Hierarchical Vision-Language Model with Multi-Level Feature Alignment and Visually Enhanced Language-Guided Reasoning for EEG Image-Based Sleep Stage Prediction

Anonymous ACL submission

Abstract

Single-channel electroencephalography (EEG) plays a vital role in evaluating sleep quality and diagnosing sleep disorders, making sleep stage classification using EEG an essential task in clinical practice. Traditional machine learning methods rely heavily on prior knowledge and handcrafted feature extraction, while deep learning approaches still face limitations in modeling frequency-domain features. Recently, Vision-Language Models (VLMs) have made significant progress in the medical domain. However, they still perform poorly when applied to physiological waveform data, especially EEG signals. These challenges mainly stem from their limited visual understanding and insufficient reasoning capability.To address this, we propose a hierarchical vision-language model that integrates multi-level feature alignment with visually enhanced language-guided reasoning to improve performance on sleep stage classification using EEG. Our approach introduces a specialized visual enhancement module that utilizes intermediate-layer outputs to construct high-level visual tokens, enabling the extraction of deep semantic information from EEG images. Subsequently, a multi-level feature alignment mechanism is employed to fuse these high-level tokens with low-level visual tokens extracted by CLIP, enhancing the VLM's image-processing capabilities in this context. In addition, by incorporating a Chain-of-Thought (CoT) reasoning strategy, the complex medical inference process is decomposed into interpretable logical steps, effectively simulating expert decision-making. Experimental results demonstrate that the proposed method significantly improves both the accuracy and interpretability of VLMs in sleep stage classification using EEG.

1 Introduction

Sleep plays a vital role in maintaining brain function and overall physiological health (Czeisler, 2015). Accurate assessment of sleep quality not only reflects an individual's health status but also serves as a critical basis for diagnosing and treating sleep-related disorders (Vatankhah et al., 2010; Brignol et al., 2012; Zhu et al., 2014). Currently, the American Academy of Sleep Medicine (AASM) standards (Berry et al., 2012) are widely adopted for sleep stage scoring. Among various physiological signals, EEG is widely regarded as the most informative and commonly used modality for sleep stage classification (Kayikcioglu et al., 2015; Alickovic and Subasi, 2018; An et al., 2021), as it captures rich physiological and pathological information and clearly differentiates between sleep stages (Li et al., 2015; Manjunath and Sathyanarayana, 2024).

Waveform morphology and frequency composition are central to EEG-based sleep stage classification. Sleep experts rely on identifying characteristic waveforms—such as alpha, beta, and theta rhythms—within each 30-second epoch to determine sleep stages. However, sleep stage classification is guided by complex clinical criteria, making it a labor-intensive, time-consuming process that is prone to inter-rater variability.

To alleviate these limitations, various automatic sleep stage classification approaches have been proposed. However, traditional machine learning methods (Phan et al., 2013; Seifpour et al., 2018; Satapathy et al., 2022; Arslan et al., 2023) remain heavily dependent on prior knowledge and manual feature extraction, making the process complex and inefficient. In contrast, deep learning approaches (Nie et al., 2021; Eldele et al., 2021; Zhang et al., 2023; Pham and Mouček, 2023) have shown promise in extracting meaningful representations from EEG signals. Nevertheless, they often struggle to capture fine-grained distinctions—particularly between physiologically similar stages such as N1 and REM—leading to suboptimal classification performance.

Recently, VLMs (Achiam et al., 2023; Liu

et al., 2023, 2024; Bai et al., 2023; Wang et al., 2024) have demonstrated remarkable capabilities in general-purpose tasks by leveraging joint visual-textual representations. While their application in the medical domain has garnered growing interest, their performance remains notably limited when dealing with physiological waveform data—particularly EEG—due to insufficient capacity for fine-grained visual perception, effective image processing, and domain-specific reasoning (Wu et al., 2023; Abdullahi et al., 2024; Kaczmarczyk et al., 2024). These challenges restrict the effectiveness of VLMs in complex clinical applications such as EEG-based sleep stage classification.

To address these challenges, we introduce a hierarchical vision-language framework tailored to EEG image representations. Specifically, we augment the visual encoder with a visual enhancement module that extracts intermediate-level representations and transforms them into high-level visual tokens, enabling the model to capture both fine-grained visual details and abstract semantic information from EEG image representations. These high-level semantic representations are then aligned and integrated with low-level visual features extracted by CLIP through a multi-level feature alignment mechanism, facilitating multi-scale perception and bridging semantic gaps across hierarchical representations. On the language side, we incorporate a CoT prompting strategy to guide the model through structured, step-wise reasoning, simulating the expert decision-making process. This integrated architecture empowers the model to make accurate and interpretable predictions, particularly for ambiguous stages such as N1 and REM.

The key contributions of our work are summarized as follows.

1. We propose a novel hierarchical VLM that combines multi-level feature alignment and visually enhanced language-guided reasoning, specifically designed for EEG image-based sleep stage classification.

2. We design a visual enhancement module that constructs high-level visual features from intermediate-layer features, enabling the model to capture deep semantic information from EEG signals.

3. We introduce a multi-level feature alignment mechanism to effectively fuse visual tokens from different levels, thereby enhancing the model's image processing and feature representation capabilities.

4. By employing Chain-of-Thought (CoT) reasoning, we simplify complex inference tasks, improving the transparency and accuracy of the model's decision-making while effectively simulating the step-by-step judgment of human experts.

5. Experimental results show that our method improves the classification of challenging sleep stages (*e.g.*, N1 and REM), enhancing both accuracy and interpretability. This advancement demonstrates the potential of VLMs in EEG-based sleep stage classification and suggests their applicability to other complex medical tasks.

2 Related Work

2.1 Traditional and Deep Learning for EEG Sleep Stage Classification

Traditional approaches to automatic sleep stage classification primarily rely on handcrafted features extracted from time-, frequency-, or time-frequency domains of EEG signals. These features are typically fed into classical machine learning algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), or Random Forests (RF) (Alickovic and Subasi, 2018; Aboalayon et al., 2016). For example,(Agarwal and Gotman, 2001) developed a rule-based system with expert-designed features, while(Park et al., 2000) proposed a hybrid model combining symbolic reasoning with neural networks. Although these methods can achieve reasonable accuracy, they often suffer from limited generalizability and require extensive domain expertise for feature engineering.

Recent advances in deep learning have enabled end-to-end models that learn hierarchical features directly from raw EEG data. Architectures based on Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers have achieved state-of-the-art performance. For instance, DeepSleepNet (Supratak et al., 2017) adopted a CNN-RNN hybrid to model both spatial and temporal patterns from single-channel EEG, while (Phan et al., 2018) introduced a joint classification-prediction CNN to exploit sequential context. More recent designs, such as SleepEEGNet (Mousavi et al., 2019), incorporate attention

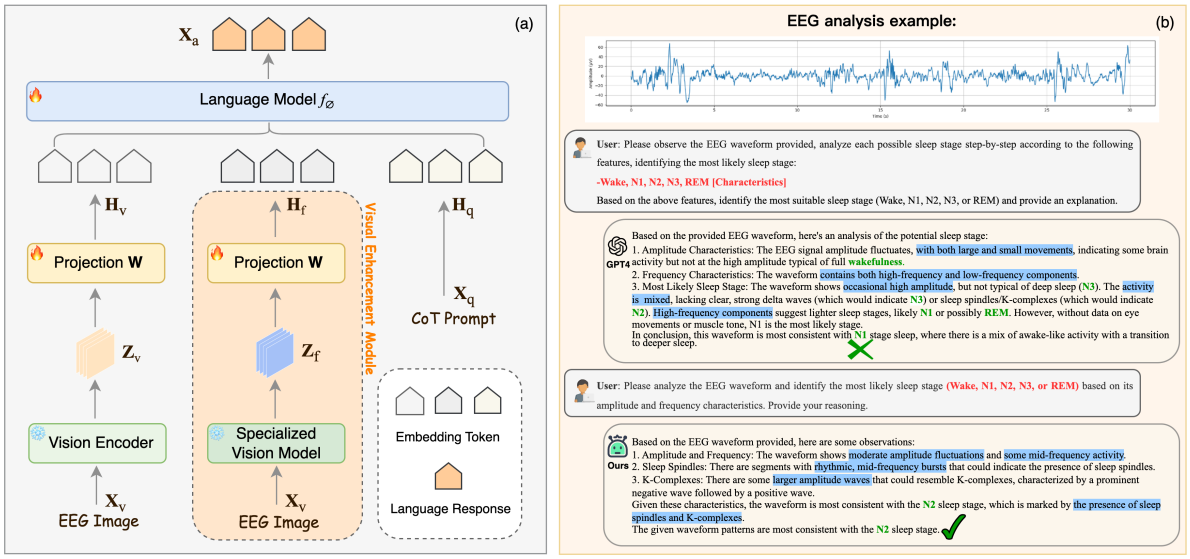

Figure 1: Model Architecture and EEG Analysis Example: (a) Proposed Model Architecture; (b) EEG Analysis Example.

mechanisms and multi-resolution processing to better capture temporal and spectral dynamics. Nevertheless, these models still face challenges in distinguishing physiologically similar stages like N1 and REM, due to subtle and overlapping signal characteristics. Furthermore, their limited interpretability and underutilization of frequency-domain priors constrain clinical trust and deployment, especially in borderline or pathological cases.

2.2 Vision-Language Models in the Medical Domain: Opportunities and Challenges

Recent advances in large-scale multimodal models, such as GPT-4V, LLaVA, and Qwen-VL, have significantly advanced the field of VLMs (Liu et al., 2023; Achiam et al., 2023; Wang et al., 2024). These models have achieved state-of-the-art performance in tasks including image captioning, visual question answering (VQA), and multimodal reasoning. Increasingly, VLMs are being adapted for medical applications such as radiology report generation, digital pathology, and biomedical image analysis (Radford et al., 2021; Li et al., 2022; Liang et al., 2024; Lu et al., 2024).

However, the application of VLMs to physiological waveform data—particularly EEG—remains underexplored. The high visual complexity of EEG-based spectrograms limits the effectiveness of general-purpose models such as CLIP (Ferrante et al., 2024). Current VLMs also struggle with capturing fine-grained details essential for clinical interpretation and lack the domain-specific inductive biases and interpretability required in high-stakes medical contexts. This limitation is especially evident in EEG-based sleep stage classification, where robust visual understanding and clinical transparency are crucial for real-world adoption (Stiglic et al., 2020).

3 Methodology

3.1 Overview

The proposed method integrates a vision encoder, a language model, and a visual enhancement module, as illustrated in Fig. 1(a). Considering the performance and computational cost of VLMs, we conducted experiments based on the LLaVA-1.5 13B model to validate the effectiveness of the proposed strategy.

In our method, the input consists of an EEG image X_v and a CoT prompt (Wei et al., 2022) X_q. The EEG image X_v is first processed by a pre-trained CLIP vision encoder (ViT-L/14) (Radford et al., 2021) to extract low-level visual features $Z_v = g(X_v)$. Simultaneously, the image is passed through a specialized vision model to obtain high-level semantic features $Z_f = \psi(X_v)$. These hierarchical representations, Z_v and Z_f, are transformed into language embedding tokens H_v and H_f through a shared projection layer W, consisting of two MLP layers, as follows:

$$H_v = W \cdot Z_v, \quad H_f = W \cdot Z_f, \tag{1}$$

where $Z_v = g(X_v), \quad Z_f = \psi(X_v)$. Then, H_f

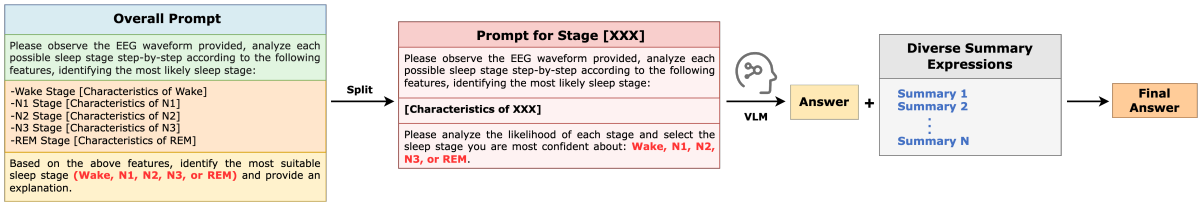

Figure 2: CoT-Guided Multi-Step EEG Sleep Stage Analysis Generation

is passed through the multi-level feature alignment function $H(\cdot)$ to generate the final feature embedding token $H'_f = H(H_f)$, which is then passed, along with the visual embedding token H_v and the text embedding token H_q derived from processing the CoT prompt, into the language model f_ϕ to generate the final language response X_a:

$$H'_f = H(H_f), X_a = f(H_v, H'_f, H_q) \quad (2)$$

3.2 Visual Enhancement Module

According to the findings of LLaVolta (Chen et al., 2024), VLMs still struggle with effectively representing and processing visual information. The intricate details of EEG pose significant challenges for VLMs when handling such tasks. To address this issue, we designed a visual enhancement module that captures high-level semantic representations from EEG images, thereby enhancing the VLM's visual understanding and processing capabilities.

We use a modified ResNet-18 (He et al., 2016) architecture as the benchmark visual enhancement module. The modifications to the standard ResNet-18 are as follows:

Modification to the final convolutional layer: The output channels of the last convolutional layer are increased from 512 to 1024 to align with the dimensionality of the low-level visual features Z_v.

Adjustment to the batch normalization layer: The batch normalization layer is updated to match the new output channel size of 1024.

Addition of a 1x1 convolution in the downsampling component: To ensure channel size matching between residual connections, a 1x1 convolution is added to the downsampling component, increasing the input channels from 512 to 1024.

Fully connected layer for classification: After modification, the feature map is flattened and passed through a fully connected layer for classification.

$$y = W \cdot \text{Flattened Features} \quad (3)$$

These modifications yield intermediate features Z_f, aligned with the low-level visual features Z_v, immediately before the classification layer. Z_f preserves fine-grained details and global semantics, making it suitable for alignment with text or other modalities. It can be fed as fixed-dimension tokens into the VLMs for further processing.

Finally, Z_f is passed through a shared mapping layer to generate the preliminary feature embedding token H_f, which is subsequently used to enhance the visual representation and understanding capabilities of the VLMs.

3.3 Multi-Level Feature Alignment

Through the aforementioned method, we obtain hierarchical feature embeddings H_v and H_f. However, how to effectively construct hierarchical embedding tokens to enhance the visual processing capabilities of the VLMs remains a challenge. To address this, we propose the following approach:

$$H'_f = H_v + \text{Expand}(H_f) \quad (4)$$

where $\text{Expand}(H_f)$ replicates H_f across the patch dimension to match the size of H_v. This expanded H_f is then added element-wise to H_v to produce the final feature embedding token H'_f. This operation defines the multi-level feature alignment function $H(\cdot)$.

This method enables the model to process local regions while integrating fine-grained visual information and global semantic priors, thereby enhancing its ability to process EEG images and represent features.

3.4 Stage-Wise CoT for EEG Sleep Stage Classification

Although hierarchical representation learning significantly enhances the visual understanding capabilities of VLMs, their performance on complex clinical reasoning tasks remains limited. This limitation is especially evident in EEG sleep stage classification, where subtle physiological differences—particularly between stages such as N1

Table 1: Performance Comparison of Different Approaches on the Sleep-EDFx Dataset

Method	Overall Results			F1-score for each class				
	Accuracy	MF1	Kappa	Wake	N1	N2	N3	REM
LLaVA-1.5	0.219	0.152	0.023	0.000	0.333	0.248	0.000	0.177
Resnet-18	0.752	0.756	0.690	0.795	0.637	0.842	0.937	0.567
ConvNeXt-Base	0.813	0.818	0.760	0.835	0.715	0.876	0.905	0.761
Ours-R18	0.792	0.797	0.740	0.839	0.654	0.859	0.944	0.688
Ours-CNxBase	0.811	0.816	0.763	0.851	0.717	0.846	0.905	0.760

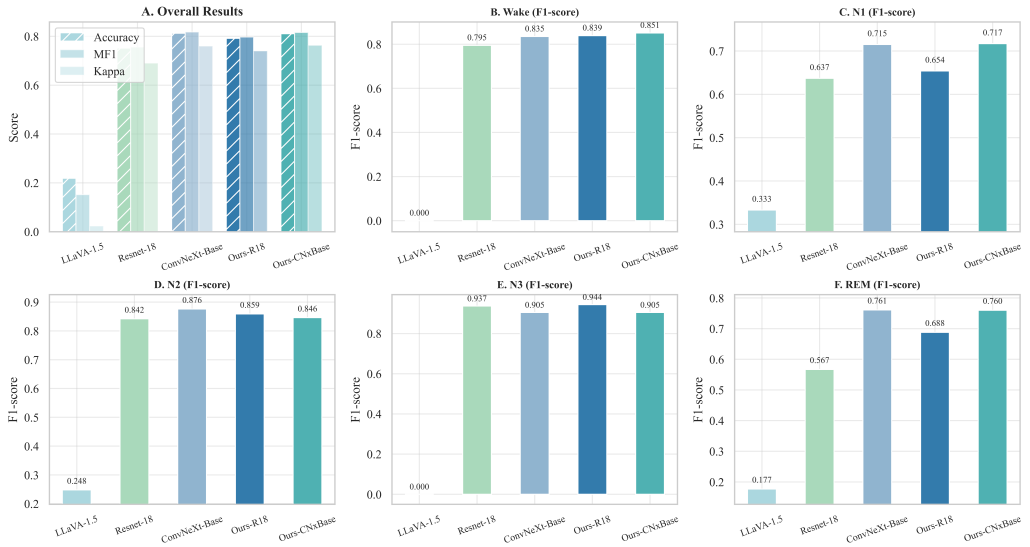

Figure 3: Overall and Per-Stage Classification Performance on the Sleep-EDFx Dataset

and REM—require expert-like, stage-specific judgment.

To address this, we propose a Stage-Wise CoT prompting strategy that breaks down the global sleep stage classification task into a series of focused, interpretable sub-tasks, as illustrated in Fig. 2. Rather than directly prompting a VLM (*e.g.*, GPT-4) with an overall CoT instruction—which often results in vague or inconsistent outputs, as shown in Fig. 1(b)—we decompose the task into sub-CoT prompts, each tailored to a specific sleep stage (*e.g.*, Wake, N1, N2, N3, REM). Each prompt emphasizes the relevant waveform features and frequency–amplitude patterns, enabling the model to conduct targeted, stage-specific reasoning.

Each of these sub-prompts is processed independently by the VLM to generate preliminary stage-level analyses. To further enhance the consistency and robustness of the output, we combine the model's intermediate answers with diverse summary expressions to construct a coherent and interpretable final answer.

This multi-step prompting mechanism not only improves classification accuracy—especially for ambiguous stages—but also more closely simulates the step-by-step analytical process of human experts, thereby enhancing both the transparency and clinical reliability of the model's decision-making.

4 Experiments

4.1 Data Collection and Evaluation Metrics

A band-pass Butterworth filter (1st order) was applied to retain EEG data within the 0.5-35Hz range using the Fpz-Cz channel. The filtered data was then visualized as 30-second EEG images, sourced from the Sleep-EDFx dataset[1]. To reduce the cost of generating a large amount of CoT data, 1300 examples from each class of the visualized data were selected for answer generation, resulting in a total of 5119 valid analysis results, with the following distribution: Wake: 1175, N1: 1186, N2: 757, N3: 836, REM: 1165. For each class, 75 examples were allocated for testing, with the remaining data used

[1] https://physionet.org/content/sleep-edfx/1.0.0/

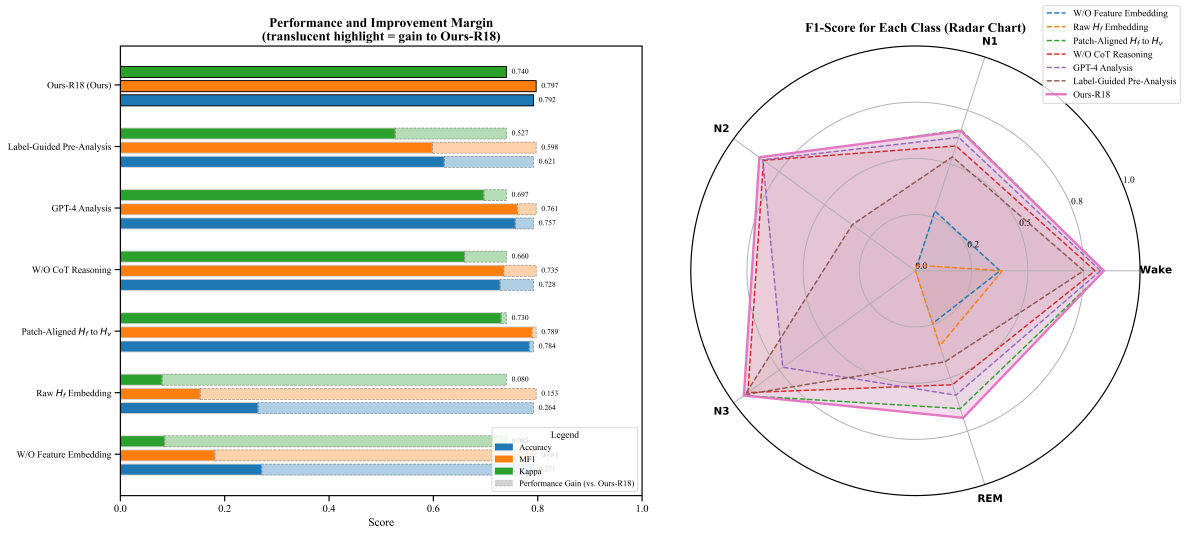

Figure 4: Ablation Study of Feature Embedding and Reasoning Strategies

Table 2: Ablation Study: Exploring Embedding and Reasoning Strategies

Configurations	Overall Results			F1-score for each class				
	Accuracy	MF1	Kappa	Wake	N1	N2	N3	REM
W/O Feature Embedding	0.271	0.181	0.085	0.377	0.280	0.000	0.000	0.247
Raw H_f Embedding	0.264	0.153	0.080	0.387	0.026	0.000	0.000	0.351
Patch-Aligned H_f to H_v	0.784	0.789	0.730	0.841	0.659	0.857	0.944	0.645
W/O CoT Reasoning	0.728	0.735	0.660	0.800	0.584	0.836	0.922	0.533
GPT-4 Analysis	0.757	0.761	0.697	0.824	0.624	0.840	0.730	0.582
Label-Guided Pre-Analysis	0.621	0.598	0.527	0.749	0.533	0.348	0.937	0.426
Ours-R18	0.792	0.797	0.740	0.839	0.654	0.859	0.944	0.688

for training.

The performance of the proposed method in sleep stage prediction was evaluated using a comprehensive set of metrics. Individual F1-scores were calculated for each sleep stage, while overall performance was assessed through accuracy (ACC), kappa (κ), and macro-averaged F1-score (MF1), providing a balanced evaluation across all stages.

4.2 Implementation Details

We used a customized version of ResNet-18 as the visual enhancement module and the pre-trained LLaVA-1.5 13B as the backbone of the VLM, applying LoRA fine-tuning to the entire model. LLaVA-1.5 was trained for 2 epochs with a learning rate of 3e-4 and a temperature of 0.1, while other hyperparameters were kept default. ResNet-18 was independently trained for 30 epochs with a learning rate of 5e-4 and a batch size of 8 on an NVIDIA GeForce RTX 4090 GPU to generate Z_f, then integrated into LLaVA and jointly trained on a single NVIDIA A100 GPU.

4.3 Main Results

To verify the effectiveness of our method, we compare it with three baselines: LLaVA-1.5 13B, the visual enhancement module used alone as a classifier, and a variant of our model in which the visual enhancement module is replaced by ConvNeXt (Liu et al., 2022) for the EEG sleep stage classification task. The results, as reported in Table 1 and Fig. 3, lead to the following observations: **1) Our method significantly boosts the performance of VLM in EEG sleep stage classification.** Ours-R18 and Ours-CNxBase outperform the baseline LLaVA-1.5 by a large margin in all three overall metrics—Accuracy, MF1, and Kappa—demonstrating the effectiveness of our model architecture in handling biomedical signal interpretation tasks. **2) Incorporating an effective visual enhancement module contributes to consistent performance gains.** The comparison between Ours-R18 and

Table 3: Performance Evaluation on External EEG Dataset

Method	Overall Results			F1-score for each class				
	Accuracy	MF1	Kappa	Wake	N1	N2	N3	REM
Resnet-18	0.674	0.675	0.592	0.646	0.591	0.828	0.819	0.489
Patch-Aligned-R18	0.710	0.717	0.638	0.660	0.601	0.832	0.868	0.623
ConvNeXt-Base	0.702	0.710	0.627	0.712	0.610	0.745	0.819	0.662
Ours-R18	0.751	0.756	0.689	0.697	0.638	0.873	0.886	0.688
Ours-CNxBase	0.719	0.722	0.649	0.752	0.632	0.641	0.857	0.727

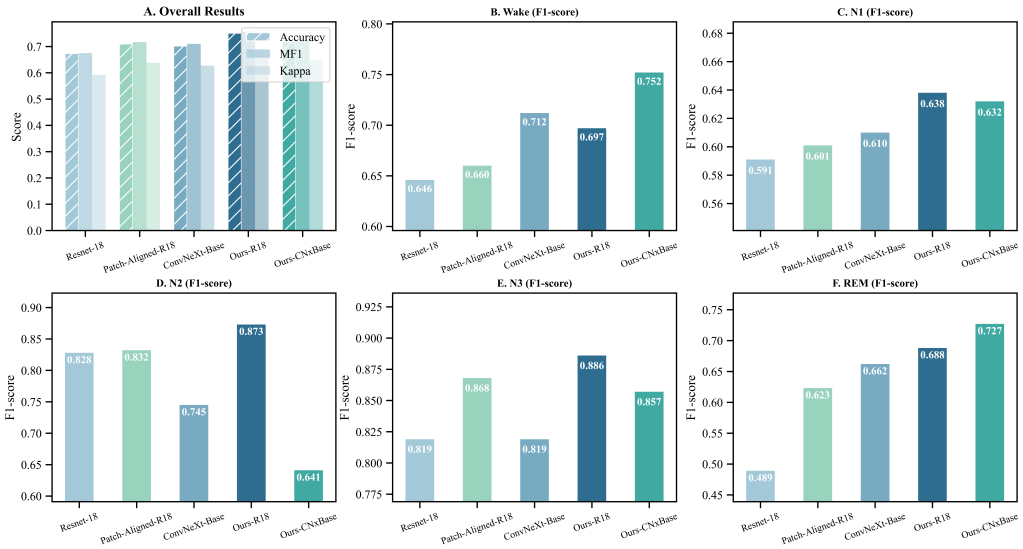

Figure 5: Overall and Per-Stage Classification Performance on the External EEG Dataset

Ours-CNxBase demonstrates that introducing a well-designed visual enhancement module within the vision-language framework leads to improved VLM's performance. This highlights the benefit of leveraging high-level visual tokens derived from intermediate-layer features for enhanced representation and reasoning. **3) The most notable improvements occur in stages with high ambiguity, such as Wake, N1, and REM.** As shown in Fig.6, Table1, and Fig. 3, although these stages share similar signal characteristics, Ours-R18 achieves substantial performance gains across all three. This confirms the model's ability to simulate expert-like stage discrimination through structured CoT reasoning and enhanced visual representation. **4) Our model demonstrates strong generalization capability with limited high-quality training data.** Despite being trained on a relatively small dataset, our approach outperforms or matches the performance of strong CNN-based backbones such as ResNet-18 across multiple sleep stages. In contrast, Ours-CNxBase achieves results comparable to ConvNeXt-Base, suggesting that our strategy—especially when integrated with an effective visual backbone like ResNet—has the potential to break through the performance ceiling typically observed in conventional classification models.

4.4 Ablation Study

Table 2 and Fig. 4 present the impact of different ablation configurations on model performance, summarized as follows: **1) W/O Feature Embedding** and **Raw H_f Embedding** both result in significant performance degradation, highlighting the importance of an effective multi-level feature alignment mechanism for enhancing VLM performance. **2)** Applying **Patch-Aligned H_f to H_v** leads to notable improvements, validating the necessity of aligning H_v to emphasize the fine-grained details and global semantic information in H_f. **3)** The **W/O CoT Reasoning** configuration causes a noticeable performance drop, demonstrating the critical role of CoT-guided reasoning in enhancing model interpretability and decision quality. **4)** The **GPT-4 Analysis** setup confirms the independent contribution of GPT-generated CoT reasoning,

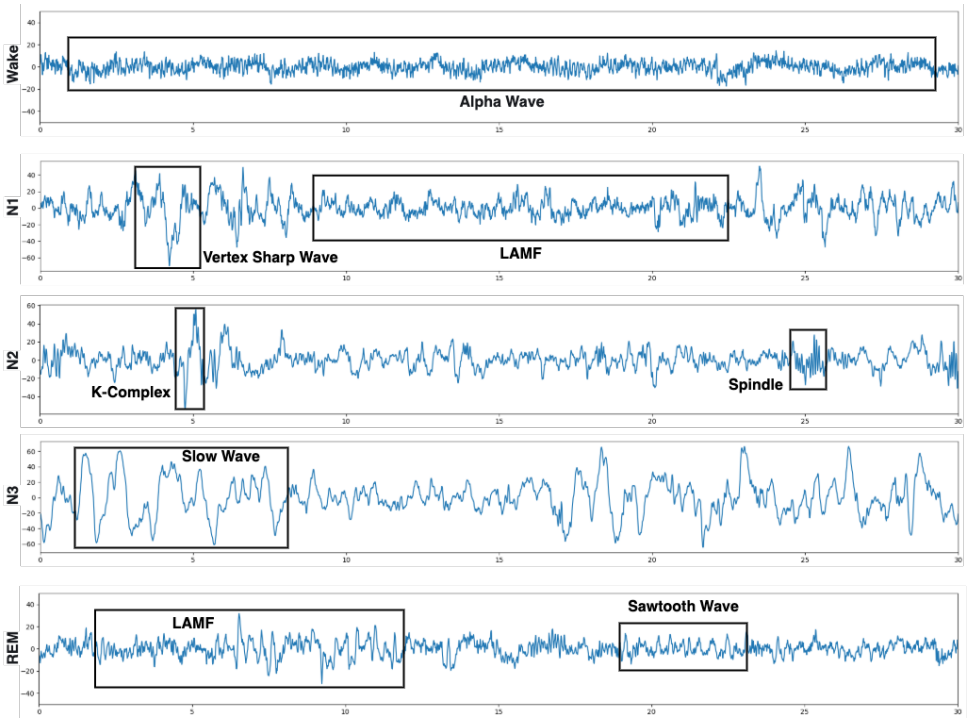

Figure 6: Typical EEG characteristics across sleep stages: Wake - Alpha Waves; N1 - Low Amplitude Mixed Frequency (LAMF: Alpha, Beta) and Vertex Sharp Waves; N2 - K-Complexes and Sleep Spindles; N3 - Slow Waves; REM - LAMF (Beta, Theta) and Sawtooth Waves.

while **Label-Guided Pre-Analysis** underperforms it, suggesting that injecting label information prior to CoT reasoning may interfere with structured inference rather than improve it. **5)** The final method (**Ours-R18**), which integrates hierarchical representation learning with optimized CoT prompting, achieves the best overall performance, validating the effectiveness of our overall framework.

4.5 External Validation

To validate our model's generalization in EEG-based sleep stage classification, we tested it on an external C4-M1 channel EEG dataset from a local hospital, applying the same preprocessing strategy and using 250 samples per class. The results in Table 3 and Fig. 5 show: **1) Ours-R18** and **Ours-CNxBase** exhibit strong generalization, especially in REM, with high Kappa values indicating stability across datasets. **2) Patch-Aligned-R18** lags behind **Ours-R18**, confirming that our multi-level alignment mechanism—designed to support hierarchical representation learning—plays a critical role in capturing both fine-grained and high-level semantic features. **3) Ours-R18** outperforms **Ours-CNxBase**, highlighting the importance of an effective visual enhancement module. **4)** Our method significantly improves upon **ResNet-18** and **ConvNeXt-Base**, further demonstrates its ability to surpass the performance ceiling of conventional classification models.

5 Conclusion

In this study, we present a hierarchical vision-language framework that enhances EEG image-based sleep stage classification through multi-level feature alignment and visually enhanced language-guided reasoning. The method incorporates a visual enhancement module to extract high-level semantic representations from intermediate visual features, which are fused with low-level CLIP representations via a multi-level alignment mechanism, while CoT reasoning guides interpretable, step-wise inference that simulates expert decision-making, thereby enhancing the visual understanding and reasoning capabilities of VLMs. Experimental results demonstrate the superior performance and strong generalization ability of the method across various datasets. We hope this work offers new insights into applying VLMs to clinically relevant tasks involving physiological signal interpretation and inspires further research into their broader applications in healthcare.

Limitations

While our proposed framework shows promising performance in EEG-based sleep stage classification and achieves substantial improvements in challenging stages such as N1 and REM, there are still several areas for further refinement. First, the effectiveness of the visual enhancement module plays a key role in overall performance, and future work may explore more generalized and adaptive designs to improve robustness across settings. Second, the current multi-level feature alignment strategy introduces some computational overhead; developing more lightweight alignment mechanisms could enhance scalability, especially in resource-constrained environments. Lastly, although the proposed method performs well on standard benchmarks, broader validation on diverse datasets and medical tasks would further support its generalizability and practical applicability.

References

Tassallah Abdullahi, Ritambhara Singh, Carsten Eickhoff, and 1 others. 2024. Learning to make rare and complex diagnoses with generative ai assistance: qualitative study of popular large language models. *JMIR Medical Education*, 10(1):e51391.

Khald Ali I Aboalayon, Miad Faezipour, Wafaa S Almuhammadi, and Saeid Moslehpour. 2016. Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation. *Entropy*, 18(9):272.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rajeev Agarwal and Jean Gotman. 2001. Computer-assisted sleep staging. *IEEE Transactions on Biomedical Engineering*, 48(12):1412–1423.

Emina Alickovic and Abdulhamit Subasi. 2018. Ensemble svm method for automatic sleep stage classification. *IEEE Transactions on Instrumentation and Measurement*, 67(6):1258–1265.

Panfeng An, Zhiyong Yuan, and Jianhui Zhao. 2021. Unsupervised multi-subepoch feature learning and hierarchical classification for eeg-based sleep staging. *Expert Systems with Applications*, 186:115759.

E Enes Arslan, Ayşe Seçkinsoy, and Mehmet Feyzi Akşahin. 2023. Sleep stages classification via eeg signals using quadratic support vector machine (svm) algorithm. In *2023 Medical Technologies Congress (TIPTEKNO)*, pages 1–4. IEEE.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, Carole Marcus, Bradley V Vaughn, and 1 others. 2012. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176(2012):7.

Arnaud Brignol, Tarik Al-Ani, and Xavier Drouot. 2012. Eeg-based automatic sleep-wake classification in humans using short and standard epoch lengths. In *2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*, pages 276–281. IEEE.

Jieneng Chen, Luoxin Ye, Ju He, Zhao-Yang Wang, Daniel Khashabi, and Alan Yuille. 2024. Efficient large multi-modal models via visual context compression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Charles A Czeisler. 2015. Duration, timing and quality of sleep are each vital for health, performance and safety. *Sleep Health: Journal of the National Sleep Foundation*, 1(1):5–8.

Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818.

Matteo Ferrante, Tommaso Boccato, Stefano Bargione, and Nicola Toschi. 2024. Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models. *Computers in Biology and Medicine*, 178:108701.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Robert Kaczmarczyk, Theresa Isabelle Wilhelm, Ron Martin, and Jonas Roos. 2024. Evaluating multimodal ai in medical diagnostics. *npj Digital Medicine*, 7(1):205.

Temel Kayikcioglu, Masoud Maleki, and Kubra Eroglu. 2015. Fast and accurate pls-based classification of eeg sleep using single channel data. *Expert Systems with Applications*, 42(21):7825–7830.

Hui Li, Datian Ye, and Cheng Peng. 2015. Development and design of portable sleep electroencephalogram monitoring system. *Sheng wu yi xue Gong Cheng xue za zhi= Journal of Biomedical Engineering= Shengwu Yixue Gongchengxue Zazhi*, 32(3):548–52.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llavanext: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, and 1 others. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874.

Shashank Manjunath and Aarti Sathyanarayana. 2024. Detection of sleep oxygen desaturations from electroencephalogram signals. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE.

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

Haodong Nie, Shikui Tu, and Lei Xu. 2021. Recsleepnet: An automatic sleep staging model based on feature reconstruction. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1458–1461. IEEE.

Haejeong Park, KwangSuk Park, and Do-Un Jeong. 2000. Hybrid neural-network and rule-based expert system for automatic sleep stage scoring. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No. 00CH37143)*, volume 2, pages 1316–1319. IEEE.

Duc Thien Pham and Roman Mouček. 2023. Automatic sleep stage classification by cnn-transformer-lstm using single-channel eeg signal. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2559–2563. IEEE.

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

Huy Phan, Quan Do, The-Luan Do, and Duc-Lung Vu. 2013. Metric learning for automatic sleep stage classification. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5025–5028. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Santosh Kumar Satapathy, Shrey Thakkar, Ayushi Patel, Dharvi Patel, and Divya Patel. 2022. An effective eeg signal-based sleep staging system using machine learning techniques. In *2022 IEEE 6th Conference on Information and Communication Technology (CICT)*, pages 1–6. IEEE.

Saman Seifpour, Hamid Niknazar, Mohammad Mikaeili, and Ali Motie Nasrabadi. 2018. A new automatic sleep staging system based on statistical behavior of local extrema using single channel eeg signal. *Expert Systems with Applications*, 104:277–293.

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379.

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

Maryam Vatankhah, Mohammad-R Akbarzadeh-T, and Ali Moghimi. 2010. An intelligent system for diagnosing sleep stages using wavelet coefficients. In *The 2010 international joint conference on neural networks (IJCNN)*, pages 1–5. IEEE.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

C Wu, J Lei, Q Zheng, W Zhao, W Lin, X Zhang, X Zhou, Z Zhao, Y Zhang, Y Wang, and 1 others. 2023. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. arxiv. 10.48550. *arXiv preprint arXiv.2310.09909*.

Yongqing Zhang, Wenpeng Cao, Lixiao Feng, Manqing Wang, Tianyu Geng, Jiliu Zhou, and Dongrui Gao. 2023. Shnn: A single-channel eeg sleep staging model based on semi-supervised learning. *Expert Systems with Applications*, 213:119288.

Guohun Zhu, Yan Li, and Peng Wen. 2014. Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal. *IEEE journal of biomedical and health informatics*, 18(6):1813–1821.