BRAVISTA: A VISUAL LANGUAGE FRAMEWORK FOR MULTI-TASK EEG DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

A key challenge in electroencephalography (EEG) decoding is learning representations that generalize across diverse cognitive tasks, subjects, and recording conditions. While EEG provides a non-invasive window into brain activity across a wide range of applications, its datasets are often limited in scale and highly heterogeneous, making multi-task learning difficult. Recent advances in foundation models have improved EEG decoding performance, yet a fundamental open question remains: how to effectively interface neural signals with these models to enable multi-task learning across datasets. To investigate this question, we introduce **BraVista**, a visual-language framework that encodes multichannel EEG signals as structured images and enables multi-task learning through instruction-conditioned vision-language models (VLMs). Our approach relies on continual post-training of a general-domain VLM, leveraging its visual and linguistic priors to adapt to neural signals without an additional large-scale EEG self-supervised pretraining stage. We evaluate BraVista across four EEG datasets spanning multiple BCI tasks and problem settings, and use this setup to identify the factors that enable effective multi-task EEG learning. First, we show that the right choice of EEG-to-image representation is critical for adapting general-domain VLMs to EEG. Second, we demonstrate that multi-task learning improves decoding performance in low-data regimes, indicating effective knowledge sharing across tasks. Third, a controlled comparison with the same vision encoder and task-specific classifier heads shows that the language model contributes beyond visual feature extraction. Finally, through controlled perturbations of the EEG signal, we observe a gradual performance degradation under increasing noise and task-specific degradation when individual frequency bands are masked, suggesting that the model relies on EEG-relevant information rather than superficial visual patterns. Together, these results establish BraVista as a framework for multi-task EEG decoding via visual representations. Our findings show that visual representations can serve as a viable interface for multi-task EEG learning when paired with appropriate representation and post-training strategies, highlighting the potential of general-domain foundation models as a scalable basis for EEG signal decoding.

1 INTRODUCTION

Electroencephalography (EEG) is a non-invasive, high-temporal-resolution recording of brain activity that underpins clinical and cognitive applications such as epilepsy diagnosis (Ahmad et al., 2023; Wong et al., 2023), sleep staging (Yang et al., 2021; Wang et al., 2024b), affective computing (Jenke et al., 2014; Dadebayev et al., 2022), and mental stress detection (Zyma et al., 2019). EEG signals are noisy, non-stationary, and highly variable across subjects and acquisition setups (Lotte et al., 2018), which makes it difficult to learn representations that generalize across datasets and support multi-task inference.

To address the lack of generalizable EEG representations, recent work has turned to foundation models inspired by advances in natural language processing and computer vision. These approaches train large-scale models on diverse EEG corpora (Jiang et al., 2024; Wang et al., 2025; 2024a; El Ouahidi et al., 2025; Zhou et al., 2025), improving robustness over supervised task-specific models. However, they remain task-specific at inference: each downstream task requires its own fine-tuned model and classifier head, which precludes unified multi-task prediction within a single model.

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

Recent advances in large language models (LLMs) enable unified multi-task prediction through instruction conditioning (Brown et al., 2020; Touvron et al., 2023), and multimodal LLMs extend this to images and time series via modality-specific encoders (Liu et al., 2023; Jin et al., 2024; Zhong et al., 2025; Zhang et al., 2025). Recent EEG work couples EEG encoders with language models to cast EEG tasks as instruction-conditioned prediction (Jiang et al., 2025b; Camaret Ndir et al., 2025; Lu et al., 2025; Zeng et al., 2025), but relies on EEG-specific pretraining to align neural signals with language, making performance dependent on the scale and diversity of the EEG pretraining corpus (Jiang et al., 2025a).

In parallel, recent work in time-series modeling has shown that rendering signals as images allows pretrained vision encoders to be reused effectively for forecasting and classification, suggesting that visual representations can serve as an effective interface for modeling temporal data (Jin et al., 2024; Zhong et al., 2025; Zhang et al., 2025; He et al., 2026). While these methods have shown promise across a range of time-series tasks, their application to unified EEG decoding remains unexplored.

Contributions To address these challenges, we introduce **BraVista**, a unified framework that represents multichannel EEG signals as structured visual inputs and leverages pretrained VLMs for instruction-conditioned multi-task decoding. By reformulating EEG decoding as an instruction-following visual understanding problem, our approach enables adaptation of general-domain VLMs to neural signals without an additional large-scale EEG self-supervised pretraining stage. BraVista is supervisedly adapted to the EEG datasets through continued post-training (finetuning); the distinction is from the conventional two-stage pipeline of EEG foundation models, which first perform large-scale EEG self-supervised pretraining before downstream fine-tuning. Specifically, we convert EEG signals into time–frequency images and pair them with task-specific textual instructions, allowing a single model to perform diverse EEG decoding tasks within a unified framework. Our main contributions are summarized as follows:

- We reformulate EEG decoding as an instruction-conditioned visual understanding problem by representing multichannel EEG signals as structured visual inputs paired with natural language task descriptions.
- We introduce BraVista, a unified framework that adapts a general-domain VLM to EEG via continued post-training, enabling multi-task EEG decoding within a single model.
- Using this framework, we investigate the role of representation design, instruction conditioning, the language model, and multi-task learning, showing that (i) EEG-to-image representation critically impacts transferability, (ii) multi-task learning improves generalization in small datasets, (iii) the language model contributes beyond visual feature extraction, and (iv) model behavior is governed by alignment between instructions and task definitions.
- We evaluate our approach on four EEG datasets spanning multiple BCI tasks, demonstrating strong performance and highlighting the potential of vision–language models as a flexible and scalable foundation for unified EEG decoding.

2 Related Works

Task-Specific and Foundation EEG Models Deep learning-based EEG models include convolutional or recurrent neural networks (Schirrmeister et al., 2017; Lawhern et al., 2018; Wang et al., 2018; Roy et al., 2019; Phan et al., 2019; Yang et al., 2021; Ding et al., 2023; Jing et al., 2023), CNN-LSTM architectures (Zhang et al., 2019; Wang et al., 2023b; Li et al., 2022b), Transformers (Song et al., 2021; Peh et al., 2022), and CNN-Transformer architectures (Song et al., 2022; Phan et al., 2022). While these approaches achieve strong performance on their target tasks, they remain specialized and typically require separate training or extensive tuning for each new dataset or application. Motivated by the need for more general EEG representations, recent work has shifted toward EEG foundation models trained on diverse datasets (Kostas et al., 2021; Yi et al., 2023; Yang et al., 2023; Jiang et al., 2024; Wang et al., 2025; Zhou et al., 2025). These models learn task-agnostic representations that are adapted to each downstream task through task-specific fine-tuning. BIOT (Yang et al., 2023) introduced a unified transformer-based framework for biosignals that enables cross-modality learning. LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), and CSBrain (Zhou et al., 2025) further advanced this direction through large-scale EEG pretraining, achieving strong performance across multiple tasks after task-specific fine-tuning. None of these models, however, supports unified multi-task decoding within a single model.

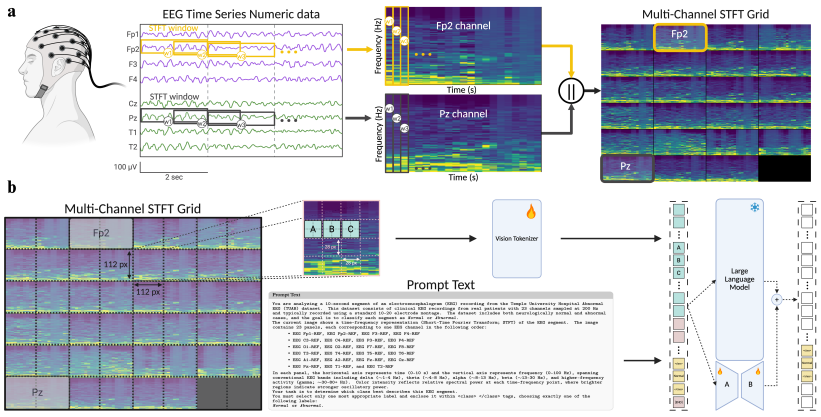


Figure 1: **BraVista framework for multi-task EEG decoding. (a)** Raw EEG time-series from multiple scalp electrodes are transformed into time–frequency representations using the short-time Fourier transform (STFT). The resulting per-channel spectrograms are stacked to form a multi-channel grid, which serves as EEG visual representation. **(b)** Multichannel EEG signals are converted into unified STFT images and processed by the vision tokenizer, which is fine-tuned to adapt to EEG STFT representations. A task-specific textual prompt describing the dataset and decoding prompt is tokenized and provided as linguistic context. The joint visual-textual representation is decoded by a language model adapted via LoRA for instruction-conditioned EEG decoding.

Large Language Models for EEG Recent work has explored adapting large language models (LLMs) for EEG decoding. For example, EEG-GPT (Kim et al., 2024) converts EEG features into textual descriptions to condition language models. EEG-CLIP (Camaret Ndir et al., 2025) adopts contrastive alignment to learn a joint embedding space between EEG signals and text using a CLIP-style objective. More recently, NeuroLM (Jiang et al., 2025b), UniMind (Lu et al., 2025), and WaveMind (Zeng et al., 2025) enable multi-task EEG decoding within an instruction-tuned LLM by training a neural tokenizer to map neural signals into tokens compatible with language models. While these approaches demonstrate the potential of LLMs for unified EEG modeling, two key challenges remain: (1) they require massive EEG-specific pre-training, and (2) there is still a large performance gap compared to specialized EEG models. This highlights that identifying an effective interface between EEG signals and language models remains an open challenge.

3 Methodology

In this section, we present **BraVista**, our multimodal framework for EEG decoding, illustrated in Fig. 1. The key idea of BraVista is to use visual representations as an interface that aligns EEG signals with the inductive biases of pretrained vision–language models. Given a multichannel EEG signal $\mathbf{X} \in \mathbb{R}^{C\times T}$, we first transform it into a structured visual representation using a time–frequency mapping. This representation is paired with a task-specific textual instruction, and both modalities are processed jointly by a pretrained VLM. The model is adapted to EEG through a continued post-training strategy described in Sec. 3.2.

3.1 EEG Visual Instruction Dataset

We construct an EEG visual instruction dataset by mapping EEG signals to images and pairing them with natural language task descriptions. Each sample consists of a visual representation of EEG and a corresponding instruction–response pair, enabling instruction-conditioned multi-task learning.

EEG Time–Frequency Representation. We begin by formalizing the input EEG signals as

$$\mathbf{X} \in \mathbb{R}^{C\times T}, \tag{1}$$

where C denotes the number of electrodes and T the number of time samples. The signal is partitioned into segments of length L, yielding $n = \lfloor T/L \rfloor$ samples per recording. Each segment is transformed

into a time–frequency representation using the Short-Time Fourier Transform (STFT), producing a spectrogram for each channel. These spectrograms capture the joint temporal–spectral structure of EEG signals, which is critical for representing non-stationary neural dynamics.

Multi-Channel STFT Visual Construction. To form a unified visual input, we arrange channel-wise spectrograms into a structured grid (Fig. 1a). This results in an image-like representation that preserves discriminative time–frequency structure and provides a two-dimensional representation that allows vision-based models to capture temporal–spectral correlations (Li et al., 2018; 2022a). This representation is well suited for EEG analysis (Schirrmeister et al., 2017; Craik et al., 2019; Wang et al., 2023a), as EEG signals exhibit intrinsically non-stationary temporal-spectral dynamics, with task-relevant neural patterns being encoded jointly across time and frequency (Makeig et al., 2002).

Vision Tokenizer. The EEG time-frequency image is then processed by the vision encoder of `Qwen3-VL`, which utilizes SigLIP-2 architecture (Bai et al., 2025; Tschannen et al., 2025). We initialize from this pretrained vision encoder and adapt it to EEG via continued post-training, enabling the model to learn EEG-specific representation. The input image is partitioned into non-overlapping 14×14 pixel patches and encoded through stacked windowed attention layers, with a window size of 112×112, interleaved with full attention layers. Subsequently, spatially adjacent groups of four patches are concatenated and projected into a sequence of visual tokens using an MLP-based vision–language merger (Bai et al., 2025; Liu et al., 2023). The input resolution is carefully chosen such that each attention window operates within a single EEG channel, enabling effective modeling of channel-specific temporal–spectral patterns, while interleaving windowed and full attention layers allows hierarchical aggregation across windows, enabling capturing of both within-channel and cross-channel correlations. (See Appendix D for dataset-specific image generation details.)

EEG Instruction Construction. For each EEG sample, we construct structured instruction–response pairs that describe the decoding task in natural language. Our instructions provide contextual information about the EEG signal including interpretation of the visual representation (e.g., time–frequency axes and spectral bands) as well as task specification. This contextualization guides the model to interpret the input image as a neural signal rather than a generic visual pattern. The instruction concludes with an explicit task specification and the response is defined by the corresponding ground-truth label. This design enables casting EEG decoding as an instruction-following problem and aligns with the training paradigm of VLMs, allowing the model to leverage its pretrained language understanding to perform instruction-conditioned prediction over EEG representations (see Appendix D for task-specific instruction details).

3.2 BraVista Framework

We introduce BraVista, a unified Vision–Language framework for multi-task EEG decoding that formulates diverse BCI tasks as instruction-conditioned multimodal prediction problems. By representing EEG signals as visual time-frequency representations and jointly modeling them with task-specific textual instructions, BraVista enables a single model to perform heterogeneous EEG decoding tasks within a shared architecture. To adapt a general-purpose VLM to EEG data, we perform continued post-training on the `Qwen3-VL-2B` model (Bai et al., 2025), specifically selected to leverage its robust pre-trained visual understanding capabilities as a foundation for EEG domain adaptation. We keep the original model architecture, which consists of three components: a vision encoder, a vision–language merger, and a language model decoder. The training curriculum of BraVista is designed to establish a foundational understanding of EEG data and use instruction-following capability to enable multi-task EEG decoding as illustrated in Figure 1(b).

Adaptation Strategy for EEG Alignment Adapting pretrained VLMs to EEG decoding requires addressing the substantial domain shift between natural images and EEG time–frequency representations. To this end, we adopt a targeted post-training strategy in which different components of the pretrained model are adapted in distinct ways. The vision encoder is fully fine-tuned to capture task-relevant temporal–spectral patterns in EEG data. This allows the model to align its visual representations with the structure of EEG signals while leveraging pretrained visual priors.

To adapt the language model while preserving its pretrained linguistic and generative capabilities, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) training strategy, targeting the attention and MLP layers of the decoder. LoRA introduces trainable low-rank matrices into the query, key, value,

and output (QKVO) projection layers of each attention block and MLP layers. For a pretrained linear projection with weight matrix $W \in \mathbb{R}^{d\times d}$, LoRA introduces a trainable low-rank update:

$$W' = W + \alpha BA, \tag{2}$$

where $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$ are trainable matrices with rank $r \ll d$, and α is a scaling factor. During training, the original weights W remain frozen and only the LoRA parameters $\{A, B\}$ are optimized. This formulation constrains task-specific updates to a low-dimensional subspace, enabling parameter-efficient adaptation while mitigating catastrophic forgetting (Hu et al., 2022).

Optimization for multi-task EEG decoding We formulate multi-task EEG decoding as a conditional generation problem over multimodal inputs. Each input consists of an EEG image $\mathcal{I}_{\text{EEG}}$ and a textual instruction $\mathbf{q} = \{q_1, \ldots, q_M\}$. Given these multimodal inputs, we define **BraVista** as π_θ parameterized by θ as an autoregressive conditional distribution over the output sequence $\mathbf{y} = \{y_1, \ldots, y_K\}$:

$$\pi_\theta(\mathbf{y} \mid \mathcal{I}_{\text{EEG}}, \mathbf{q}) = \prod_{k=1}^{K} \pi_\theta(y_k \mid y_{<k}, \mathcal{I}_{\text{EEG}}, \mathbf{q}), \tag{3}$$

where K is the length of the generated output sequence. This formulation allows the model to condition its predictions jointly on EEG visual features, task-level textual context and prior outputs. To learn the conditional distribution in Eq. 3, we use supervised fine-tuning (SFT) objective:

$$\mathcal{L} = -\sum_{k=1}^{K} \log p_\theta(y_k \mid y_{<k}, \mathcal{I}_{\text{EEG}}, \mathbf{q}). \tag{4}$$

This objective trains the model to generate task-specific outputs conditioned jointly on the EEG visual input and the textual prompt, consistent with instruction-following training in LLMs (Ouyang et al., 2022; Touvron et al., 2023). During training, gradients are propagated through the vision encoder and the LoRA parameters of the language model, enabling coordinated adaptation of visual and linguistic representations for multi-task EEG decoding. Implementation details are provided in Appendix C.2.

4 Experiments

4.1 Evaluation Tasks and Datasets

We evaluate our approach across four EEG datasets spanning diverse recording conditions, data scales, and cognitive and clinical tasks. This selection enables a systematic assessment of model behavior under heterogeneous settings, including variations in acquisition protocols, label structures, and task objectives. Dataset statistics are summarized in Table A.2 (Appendix D).

- **Abnormal detection:** TUAB (Obeid & Picone, 2016) is a large-scale clinical EEG dataset for abnormality detection, where recordings are labeled as normal or abnormal.
- **Emotion classification:** SEED (Zheng & Lu, 2015) is an emotion recognition dataset containing EEG recordings from subjects watching emotion-eliciting videos, with three emotion categories: positive, negative, and neutral.
- **Sleep staging:** HMC (Alvarez-Estevez & Rijsman, 2021) is designed for automatic sleep staging where EEG recordings are labeled as five sleep stages (Wake, NREM1, NREM2, NREM3, and REM).
- **Cognitive workload:** EEGMAT (Zyma et al., 2019) contains EEG data from subjects during and before a mental arithmetic task, and the recordings are labeled as high or low workload.

To ensure fair and directly comparable evaluation, we adopt the preprocessing pipelines and train/validation/test splits used in prior works (Jiang et al., 2024; 2025b; Zhou et al., 2025), enabling direct comparison with reported results. For datasets with different EEG segmentation protocols (e.g., EEGMAT), we restrict comparisons to methods evaluated under consistent settings. Furthermore, for the SEED dataset, we explicitly exclude foundation models whose pre-training data includes SEED to prevent data leakage; in addition, none of the compared EEG foundation models reports using HMC, EEGMAT, or TUAB in its pretraining corpus. Baseline results are taken from the original publications under these standardized protocols, following common practice for EEG foundation models (Wang et al., 2025; Zhou et al., 2025), while all BraVista results are obtained by us. We report balanced

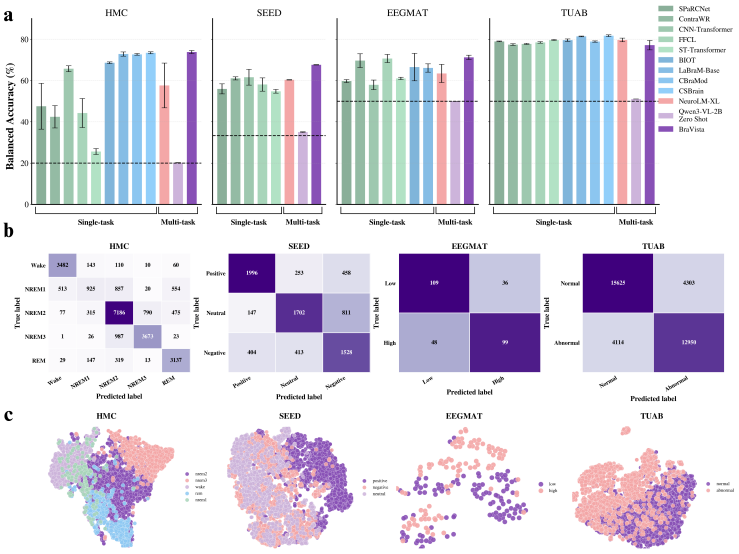


Figure 2: **BraVista multi-task EEG decoding performance across diverse BCI tasks (a)** Balanced accuracy (%) of BraVista compared to supervised models, EEG foundation models, and zero-shot VLM across four datasets (HMC, SEED, EEGMAT, TUAB). Error bars denote standard deviation; dashed lines denote chance level. **(b)** Confusion matrices for BraVista under the multi-task setting. The matrices exhibit clear diagonal structure across datasets. **(c)** t-SNE visualization of the latent embeddings learned by BraVista, where individual points are color-coded according to their class labels within each dataset.

accuracy as the primary evaluation metric to account for class imbalance, computing accuracy per class and averaging across classes. We additionally report weighted F1 scores for ablation studies to capture precision–recall trade-offs under imbalanced label distributions. All results are presented as mean $\pm$ standard deviation over three runs with different random seeds.

4.2 BraVista Enables Multi-Task EEG Decoding Across Diverse BCI Tasks

We evaluate BraVista in a unified multi-task setting across four EEG datasets spanning diverse BCI tasks. We compare BraVista with supervised models, EEG foundation models, and recent LLM-based approaches. As shown in Fig. 2a, BraVista consistently achieves strong performance while operating as a single unified model on all datasets. In particular, it outperforms the most directly comparable multi-task baseline (Jiang et al., 2025b) on the HMC, SEED, and EEGMAT datasets, and remains competitive on TUAB. Importantly, BraVista substantially surpasses the zero-shot performance of the base `Qwen3-VL-2B` across all datasets, highlighting that our post-training effectively aligns visual-linguistic priors with neural signal characteristics. Notably, BraVista achieves this alignment without an additional large-scale EEG self-supervised pretraining stage. The single-task models, each fine-tuned separately per dataset, are included for context.

We further analyze the class-level behavior of BraVista using confusion matrices (Fig. 2b). The strong diagonal structure indicates that the model maintains class-wise discrimination without bias toward majority classes. The main confusions coincide with known task difficulty: NREM1 vs. NREM2 in HMC, whose distinction relies on transient features such as sleep spindles and K-complexes that may be absent from a single epoch (Malhotra, 2024), and Neutral vs. Negative in SEED, which share similar beta- and gamma-band patterns (Zheng & Lu, 2015). This suggests that the errors stem from intrinsic task difficulty rather than model failure.

Finally, to evaluate how well the model understands EEG images, we visualize the learned latent embeddings of BraVista using t-SNE. The embeddings of BraVista (Fig. 2c) mirror this structure: classes with higher confusion (e.g., NREM1 and NREM2 in HMC; Negative and Neutral in SEED) show greater overlap, while well-separated clusters correspond to more easily distinguishable classes.

These results suggest that BraVista learns structured and semantically meaningful representations from our EEG visual representations.

4.3 The Language Model Contributes Beyond Visual Feature Extraction

To test whether BraVista benefits from the language model or acts mainly as a large image classifier, we train a multi-task baseline that uses the identical pretrained vision encoder, jointly trained with four task-specific linear classifier heads (one per dataset) in place of the language-model decoder (Fig. 3). BraVista outperforms this baseline on all four datasets.

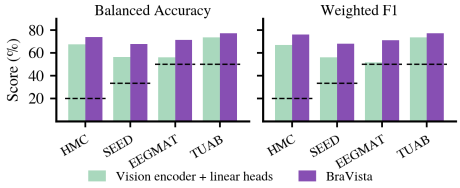


Figure 3: **Isolating the language model.** Same vision encoder with linear heads vs. BraVista.

4.4 Multi-Task Training Improves EEG Decoding

We study the impact of multi-task continued post-training by comparing BraVista with its single-task counterpart. Multi-task training consistently matches or improves performance across all datasets, with the largest gains observed on SEED and EEGMAT (Fig. 4). These datasets contain fewer training samples (Tab. A.2), indicating that multi-task learning is particularly beneficial in low-data regimes. On larger datasets such as HMC and TUAB, performance remains comparable between single-task and multi-task settings, showing that training a unified model across tasks does not sacrifice performance on larger datasets. Overall, these results suggest that multi-task post-training improves EEG representation learning by enabling the model to leverage more data with training across different tasks, leading to better generalization in low-data settings without sacrificing performance on larger datasets.

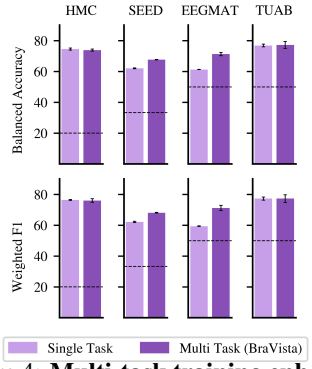


Figure 4: **Multi-task training enhances EEG decoding.** Single-task vs. multi-task (BraVista); balanced accuracy (top) and weighted F1 (bottom).

4.5 Ablation Studies

To characterize the mechanisms enabling effective multi-task EEG decoding in BraVista, we examine the role of (1) EEG-to-image representation and (2) instruction conditioning. We additionally analyze (3) post-training strategy (Appendix F) and (4) model scale (Appendix G) to understand whether performance is driven by model capacity or by the proposed interface.

4.5.1 EEG-to-Image Representation Significantly Impacts EEG Decoding

We first study how the choice of visual representation affects the ability of VLMs to interpret EEG signals. We compare three EEG-to-image mappings: short-time Fourier transform (STFT) spectrograms, raw time-series plots, and topographic maps (Fig. 5a). The choice of representation has a substantial impact on performance. Time-frequency representations achieve the highest performance in both balanced accuracy and weighted F1 on HMC, SEED, and EEGMAT and are comparable to time-series plots on TUAB, while topographic representations result in noticeable degradation (Fig. 5b). This result highlights the importance of preserving time–frequency structure when adapting vision–language models to EEG data: time-series plots retain temporal information and topographic maps capture spatial distributions, but neither provides a joint representation of temporal and spectral dynamics.

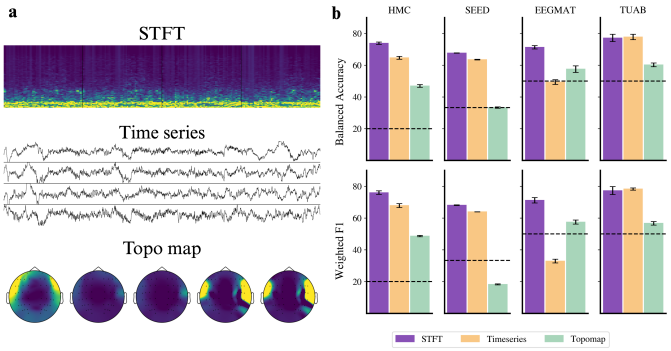


Figure 5: **Impact of EEG-to-image representation on decoding performance. (a)** Examples of different visual representations of EEG signals, including STFT spectrograms, raw time-series plots, and topographic maps. **(b)** Performance comparison across representations on four datasets using balanced accuracy (top) and weighted F1 (bottom).

4.5.2 Instruction Conditioning Determines Task and Label Space

Next, we analyze how BraVista uses instructions to specify both the decoding task and its associated label space by introducing controlled variations in the input prompts (see Appendix E). We first examine the effect of modifying the task description phrasing while preserving the correct label set. Simplifying the instruction text or shuffling the order of label options has minimal impact on performance, indicating that the model does not rely on superficial wording or positional biases, and instead focuses on the semantic content of the instruction; EEGMAT, the smallest dataset, is more sensitive to both changes, and label shuffling also reduces performance on TUAB. Because the simplified instruction removes the dataset name and task-specific metadata while keeping the label options, it also tests whether the model relies on dataset identity: balanced accuracy changes only marginally on HMC , SEED, and TUAB, indicating that the visual EEG representation, rather than dataset identity, is the dominant predictive signal. Next, we vary the label space provided in the instruction. Expanding the instruction to include label options from multiple datasets degrades performance, most strongly on SEED and EEGMAT, suggesting that ambiguity in the output space affects prediction quality. When the correct label set is removed entirely, performance drops to (near) zero across all datasets. This behavior indicates that BraVista keeps its instruction following after continued post-training and selects only among the options offered in the instruction.

Taken together, these results show that BraVista does not ignore or memorize instructions, but uses them to determine both the task and the set of valid outputs.

4.6 BraVista Relies on EEG Signal Structure

To assess whether BraVista relies on EEG information encoded in the visual representation, we evaluate its behavior under controlled noise corruption by progressively adding noise to the input signals at different signal-to-noise ratios (20 dB, 10 dB, 5 dB, and 0 dB). The STFT visualizations illustrate the gradual loss of time–frequency structure as noise increases (Fig. 6a). As noise increases, performance degrades consistently across all datasets for both balanced accuracy and weighted F1. Importantly, this degradation is gradual rather than abrupt, and performance remains above chance even under severe noise conditions (Fig. 6b). The gradual performance decay suggests that predictions depend on the quality of the underlying EEG signal rather than superficial visual patterns.

To identify *which* signal content the model uses, we mask individual EEG frequency bands before constructing the STFT images (Fig. 6c). The resulting degradation is task-specific, which a generic visual shortcut would not produce. HMC is most sensitive to delta masking, which reduces balanced accuracy to chance level „ consistent with the central role of slow-wave activity in sleep staging (Malhotra, 2024). SEED is most sensitive to beta masking, in line with reports that beta and gamma activity carry emotion-discriminative information (Zheng & Lu, 2015). EEGMAT is most sensitive to gamma and alpha masking. TUAB is most sensitive to beta and alpha masking. These task-dependent

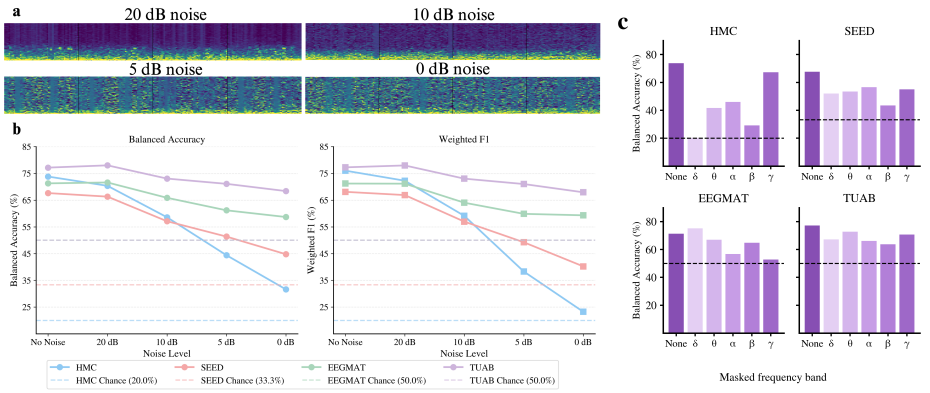


Figure 6: **BraVista relies on EEG signal structure under input corruption. (a)** Examples of STFT representations for HMC dataset under increasing noise levels (20 dB to 0 dB). **(b)** Performance across noise levels for four datasets, evaluated using balanced accuracy (left) and weighted F1 (right). **(c)** Balanced accuracy when individual EEG frequency bands (δ, θ, α, β, γ) are masked before constructing the STFT images; "None" is the unmasked input. Dashed lines denote chance level.

profiles indicate that BraVista relies on EEG-relevant spectral information rather than superficial visual patterns.

5 Discussion

In this work we present BraVista, a framework that reformulates EEG decoding by interfacing neural signals with general-domain vision–language models. By transforming multichannel EEG recordings into time–frequency images and pairing them with task-specific instructions, we enable unified multi-task decoding within a single model, eliminating the need for task-specific architectures or an additional large-scale EEG self-supervised pretraining stage. This framework shifts the focus from designing specialized EEG models to designing effective interfaces between neural signals and general-domain foundation models.

Through this framework, we highlight several key insights. First, the choice of EEG-to-image representation is critical: preserving time–frequency structure is essential for enabling vision–language models to extract meaningful patterns from EEG signals. Second, multi-task continued post-training benefits low-data tasks. Third, the language model contributes beyond visual feature extraction. Fourth, through frequency-band masking, noise experiments, and class-wise error analysis, we show that BraVista's predictions are governed by underlying EEG-relevant spectral features rather than superficial visual artifacts.

Limitations and future work. Despite these promising results, several limitations point to important directions for future work. Our approach relies on a single representation modality (time–frequency images), which may not fully capture the spatial, temporal, and relational structure of EEG signals; exploring hybrid or multi-view representations that combine complementary signal perspectives could improve performance. Furthermore, while our framework enables instruction-conditioned multi-task decoding, it currently focuses on producing class labels without explicitly leveraging the reasoning capabilities of VLMs. An important direction is to extend BraVista toward richer outputs, such as intermediate explanations and multi-step reasoning over EEG segments. In particular, reinforcement-learning-based post-training could be explored to encourage structured reasoning grounded in the EEG input, using task-specific or verifiable rewards to optimize not only prediction accuracy but also the intermediate reasoning process. Such approaches could further enable the model to integrate complementary contextual information, such as subject-specific metadata or clinical history, when addressing more complex tasks. Together, these extensions could move BraVista beyond classification toward more interpretable and interactive EEG analysis, while further exploiting the capabilities of large multimodal models. Overall, our findings suggest that representation design, combined with general-domain foundation models, provides a promising path toward scalable and unified EEG decoding.

Ethics statement

This work uses only publicly available, de-identified EEG datasets (HMC, SEED, EEGMAT, and TUAB), each collected under the ethical approval and consent procedures described by its original authors; no new human-subject data were collected. BraVista is a research prototype and is not validated for clinical decision-making. Models trained on clinical corpora may inherit biases in patient populations and annotation practices, and any clinical use would require prospective validation.

Reproducibility statement

The preprocessing pipelines, data splits, STFT parameters, image construction, and full instruction prompts for every dataset are described in Appendices A and D. Training details and all hyperparameters are given in Appendix C.2; and the evaluation metrics are defined in Appendix C.3. All reported results are averaged over three random seeds. The code and final weights of the models will be shared upon acceptance.

AI use statement

In this work, we used generative AI tools for editing (grammar, spelling, and word choice). We have not used generative AI tools for research ideation, experimental design, data analysis, or generating results. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

References

Ijaz Ahmad, Xin Wang, Danish Javeed, Prabhat Kumar, Oluwarotimi Williams Samuel, and Shixiong Chen. A hybrid deep learning approach for epileptic seizure detection in EEG signals. *IEEE Journal of Biomedical and Health Informatics*, 2023. doi: 10.1109/JBHI.2023.3265983.

Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one*, 16(8):e0256111, 2021.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Tidiane Camaret Ndir, Robin T Schirrmeister, and Tonio Ball. Eeg-clip: Learning eeg representations from natural language descriptions. *Frontiers in Robotics and AI*, 12:1625731, 2025.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

Didar Dadebayev, Wei Wei Goh, and Ee Xion Tan. Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4385–4401, 2022.

Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2023.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. *arXiv preprint arXiv:2510.21585*, 2025.

Zelin He, Sarah Alnegheimish, and Matthew Reimherr. Harnessing vision-language models for time series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21690–21698, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Robert Jenke, Angelika Peer, and Martin Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective computing*, 5(3):327–339, 2014.

Linxing Preston Jiang, Shirui Chen, Emmanuel Tanumihardja, Xiaochuang Han, Weijia Shi, Eric Shea-Brown, and Rajesh PN Rao. Data heterogeneity limits the scaling effect of pretraining neural data transformers. *bioRxiv*, pp. 2025–05, 2025a.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *International Conference on Learning Representations*, 2025b.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 100(17): e1750–e1762, 2023.

Jonathan W Kim, Ahmed Alaa, and Danilo Bernardo. Eeg-gpt: exploring capabilities of large language models for eeg classification and interpretation. *arXiv preprint arXiv:2401.18006*, 2024.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Chengfan Li, Yueyu Qi, Xuehai Ding, Junjuan Zhao, Tian Sang, and Matthew Lee. A deep learning method approach for sleep stage classification with eeg spectrogram. *International journal of environmental research and public health*, 19(10):6322, 2022a.

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. *Biomedical signal processing and control*, 72:103342, 2022b.

Jinpeng Li, Zhaoxiang Zhang, and Huiguang He. Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cognitive Computation*, 10(2):368–380, 2018.

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

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

Weiheng Lu, Chunfeng Song, Jiamin Wu, Pengyu Zhu, Yuchen Zhou, Weijian Mai, Qihao Zheng, and Wanli Ouyang. UniMind: Unleashing the power of LLMs for unified multi-task brain decoding. *arXiv preprint arXiv:2506.18962*, 2025.

Scott Makeig, Marissa Westerfield, T-P Jung, Sonia Enghoff, Jeanne Townsend, Eric Courchesne, and Terrence J Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295(5555): 690–694, 2002.

Raman K Malhotra. AASM scoring manual 3: a step forward for advancing sleep care for patients with obstructive sleep apnea. *Journal of Clinical Sleep Medicine*, 20(5):835–836, 2024. doi: 10.5664/jcsm.11040.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602. IEEE, 2022.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

Huy Phan, 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.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations*, 2023a.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024a.

Jialing Wang, Shiwei Cheng, Jieming Tian, and Yuefan Gao. A 2d cnn-lstm hybrid algorithm using time series segments of eeg data for motor imagery classification. *Biomedical Signal Processing and Control*, 83:104627, 2023b.

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

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025.

Ping Wang, Aimin Jiang, Xiaofeng Liu, Jing Shang, and Li Zhang. Lstm-based eeg classification in motor imagery tasks. *IEEE transactions on neural systems and rehabilitation engineering*, 26(11): 2086–2095, 2018.

Sheng Wong, Anj Simmons, Jessica Rivera-Villicana, Scott Barnett, Shobi Sivathamboo, Piero Perucca, Zongyuan Ge, Patrick Kwan, Levin Kuhlmann, Rajesh Vasa, Kon Mouzakis, and Terence J O'Brien. EEG datasets for seizure detection and prediction—a review. *Epilepsia Open*, 8(2): 252–267, 2023.

Chaoqi Yang, Danica Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised eeg representation learning for automatic sleep staging. *arXiv preprint arXiv:2110.15278*, 2021.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260, 2023.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

Ziyi Zeng, Zhenyang Cai, Yixi Cai, Xidong Wang, Junying Chen, Rongsheng Wang, Yipeng Liu, Siqi Cai, Benyou Wang, Zhiguo Zhang, and Haizhou Li. WaveMind: Towards a conversational EEG foundation model aligned to textual and visual modalities. *arXiv preprint arXiv:2510.00032*, 2025.

Junru Zhang, Lang Feng, Xu Guo, Yuhan Wu, Yabo Dong, and Duanqing Xu. Timemaster: Training time-series multimodal llms to reason via reinforcement learning. *arXiv preprint arXiv:2506.13705*, 2025.

Ruilong Zhang, Qun Zong, Liqian Dou, and Xinyi Zhao. A novel hybrid deep learning scheme for four-class motor imagery classification. *Journal of neural engineering*, 16(6):066004, 2019.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting. In *International Conference on Machine Learning*, 2025.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems*, 2025.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A TIME-FREQUENCY REPRESENTATION PARAMETERS

For each channel $c \in C$, the STFT is computed as

$$X_c(\tau, \omega) = \sum_t x_c(t)\, w(t-\tau)\, e^{-j\omega t}, \tag{5}$$

where $w(\cdot)$ denotes a window function and τ indexes the window center. We use a window size of 400 samples (2 seconds) with an overlap of 300 samples (1.5 seconds). We retain the magnitude spectrogram

$$S_c(\tau, f) = |X_c(\tau, 2\pi f)|\,, \tag{6}$$

which captures localized spectral energy over time. The frequency axis is also truncated to a fixed upper bound of 100 Hz to focus on physiologically relevant EEG bands and to suppress high-frequency noise.

B BASELINES

Here, we categorize the baseline methods used for our comparisons:

Task-specific supervised Models:

- **SPaRCNet** (Jing et al., 2023) is a deep neural network based on one-dimensional CNNs with dense residual connections for EEG decoding.
- **ContraWR** (Yang et al., 2021) is a CNN-based model that first transforms EEG signals into multi-channel spectrogram representations and then applies a ResNet-style two-dimensional CNN to extract discriminative features from the spectrograms.
- **CNN-Transformer** (Peh et al., 2022) utilizes CNNs to extract local features and transformer layers to model global dependencies in EEG signals.
- **FFCL** (Li et al., 2022b) employs parallel CNN and LSTM layers, where the CNN captures spatial features and the LSTM models temporal dynamics; features from both layers are then fused for classification.
- **ST-Transformer** (Song et al., 2021) is a transformer-based architecture that uses attention mechanisms to learn spatial and temporal representations from EEG signals.

Foundation Models:

- **BIOT** (Yang et al., 2023) is a transformer-based model for learning general-purpose biosignal representations using a mix of supervised and unsupervised pretraining. The pretrained model supports up to 18 EEG channels as input. For recordings with more than 18 channels, a 1×1 convolution is used to reduce the input to 18 channels.
- **LaBraM** (Jiang et al., 2024) is a large-scale EEG foundation model pretrained in an unsupervised manner on over 2,500 hours of diverse EEG data. It learns universal EEG representations by segmenting raw signals into channel-wise patches and employing vector-quantized neural spectrum prediction as a tokenizer, with a Transformer backbone capturing both temporal and spatial features for downstream applicability.
- **CBraMod** (Wang et al., 2025) is an EEG foundation model that learns generic representations via patch-based masked EEG reconstruction. It employs a criss-cross transformer to jointly model spatial and temporal dependencies across EEG patches in parallel, along with an asymmetric convolutional positional encoding to capture diverse spatial–temporal signal formats.
- **CSBrain** (Zhou et al., 2025) is a cross-subject brain foundation model that captures shared neural patterns across individuals through large-scale pretraining, aiming to enhance robustness and adaptability to downstream EEG analysis tasks.

Multi-task LLMs:

- **NeuroLM** (Jiang et al., 2025b) is a multi-task EEG foundation model that formulates EEG signals as discrete tokens by learning a neural tokenizer and passes the tokens to a large language model architecture. It is pre-trained on large-scale EEG corpus to learn generic representations and is fine-tuned using instruction tuning.

C EXPERIMENTAL SETTINGS

C.1 IMPLEMENTATION DETAILS

We trained BraVista for 1 epoch with an effective batch size of 128, using a local batch size of 8 and gradient accumulation steps of 4. Training is performed on 4 NVIDIA RTX A6000 GPUs for 12 hours. We apply a linear warm-up over the first 3% of total training steps, followed by a cosine learning rate decay schedule. The language model decoder is optimized using the AdamW optimizer with a peak learning rate of 1×10^{-4}, while the vision encoder is trained with a learning rate of 2×10^{-6}.

C.2 HYPERPARAMETER SETTINGS

Table A.1: Post-training hyperparameters.

Hyperparameters	**Value**
Base Model	Qwen3-VL-2B-Instruct
Training mode	LLM-Lora + Vision Encoder
Batch size per device	8
Number of devices	4
Global batch size	128
Gradient accumulation steps	4
Epochs	1
Learning rate (LLM/LoRA parameters)	1×10^{-4}
Learning rate (Merger)	1×10^{-5}
Learning rate (Vision)	2×10^{-6}
Weight decay	0.1
Warmup ratio	0.03
LR scheduler	Cosine
Freeze vision tower	False
Freeze LLM	True
Tune merger	False
LoRA rank (r)	32
LoRA scaling factor (α)	64
LoRA dropout	0.05

The hyperparameter settings in Table A.1 correspond to the multi-task BraVista. For the single-task using BraVista framework, we adjust the batch size and LoRA configuration on a per-dataset basis to account for differences in dataset size and task complexity. Specifically, we use a batch size of 32 for the Mental Arithmetic dataset, 128 for SEED, HMC, and TUAB. For LoRA adaptation, we employ a smaller configuration with rank $r = 8$ and scaling factor $\alpha = 8$ for the Mental Arithmetic, while using the same LoRA settings as the multi-task model for SEED, HMC, and TUAB. All other hyperparameters are kept identical unless otherwise specified. The single-task model on EEGMAT uses a smaller configuration tuned to limit overfitting on this small dataset. To rule out configuration differences as the source of the multi-task gain, we retrain the single-task EEGMAT model with exactly the multi-task hyperparameters. Its balanced accuracy drops to 52.38%, below both the tuned single-task model (61.28%) and BraVista (71.26%), so the multi-task gain is not explained by the training configuration.

C.3 METRICS

In this section, we describe the evaluation metrics used throughout the paper. We report evaluation metrics that account for class imbalance and task heterogeneity, which are common characteristics of EEG datasets. In many EEG classification settings, label distributions are imbalanced across classes, making standard accuracy insufficient to fully capture model performance. Following prior work in EEG foundation models (Jiang et al., 2024), we adopt metrics that provide a more balanced and informative evaluation.

- **Balanced Accuracy** is computed as the average recall across all classes. It accounts for imbalanced label distributions by giving equal weight to each class and is used for both binary and multi-class classification tasks. This metric is particularly suitable for EEG classification tasks with imbalanced label distributions, as it prevents performance on majority classes from dominating the overall score.
- **Weighted F1** is the average of per-class F1 scores, weighted by the number of samples in each class. It captures both precision and recall while reflecting the overall data distribution, providing a complementary view of performance that emphasizes effectiveness on more prevalent classes.

D DATASET DETAILS

In this section, we provide descriptions of the datasets, pre-processing procedures, data splitting protocols and the generated image and instruction for each dataset. Dataset statistics are summarized in Table A.2.

Table A.2: Overview of BCI datasets and tasks used for evaluation.

Dataset	BCI Task	#Channels	Sampling Rate	Duration	#Samples	Label
HMC	Sleep Stage Classification	4	256 Hz	30 s	137,243	5-class
SEED	Emotion Recognition	62	1000 Hz	4 s	38,475	3-class
EEGMAT	Cognitive Workload Classification	19	500 Hz	4 s	2,088	2-class
TUAB	Abnormal EEG Detection	23	256 Hz	10 s	409,455	2-class

D.1 SLEEP STAGE CLASSIFICATION

The Haaglanden Medisch Centrum (HMC) (Alvarez-Estevez & Rijsman, 2021) sleep staging dataset is a large EEG corpus collected for automated sleep stage classification. It contains overnight EEG recordings from 151 subjects with segments labeled according to five standard sleep stages: *Wake*, *NREM1*, *NREM2*, *NREM3*, and *REM*. EEG signals were recorded from four channels (F4–M1, C4–M1, O2–M1, and C3–M2) using a clinical 10–20 montage and sampled at 256Hz. Following NeuroLM (Jiang et al., 2025b) pre-processing settings, the raw signals are bandpass filtered between 0.1Hz and 75Hz, notch filtered at 50Hz to mitigate power-line interference, resampled to 200Hz, and normalized by dividing signal amplitudes by 100. For data division, the first 100 subjects are used for training, the next 25 subjects for validation, and the final 26 subjects for testing.

Each preprocessed EEG segment is converted into an STFT image and paired with a dataset-specific natural language prompt for sleep stage classification. An example STFT image for the Sleep staging dataset is shown in Figure A.1. The STFT representations of the EEG channels are arranged in a grid according to the channel ordering provided by the dataset. In particular, the 4 channels are placed into an 1×4 grid, where each channel has a resolution of 224×448 resulting in a final image resolution of 224×1792.

EEG Sleep Stage Classification Prompt

```
<image> You are analyzing a 30-second segment from a full-night
electroencephalogram (EEG) recording from the Haaglanden Medisch Centrum
(HMC) sleep staging dataset.
This dataset contains EEG recordings from 151 subjects, with segments
corresponding to one of the following 5 sleep stages:
Wake (Conscious awareness), NREM1 (Sleep onset), NREM2 (Light stable
sleep), NREM3 (Deep restorative sleep), and REM (Dream associated sleep).
```

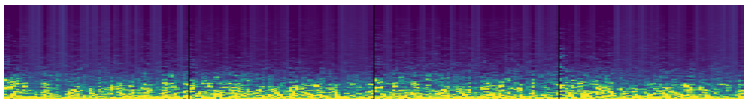


Figure A.1: Sample HMC STFT image

```
Recordings include EEG sampled at 200 Hz recorded from 4 channels:  EEG
F4-M1, EEG C4-M1, EEG O2-M1, and EEG C3-M2.
The current image shows a time-frequency representation (Short-Time
Fourier Transform; STFT) of the EEG segment.
The image contains four panels (left to right):  EEG F4-M1, EEG C4-M1,
EEG O2-M1, and EEG C3-M2.
In each panel, the horizontal axis represents time (0-30 s) and the
vertical axis represents frequency (0-100Hz), spanning conventional EEG
bands including delta (∼1-4 Hz), theta (∼4-8 Hz), alpha (∼8-13 Hz),
beta (∼13-30 Hz), and higher-frequency activity (gamma; ∼30-80+ Hz).
Color intensity reflects relative spectral power at each time-frequency
point, where brighter regions indicate stronger oscillatory power.
Your task is to determine the sleep stage associated with this 30-second
EEG segment.
You MUST only select ONE most appropriate label and enclose it within
<class> </class> tags, choosing exactly one of the following labels:
"Wake", "NREM1", "NREM2", "NREM3", or "REM".
```

D.2 Emotion Recognition

SEED (SJTU Emotion EEG Dataset) (Zheng & Lu, 2015) comprises EEG recordings from 15 subjects watching emotion-eliciting Chinese film clips corresponding to three affective states: *positive*, *neutral*, and *negative*. EEG signals were recorded using 62 channels at a sampling rate of 1000 Hz across three experimental sessions per subject, with each session containing 15 trials evenly distributed across the three emotion categories. Following NeuroLM (Jiang et al., 2025b) pre-processing settings, the EEG signals are bandpass filtered between 0.1 Hz and 75 Hz, notch filtered at 50 Hz to remove power-line interference, resampled to 200 Hz, and normalized by dividing signal amplitudes by 100. For data splitting, we follow a chronological trial-based protocol: the 15 trials are split into training, validation, and test sets using a 9:3:3 ratio, respectively, and all sessions are merged to form the final splits.

Each preprocessed EEG segment is converted into an STFT image and paired with a dataset-specific natural language prompt for emotion classification. An example STFT image for the SEED dataset is shown in Figure A.2. The STFT representations of the EEG channels are arranged in a grid according to the channel ordering provided by the dataset. In particular, the 62 channels are placed into an 8×8 grid, where each channel has a resolution of 112×112 resulting in a final image resolution of 896×896.

EEG Emotion Recognition Prompt

```
<image> You are analyzing a 4-second segment from an electroencephalogram
(EEG) recording from the SJTU Emotion EEG Dataset (SEED).
This dataset contains EEG recordings collected from multiple subjects
while they watched emotionally evocative video clips.
Each EEG segment corresponds to one of the following emotional states:
Positive (pleasant or high-valence emotion), Neutral (emotionally calm or
baseline), or Negative (unpleasant or low-valence emotion).
Recordings include EEG signals sampled at 200 Hz and recorded from 62
channels.
The current image shows a time-frequency representation (Short-Time
Fourier Transform; STFT) of the EEG segment.
The image contains 62 panels corresponding to each of the EEG channels.
In each panel, the horizontal axis represents time (seconds) and the
vertical axis represents frequency (0-100 Hz), spanning standard EEG
bands including delta (∼1-4 Hz), theta (∼4-8 Hz), alpha (∼8-13 Hz),
beta (∼13-30 Hz), and higher-frequency activity (gamma; ∼30-80+ Hz).
Color intensity reflects relative spectral power at each time-frequency
point, where brighter regions indicate stronger oscillatory power.
Your task is to determine the emotional state associated with this EEG
segment.
You MUST only select ONE most appropriate label and enclose it within
<class> </class> tags, choosing exactly one of the following labels:
"Positive", "Neutral", or "Negative".
```

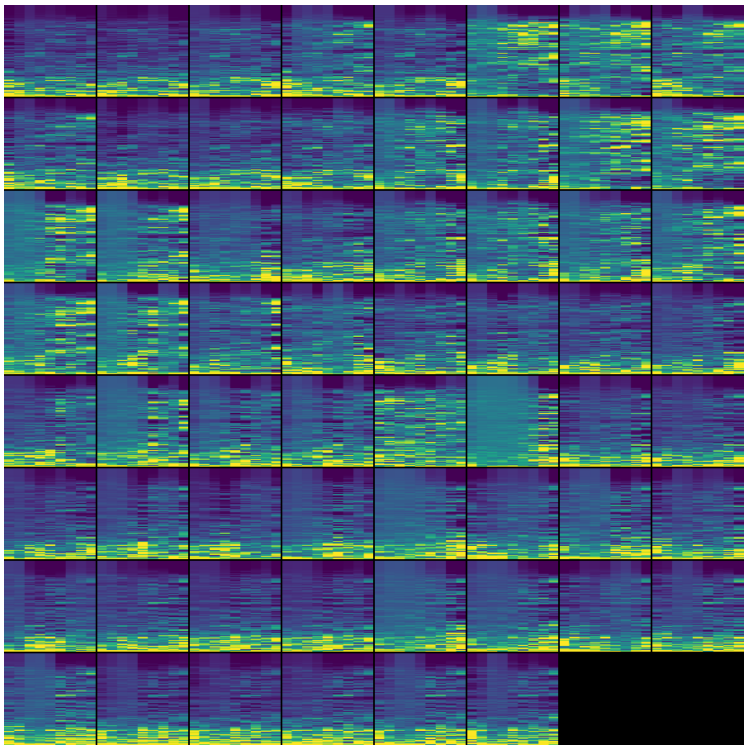


Figure A.2: Sample SEED STFT image

D.3 Mental Arithmetic Task

The EEG During Mental Arithmetic Tasks dataset (EEGMAT) (Zyma et al., 2019) comprises EEG recordings from 36 subjects performing mental arithmetic tasks and resting-state conditions. EEG signals were recorded using 19 channels arranged according to the international 10–20 system and sampled at 500 Hz. During the task condition, subjects performed serial subtraction task, inducing a high cognitive workload, while the resting condition corresponds to a low workload state. Each recording consists of continuous EEG data, which we segment into non-overlapping 4-second windows for classification. The EEG signals are are downsampled to 200Hz, bandpass filtered between 0.1 Hz and 75 Hz, notch filtered at 50 Hz to remove power-line interference, and normalized by dividing signal amplitudes by 100. We adopt a subject-independent split, using subjects 0–25 for training, 26–30 for validation, and 31–35 for testing.

Each preprocessed EEG segment is converted into an STFT image and paired with a dataset-specific natural language prompt for mental arithmetic detection. An example STFT image for the Mental Arithmetic dataset is shown in Figure A.3. The STFT representations of the EEG channels are arranged in a grid according to the channel ordering provided by the dataset. In particular, the 19 channels are placed into an 5×4 grid, where each channel has a resolution of 112×224 resulting in a final image resolution of 560×896.

EEG Mental Arithmetic Detection Prompt

```
<image> You are analyzing a 4-second segment of an electroencephalogram
(EEG) recording from the EEG During Mental Arithmetic Tasks (Zyma et al.,
2019) dataset.
This dataset contains EEG recordings from 36 subjects from 19 channels
sampled at 200 Hz and typically recorded using a standard 10-20 electrode
montage.
Each EEG segments corresponds either to serial subtraction tasks (high
cognitive workload) or to resting-state EEG (low cognitive workload).
The current image shows a time-frequency representation (Short-Time
Fourier Transform; STFT) of the EEG segment.
```

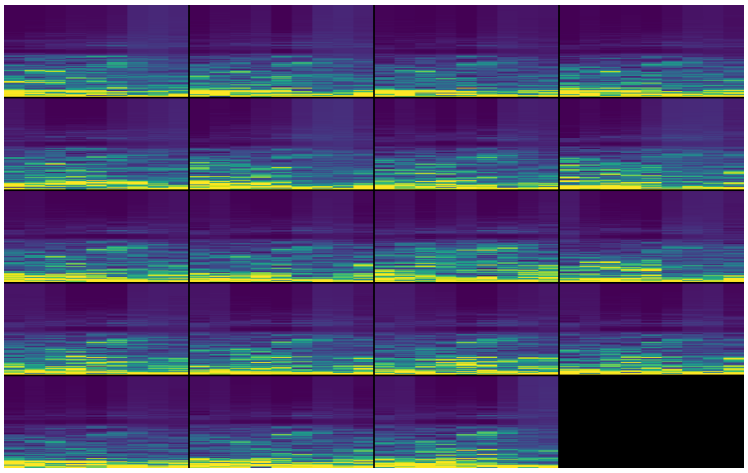


Figure A.3: Sample Workload STFT image

```
The image contains 19 panels each correspondeing to one EEG channel in
the following order:
EEG Fp1, EEG Fp2, EEG F3, EEG F4,
EEG F7, EEG F8, EEG T3, EEG T4,
EEG C3, EEG C4, EEG T5, EEG T6,
EEG P3, EEG P4, EEG O1, EEG O2,
EEG Fz, EEG Cz, EEG Pz.
In each panel, the horizontal axis represents time (0-4 s) and the
vertical axis represents frequency (0-100Hz), spanning conventional EEG
bands including delta (∼1-4 Hz), theta (∼4-8 Hz), alpha (∼8-13 Hz),
beta (∼13-30 Hz), and higher-frequency activity (gamma; ∼30-80+ Hz).
Color intensity reflects relative spectral power at each time-frequency
point, where brighter regions indicate stronger oscillatory power.
Your task is to determine the cognitive workload level associated with
this EEG segment.
You MUST only select ONE most appropriate label and enclose it within
<class> </class> tags, choosing exactly one of the following labels:
"Low" or "High".
```

D.4 Abnormal Detection

The Temple University Hospital Abnormal EEG (TUAB) dataset (Obeid & Picone, 2016) is a large-scale clinical EEG corpus designed for abnormality detection in EEG recordings. It contains EEG data collected from patients in a clinical setting, where each recording is labeled as either *normal* or *abnormal* based on expert neurological assessment. EEG signals are recorded using 23 channels following the standard 10–20 electrode montage and sampled at 256 Hz. For this task, continuous EEG recordings are segmented into non-overlapping 10-second clips, each inheriting the label of the corresponding recording. The EEG signals are downsampled to 200Hz, bandpass filtered between 0.1 Hz and 75 Hz, notch filtered at 50 Hz to remove power-line interference, and normalized by dividing signal amplitudes by 100. Since the original dataset provides an official split into training and test sets at the patient level, we further divide the training set into training and validation subsets using an 80%:20% ratio.

Each preprocessed EEG segment is converted into an STFT image and paired with a dataset-specific natural language prompt for abnormal detection. An example STFT image for the TUAB dataset is shown in Figure A.4. The STFT representations of the EEG channels are arranged in a grid according to the channel ordering provided by the dataset. In particular, the 23 channels are placed into an 6×4 grid, where each channel has a resolution of 112×224 resulting in a final image resolution of 672×896.

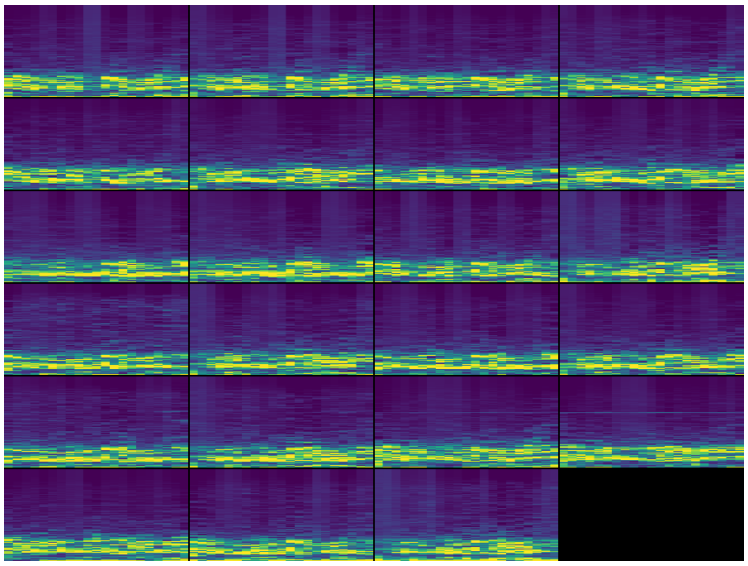


Figure A.4: Sample TUAB STFT image

EEG Abnormal Detection Prompt

```
<image> You are analyzing a 10-second segment of an electroencephalogram
(EEG) recording from the Temple University Hospital Abnormal EEG (TUAB)
dataset.
This dataset consists of clinical EEG recordings from real patients from
23 channels sampled at 200 Hz and typically recorded using a standard
10-20 electrode montage.
The dataset includes both neurologically normal and abnormal cases, and
the goal is to classify each segment as "Normal" or "Abnormal".
The current image shows a time-frequency representation (Short-Time
Fourier Transform; STFT) of the EEG segment.
The image contains 23 panels each correspondeing to one EEG channel in
the following order:
EEG Fp1-REF, EEG Fp2-REF, EEG F3-REF, EEG F4-REF,
EEG C3-REF, EEG C4-REF, EEG P3-REF, EEG P4-REF,
EEG O1-REF, EEG O2-REF, EEG F7-REF, EEG F8-REF,
EEG T3-REF, EEG T4-REF, EEG T5-REF, EEG T6-REF,
EEG A1-REF, EEG A2-REF, EEG Fz-REF, EEG Cz-REF,
EEG Pz-REF, EEG T1-REF, and EEG T2-REF.
In each panel, the horizontal axis represents time (0-10 s) and the
vertical axis represents frequency (0-100Hz), spanning conventional EEG
bands including delta (∼1-4 Hz), theta (∼4-8 Hz), alpha (∼8-13 Hz),
beta (∼13-30 Hz), and higher-frequency activity (gamma; ∼30-80+ Hz).
Color intensity reflects relative spectral power at each time-frequency
point, where brighter regions indicate stronger oscillatory power.
Your task is to determine which class best describes this EEG segment.
You MUST only select ONE most appropriate label and enclose it within
<class> </class> tags, choosing exactly one of the following labels:
"Normal" or "Abnormal".
```

E Instruction Ablation Details

In this section, we provide detailed descriptions of the instruction variations used in the ablation study. These experiments are designed to isolate how instruction formulation and label specification influence model behavior.

Each input consists of an EEG-derived visual representation paired with a textual instruction. We systematically modify the instruction to evaluate the role of (i) task description phrasing and (ii) label space definition. We test five variations here:

- **ORG+ORG (Baseline).** This setting uses the original task-specific instruction with the correct label options for the corresponding dataset. It serves as the reference condition for all comparisons.

- **Simple+ORG.** The task-specific instruction is replaced with a simplified generic description: *"This is an image of the STFT of an EEG signal."* while retaining the correct label options. This setting evaluates whether detailed task descriptions are necessary, or whether the model can rely on the label space and visual input alone at the inference time.
- **ORG+ORG (Shuffled).** The original instruction is preserved, but the order of label options is randomly shuffled. This setting tests whether the model is sensitive to positional biases in the instruction or relies on the semantic meaning of the labels.
- **ORG+ALL.** The instruction includes the original task description but expands the label space to include all label options from all datasets. This introduces ambiguity in the output space and evaluates whether the model can still identify the correct task within a larger set of possible labels.
- **ORG+ALL (excluding ORG).** The instruction includes label options from all datasets except those corresponding to the correct task. This creates a mismatched label space in which the correct answer is not present. This setting evaluates whether the model strictly follows the instruction-defined label space, even when it conflicts with the input signal.

These variations allow us to disentangle the role of instruction phrasing, label ordering, and label space definition in guiding model predictions. In particular, they test whether the model relies on semantic alignment between the EEG input and the instruction-defined output space.

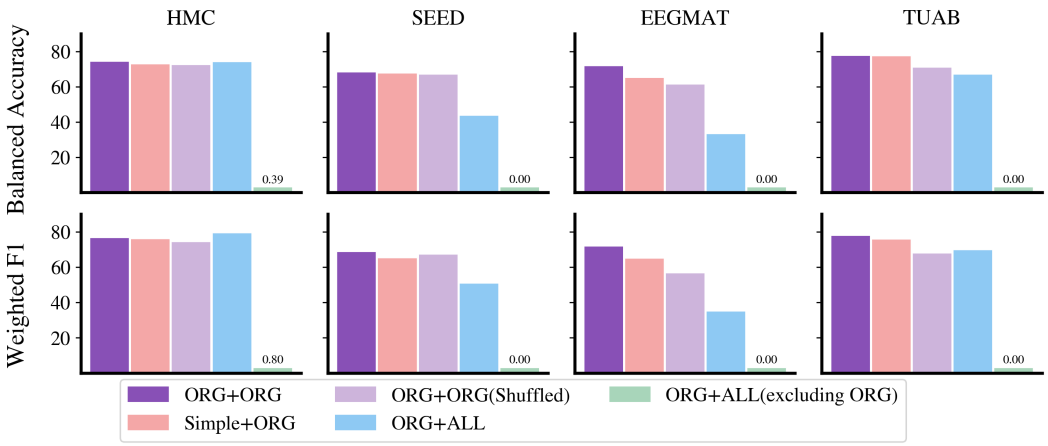


Figure A.5: **Instruction robustness analysis.** Performance of BraVista under different instruction formulations and training configurations across four datasets, evaluated using balanced accuracy (top) and weighted F1 (bottom). Results show that the model is robust to task description variations but degrades significantly when the correct label options are removed from the instruction, highlighting the importance of meaningful instruction.

F Effective EEG Decoding Requires Joint Adaptation of Vision and Language Components

Fig A.6 presents an ablation study evaluating the contribution of different trainable components in the vision–language model. We vary whether the vision encoder, vision–language merger, and language model decoder (via LoRA) are trained or frozen, and report balanced accuracy across all datasets.

We observe that effective EEG decoding requires both the vision encoder and the language model decoder to be trainable. Configurations in which either component is frozen result in a substantial drop in performance, indicating that adaptation of both the visual representation and the language model is critical. Training the vision–language merger in isolation leads to consistently poor performance, suggesting that it is insufficient for aligning EEG signals with downstream tasks. Fully unfreezing all

components yields comparable performance to BraVista, but does not provide consistent improvements and can introduce higher standard deviation suggesting less stable setup for continued post training.

Overall, these results indicate that the BraVista configuration training, which is the vision encoder and language model decoder while keeping the vision–language merger frozen, provides a stable and effective trade-off between adaptability and generalization.

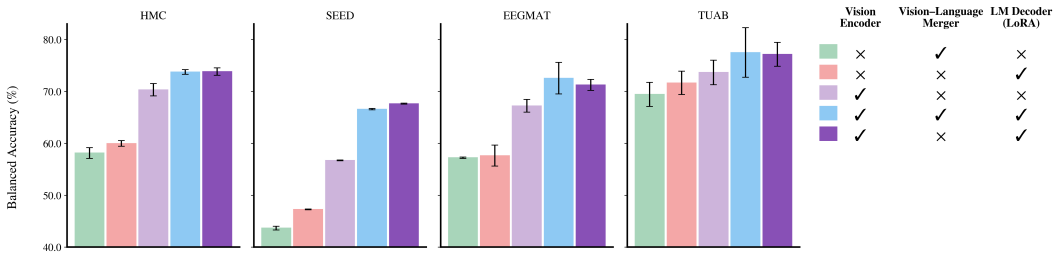


Figure A.6: **Ablation of training configurations.** Balanced accuracy across datasets under different training strategies, varying whether the vision encoder, vision–language merger, and language model decoder are trainable. Results show that jointly adapting the vision encoder and language model is critical, while training the merger alone is insufficient.

G Increasing Language Model Size Does Not Improve EEG Decoding Performance

We analyze the effect of language model scale on EEG decoding performance by evaluating BraVista with different backbone sizes (2B, 3B, and 7B parameters). While minor improvements are observed when increasing model size, these gains are marginal. This suggests that scaling the language model alone does not substantially improve EEG decoding performance in our setting and performance appears to be primarily governed by the quality of the visual representation and the effectiveness of cross-modal alignment, rather than increased language model capacity (Fig. A.7).

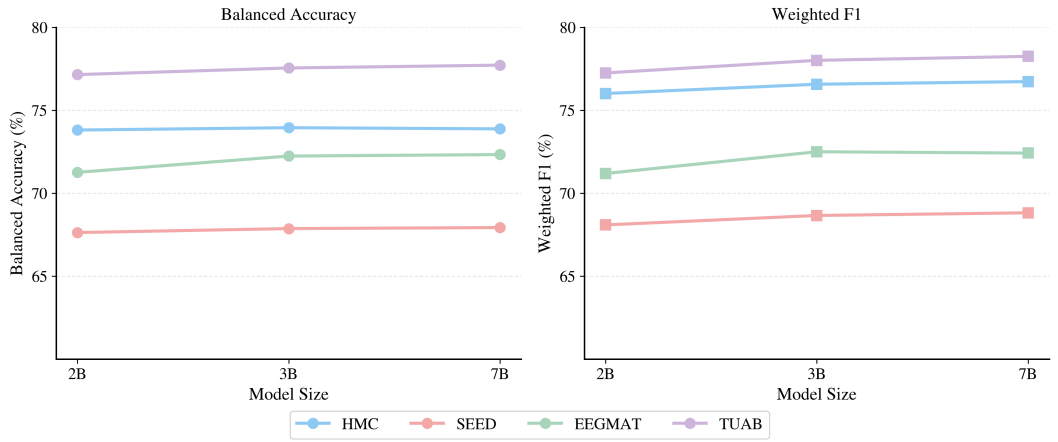


Figure A.7: **Scaling analysis of BraVista.** Performance across different language model sizes (2B, 3B, 7B) evaluated using balanced accuracy (left) and weighted F1 (right). Results show minimal performance variation across scales, indicating that increasing language model capacity alone does not lead to significant improvements in EEG decoding.