gallusser, Gabriel Krummenacher, Christian Baumann, Antoine Adamantidis, et al. Spindle: End-to-end learning from eeg/emg to extrapolate animal sleep scoring across experimental settings, labs and species. *PLoS computational biology*, 15(4):e1006968, 2019.

[23] Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020.

[24] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[25] Shreyasi Pathak, Changqing Lu, Sunil Belur Nagaraj, Michel van Putten, and Christin Seifert. Stqs: Interpretable multi-modal spatial-temporal-sequential model for automatic sleep scoring. *Artificial Intelligence in Medicine*, 114:102038, 2021.

[26] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 1452–1455. IEEE, 2018.

[27] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 453–456. IEEE, 2018.

[28] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Joint classification and prediction cnn framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering*, 66(5):1285–1296, 2018.

[29] 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.

[30] 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*, 2021.

[31] 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.

[32] Jathurshan Pradeepkumar, Mithunjha Anandakumar, Vinith Kugathasan, Dhinesh Suntharalingham, Simon L Kappel, Anjula C De Silva, and Chamira US Edussooriya. Towards interpretable sleep stage classification using cross-modal transformers. *arXiv preprint arXiv:2208.06991*, 2022.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[34] Justus TC Schwabedal, Daniel Sippel, Moritz D Brandt, and Stephan Bialonski. Automated classification of sleep stages and eeg artifacts in mice with deep learning. *arXiv preprint arXiv:1809.08443*, 2018.

[35] Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. Self-distillation from the last mini-batch for consistency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11943–11952, 2022.

[36] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. A convolutional neural network for sleep stage scoring from raw single-channel eeg. *Biomedical Signal Processing and Control*, 42:107–114, 2018.

[37] Jens B Stephansen, Alexander N Olesen, Mads Olsen, Aditya Ambati, Eileen B Leary, Hyatt E Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature communications*, 9(1):5229, 2018.

[38] 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.

[39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] Albert Vilamala, Kristoffer H Madsen, and Lars K Hansen. Deep convolutional neural networks for interpretable analysis of eeg sleep stage scoring. In *2017 IEEE 27th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2017.

[43] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[44] Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, and Furu Wei. Distilled dual-encoder model for vision-language understanding. *arXiv preprint arXiv:2112.08723*, 2021.

[45] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2019.

[46] Dingwen Zhang, Guohai Huang, Qiang Zhang, Jungong Han, Junwei Han, and Yizhou Yu. Cross-modality deep feature learning for brain tumor segmentation. *Pattern Recognition*, 110:107562, 2021.

[47] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.

[48] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[49] Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, and Cuihua Li. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15520–15528, 2021.

[50] Zhilu Zhang and Mert Sabuncu. Self-distillation as instance-specific label smoothing. *Advances in Neural Information Processing Systems*, 33:2184–2195, 2020.

[51] Haoran Zhao, Xin Sun, Junyu Dong, Zihe Dong, and Qiong Li. Knowledge distillation via instance-level sequence learning. *Knowledge-Based Systems*, 233:107519, 2021.