DUALBRAIN: SPHERICAL GEOMETRY AND BIDIRECTIONAL DYNAMICS FOR EEG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) foundation models have recently emerged to overcome the limitations of task-specific decoders, pretraining general-purpose representations that transfer across diverse downstream tasks. However, these architectures do not explicitly maintain both spatial-to-temporal and temporal-to-spatial processing orders as separate pathways within each stage. DualBrain makes this distinction explicit and combines the resulting representations. As a result, existing models fall short of capturing both aspects of spatiotemporal dependency: the temporal evolution of spatial patterns, and the spatial distribution and interaction of temporal patterns. We introduce **DualBrain**, an EEG foundation model that explicitly models these bidirectional spatiotemporal dependencies through a **Multi-view Joint Dynamics Encoder (MJDE)**, comprising a topographic dynamics branch and a temporal dynamics branch that respectively capture the temporal evolution of spatial patterns and the spatial interactions of temporal patterns, with their complementary representations adaptively integrated through gated fusion. To further ground the model in the physical geometry of the scalp, DualBrain incorporates **Spherical Harmonic Positional Encoding (SHPE)**, which jointly represents electrode positions across both angular dimensions using spherical harmonic bases, enabling structured, montage-independent positional grounding at multiple spatial scales. We pretrain DualBrain on TUH-EEG, the largest currently available single EEG dataset, using a masked autoencoding objective, and evaluate it on 7 tasks across 11 datasets spanning heterogeneous channel configurations and recording conditions. DualBrain achieves the best average performance compared to existing EEG foundation models, with particularly strong gains on seizure detection and mental state identification. These results suggest that explicitly modeling bidirectional spatiotemporal dependencies is a key ingredient for building general-purpose EEG representations.

1 INTRODUCTION

Decoding cognitive states from brain activity is central to neuroscience and brain–computer interfaces (BCIs). Electroencephalography (EEG) noninvasively records scalp electrical activity with high temporal resolution and low cost (Wolpaw et al., 2002; Nicolas-Alonso & Gomez-Gil, 2012) supporting applications ranging from sleep-stage classification and emotion recognition to motor imagery, seizure detection, and abnormal EEG detection (Aboalayon et al., 2016; Li et al., 2022; Ang & Guan, 2016). Traditional EEG analysis has examined spatiotemporal activity from two complementary perspectives: the spatial distribution of channel-wise temporal features and the temporal evolution of multichannel spatial patterns (Brunet et al., 2011; Michel & Murray, 2012). The former extracts handcrafted temporal features from channel-wise ERP, PSD, and ERSP analyses to examine their spatial distributions (Sur & Sinha, 2009; Alam et al., 2020; Possti et al., 2021). The latter characterizes scalp electric fields and examines their temporal dynamics through global field power, successive-map dissimilarity, and microstate segmentation (Brunet et al., 2011; Pascual-Marqui et al., 1995; Van de Ville et al., 2010).

Deep architectures, including CNNs (Simonyan & Zisserman, 2014), LSTMs (Hochreiter & Schmidhuber, 1997), and GNNs (Scarselli et al., 2008), subsequently enabled models to learn features directly from raw signals, giving rise to task-specific models such as DeepSleepNet (Supratak et al., 2017), DGCNN (Song et al., 2018), and EEG Conformer (Song et al., 2022). However,

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.

these models are trained specifically for each task and dataset, and variation in signal length, sampling frequency, and electrode montage across datasets fundamentally restricts their generalization. Foundation models have emerged to address this generalization by adapting Transformer architectures (Vaswani et al., 2017) from natural language processing and masked autoencoder (MAE)-based self-supervised learning from computer vision (He et al., 2022), pretraining general-purpose representations on large-scale unlabeled EEG that transfer across downstream tasks.

Recent EEG foundation models have sought to learn transferable representations from large-scale unlabeled data. Existing EEG foundation models do not adequately capture heterogeneous spatiotemporal sequential patterns. BIOT treats channel-wise temporal patches as a flattened token sequence (Yang et al., 2023a), whereas LaBraM represents EEG patches as neural codebook tokens and processes the resulting sequence with a standard Transformer (Jiang et al., 2024); both capture contextual dependencies, but the distinct roles and compound interactions of spatiotemporal patterns remain implicit. CBraMod separates spatial and temporal attention explicitly, but applies them in parallel to the same input, preventing patterns formed along one axis from sequentially conditioning the other (Wang et al., 2025). CSBrain introduces sequential spatiotemporal processing to capture cross-axis dependencies, but models only how temporally encoded patterns are distributed across brain regions, leaving the temporal evolution of spatial patterns unexplored (Zhou et al., 2025). Building on the two complementary perspectives of conventional EEG analysis (Kappenman et al., 2021; Michel & Koenig, 2018), we address this gap by explicitly combining temporal-to-spatial and spatial-to-temporal processing pathways, respectively modeling the spatial organization of temporal features and the temporal evolution of spatial patterns within a unified encoder.

Beyond spatiotemporal modeling, existing EEG foundation models also fall short in representing electrode geometry. Although positional information has grown progressively richer across this line of work, electrode geometry over the scalp surface remains insufficiently represented in a structured form. BIOT and LaBraM rely on channel identities or learned embeddings without encoding physical geometry (Yang et al., 2023a; Jiang et al., 2024); CBraMod captures relative positional relationships through input-dependent ACPE, but without grounding them in actual electrode locations (Wang et al., 2025). REVE narrows this gap by jointly encoding physical electrode coordinates and temporal positions via 4D Fourier positional encoding (El Ouahidi et al., 2025). SPE adopts spherical coordinates, but applies multi-frequency sinusoidal encoding independently to each angular dimension, without modeling their joint interaction (Yuce & Stober, 2026). As a result, no existing positional representation jointly captures structured spatial patterns over the spherical scalp surface—a geometric prior with direct relevance to how EEG signals are physically generated and measured. To close this gap, we ground electrode positions in scalp geometry using spherical harmonic bases that jointly represent both angular dimensions, capturing structured scalp topology across multiple spatial scales independently of electrode naming conventions.

We introduce **DualBrain**, an EEG foundation model that unifies these two solutions based with MAE. The dual-pathway design described above is formalized as the **Multi-view Joint Dynamics Encoder (MJDE)**, comprising a **Topographic Dynamics Branch** and a **Temporal Dynamics Branch** whose complementary representations are adaptively integrated through input-dependent gated fusion. The geometric grounding described above is formalized as **Spherical Harmonic Positional Encoding (SHPE)**. Together, MJDE and SHPE enable DualBrain to model spatiotemporal dynamics more comprehensively and ground its representations more faithfully in scalp geometry than existing EEG foundation models. Our contributions are as follows.

- We propose the **Multi-view Joint Dynamics Encoder (MJDE)**, comprising a **topographic dynamics branch** that models the temporal evolution of spatial patterns and a **temporal dynamics branch** that models the spatial interactions of temporal patterns, with their complementary representations adaptively combined through input-dependent gated fusion.
- We propose the **Spherical Harmonic Positional Encoding (SHPE)**, representing electrode positions via spherical harmonics to capture the spherical geometry of the scalp at multiple spatial scales. Because this representation is independent of electrode names, it applies consistently across varying channel counts and montage configurations.
- We pretrain the resulting model, **DualBrain**, on TUH-EEG, the largest currently available single EEG dataset, and evaluate it on 7 tasks across 11 datasets spanning heterogeneous channel configurations and spatiotemporal characteristics. DualBrain achieves the highest reported mean B-Acc across the 11 evaluated datasets.

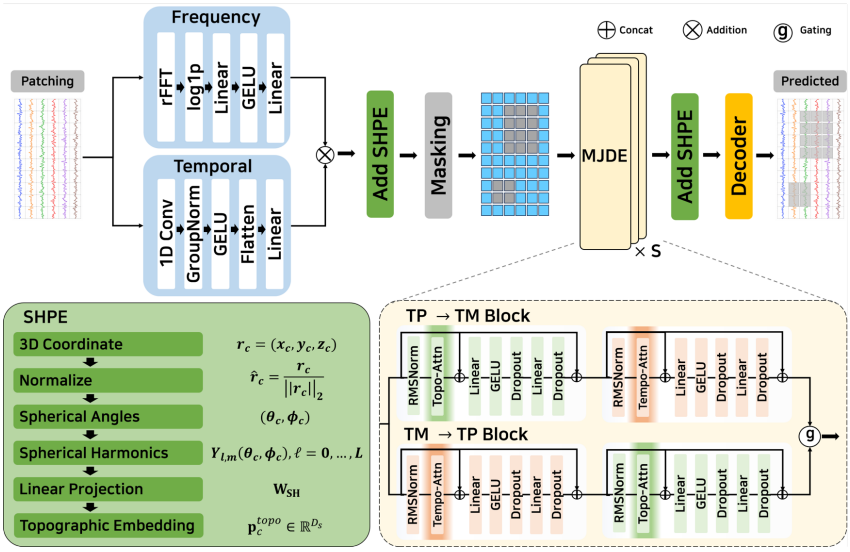

Figure 1: Overview of DualBrain. Multi-channel EEG signals are divided into temporal patches and encoded using temporal and frequency features, augmented with Spherical Harmonic Positional Encoding (SHPE). The resulting tokens are processed by the Multi-view Joint Dynamics Encoder (MJDE), which stacks three dual-branch stages modeling temporal-to-topographic (Temp→Topo) and topographic-to-temporal (Topo→Temp) interactions, with their outputs adaptively combined through a learnable, input-dependent gate. During pretraining, a lightweight decoder reconstructs the masked EEG patches from encoded visible representations and learnable mask tokens.

2 METHODS

We introduce **DualBrain**, an EEG foundation model that learns heterogeneous spatiotemporal dependencies from two complementary views. DualBrain uses masked autoencoding: preprocessed EEG is divided into patches, converted into tokens, and augmented with **SHPE**. All channel-time positions remain in a dense grid; masked positions are zeroed, excluded as attention keys, and kept zero in the **MJDE** outputs. A lightweight decoder reconstructs masked raw EEG patches from the encoded representations and a learnable mask token, each augmented with positional encodings.

2.1 PATCHING & TOKENIZATION

Given multichannel preprocessed EEG recordings $\mathbf{X} \in \mathbb{R}^{B\times C\times N}$, where B, C, and N denote the batch size, number of EEG channels, and number of samples per channel, respectively, we divide each channel into $T = \frac{N}{P}$ non-overlapping 1-second patches, each containing P samples (i.e., P equals the sampling rate). The patch from channel $c \in \{1, \dots, C\}$ at temporal index $t \in \{1, \dots, T\}$ is denoted as $\mathbf{x}_{c,t} \in \mathbb{R}^P$. A shared tokenizer is applied independently to each channel–time patch.

Time–Frequency Patch Tokenizer. We employ a convolutional patch encoder that independently processes each EEG patch $\mathbf{x}_{c,t}$ using a sequence of 1-D convolutional blocks, each followed by Group Normalization and GELU activation. The resulting features are flattened and linearly projected to obtain a temporal representation $\mathbf{e}^{\text{temp}}_{c,t} \in \mathbb{R}^{D/2}$. In parallel, we apply a real-valued FFT (rFFT) to each patch and transform its log-magnitude spectrum, $\log(1+|\,\text{rFFT}(\mathbf{x}_{c,t})|)$ using a two-layer MLP with GELU activation to obtain a frequency representation $\mathbf{e}^{\text{freq}}_{c,t} \in \mathbb{R}^{D/2}$. The temporal and frequency representations are concatenated to form the patch token:

$$\mathbf{e}_{c,t} = \left[\mathbf{e}^{\text{temp}}_{c,t} \| \mathbf{e}^{\text{freq}}_{c,t}\right], \qquad \mathbf{e}_{c,t} \in \mathbb{R}^D. \tag{1}$$

Collecting $\mathbf{e}_{c,t}$ over all channel-time positions yields the token set $\mathbf{E} \in \mathbb{R}^{B\times C\times T\times D}$.

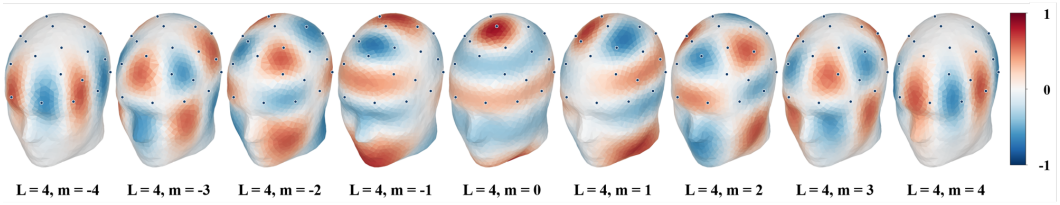

Figure 2: Real spherical harmonics on the scalp at degree $\ell = 4$ ($m = -4, \ldots, 4$), illustrating SHPE's fine-scale spatial patterns. The full 25-component basis ($\ell \leq 4$) is shown in Appendix E.2.

2.2 Spherical Harmonic Positional Encoding (SHPE)

Existing positional encodings based on electrode identities or Cartesian coordinates do not explicitly incorporate the spherical geometry of the scalp. Although spherical-coordinate-based encoding better reflects scalp geometry, independently encoding angular coordinates does not directly capture spatial patterns jointly defined over the spherical surface. Therefore, we propose **Spherical Harmonic Positional Encoding (SHPE)**, which represents electrode locations using spherical harmonic basis functions at multiple spatial frequencies and incorporates temporal embeddings.

For each electrode c, let $\mathbf{r}_c = (x_c, y_c, z_c)$ denote its Cartesian coordinate. We normalize the coordinate onto the unit sphere, $\hat{\mathbf{r}}_c = (\hat{x}_c, \hat{y}_c, \hat{z}_c) = \mathbf{r}_c / \|\mathbf{r}_c\|_2$, and convert it into spherical angles (θ_c, ϕ_c):

$$\theta_c = \arccos(\hat{z}_c), \qquad \phi_c = \operatorname{atan2}(\hat{y}_c, \hat{x}_c). \tag{2}$$

The position of each electrode is then represented using real spherical harmonic basis functions. For degree $\ell \geq 0$ and order m with $|m| \leq \ell$, the basis is defined as

$$N_\ell^{|m|} = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-|m|)!}{(\ell+|m|)!}}, \qquad Y_{\ell,m}(\theta,\phi) = \begin{cases} \sqrt{2} N_\ell^m P_\ell^m(\cos\theta)\cos(m\phi), & m > 0, \\ N_\ell^0 P_\ell^0(\cos\theta), & m = 0, \\ \sqrt{2} N_\ell^{|m|} P_\ell^{|m|}(\cos\theta)\sin(|m|\phi) & m < 0. \end{cases} \tag{3}$$

where $P_\ell^{|m|}(\cdot)$ denotes the associated Legendre polynomial and $N_\ell^{|m|}$ is the normalization factor. Unlike independently encoding the two spherical angles, each basis function jointly depends on (θ, ϕ) and therefore represents a spatial pattern defined over the spherical surface. Moreover, increasing ℓ corresponds to progressively finer spatial variations, allowing SHPE to represent scalp topology at multiple spatial scales.

We construct the spherical representation of electrode c by concatenating all basis functions for $\ell \in \{0, \ldots, L\}$ and $m \in \{-\ell, \ldots, \ell\}$:

$$\mathbf{y}_c = [Y_{\ell,m}(\theta_c, \phi_c)]_{\ell,m} \in \mathbb{R}^{(L+1)^2}. \tag{4}$$

The resulting spherical representation is linearly projected to obtain the topographic positional embedding,

$$\mathbf{p}_c^{\text{topo}} = \mathbf{W}_{\text{SH}}\, \mathbf{y}_c, \qquad \mathbf{p}_c^{\text{topo}} \in \mathbb{R}^{D/2}. \tag{5}$$

Following sinusoidal positional encoding (Vaswani et al., 2017), we encode each temporal index t using sine and cosine functions at multiple frequencies to obtain $\mathbf{p}_t^{\text{temp}} \in \mathbb{R}^{D/2}$. To jointly encode electrode location and temporal position, the topographic and temporal embeddings are concatenated and transformed to obtain the positional representation:

$$\mathbf{p}_{c,t} = \text{RMSNorm}\big(\text{GELU}\big(\big[\mathbf{p}_c^{\text{topo}} \,\|\, \mathbf{p}_t^{\text{temp}}\big]\big)\big), \qquad \mathbf{p}_{c,t} \in \mathbb{R}^D. \tag{6}$$

The resulting positional representation is added to the tokenizer output $\mathbf{e}_{c,t}$ for all tokens:

$$\mathbf{z}_{c,t} = \mathbf{e}_{c,t} + \mathbf{p}_{c,t}, \qquad \mathbf{z}_{c,t} \in \mathbb{R}^D. \tag{7}$$

SHPE generates positional encodings for any electrode layout without a predefined channel set, using electrode coordinates to capture scalp geometry more specifically than sinusoidal encodings of Cartesian coordinates.

2.3 MASKING

For each masking block, we randomly select a seed electrode and sample an angular radius $r \sim \mathcal{U}(35^\circ, 75^\circ)$ on the normalized scalp sphere. All electrodes whose angular distance from the seed falls within this radius are masked over τ contiguous temporal patches, where $\tau \sim \mathcal{U}\{2, \ldots, 10\}$. We repeatedly sample such blocks and take their union until a target masking ratio of 55% is reached. If the final block exceeds the remaining masking budget, only the required number of patches closest to its spatiotemporal center are retained.

The selected masked tokens are set to zero vector prior to the encoder. Let $m_{c,t} \in \{0, 1\}$ denote the masking indicator, where $m_{c,t} = 1$ indicates a masked patch. MJDE input is defined as

$$\widetilde{\mathbf{z}}_{c,t} = (1 - m_{c,t})\, \mathbf{z}_{c,t}, \qquad m_{c,t} \in \{0, 1\}, \qquad \widetilde{\mathbf{z}}_{c,t} \in \mathbb{R}^D. \tag{8}$$

2.4 MULTI-VIEW JOINT DYNAMICS ENCODER (MJDE)

Given the token grid, MJDE models EEG from two complementary perspectives: how temporal activity patterns are distributed across the scalp, and how spatial brain activity changes over time. In the **Temp→Topo pathway**, temporal attention first captures the temporal activity pattern within each electrode. Topographic attention then examines how these temporal patterns are distributed across different scalp regions. This follows a common EEG analysis perspective that first characterizes channel-wise temporal dynamics and then examines their spatial organization (Kappenman et al., 2021; Barry et al., 2007; McFarland et al., 2000). In the **Topo→Temp pathway**, topographic attention first captures the spatial pattern of brain activity across electrodes for each t. Temporal attention then tracks how this spatial pattern changes over time. This is similar to EEG topographic analyses that study the temporal evolution of scalp activity patterns (Michel & Koenig, 2018; Lehmann et al., 1987; Van de Ville et al., 2010).

These two pathways provide complementary views of the same EEG signal. At each stage, their representations are combined through a learnable gate, and the fused features are further processed by both pathways in subsequent stages. This repeated process progressively integrates and refines the two views, with the pathways and fusion gates jointly optimized during training.

Collecting tokens $\widetilde{\mathbf{z}}_{c,t}$ on all electrodes and temporal indices produces the MJDE input, $\widetilde{\mathbf{Z}} \in \mathbb{R}^{B \times C \times T \times D}$. Given $\widetilde{\mathbf{Z}}$, topographic attention (Topo-Attn) applies multi-head self-attention along the electrode axis independently for each temporal index, whereas temporal attention (Temp-Attn) applies multi-head self-attention along the temporal axis independently for each electrode.

MJDE stacks S encoding stages, each consisting of two pathways with reversed attention order. Let $\mathbf{Z}_{i-1}$ denote the input to stage i, with $\mathbf{Z}_0 = \widetilde{\mathbf{Z}}$. We denote the transformer block containing Temp-Attn as $\mathcal{TM}_i$ and the block containing Topo-Attn as $\mathcal{TP}_i$. The two pathways are computed as

$$\mathbf{H}_i^{\text{Topo}\to\text{Temp}} = \mathcal{TM}'_i\left(\mathcal{TP}'_i\left(\mathbf{Z}_{i-1}\right)\right). \tag{9}$$

$$\mathbf{H}_i^{\text{Temp}\to\text{Topo}} = \mathcal{TP}_i\left(\mathcal{TM}_i\left(\mathbf{Z}_{i-1}\right)\right), \tag{10}$$

The primed blocks $\mathcal{TP}'_i$ and $\mathcal{TM}'_i$ share the same architecture as their unprimed counterparts but use separate parameters for the reverse path. Each block uses RMSNorm, multi-head self-attention, residual connections, and a GELU-activated feed-forward network.

The two pathway representations are combined using a stage-specific, input-dependent gate. For each stage i, a gating vector is computed from both pathway representations via a linear projection followed by a sigmoid function,

$$\mathbf{G}_i = \sigma\Big(\mathbf{W}_i^g \left[\mathbf{H}_i^{\text{Topo}\to\text{Tempo}} \,\|\, \mathbf{H}_i^{\text{Tempo}\to\text{Topo}}\right] + \mathbf{b}_i^g\Big), \qquad \mathbf{G}_i \in \mathbb{R}^{B \times C \times T \times D}, \tag{11}$$

and used to combine the two pathways at each position:

$$\mathbf{Z}_i = \mathbf{G}_i \odot \mathbf{H}_i^{\text{Topo}\to\text{Temp}} + (1 - \mathbf{G}_i) \odot \mathbf{H}_i^{\text{Temp}\to\text{Topo}}. \tag{12}$$

The gating projection $(\mathbf{W}_i^g, \mathbf{b}_i^g)$ is learned independently for each stage i, allowing the balance between pathways to adapt both to the input and to the stage of processing, where $\odot$ denotes element-wise multiplication. After the final stage S, RMSNorm is applied to obtain the encoded representation $\mathbf{H}^{\text{enc}} \in \mathbb{R}^{B \times C \times T \times D}$.

2.5 Decoder and Pretraining Objective

Unlike CBraMod and CSBrain (Wang et al., 2025; Zhou et al., 2025) which use a linear reconstruction head, we adopt an MAE-style lightweight Transformer decoder (He et al., 2022), as in REVE (El Ouahidi et al., 2025). Our decoder has two Transformer blocks and an embedding dimension of $D/2$, compared with the deeper MJDE encoder of dimension D. The D-dimensional encoder outputs at visible positions are projected to $D' = D/2$. A separate decoder-specific SHPE generates D'-dimensional positional representations for all channel-time positions. These are added to the projected encoder outputs at visible positions and to a shared learnable mask token $\mathbf{q} \in \mathbb{R}^{D'}$ at masked positions. The decoder processes these tokens and predicts the P raw signal samples of each masked patch through a linear output head. We minimize the Smooth L_1 reconstruction loss over masked patches only:

$$\mathcal{L} = \frac{1}{\sum_{b,c,t} m_{b,c,t}} \sum_{b,c,t} m_{b,c,t} \, \mathrm{SmoothL1}_\beta \left(\widehat{\mathbf{x}}_{b,c,t}, \mathbf{x}_{b,c,t} \right), \tag{13}$$

where $\widehat{\mathbf{x}}_{b,c,t} \in \mathbb{R}^P$ is the decoder's reconstruction of the original EEG patch $\mathbf{x}_{b,c,t} \in \mathbb{R}^P$, $m_{b,c,t} = 1$ denotes a masked patch, and $\mathrm{SmoothL1}_\beta$ denotes the mean Smooth L_1 loss over its P signal samples, with $\beta = 0.1$.

Table 1: Downstream EEG datasets statistics after preprocessing. Samples denotes the number of segmented windows; rate denotes the original sampling rate.

Task	Dataset	Channels	Duration	Samples	Rate (Hz)	Classes
Abnormal detection	TUAB	16	10 s	409,455	256	2
Event classification	TUEV	16	5 s	113,353	256	6
Seizure detection	TUSZ	19	10 s	50,554	256	2
	CHB-MIT	16	10 s	326,993	256	2
	Siena	29	10 s	51,307	512	2
Emotion recognition	SEED-V	62	1 s	117,744	1,000	5
	FACED	30	10 s	10,332	250	9
Mental stress detection	MentalArithmetic	19	5 s	1,707	500	2
Motor imagery	PhysioNet-MI	64	4 s	9,837	160	4
Sleep staging	ISRUC	6	30 s	89,240	200	5
	HMC	4	30 s	137,243	256	5

3 Experiments

3.1 Pre-training

Pretraining Data and Preprocessing. We pretrain DualBrain on the Temple University Hospital EEG Corpus (TUEG) (Obeid & Picone, 2016). Following the CBraMod and CSBrain (Wang et al., 2025; Zhou et al., 2025), we use their processed version of the TUEG corpus. Recordings with the required 19 EEG channels are resampled to 200 Hz, band-pass filtered at 0.3–75 Hz, and notch filtered at 60 Hz. Recordings shorter than five minutes are excluded, and the first and last minute of each remaining recording are discarded before division into non-overlapping 30-s segments. Segments containing any sample with an absolute amplitude of at least $100\,\mu$V are removed. Before being fed to the model, the retained signals are divided by $100\,\mu$V, placing their amplitudes within $(-1, 1)$. The resulting corpus contains 1,109,545 segments, equivalent to 9,246 hours of EEG.

Pretraining Setup. DualBrain was pretrained on four NVIDIA A100 GPUs using distributed data parallelism. We trained for 40 epochs with a global batch size of 512 using AdamW, an initial learning rate of 5×10^{-4}, and a weight decay of 0.05. The learning rate followed a cosine schedule with a minimum of 10^{-5}. The embedding dimension was $D = 200$, and geometry-based masking selected 55% of the spatiotemporal patches. We used Smooth L_1 reconstruction loss with $\beta = 0.1$. Additional details are provided in the Appendix Section B.

Table 2: Comparison with existing EEG models across 11 downstream datasets spanning 7 tasks. Best results are **bold**; second-best are underlined.

Dataset	Metric	EEGNet	ContraWR	ST-Trans.	BIOT	LaBraM	CBraMod	CSBrain	REVE	CodeBrain	**DualBrain**
TUAB	B-Acc	0.7642	0.7746	0.7966	0.7959	0.8140	0.8289	0.8172	**0.8315**	0.8294	0.8100
	AUROC	0.8412	0.8456	0.8707	0.8815	0.9022	0.9227	0.8957	**0.9245**	0.9030	0.8905
TUEV	B-Acc	0.3876	0.4384	0.3984	0.5281	0.6409	0.6671	0.6903	0.6759	0.6428	**0.6978**
	F1-W	0.6539	0.6893	0.6823	0.7492	0.8312	0.8342	0.8333	**0.8451**	0.8362	0.8430
TUSZ	B-Acc	0.5366†	0.8466†	0.8533†	0.8743†	0.9246†	0.8808†	0.8734†	0.9197†	0.8934†	**0.9279**
	AUROC	0.9104†	0.9171†	0.8533†	0.9437†	**0.9862**†	0.9498†	0.9418†	0.9826†	0.9550†	0.9852
CHB-MIT	B-Acc	0.5658	0.6344	0.5915	0.7068	0.7075	0.7398	0.7262	0.7450†	0.7273	**0.8261**
	AUROC	0.8048	0.8097	0.8237	0.8761	0.8679	0.8892	0.8915	**0.9301**†	0.8961	0.9180
Siena	B-Acc	0.7487	0.6546	0.7527	0.7352	0.7082	0.7317	0.7662	0.7878†	0.6122†	**0.8410**
	AUROC	0.8687	0.7819	0.8884	0.9029	0.8814	0.9038	0.9076	0.9017†	0.8207†	**0.9354**
SEED-V	B-Acc	0.2961	0.3546	0.3052	0.3837	0.3976	0.4091	0.4197	0.3871†	0.4137	**0.4199**
	F1-W	0.2749	0.3544	0.2833	0.3856	0.3974	0.4101	**0.4280**	0.3914†	0.4235	0.4257
FACED	B-Acc	0.4090	0.4887	0.4810	0.5118	0.5273	0.5509	0.5752	0.5646	0.5941	**0.6082**
	F1-W	0.4124	0.4884	0.4795	0.5136	0.5288	0.5618	0.5796	0.5659	0.5953	**0.6124**
MentalArithmetic	B-Acc	0.6770	0.6631	0.6631	0.6875	0.6909	0.7256	0.7558	0.7660	0.7514	**0.7938**
(M-AT)	AUROC	0.7321	0.7332	0.7132	0.7536	0.7721	0.7905	0.8478	0.8450	0.8707	**0.8769**
PhysioNet-MI	B-Acc	0.5814	0.5892	0.6035	0.6153	0.6173	0.6417	0.6304	0.6480	0.6051†	**0.6503**
(P-MI)	F1-W	0.5796	0.5918	0.6053	0.6158	0.6177	0.6427	0.6308	0.6484	0.6062†	**0.6510**
ISRUC	B-Acc	0.7154	0.7402	0.7381	0.7527	0.7633	0.7865	**0.7925**	0.7819	0.7835	0.7883
	F1-W	0.7513	0.7610	0.7681	0.7790	0.7810	0.8011	0.7990	0.8005	0.8020	**0.8068**
HMC	B-Acc	0.6534	0.4242	0.2559	0.6862	0.7277	0.7269	0.7345	0.7401	0.7210†	**0.7414**
	F1-W	0.6536	0.2987	0.1428	0.7091	0.7454	0.7395	0.7506	**0.7638**	0.7388†	0.7604
Average	B-Acc	0.5759	0.6008	0.5854	0.6616	0.6836	0.6990	0.7074	0.7134	0.6885	**0.7368**

† Results reproduced using the publicly released code and pretrained weights.

3.2 Downstream Tasks

Datasets. We evaluate DualBrain on 11 downstream EEG datasets, including abnormal EEG detection (TUAB (Obeid & Picone, 2016)), EEG event classification (TUEV (Obeid & Picone, 2016)), seizure detection (TUSZ (Obeid & Picone, 2016), CHB-MIT (Guttag, 2010) and Siena (Detti, 2020)), emotion recognition (SEED-V (Liu et al., 2021) and FACED (Chen et al., 2023)), mental-state classification (MentalArithmetic (Zyma et al., 2019)), motor imagery classification (PhysioNet-MI (Schalk, 2009)), sleep staging (ISRUC (Khalighi et al., 2016) and HMC (Alvarez-Estevez & Rijsman, 2022)). For each dataset, we follow the preprocessing established in prior EEG foundation models.

Baselines. We compare DualBrain with both 3 task-specific EEG models and 6 pretrained EEG foundation models. The task-specific baselines include EEGNet Lawhern et al. (2018), ContraWR Yang et al. (2023b), and ST-Transformer Song et al. (2021). The EEG foundation model baselines include BIOT Yang et al. (2023a), LaBraM Jiang et al. (2024), CBraMod Wang et al. (2025), CSBrain Zhou et al. (2025), REVE El Ouahidi et al. (2025), and CodeBrain Ma et al. (2026).

Evaluation Metric. DualBrain is fully fine-tuned on each dataset following the corresponding data splits and evaluation protocols. For binary classification tasks, we report balanced accuracy (B-Acc), AUROC and AUPRC; for multi-class classification tasks, we report B-Acc, weighted F1-score (F1-W), and Cohen's Kappa. The best checkpoint is selected based on validation B-Acc. Our experiments and reproduced baselines are repeated over five random seeds, and we report the mean and standard deviation in the Appendix Section C.3 - C.9.

4 Results

4.1 Comparison with Existing EEG Models

Table 2 compares DualBrain with 3 task-specific baselines (EEGNet, ContraWR, ST-Transformer) and 6 EEG foundation models (BIOT, LaBraM, CBraMod, CSBrain, REVE, CodeBrain) across 11 datasets spanning 7 tasks. DualBrain achieves the best performance on mean B-Acc performance (0.7368), surpassing the strongest baseline, REVE, by +2.34 percentage points. With only 6.33M

parameters compared with REVE-Base's 69M, DualBrain delivers stronger average performance across diverse EEG tasks.

DualBrain shows its largest gains in seizure detection, improving B-Acc by 8.11 percentage points on CHB-MIT (0.8261 vs. 0.7450, REVE) and 5.32 on Siena (0.8410 vs. 0.7878, REVE), alongside AUROC gains. These results are consistent with MJDE's modeling of evolving spatial patterns and spatially distributed temporal activity. DualBrain also leads B-Acc on TUEV (0.6978), TUSZ (0.9279), SEED-V (0.4199), FACED (0.6082), MentalArithmetic (0.7938), PhysioNet-MI (0.6503), and HMC (0.7414). On ISRUC, it achieves the highest weighted F1 (0.8068), while CSBrain leads B-Acc. TUAB remains a limitation, with REVE leading B-Acc and AUROC. See Appendix Sections C.3-C.9 for further analysis.

4.2 Ablation Study

Table 3: Comparison of encoder architectures from prior EEG foundation models across downstream EEG datasets. Results are B-Acc (mean ± SD).

Dataset	LaBraM	CBraMod	CSBrain	DualBrain
TUAB	0.8033 ± 0.0075	0.8025 ± 0.0025	0.7979 ± 0.0096	$\mathbf{0.8100 \pm 0.0073}$
TUEV	0.6881 ± 0.0153	0.6658 ± 0.0163	0.6417 ± 0.0086	$\mathbf{0.6978 \pm 0.0104}$
TUSZ	0.9111 ± 0.0113	0.8964 ± 0.0200	0.8839 ± 0.0061	$\mathbf{0.9279 \pm 0.0187}$
CHB-MIT	$\mathbf{0.8326 \pm 0.0185}$	0.8101 ± 0.0419	0.7795 ± 0.0517	0.8261 ± 0.0348
Siena	$\mathbf{0.8698 \pm 0.0217}$	0.8538 ± 0.0229	0.8660 ± 0.0110	0.8410 ± 0.0114
SEED-V	0.4177 ± 0.0039	0.4050 ± 0.0052	0.4133 ± 0.0033	$\mathbf{0.4199 \pm 0.0034}$
FACED	0.5952 ± 0.0077	0.5675 ± 0.0062	0.5444 ± 0.0056	$\mathbf{0.6082 \pm 0.0052}$
M-AT	0.7542 ± 0.0363	0.7667 ± 0.0678	0.6958 ± 0.0723	$\mathbf{0.7938 \pm 0.0393}$
P-MI	$\mathbf{0.6503 \pm 0.0079}$	0.6332 ± 0.0094	0.6391 ± 0.0021	$\mathbf{0.6503 \pm 0.0034}$
ISRUC	$\mathbf{0.7934 \pm 0.0106}$	0.7746 ± 0.0126	0.7662 ± 0.0022	0.7883 ± 0.0033
HMC	$\mathbf{0.7426 \pm 0.0014}$	0.7289 ± 0.0019	0.7134 ± 0.0042	0.7414 ± 0.0042
Avg.	0.7326	0.7186	0.7037	**0.7368**

Encoder Comparison. We compare **MJDE** with the encoders of LaBraM, CBraMod, and CSBrain. Across all 11 datasets with 4 encoders, DualBrain achieves the highest mean B-Acc (0.7368), compared with LaBraM (0.7326), CBraMod (0.7186), and CSBrain (0.7037). It outperforms CBraMod and CSBrain each on 10 of the 11 datasets, with mean advantages of 1.82 and 3.31 percentage points. These results show that explicitly modeling both orders of spatial and temporal dependency provides a substantial advantage over the other two structured EEG encoders across the evaluated tasks.

DualBrain leads the full-attention LaBraM on 6 of the 11 datasets and lower B-Acc on 4, with nearly identical scores on P-MI. Despite using axis-wise attention, DualBrain achieves comparable overall performance to LaBraM while outperforming the other axis-wise encoders, CBraMod and CSBrain. The explicit separation of the $\mathcal{TM} \rightarrow \mathcal{TP}$ and $\mathcal{TP} \rightarrow \mathcal{TM}$ pathways facilitates targeted ablations of each processing order. It also allows us to examine how their representations are combined at each stage across tasks.

Table 4: Ablation study of the dual-path encoder across downstream EEG datasets. Results are B-Acc (mean ± SD).

Dataset	$\mathcal{TM} \rightarrow \mathcal{TP}$	$\mathcal{TP} \rightarrow \mathcal{TM}$	Dual fixed	DualBrain
TUAB	0.8059 ± 0.0109	$\mathbf{0.8164 \pm 0.0050}$	0.8101 ± 0.0081	0.8100 ± 0.0073
TUEV	0.6879 ± 0.0116	0.6835 ± 0.0119	0.6865 ± 0.0089	$\mathbf{0.6978 \pm 0.0104}$
TUSZ	0.9202 ± 0.0073	0.9085 ± 0.0107	0.9102 ± 0.0118	$\mathbf{0.9279 \pm 0.0187}$
CHB-MIT	0.8248 ± 0.0264	0.7887 ± 0.0347	0.8074 ± 0.0208	$\mathbf{0.8261 \pm 0.0348}$
Siena	0.8390 ± 0.0249	$\mathbf{0.8694 \pm 0.0099}$	0.8182 ± 0.0200	0.8410 ± 0.0114
SEED-V	0.4050 ± 0.0078	0.4124 ± 0.0052	0.4116 ± 0.0088	$\mathbf{0.4199 \pm 0.0034}$
FACED	0.5851 ± 0.0049	$\mathbf{0.6171 \pm 0.0098}$	0.5859 ± 0.0136	0.6082 ± 0.0052
M-AT	0.7389 ± 0.0704	0.6938 ± 0.0691	0.7382 ± 0.0446	$\mathbf{0.7938 \pm 0.0393}$
P-MI	0.6363 ± 0.0041	0.6285 ± 0.0069	0.6483 ± 0.0034	$\mathbf{0.6503 \pm 0.0034}$
ISRUC	0.7826 ± 0.0121	$\mathbf{0.7917 \pm 0.0029}$	0.7831 ± 0.0117	0.7883 ± 0.0033
HMC	0.7368 ± 0.0015	0.7351 ± 0.0011	$\mathbf{0.7439 \pm 0.0010}$	0.7414 ± 0.0042
Avg.	0.7239	0.7223	0.7221	**0.7368**

Encoder Ablation. We compare each processing order in isolation with two 3-stage dual-path encoders. In the Dual fixed, the $\mathcal{TP} \rightarrow \mathcal{TM}$ and $\mathcal{TM} \rightarrow \mathcal{TP}$ outputs are averaged element-wise at every stage, and the resulting representation is passed to both paths in the next stage. DualBrain instead learns how to combine the two outputs through a gate. Across the 11 datasets shared by both single-path models, DualBrain achieves the highest mean B-Acc (0.7368), exceeding $\mathcal{TM} \rightarrow \mathcal{TP}$ (0.7239) and $\mathcal{TP} \rightarrow \mathcal{TM}$ (0.7223) by 1.29 and 1.45 percentage points, respectively. It leads on 6 datasets and ranks second on the other 4. Although $\mathcal{TP} \rightarrow \mathcal{TM}$ performs best on TUAB, Siena, FACED, and ISRUC, DualBrain's consistently high ranking shows that combining the two orders works well across tasks with different preferred processing orders.

DualBrain also improves the average B-Acc over Dual fixed by 1.47 percentage points and outperforms it on 9 of the 11 datasets. Because Dual fixed retains both paths and combines them at every stage, this comparison isolates the benefit of learning their relative contributions beyond simply including both processing orders. The activation-aware analysis in Appendix Fig. 3 further shows that the learned gate mixes the two branches with low temporal variation while retaining task- and stage-specific contribution patterns. Together, these results support stage-wise adaptive fusion as a key component of the encoder.

Table 5: Comparison of positional encoding methods across downstream EEG datasets. Results are B-Acc (mean ± SD).

Dataset	w Ch-ID	w ACPE	w 4D PE	DualBrain
TUAB	0.8021 ± 0.0075	0.8019 ± 0.0083	$\mathbf{0.8103 \pm 0.0095}$	$\underline{0.8100 \pm 0.0073}$
TUEV	0.6714 ± 0.0151	0.6841 ± 0.0146	$\mathbf{0.7015 \pm 0.0129}$	$\underline{0.6978 \pm 0.0104}$
TUSZ	0.9053 ± 0.0245	0.9080 ± 0.0144	$\underline{0.9103 \pm 0.0113}$	$\mathbf{0.9279 \pm 0.0187}$
CHB-MIT	0.8102 ± 0.0370	0.7988 ± 0.0451	$\underline{0.8244 \pm 0.0308}$	$\mathbf{0.8261 \pm 0.0348}$
Siena	$\underline{0.8499 \pm 0.0181}$	0.8254 ± 0.0378	$\mathbf{0.8509 \pm 0.0200}$	0.8410 ± 0.0114
SEED-V	$\underline{0.4094 \pm 0.0053}$	0.4060 ± 0.0045	0.4079 ± 0.0084	$\mathbf{0.4199 \pm 0.0034}$
FACED	0.6010 ± 0.0066	$\mathbf{0.6108 \pm 0.0066}$	0.5906 ± 0.0130	$\underline{0.6082 \pm 0.0052}$
M-AT	$\mathbf{0.8056 \pm 0.0326}$	0.7688 ± 0.0750	0.7125 ± 0.0083	$\underline{0.7938 \pm 0.0393}$
P-MI	0.6295 ± 0.0039	0.6422 ± 0.0099	$\mathbf{0.6519 \pm 0.0043}$	$\underline{0.6503 \pm 0.0034}$
ISRUC	0.7806 ± 0.0084	$\mathbf{0.7997 \pm 0.0076}$	0.7835 ± 0.0082	$\underline{0.7883 \pm 0.0033}$
HMC	0.7315 ± 0.0045	$\underline{0.7409 \pm 0.0026}$	0.7368 ± 0.0037	$\mathbf{0.7414 \pm 0.0042}$
Avg.	$\underline{0.7270}$	0.7261	0.7255	$\mathbf{0.7368}$

Positional Encoding Comparison. We replaced SHPE in DualBrain with 3 positional encoding methods from prior work: channel-ID embeddings used in BIOT and LaBraM, ACPE (Wang et al., 2025), and 4D PE (El Ouahidi et al., 2025). More details are provided in the Appendix Section D.2. No alternative positional encoding consistently performs best across tasks. Channel-ID embeddings lead on M-AT, while ACPE leads on FACED and ISRUC. The 4D PE achieve the highest BAcc on TUAB, TUEV, Siena, and P-MI. However, channel-ID embeddings trail the best B-Acc on P-MI by 2.24 percentage points. The corresponding gaps are 2.55 points for ACPE on Siena and 9.31 points for 4D PE on M-AT. In contrast, SHPE achieves the highest mean B-Acc across the 11 datasets (0.7368). It remains within 1.18 percentage points of the best result on every dataset. These findings supports consistent performance across the evaluated electrode montages

5 Conclusion

We proposed DualBrain, an EEG foundation model designed for transferable representation learning across heterogeneous tasks and electrode configurations. DualBrain combines the Multi-view Joint Dynamics Encoder (MJDE), which models EEG through complementary topography-to-temporal and temporal-to-topography pathways, with Spherical Harmonic Positional Encoding (SHPE). This design enables the model to preserve both ordered spatiotemporal dependencies and continuous electrode geometry when adapting to downstream configurations not observed during pretraining. Across 11 datasets spanning 7 clinical and BCI tasks, DualBrain achieves state-of-the-art performance on the majority of the evaluated benchmarks, demonstrating broad transferability across downstream settings. These results emphasize that the order of topography and temporal integration should be explicitly considered in EEG foundation models.

AI USE STATEMENT

In this work, we used generative AI tools to refine hypotheses, provide feedback on research methodology and experiments, refine methods, interpret results, assist with translation, and support qualitative and thematic data analysis. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide ingredients for proving mathematical claims, assist in writing proofs, or clean and reformat datasets. Survey and interview question formulation, scientific figure or image generation, and transcription of research recordings were not applicable to this work. Additionally, we used generative AI tools to refactor code, edit manuscript text, summarize and search the literature, and improve readability. We reviewed all AI-assisted work. We checked literature-based claims against original sources, verified AI-assisted code and reported results, and reviewed and finalized all AI-assisted text. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study used existing public EEG datasets to develop and evaluate a foundation model. We did not collect new data from human participants. We used the datasets in accordance with their respective access and usage conditions and report results in aggregate.

REPRODUCIBILITY STATEMENT

To preserve anonymity during peer review, we do not provide a public code link at this stage. The supplementary material includes the code with identifying and other information that could compromise the double-blind review removed. We plan to release the complete code upon acceptance.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

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

Raquib-ul Alam, Haifeng Zhao, Andrew Goodwin, Omid Kavehei, and Alistair McEwan. Differences in power spectral densities and phase quantities due to processing of eeg signals. *Sensors*, 20(21):6285, 2020.

Diego Alvarez-Estevez and Roselyne Rijsman. Haaglanden Medisch Centrum sleep staging database. *PhysioNet*, March 2022. doi: 10.13026/t79q-fr32. URL `https://doi.org/10.13026/t79q-fr32`. Version 1.1.

Kai Keng Ang and Cuntai Guan. Eeg-based strategies to detect motor imagery for control and rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(4): 392–401, 2016.

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 2390–2397. IEEE, 2008.

Robert J Barry, Adam R Clarke, Stuart J Johnstone, Christopher A Magee, and Jacqueline A Rushby. Eeg differences between eyes-closed and eyes-open resting conditions. *Clinical neurophysiology*, 118(12):2765–2773, 2007.

Denis Brunet, Micah M Murray, and Christoph M Michel. Spatiotemporal analysis of multichannel eeg: Cartool. *Computational intelligence and neuroscience*, 2011(1):813870, 2011.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Paolo Detti. Siena Scalp EEG Database. *PhysioNet*, August 2020. doi: 10.13026/5d4a-j060. URL https://doi.org/10.13026/5d4a-j060. Version 1.0.0.

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. *Advances in Neural Information Processing Systems*, 2025. URL https://brain-bzh.github.io/reve/.

John Guttag. CHB-MIT Scalp EEG Database. *PhysioNet*, June 2010. doi: 10.13026/C2K01R. URL https://doi.org/10.13026/C2K01R. Version 1.0.0.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Ole Jensen and Claudia D Tesche. Frontal theta activity in humans increases with memory load in a working memory task. *European journal of Neuroscience*, 15(8):1395–1399, 2002.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Emily S Kappenman, Jaclyn L Farrens, Wendy Zhang, Andrew X Stewart, and Steven J Luck. Erp core: An open resource for human event-related potential research. *NeuroImage*, 225:117465, 2021.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2015.10.013. URL https://www.sciencedirect.com/science/article/pii/S0169260715002734.

Wolfgang Klimesch. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29(2-3):169–195, 1999.

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.

Dietrich Lehmann, Hisaki Ozaki, and Ivan Pál. Eeg alpha map series: brain micro-states by space-oriented adaptive segmentation. *Electroencephalography and clinical neurophysiology*, 67(3): 271–288, 1987.

Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, and Pekka Marttinen. Eeg based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4):1–57, 2022.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Bridging decoupled tokenizer and multi-scale architecture for eeg foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 83526–83574, 2026.

Dennis J McFarland, Laurie A Miner, Theresa M Vaughan, and Jonathan R Wolpaw. Mu and beta rhythm topographies during motor imagery and actual movements. *Brain topography*, 12(3): 177–186, 2000.

Christoph M Michel and Thomas Koenig. Eeg microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *Neuroimage*, 180:577–593, 2018.

Christoph M Michel and Micah M Murray. Towards the utilization of eeg as a brain imaging tool. *Neuroimage*, 61(2):371–385, 2012.

Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *sensors*, 12(2):1211–1279, 2012.

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

R.D. Pascual-Marqui, C.M. Michel, and D. Lehmann. Segmentation of brain electrical activity into microstates: model estimation and validation. *IEEE Transactions on Biomedical Engineering*, 42 (7):658–665, 1995. doi: 10.1109/10.391164.

Daniel Possti, Firas Fahoum, Ronen Sosnik, Nir Giladi, Jeffrey M Hausdorff, Anat Mirelman, and Inbal Maidan. Changes in the eeg spectral power during dual-task walking with aging and parkinson's disease: initial findings using event-related spectral perturbation analysis. *Journal of neurology*, 268(1):161–168, 2021.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Gerwin Schalk. EEG Motor Movement/Imagery Dataset. *PhysioNet*, September 2009. doi: 10.13026/C28G6P. URL `https://doi.org/10.13026/C28G6P`. Version 1.0.0.

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.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541, 2018.

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.

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

Shravani Sur and Vinod Kumar Sinha. Event-related potential: An overview. *Industrial psychiatry journal*, 18(1):70, 2009.

Dimitri Van de Ville, Juliane Britz, and Christoph M Michel. Eeg microstate sequences in healthy humans at rest reveal scale-free dynamics. *Proceedings of the National Academy of Sciences*, 107 (42):18179–18184, 2010.

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

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*, volume 2025, pp. 75310–75346, 2025.

Desirée Weber, Stephan Hertweck, Hisham Alwanni, Lukas DJ Fiederer, Xi Wang, Fabian Unruh, Martin Fischbach, Marc Erich Latoschik, and Tonio Ball. A structured approach to test the signal quality of electroencephalography measurements during use of head-mounted displays for virtual reality applications. *Frontiers in neuroscience*, 15:733673, 2021.

B. M. Wingeier, P. L. Nunez, and R. B. Silberstein. Spherical harmonic decomposition applied to spatial-temporal analysis of human high-density electroencephalogram. *Phys. Rev. E*, 64:051916, Oct 2001. doi: 10.1103/PhysRevE.64.051916. URL https://link.aps.org/doi/10.1103/PhysRevE.64.051916.

Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

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

Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI*, 2(1):e46769, 2023b.

Ayse Betul Yuce and Sebastian Stober. Benchmarking positional encoding strategies for transformer-based eeg foundation models. *arXiv preprint arXiv:2605.29754*, 2026.

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.

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. *arXiv preprint arXiv:2506.23075*, 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 Related Work

A.1 EEG-specific Neural Architectures

EEG decoding has evolved from handcrafted spectral and spatial features with conventional classifiers (Ang et al., 2008) toward neural architectures that learn task-relevant representations. DeepConvNet and ShallowConvNet were presented in the same study (Schirrmeister et al., 2017). DeepConvNet uses a deeper convolutional architecture for EEG decoding. ShallowConvNet is designed to extract band-power features. EEGNet introduced spatial and separable convolutions in depth in a compact architecture evaluated (Lawhern et al., 2018). DGCNN modeled relationships among electrodes using a learnable graph (Song et al., 2018), while EEG Conformer combined temporal–spatial convolution with self-attention to capture local features and longer-range dependencies (Song et al., 2022). Inspired by advances in large language models (Vaswani et al., 2017) and self-supervised learning (Chen et al., 2020), recent EEG foundation models have adopted large-scale pretraining to learn representations that can be adapted to diverse downstream tasks (Yang et al., 2023a; Jiang et al., 2024; Wang et al., 2025).

A.2 Positional Encoding for Heterogeneous EEG Montages

Electrode positions provide important spatial information for EEG modeling, yet montages differ in channel identity, number, and placement. BIOT and LaBraM use learned channel or spatial embeddings to distinguish electrodes (Yang et al., 2023a; Jiang et al., 2024). Such embeddings accommodate different combinations of known channels, but do not explicitly encode physical distances between electrodes or directly assign a geometry-based representation to a previously unseen location. CBraMod instead generates asymmetric conditional positional encodings from neighboring patches along the channel and time axes, allowing adaptation to varying EEG formats (Wang et al., 2025). However, proximity along the ordered channel axis does not necessarily correspond to anatomical proximity on the scalp.

Geometry-aware approaches incorporate electrode coordinates more directly. REVE encodes continuous electrode coordinates and patch time as (x, y, z, t) using multi-frequency Fourier features together with a learnable adjustment, supporting variable electrode arrangements (El Ouahidi et al., 2025). Spherical positional encoding (SPE) represents electrode azimuth and inclination through separate multi-frequency sinusoidal mappings (Yuce & Stober, 2026). Although both approaches use spatial coordinates, neither constructs a joint basis organized by spatial frequency over the sphere. Spherical harmonics provide such a basis for continuous spherical surfaces and have previously been applied to the analysis of EEG scalp topographies (Wingeier et al., 2001). This motivates our use of spherical harmonic positional encoding to represent electrode geometry across heterogeneous montages.

A.3 Spatiotemporal EEG Encoders

EEG's spatiotemporal structure can be examined from two complementary perspectives. EEG microstate analysis identifies recurring scalp topographies and characterizes their durations and transitions, describing activity as a sequence of spatial states (Michel & Koenig, 2018). Conversely, event-related potential (ERP) analysis examines event waveforms at individual electrodes and compares the amplitude and timing of components such as P3 across scalp locations (Kappenman et al., 2021). Spectral analysis likewise estimates channel-wise power and maps its spatial distribution, as illustrated by the posterior increase in alpha power during eye closure (Weber et al., 2021).

These perspectives motivate two possible orders for integrating EEG features. Topo-to-tempo modeling first captures relationships across electrodes at each time interval and then tracks how these spatial representations evolve. Tempo-to-topo modeling first extracts temporal features within each electrode and then relates the resulting representations across electrodes.

Existing EEG foundation models integrate these dimensions in different ways. BIOT and LaBraM apply attention to channel–time token sequences (Yang et al., 2023a; Jiang et al., 2024), while CBraMod models spatial and temporal dependencies through parallel attention (Wang et al., 2025). CSBrain instead applies inter-window temporal attention followed by inter-region spatial attention, corresponding to a temporal-to-spatial processing order (Zhou et al., 2025). However, whether ex-

plicitly combining both integration orders provides complementary benefits across EEG tasks remains insufficiently examined.

B PRETRAINING DETAILS

To facilitate reproducibility, we provide a complete specification of the architectural, optimization, and training hyperparameters used at our experiments in Table. 6.

Table 6: Pretraining data and Hyperparameter of DualBrain.

MODULE	HYPERPARAMETER	DualBrain
Data	Pretraining dataset	TUEG, 1,109,545 unlabeled segments
	Sampling rate	200 Hz
	Number of channels	19
	Segment duration	30 s
	Frequency range	0.3 - 75 Hz
Patch	Patch duration	1 s
	Patch size	200 samples
	Number of patches	$19 \times 30 = 570$
	Patch overlap	None
Tokenizer	Token dimension	200
	CNN channels	$1 \to 64 \to 128 \to 200$
	Convolution kernel	[15, 7, 3]
	Convolution strides	[5, 3, 2]
	Temporal dimension	100
	Frequency dimension	100
	Frequency representation	101-bin log-amplitude rFFT representation
	Feature fusion	Concatenation
	Dropout	0.1
Positional encoding	Topography encoding	25 real spherical harmonic bases, $l = 0, \ldots, 4$
	Temporal encoding	Sinusoidal encoding
	Encoder allocation	100 topo + 100 tempo dimensions
	Decoder allocation	50 topo + 50 tempo dimensions
	Fusion	Concatenation followed by GELU and RMSNorm
Encoder	Number of stages	3
	Hidden dimension	200
	FFN dimension	800
	Attention heads	4 spatial and 4 temporal heads
	Normalization	Pre-norm RMSNorm
	Activation	GELU
	Dropout	0.1
Fusion	Fusion method	Feature-wise learned sigmoid gate
	Gate initialization	Equal 0.5 : 0.5
Masking	Mask ratio	0.55
	Topography block size	$35^\circ - 75^\circ$
	Temporal block size	2-10 contiguous patches
	Encoder input	Dense token grid with masked positions zeroed and excluded from attention
Decoder	Number of layers	2
	Hidden dimension	100
	FFN dimension	400
	Attention heads	4
	Mask token	Learnable
Objective	Reconstruction target	Masked raw EEG waveform patches
	Loss function	Smooth L1 loss
	Smooth L1 beta	0.1

Table 7: Common hyperparameters used for downstream fine-tuning of DualBrain.

HYPER PARAMETER	VALUE
Initialization	Epoch-40 pretrained checkpoint
Optimizer	AdamW
Adam coefficients	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Adam epsilon	1×10^{-8}
Learning-rate scheduler	Per-update cosine annealing
Minimum learning rate	1×10^{-6}
Gradient clipping	Global norm 1.0
Classification selector	Validation balanced accuracy

C Downstream Tasks Details

C.1 Downstream Tasks

To ensure reproducibility, we report all hyperparameters used for downstream fine-tuning. The common optimization settings are summarized in Table 7, and the dataset-specific settings are provided in Table 9. None of the compared methods reported results on TUSZ. We therefore reproduced all TUSZ baselines using the publicly released code and pretrained weights. We also reproduced REVE on SEED-V, CHB-MIT, and Siena. CodeBrain was additionally reproduced on Siena, PhysioNet-MI, and HMC. All reproduced results are marked with a dagger (†) in the corresponding tables; unmarked baseline results are reported in prior studies.

Table 8: Baseline models complexity comparison.

METHOD	PARAMS (M)	FLOPS (G)
BIOT	3.2	0.483
LaBraM-Base	5.8	0.483
LaBraM-Large	46.0	3.060
LaBraM-Huge	369.0	22.800
CBraMod	4.0	0.319
CSBrain	8.8	8.450
REVE-Small	12.0	19.100
REVE-Base	69.0	104.440
REVE-Large	408.0	531.020
CodeBrain	15.2	8.740
DualBrain (Ours)	6.33	8.729

EEGNet (Lawhern et al., 2018) addresses the need for an EEG classifier that works across different BCI paradigms without requiring a separate architecture for each task. It uses depthwise and separable convolutions to build a compact network around established EEG feature-extraction principles.

ContraWR (Yang et al., 2023b) learns representations from unlabeled EEG for automatic sleep staging. Its contrastive objective uses global statistics of the dataset as a "world" reference to distinguish EEG signals associated with different sleep stages, particularly when labels are scarce or recordings are noisy.

ST-Transformer (Song et al., 2021) addresses the limited ability of convolutional models to capture global EEG dependencies. It spatially filters the signal, applies attention across the feature-channel dimension, and then attends over segments along the time dimension for EEG classification.

BIOT (Yang et al., 2023a) targets cross-dataset learning when biosignals differ in their channels, recording lengths, or missing values. It divides each channel into fixed-length segments and forms

unified "biosignal sentences," adding channel and relative-position embeddings to retain spatial and temporal information.

LaBraM (Jiang et al., 2024) enables pretraining across EEG datasets with different electrode counts, recording lengths, and tasks by dividing signals into channel patches. It trains a vector-quantized tokenizer through neural spectrum prediction, then pretrains a Transformer to predict the neural codes of masked patches.

CBraMod (Wang et al., 2025) treats spatial and temporal dependencies as distinct structures rather than modeling all EEG patches together. Its criss-cross Transformer uses parallel spatial and temporal attention, while asymmetric conditional positional encoding accommodates different EEG formats; pretraining reconstructs masked EEG patches.

CSBrain (Zhou et al., 2025) addresses the variation in EEG patterns across temporal and spatial scales. Cross-scale spatiotemporal tokenization aggregates features from local time windows and anatomical regions, and structured sparse attention models dependencies between those windows and regions.

REVE (El Ouahidi et al., 2025) is designed to transfer across EEG recordings collected with different electrode arrangements and durations. Its four-dimensional positional encoding permits arbitrary electrode layouts and signal lengths, while masked autoencoding supports pretraining across datasets with different recording setups.

CodeBrain (Ma et al., 2026) addresses EEG representations that insufficiently distinguish local neural events from long-range dependencies. Its temporal-frequency dual tokenizer maps the two signal domains to discrete codes, and its multi-scale architecture combines structured global convolution with sliding-window attention.

C.2 Downstream Hyperparameters Settings

We use a common fine-tuning configuration of 50 epochs, a batch size of 64, a learning rate of 10^{-4}, a weight decay of 10^{-2}, and a dropout rate of 0.1. Dataset-specific adjustments are made to accommodate differences in input structure and training behavior. Table 9 summarizes the settings used for each dataset.

Table 9: Hyperparameters of downstream tasks for fine-tuning .

Dataset	Epochs	Batch × Accum.	LR	WD	Dropout
TUAB	20	64×8	5e−4	5e−2	0.3
TUEV	50	64×1	1e−4	1e−2	0.3
TUSZ	50	64×1	2e−4	5e−3	0.3
CHB-MIT	20	64×1	1e−4	1e−2	0.1
PhysioNet-MI	50	64×1	1e−4	2e−2	0.1
FACED	50	64×1	1e−4	1e−2	0.1
SEED-V	50	64×1	1e−4	1e−2	0.1
Mental Arithmetic	50	64×1	1e−4	2e−2	0.1
ISRUC	50	2×32	1e−4	1e−2	0.1
HMC	50	64×1	5e−5	5e−3	0.3
Siena	50	64×1	2.5e−5	5e−3	0.1

C.3 Abnormal Detection

TUAB v3.0.1 is a clinical EEG recordings collected at Temple University Hospital from patients referred for neurological evaluation (Obeid & Picone, 2016). Recordings were acquired using the international 10-20 system, and labeled as normal or abnormal based on the neurologist's report. The dataset contains 2,993 sessions from 2,329 subjects: 2,717 sessions from 2,076 subjects for training and 276 sessions from 253 subjects for testing, with no subject overlap between the partitions. Each session is represented by a clinically relevant recording longer than 15 min. We derive 16 longitudinal bipolar channels, resample the signals to 200 Hz, apply 0.3–75 Hz band-pass and 60-Hz notch filtering, and divide them into non-overlapping 10-s windows. The official training subjects are further divided into training and validation sets at an 80:20 ratio, yielding 297,103/ 75,407/ 36,945 training/validation/test windows.

Table 10: TUAB (binary classification).

Method	B-Acc	AUPRC	AUROC
EEGNet	0.7642 ± 0.0036	0.8299 ± 0.0043	0.8412 ± 0.0031
ContraWR	0.7746 ± 0.0041	0.8421 ± 0.0104	0.8456 ± 0.0074
ST-Transformer	0.7966 ± 0.0023	0.8521 ± 0.0026	0.8707 ± 0.0019
BIOT	0.7959 ± 0.0057	0.8792 ± 0.0023	0.8815 ± 0.0043
LaBraM	0.8140 ± 0.0019	0.8965 ± 0.0016	0.9022 ± 0.0009
CBraMod	0.8289 ± 0.0022	0.9258 ± 0.0008	0.9227 ± 0.0011
CSBrain	0.8172 ± 0.0043	0.9005 ± 0.0066	0.8957 ± 0.0046
REVE	$\mathbf{0.8315} \pm 0.0014$	$\mathbf{0.9281} \pm 0.0009$	$\mathbf{0.9245} \pm 0.0013$
CodeBrain	0.8294 ± 0.0013	0.9100 ± 0.0006	0.9030 ± 0.0009
Ours	0.8100 ± 0.0073	0.8977 ± 0.0084	0.8905 ± 0.0094

Our treatment of bipolar inputs may contribute to the TUAB performance gap. Channel-ID embeddings and ACPE do not require explicit electrode coordinates, although they can still be affected by montage changes. Coordinate-based encodings, including SHPE and REVE's positional encoding, additionally require a spatial interpretation of each input channel. We represent each bipolar derivation by its electrode midpoint, which does not explicitly encode the pair's orientation or separation. Fine-tuning may partly compensate by learning from the bipolar signals, but we have not established how effectively it does so.

The comparison with REVE also involves differences in input construction. Its public TUAB configuration specifies 21 electrode channels, whereas we use 16 bipolar derivations. These differences may contribute to the performance gap, but our current experiments do not separate their effects from those of the model. We plan controlled comparisons of montage and preprocessing, followed by an evaluation of explicit electrode-pair representations for bipolar inputs.

C.4 Event Classification

Table 11: TUEV (six-class classification).

Method	B-Acc	Kappa	F1-w
EEGNet	0.3876 ± 0.0143	0.3577 ± 0.0155	0.6539 ± 0.0120
ContraWR	0.4384 ± 0.0349	0.3912 ± 0.0237	0.6893 ± 0.0136
ST-Transformer	0.3984 ± 0.0228	0.3765 ± 0.0306	0.6823 ± 0.0190
BIOT	0.5281 ± 0.0225	0.5273 ± 0.0249	0.7492 ± 0.0082
LaBraM	0.6409 ± 0.0065	0.6637 ± 0.0093	0.8312 ± 0.0052
CBraMod	0.6671 ± 0.0107	0.6772 ± 0.0096	0.8342 ± 0.0064
CSBrain	0.6903 ± 0.0059	0.6833 ± 0.0047	0.8333 ± 0.0057
REVE	0.6759 ± 0.0229	0.6783 ± 0.0199	$\mathbf{0.8451} \pm 0.0129$
CodeBrain	0.6428 ± 0.0062	0.6912 ± 0.0101	0.8362 ± 0.0048
Ours	$\mathbf{0.6978 \pm 0.0104}$	$\mathbf{0.6979 \pm 0.0100}$	0.8430 ± 0.0045

TUEV. The Temple University Hospital EEG Events Corpus (TUEV) is derived from routine clinical EEG recordings collected at Temple University Hospital and contains manually annotated EEG events from 370 patients (Obeid & Picone, 2016). TUEV is a six-class event-classification task. The classes are: spike and sharp wave (SPSW), a transient epileptiform waveform; generalized periodic epileptiform discharge (GPED), a repetitive discharge distributed over both hemispheres; periodic lateralized epileptiform discharge (PLED), a repetitive discharge predominantly localized to one hemisphere; eye movement (EYEM); non-physiological or physiological artifact (ARTF); and ordinary background activity (BCKG). The recordings are preprocessed to 16 bipolar derivations and resampled from their nominal 256-Hz acquisition rate to 200 Hz. Each annotated event is represented by a 5-s window, resulting in 113,353 samples. Subjects in the official training partition are sorted and divided into 80% training and 20% validation subsets, while the official evaluation partition is used as the test set.

On TUEV, DualBrain achieves the highest balanced accuracy of 0.6978, exceeding CSBrain 0.6903 by 0.75 points, while obtaining a competitive weighted F1 of 0.8430, only 0.21 percentage points below REVE 0.8451. These results support the complementary use of both spatiotemporal processing orders. TUEV contains brief and spatially localized EEG events. Their classification requires both temporal waveform analysis and cross-channel spatial comparison. The strong performance suggests that DualBrain captures these complementary characteristics effectively.

C.5 SEIZURE DETECTION

Table 12: TUSZ (binary seizure detection).

Method	B-Acc	AUPRC	AUROC
EEGNet†	0.5366 ± 0.0062	0.9715 ± 0.0016	0.9104 ± 0.0086
ContraWR†	0.8466 ± 0.0149	0.9732 ± 0.0052	0.9171 ± 0.0134
ST-Transformer†	0.8533 ± 0.0268	0.9484 ± 0.0138	0.8533 ± 0.0286
BIOT†	0.8743 ± 0.0076	0.9831 ± 0.0018	0.9437 ± 0.0056
LaBraM†	0.9246 ± 0.0151	$\mathbf{0.9962 \pm 0.0015}$	$\mathbf{0.9862 \pm 0.0060}$
CBraMod†	0.8808 ± 0.0124	0.9854 ± 0.0020	0.9498 ± 0.0477
CSBrain†	0.8734 ± 0.0872	0.9843 ± 0.0120	0.9418 ± 0.0475
REVE†	0.9197 ± 0.0296	0.9951 ± 0.0042	0.9826 ± 0.0137
CodeBrain†	0.8934 ± 0.0204	0.9873 ± 0.0032	0.9550 ± 0.0108
DualBrain	$\mathbf{0.9279 \pm 0.0187}$	0.9960 ± 0.0013	0.9852 ± 0.0047

† Results reproduced using the publicly released code and pretrained weights.

TUSZ. TUSZ v2.0.6 contains clinical EEG recordings from 675 patients, with expert annotations of seizure timing, affected channels, and seizure type. Following prior seizure-detection protocols, we merge all seizure types into a single seizure class and distinguish them from background activity. We retain 19 canonical 10-20 electrodes, resample each recording to 200 Hz, and extract 50,554 non-overlapping 10-s windows. Recordings with fewer than 16 canonical electrodes are excluded, and missing electrodes are handled using a channel-validity mask. Because none of the baseline models originally reported results on TUSZ under our binary seizure-detection protocol, we reproduced all available baselines using their official implementations, released pretrained checkpoints, and reported downstream settings. All models were evaluated using the same preprocessing, subject-disjoint split, and evaluation procedure.

DualBrain achieves 0.9279 B-Acc, 0.9960 AUPRC, and 0.9852 AUROC. It ranks first in B-Acc and second in AUPRC and AUROC. Compared with LaBraM, the strongest reproduced baseline, DualBrain improves B-Acc by 0.33 percentage points. Its AUPRC and AUROC are slightly lower, by 0.02 and 0.10 points, respectively. DualBrain also exceeds CBraMod, CSBrain, and CodeBrain in B-Acc by 4.71, 5.45, and 3.45 points. These results suggest that combining the two spatiotemporal processing orders is effective for seizure detection.

Table 13: CHB-MIT (binary classification).

Method	B-Acc	AUPRC	AUROC
EEGNet	0.5658 ± 0.0106	0.1914 ± 0.0182	0.8048 ± 0.0136
ContraWR	0.6344 ± 0.0002	0.2264 ± 0.0174	0.8097 ± 0.0114
ST-Transformer	0.5915 ± 0.0195	0.1422 ± 0.0094	0.8237 ± 0.0491
BIOT	0.7068 ± 0.0457	0.3277 ± 0.0460	0.8761 ± 0.0284
LaBraM	0.7075 ± 0.0358	0.3287 ± 0.0402	0.8679 ± 0.0199
CBraMod	0.7398 ± 0.0284	0.3689 ± 0.0382	0.8892 ± 0.0154
CSBrain	0.7262 ± 0.0115	0.5164 ± 0.0449	0.8915 ± 0.0321
REVE†	0.7450 ± 0.0395	$\mathbf{0.5301} \pm 0.0601$	$\mathbf{0.9301} \pm 0.0176$
CodeBrain	0.7273 ± 0.0240	0.4377 ± 0.0288	0.8961 ± 0.0174
Ours	$\mathbf{0.8261} \pm 0.0348$	0.4213 ± 0.0172	0.9180 ± 0.0064

CHB-MIT. The CHB-MIT Scalp EEG Database contains long-term recordings from 22 pediatric patients with medically intractable epilepsy monitored at Boston Children's Hospital, including one patient recorded twice 1.5 years apart (Guttag, 2010). The recordings were acquired at 256 Hz using the international 10-20 system. We follow the established benchmark protocol using 16 bipolar derivations and 10-s seizure or non-seizure windows. This produces 326,993 windows from 21 retained subject identifiers, resampled to 200 Hz.

DualBrain achieves 0.8261 B-Acc, 0.4213 AUPRC, and 0.9180 AUROC over 5 seeds. It achieves the highest B-Acc, exceeding REVE, CBraMod, and CSBrain by 8.11, 8.63, and 9.99 percentage points, respectively. It also obtains the second-highest AUROC, 1.20 points below REVE, although its AUPRC remains lower than those of REVE, CSBrain, and CodeBrain. The large B-Acc gain is particularly meaningful because only 1,962 of the 326,993 windows contain seizure activity. Seizures often appear as abrupt, spatially localized discharges that propagate across neighboring electrodes over short time intervals. MJDE explicitly models both the temporal evolution of spatial patterns and the spatial distribution of temporal features. This complementary sequential modeling may better capture seizure propagation than encoders that model spatial and temporal dependencies jointly or in parallel.

Table 14: Siena (binary seizure detection).

Method	B-Acc	AUPRC	AUROC
EEGNet	0.7487 ± 0.0521	0.3753 ± 0.0867	0.8687 ± 0.0527
ContraWR	0.6546 ± 0.0311	0.3711 ± 0.0405	0.7819 ± 0.0596
ST-Transformer	0.7527 ± 0.0381	0.3636 ± 0.0252	0.8884 ± 0.0091
BIOT	0.7352 ± 0.0669	0.3809 ± 0.0892	0.9029 ± 0.0304
LaBraM	0.7082 ± 0.0329	0.3122 ± 0.0976	0.8814 ± 0.0328
CBraMod	0.7317 ± 0.0647	0.4107 ± 0.0720	0.9038 ± 0.0218
CSBrain	0.7662 ± 0.0471	$\mathbf{0.4871} \pm 0.0343$	0.9076 ± 0.0119
REVE†	0.7878 ± 0.0375	0.4158 ± 0.1022	0.9017 ± 0.0368
CodeBrain†	0.6122 ± 0.0400	0.3543 ± 0.0229	0.8207 ± 0.0189
Ours	$\mathbf{0.8410} \pm 0.0114$	0.4557 ± 0.0637	$\mathbf{0.9354} \pm 0.0025$

Siena. The Siena Scalp EEG Database contains long-term video-EEG recordings from 14 adult epilepsy patients collected at the Unit of Neurology and Neurophysiology of the University of Siena, Italy (Detti, 2020). The cohort includes 9 men and 5 women aged 20-71 years who were monitored while awake or asleep, with seizures identified by an expert. EEG was acquired at 512 Hz using the international 10-20 system. We retain 29 common EEG channels, resample the signals to 200 Hz, and construct 51,307 10-s seizure and non-seizure windows. Subjects PN01-PN12 are used for training and validation with a class-stratified 80:20 split within each subject, while PN13–PN14 are held out for testing.

Although DualBrain achieves strong B-Acc (0.8410) and AUROC (0.9354), its mean AUPRC is lower (0.4557) and exhibits relatively high variability across seeds. AUPRC remains between 0.4689 and 0.4951 for 4 seeds, but drops markedly to 0.3297 for one seed. This single unstable run accounts for most of the observed standard deviation (0.0637), indicating seed-specific sensitivity under the severe class imbalance rather than consistently weak precision-recall performance.

C.6 Emotion Recognition

SEED-V. SEED-V is an affective EEG dataset collected from 16 healthy participants while they watched emotion-eliciting film clips. Each participant completed 3 sessions, each comprising 15 trials from 5 emotional states: disgust, fear, sadness, neutrality, and happiness. EEG was recorded from 62 scalp electrodes at 1,000 Hz using a NeuroScan system (Liu et al., 2021). We retain the 62 EEG channels, resample them to 200 Hz, apply a 0.3-75-Hz band-pass filter, and divide each film-viewing period into non-overlapping 1-s windows, yielding 117,744 samples. Within each session, trials 1-5, 6-10, and 11-15 are assigned to training, validation, and testing, respectively. The same participants therefore appear in all partitions, while the emotion-induction trials remain disjoint.

DualBrain achieves a B-Acc of 0.4199, weighted F1 of 0.4257, and Cohen's kappa of 0.2781, closely matching CSBrain across all 3 metrics with similarly low variation over 5 seeds. This similarity may

Table 15: SEED-V (five-class classification).

Method	B-Acc	Kappa	F1-w
EEGNet	0.2961 ± 0.0102	0.1006 ± 0.0143	0.2749 ± 0.0098
ContraWR	0.3546 ± 0.0105	0.1905 ± 0.0188	0.3544 ± 0.0121
ST-Transformer	0.3052 ± 0.0072	0.1083 ± 0.0121	0.2833 ± 0.0105
BIOT	0.3837 ± 0.0187	0.2261 ± 0.0262	0.3856 ± 0.0203
LaBraM	0.3976 ± 0.0138	0.2386 ± 0.0209	0.3974 ± 0.0111
CBraMod	0.4091 ± 0.0097	0.2569 ± 0.0143	0.4101 ± 0.0108
CSBrain	0.4197 ± 0.0033	$\mathbf{0.2785 \pm 0.0034}$	$\mathbf{0.4280 \pm 0.0023}$
REVE†	0.3871 ± 0.0149	0.2331 ± 0.0173	0.3914 ± 0.0139
CodeBrain	0.4137 ± 0.0023	0.2735 ± 0.0032	0.4235 ± 0.0022
Ours	$\mathbf{0.4199} \pm 0.0034$	0.2781 ± 0.0045	0.4257 ± 0.0040

† Results reproduced using the publicly released code and pretrained weights.

reflect their shared temporal-to-topographic processing, which extracts temporal patterns within individual electrodes before comparing them across the scalp. SEED-V uses a dense 62-channel montage, where volume conduction can produce correlated and redundant signals among neighboring electrodes. Without precise electrode positions, the model may be unable to distinguish meaningful interactions between brain regions from redundant correlations caused by the close physical proximity of neighboring electrodes. SHPE provides fine-grained electrode locations through multiscale spherical harmonic bases, helping interpret channel features within their spatial context. These results suggest that temporal-to-topographic processing combined with geometry-aware positional encoding is effective for classifying short 1-s affective EEG segments.

Table 16: FACED (nine-class classification).

Method	B-Acc	Kappa	F1-w
EEGNet	0.4090 ± 0.0122	0.3342 ± 0.0251	0.4124 ± 0.0141
ContraWR	0.4887 ± 0.0078	0.4231 ± 0.0151	0.4884 ± 0.0074
ST-Transformer	0.4810 ± 0.0079	0.4137 ± 0.0133	0.4795 ± 0.0096
BIOT	0.5118 ± 0.0118	0.4476 ± 0.0254	0.5136 ± 0.0112
LaBraM	0.5273 ± 0.0107	0.4698 ± 0.0102	0.5288 ± 0.0188
CBraMod	0.5509 ± 0.0089	0.5041 ± 0.0122	0.5618 ± 0.0093
CSBrain	0.5752 ± 0.0042	0.5204 ± 0.0036	0.5796 ± 0.0031
REVE	0.5646 ± 0.0164	0.5080 ± 0.0191	0.5659 ± 0.0172
CodeBrain	0.5941 ± 0.0098	0.5406 ± 0.0084	0.5953 ± 0.0113
Ours	$\mathbf{0.6082 \pm 0.0052}$	$\mathbf{0.5578 \pm 0.0061}$	$\mathbf{0.6124 \pm 0.0058}$

FACED. The Finer-grained Affective Computing EEG Dataset (FACED) was collected from 123 healthy participants while they watched 28 emotion-eliciting video clips spanning 9 categories: amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness, and neutrality. Each trial includes fixation, video presentation, subjective ratings, and rest (Chen et al., 2023). EEG was recorded using 32 electrodes, with 92 participants sampled at 1,000 Hz and 31 at 250 Hz; the released processed signals are standardized to 250 Hz. We retain 30 scalp channels after excluding the 2 mastoid references, resample the signals to 200 Hz, and divide each trial into 3 non-overlapping 10-s windows, yielding 10,332 samples. The 123 participants are divided into 80 training, 20 validation, and 23 test participants for subject-disjoint evaluation.

DualBrain achieves a B-Acc of 0.6082, weighted F1 of 0.6124, and Cohen's kappa of 0.5578, outperforming all prior methods across the 3 metrics. FACED requires fine-grained discrimination among 9 emotional states, beyond the 3-class settings commonly examined in earlier studies. Emotion-related information is distributed nonuniformly across the scalp, and inter-electrode relationships provide important discriminative information (Zheng & Lu, 2015; Song et al., 2018). Thus, the key is not simply to use more channels, but to represent their locations and spatial relationships accurately. SHPE provides precise, fine-grained positional information for FACED's 30 scalp electrodes through multiscale spherical harmonic bases. The consistent gains across all 3 metrics suggest that

this geometry-aware representation supports the subtle spatial discrimination required for 9-class emotion recognition.

C.7 Mental Stress

Table 17: Mental Arithmetic (binary classification).

Method	B-Acc	AUPRC	AUROC
EEGNet	0.6770 ± 0.0116	0.5763 ± 0.0102	0.7321 ± 0.0108
ContraWR	0.6631 ± 0.0097	0.5787 ± 0.0164	0.7332 ± 0.0082
ST-Transformer	0.6631 ± 0.0173	0.5672 ± 0.0259	0.7132 ± 0.0174
BIOT	0.6875 ± 0.0186	0.6004 ± 0.0195	0.7536 ± 0.0144
LaBraM	0.6909 ± 0.0125	0.5999 ± 0.0155	0.7721 ± 0.0093
CBraMod	0.7256 ± 0.0132	0.6267 ± 0.0099	0.7905 ± 0.0073
CSBrain	0.7558 ± 0.0106	0.6696 ± 0.0221	0.8478 ± 0.0297
REVE	0.7660 ± 0.0355	0.7470 ± 0.0807	0.8450 ± 0.0514
CodeBrain	0.7514 ± 0.0203	0.7177 ± 0.0421	0.8707 ± 0.0209
Ours	$\mathbf{0.7938 \pm 0.0393}$	$\mathbf{0.8130 \pm 0.0365}$	$\mathbf{0.8769 \pm 0.0152}$

Mental Arithmetic. The EEG During Mental Arithmetic Tasks dataset contains recordings from 36 healthy university students collected at Taras Shevchenko National University of Kyiv. Each participant underwent a 3-min eyes-closed resting condition and a 1-min mental-arithmetic condition involving repeated subtraction of a 2-digit number from a 4-digit number. EEG was recorded at 500 Hz using a Neurocom 23-channel system with Ag/AgCl electrodes arranged according to the international 10–20 system and referenced to interconnected ear electrodes (Zyma et al., 2019). We use 19 scalp channels with the stored ear-reference derivation, resample the recordings to 200 Hz, and divide them into non-overlapping 5-s windows, yielding 1,707 resting-state and mental-arithmetic samples. We use subject-disjoint partitions comprising 28 training, 4 validation, and 4 test participants.

DualBrain achieves 0.7938 B-Acc, 0.8130 AUPRC, and 0.8769 AUROC, outperforming all baselines across the 3 metrics. Compared with the strongest baseline for each metric, it improves B-Acc over REVE by 2.78 percentage points, AUPRC over REVE by 6.60 points, and AUROC over CodeBrain by 0.62 points. Mental arithmetic imposes sustained working-memory demands that appear as frequency-specific and spatially distributed EEG responses. Frontal theta activity increases with working-memory load, while alpha and theta oscillations are established markers of cognitive and memory processing (Jensen & Tesche, 2002; Klimesch, 1999). Multi-digit subtraction also engages distributed frontoparietal activity. DualBrain's frequency branch captures these oscillatory features within individual channels, and the Temp→Topo pathway compares their distribution across electrodes. SHPE further represents the spatial organization of these responses using precise electrode locations. The strong subject-disjoint performance suggests that combining spectral features with temporal-to-topographic integration effectively captures EEG patterns associated with mental-arithmetic workload.

C.8 Motor Imagery

PhysioNet-MI. The EEG Motor Movement/Imagery Dataset contains recordings from 109 participants collected at the Wadsworth Center, New York State Department of Health. Each participant completed 14 runs comprising 2 baseline recordings and 12 motor execution or imagery recordings. EEG was acquired at 160 Hz from 64 electrodes arranged according to the international 10-10 system using BCI2000 (Schalk, 2009). We use the 6 motor-imagery runs to classify imagined movements of the left fist, right fist, both fists, and both feet. Participants performed these tasks in response to visual cues. We retain all 64 channels, apply common-average referencing, a 0.3-Hz high-pass filter, and a 60-Hz notch filter, and resample the signals to 200 Hz. We extract 4-s task epochs and exclude rest periods, yielding 9,837 samples. Subjects S001-S070, S071-S089, and S090-S109 form the training, validation, and test sets, respectively, ensuring subject-disjoint evaluation.

Table 18: PhysioNet-MI (multi-class classification).

Method	B-Acc	Kappa	F1-w
EEGNet	0.5814 ± 0.0125	0.4468 ± 0.0199	0.5796 ± 0.0115
ContraWR	0.5892 ± 0.0133	0.4527 ± 0.0248	0.5918 ± 0.0116
ST-Transformer	0.6035 ± 0.0081	0.4712 ± 0.0199	0.6053 ± 0.0075
BIOT	0.6153 ± 0.0154	0.4875 ± 0.0272	0.6158 ± 0.0197
LaBraM	0.6173 ± 0.0122	0.4912 ± 0.0192	0.6177 ± 0.0141
CBraMod	0.6417 ± 0.0091	0.5222 ± 0.0169	0.6427 ± 0.0100
CSBrain	0.6304 ± 0.0090	0.5071 ± 0.0120	0.6308 ± 0.0095
REVE	0.6480 ± 0.0140	0.5306 ± 0.0187	0.6484 ± 0.0170
CodeBrain†	0.6051 ± 0.0161	0.4734 ± 0.0215	0.6062 ± 0.0157
Ours	$\mathbf{0.6503 \pm 0.0034}$	$\mathbf{0.5337 \pm 0.0046}$	$\mathbf{0.6510 \pm 0.0036}$

† Results reproduced using the publicly released code and pretrained weights.

DualBrain achieves 0.6503 B-Acc, 0.5337 Cohen's kappa, and 0.6510 weighted F1, outperforming the strongest baseline, REVE, by 0.23, 0.31, and 0.26 percentage points, respectively. PhysioNet-MI uses a dense 64-channel montage and requires discrimination among spatially distinct motor-imagery patterns. The small but consistent gains across all 3 metrics suggest that SHPE preserves fine-grained electrode geometry, while MJDE integrates localized sensorimotor activity with its temporal evolution. This combination supports robust cross-subject decoding without relying on fixed channel identities alone.

C.9 Sleep Staging

Table 19: ISRUC (five-class sleep staging; subgroup 1, subject-independent split).

Method	B-Acc	Kappa	F1-w
EEGNet	0.7154 ± 0.0121	0.7040 ± 0.0173	0.7513 ± 0.0124
ContraWR	0.7402 ± 0.0126	0.7178 ± 0.0156	0.7610 ± 0.0137
ST-Transformer	0.7381 ± 0.0205	0.7013 ± 0.0352	0.7681 ± 0.0175
BIOT	0.7527 ± 0.0121	0.7291 ± 0.0230	0.7790 ± 0.0146
LaBraM	0.7633 ± 0.0102	0.7231 ± 0.0182	0.7810 ± 0.0133
CBraMod	0.7865 ± 0.0110	0.7442 ± 0.0152	0.8011 ± 0.0099
CSBrain	$\mathbf{0.7925 \pm 0.0030}$	0.7406 ± 0.0102	0.7990 ± 0.0091
REVE	0.7819 ± 0.0078	$\mathbf{0.7500 \pm 0.0156}$	0.8005 ± 0.0135
CodeBrain	0.7835 ± 0.0033	0.7476 ± 0.0040	0.8020 ± 0.0018
Ours	0.7883 ± 0.0033	0.7480 ± 0.0093	$\mathbf{0.8068 \pm 0.0067}$

ISRUC. We use Subgroup I of the ISRUC-Sleep dataset, which contains 1 whole-night polysomnographic recording from each of 100 subjects diagnosed with sleep-related disorders at the Sleep Medicine Centre of Coimbra University Hospital, Portugal. Sleep stages were independently annotated by 2 experts (Khalighi et al., 2016). We use 6 EEG derivations sampled at 200 Hz and divide each recording into 30-s epochs labeled as wake, N1, N2, N3, or REM. N1 represents transitional sleep, N2 is characterized by sleep spindles and K-complexes, and N3 corresponds to slow-wave sleep. REM exhibits wake-like EEG accompanied by rapid eye movements and low muscle tone. The processed dataset contains 89,240 epochs grouped into 4,462 sequences of 20 consecutive epochs. Subjects 1-80, 81-90, and 91-100 are assigned to training, validation, and testing, respectively. We use a micro-batch of 2 sequences and accumulate gradients over 32 steps, yielding an effective batch size of 64 sequences.

DualBrain achieves 0.7883 B-Acc, 0.7480 Cohen's kappa, and 0.8068 weighted F1. It obtains the highest weighted F1, exceeding CodeBrain by 0.48 percentage points, while remaining 0.42 B-Acc points below CSBrain and 0.20 kappa points below REVE. Sleep staging depends strongly on temporal and spectral patterns, including spindles, K-complexes, slow waves, and transitions between consecutive stages. DualBrain's time-frequency tokenizer captures characteristic waveforms and oscillations within each epoch, while the shared sequence-level Transformer models transitions across

consecutive epochs. The sparse 6-channel montage provides less spatial information than dense EEG datasets, which may limit the additional benefit of geometry-aware encoding. Nevertheless, the competitive performance across all 3 metrics indicates that DualBrain transfers effectively to a task dominated by temporal, spectral, and inter-epoch structure.

Table 20: HMC (five-class sleep staging).

Method	B-Acc	Kappa	F1-w
EEGNet	0.6534 ± 0.0122	0.5886 ± 0.0201	0.6536 ± 0.0168
ContraWR	0.4242 ± 0.0541	0.2340 ± 0.0554	0.2987 ± 0.0288
ST-Transformer	0.2559 ± 0.0141	0.0503 ± 0.0183	0.1428 ± 0.0122
BIOT	0.6862 ± 0.0041	0.6295 ± 0.0113	0.7091 ± 0.0147
LaBraM	0.7277 ± 0.0101	0.6813 ± 0.0053	0.7454 ± 0.0027
CBraMod	0.7269 ± 0.0041	0.6685 ± 0.0104	0.7395 ± 0.0089
CSBrain	0.7345 ± 0.0047	0.6818 ± 0.0046	0.7506 ± 0.0042
REVE	0.7401 ± 0.0075	$\mathbf{0.6982 \pm 0.0078}$	$\mathbf{0.7638 \pm 0.0074}$
CodeBrain†	0.7210 ± 0.0110	0.6645 ± 0.0033	0.7388 ± 0.0030
Ours	$\mathbf{0.7414 \pm 0.0042}$	0.6920 ± 0.0052	0.7604 ± 0.0041

† Results reproduced using the publicly released code and pretrained weights.

HMC. The Haaglanden Medisch Centrum (HMC) database contains 151 whole-night polysomnographic recordings acquired in 2018 at the HMC sleep center in the Netherlands. The cohort comprises 85 male and 66 female patients with a mean age of 53.9 ± 15.4 years. These patients were referred for polysomnography because of suspected sleep disorders. The recordings include EEG, EOG, chin EMG, ECG, and technician-scored sleep-stage annotations (Alvarez-Estevez & Rijsman, 2022). We use 4 EEG derivations, F4-M1, C4-M1, O2-M1, and C3-M2. The original 256-Hz signals are band-pass filtered between 0.1 and 75 Hz, notch-filtered at 50 Hz, and resampled to 200 Hz. The resulting 137,243 valid 30-s epochs are classified as wake, N1, N2, N3, or REM. After sorting by subject identifier, 100, 25, and 26 subjects are assigned to each training, validation, and testing.

DualBrain achieves a B-Acc of 0.7414, weighted F1 of 0.7604, and Cohen's kappa of 0.6920. It obtains the highest B-Acc, exceeding REVE by 0.13 percentage points, while remaining close to the strongest weighted-F1 result of 0.7638. HMC provides only 4 EEG channels, limiting the spatial information available to geometry-aware models. Sleep-stage discrimination therefore depends primarily on temporal and spectral patterns within each 30-s epoch, including spindles, K-complexes, slow waves, and wake-like REM activity. DualBrain's time–frequency tokenizer and temporal modeling capture these stage-specific signatures, while the complementary spatial pathway integrates the limited cross-channel information. Its competitive performance on a heterogeneous clinical cohort suggests robust transfer to sleep recordings with sparse electrode coverage.

D Additional Analysis and Ablations

Table 21: Encoder comparison across 11 downstream datasets. Results are mean ± SD over 5 seeds. Bold and underline indicate the highest and second-highest distinct means, respectively, at the reported precision. The final row averages B-Acc across datasets.

Dataset	Metric	LaBraM	CBraMod	CSBrain	DualBrain
TUAB	B-Acc	$\underline{0.8033 \pm 0.0075}$	0.8025 ± 0.0025	0.7979 ± 0.0096	$\mathbf{0.8100 \pm 0.0073}$
	AUPRC	$\underline{0.8915 \pm 0.0087}$	0.8887 ± 0.0097	0.8834 ± 0.0087	$\mathbf{0.8977 \pm 0.0084}$
	AUROC	$\underline{0.8830 \pm 0.0113}$	0.8804 ± 0.0101	0.8754 ± 0.0099	$\mathbf{0.8905 \pm 0.0094}$
TUEV	B-Acc	$\underline{0.6881 \pm 0.0153}$	0.6658 ± 0.0163	0.6417 ± 0.0086	$\mathbf{0.6978 \pm 0.0104}$
	Kappa	$\underline{0.6920 \pm 0.0246}$	0.6832 ± 0.0210	0.6710 ± 0.0172	$\mathbf{0.6979 \pm 0.0100}$
	F1-W	$\underline{0.8401 \pm 0.0116}$	0.8326 ± 0.0091	0.8231 ± 0.0098	$\mathbf{0.8430 \pm 0.0045}$
TUSZ	B-Acc	$\underline{0.9111 \pm 0.0113}$	0.8964 ± 0.0200	0.8839 ± 0.0061	$\mathbf{0.9279 \pm 0.0187}$
	AUPRC	$\underline{0.9930 \pm 0.0031}$	0.9918 ± 0.0012	0.9841 ± 0.0011	$\mathbf{0.9960 \pm 0.0013}$
	AUROC	$\underline{0.9750 \pm 0.0115}$	0.9703 ± 0.0052	0.9414 ± 0.0039	$\mathbf{0.9852 \pm 0.0047}$
CHB-MIT	B-Acc	$\mathbf{0.8326 \pm 0.0185}$	0.8101 ± 0.0419	0.7795 ± 0.0517	$\underline{0.8261 \pm 0.0348}$
	AUPRC	$\underline{0.3917 \pm 0.0218}$	0.3456 ± 0.0685	0.2952 ± 0.0726	$\mathbf{0.4213 \pm 0.0172}$
	AUROC	$\underline{0.9104 \pm 0.0103}$	0.9089 ± 0.0228	0.8999 ± 0.0333	$\mathbf{0.9180 \pm 0.0064}$
Siena	B-Acc	$\mathbf{0.8698 \pm 0.0217}$	0.8538 ± 0.0229	$\underline{0.8660 \pm 0.0110}$	0.8410 ± 0.0114
	AUPRC	0.4445 ± 0.0301	$\mathbf{0.4753 \pm 0.0250}$	0.4358 ± 0.0110	$\underline{0.4557 \pm 0.0637}$
	AUROC	$\mathbf{0.9501 \pm 0.0050}$	$\underline{0.9431 \pm 0.0014}$	0.9265 ± 0.0060	0.9354 ± 0.0025
SEED-V	B-Acc	$\underline{0.4177 \pm 0.0039}$	0.4050 ± 0.0052	0.4133 ± 0.0033	$\mathbf{0.4199 \pm 0.0034}$
	Kappa	$\underline{0.2745 \pm 0.0053}$	0.2574 ± 0.0083	0.2706 ± 0.0042	$\mathbf{0.2781 \pm 0.0045}$
	F1-W	$\underline{0.4234 \pm 0.0039}$	0.4096 ± 0.0071	0.4219 ± 0.0038	$\mathbf{0.4257 \pm 0.0040}$
FACED	B-Acc	$\underline{0.5952 \pm 0.0077}$	0.5675 ± 0.0062	0.5444 ± 0.0056	$\mathbf{0.6082 \pm 0.0052}$
	Kappa	$\underline{0.5430 \pm 0.0092}$	0.5123 ± 0.0069	0.4847 ± 0.0058	$\mathbf{0.5578 \pm 0.0061}$
	F1-W	$\underline{0.5983 \pm 0.0078}$	0.5698 ± 0.0061	0.5466 ± 0.0038	$\mathbf{0.6124 \pm 0.0058}$
M-AT	B-Acc	0.7542 ± 0.0363	$\underline{0.7667 \pm 0.0678}$	0.6958 ± 0.0723	$\mathbf{0.7938 \pm 0.0393}$
	AUPRC	$\underline{0.7823 \pm 0.0593}$	0.7515 ± 0.0364	0.6581 ± 0.0375	$\mathbf{0.8130 \pm 0.0365}$
	AUROC	$\underline{0.8690 \pm 0.0361}$	0.8474 ± 0.0311	0.8220 ± 0.0283	$\mathbf{0.8769 \pm 0.0152}$
P-MI	B-Acc	$\mathbf{0.6503 \pm 0.0079}$	0.6332 ± 0.0094	$\underline{0.6391 \pm 0.0021}$	$\mathbf{0.6503 \pm 0.0034}$
	Kappa	$\mathbf{0.5337 \pm 0.0106}$	0.5109 ± 0.0125	$\underline{0.5187 \pm 0.0028}$	$\mathbf{0.5337 \pm 0.0046}$
	F1-W	$\mathbf{0.6516 \pm 0.0080}$	0.6334 ± 0.0095	0.6392 ± 0.0021	$\underline{0.6510 \pm 0.0036}$
ISRUC	B-Acc	$\mathbf{0.7934 \pm 0.0106}$	0.7746 ± 0.0126	0.7662 ± 0.0022	$\underline{0.7883 \pm 0.0033}$
	Kappa	$\mathbf{0.7521 \pm 0.0220}$	0.7223 ± 0.0207	0.7153 ± 0.0086	$\underline{0.7480 \pm 0.0093}$
	F1-W	$\underline{0.8044 \pm 0.0202}$	0.7847 ± 0.0181	0.7789 ± 0.0053	$\mathbf{0.8068 \pm 0.0067}$
HMC	B-Acc	$\mathbf{0.7426 \pm 0.0014}$	0.7289 ± 0.0019	0.7134 ± 0.0042	$\underline{0.7414 \pm 0.0042}$
	Kappa	$\mathbf{0.6975 \pm 0.0026}$	0.6780 ± 0.0030	0.6539 ± 0.0081	$\underline{0.6920 \pm 0.0052}$
	F1-W	$\mathbf{0.7648 \pm 0.0025}$	0.7472 ± 0.0026	0.7271 ± 0.0058	$\underline{0.7604 \pm 0.0041}$
Average	B-Acc	$\underline{0.7326}$	0.7186	0.7037	$\mathbf{0.7368}$

D.1 Encoder Comparison

We compare MJDE with encoders adapted from LaBraM, CBraMod, and CSBrain to examine how different spatiotemporal modeling structures perform within a common framework. These alternatives represent global channel-time attention, parallel spatial and temporal attention, and sequential temporal and regional attention, respectively. We replace the encoder while retaining the time-frequency tokenizer, SHPE, masking strategy, reconstruction decoder, and pretraining objective. Each alternative is pretrained within this framework rather than initialized from its publicly released pretrained weights. This comparison therefore concerns adapted encoder architectures, not the original models as complete systems.

All encoders use an embedding dimension of 200 and a feed-forward hidden dimension of 800. The LaBraM-based encoder uses 12 global-attention blocks with 10 heads. It retains query-key normalization and LayerScale but omits the classification token and replaces the original tokenizer and positional embeddings with the shared modules. The CBraMod-based encoder uses 12 criss-cross blocks. Each block divides the representation into 100-dimensional spatial and temporal streams, each with 4 attention heads, and concatenates their outputs before the feed-forward sublayer. The CSBrain-based encoder uses 12 blocks with 8 heads per attention module. Inter-window temporal attention is followed by region-structured spatial attention. Its additional temporal and regional con-

volutional embedding modules are omitted. The alternatives retain their block-level LayerNorm and GELU operations. Attention and feature dropout are set to 0.1 during pretraining. Masked positions are excluded from attention keys, and their output representations are zeroed while retaining the dense channel-time grid.

The encoder-only parameter counts are 5,794,960 for LaBraM, 4,831,200 for CBraMod, 8,208,000 for CSBrain, and 6,027,200 for MJDE. These counts exclude the shared tokenizer, positional encoding, decoder, and downstream head. The comparison matches the nominal number of blocks, but not the exact parameter budget or the number of attention operations. Each encoder is pretrained for 40 epochs with a masking ratio of 0.55 and pretraining seed 42. The shared decoder has 2 layers, embedding dimension 100, 4 attention heads, and feed-forward dimension 400. Reported means and SDs therefore describe fine-tuning variability from a single pretrained checkpoint per encoder, rather than variability across 5 pretraining runs.

Table 21 shows that DualBrain achieves the highest mean B-Acc of 0.7368 across 11 datasets. LaBraM, CBraMod, and CSBrain obtain 0.7326, 0.7186, and 0.7037, respectively. DualBrain exceeds CBraMod and CSBrain in B-Acc on 10 datasets each. It leads LaBraM on 6 datasets, trails on 4, and ties on P-MI at the reported precision. DualBrain also leads all 3 metrics on TUAB, TUEV, TUSZ, SEED-V, FACED, and M-AT. On TUSZ, its gains over the strongest alternative are 1.68 percentage points in B-Acc, 0.30 points in AUPRC, and 1.02 points in AUROC. On FACED, the corresponding gains are 1.30 points in B-Acc, 1.48 points in kappa, and 1.41 points in weighted F1.

The remaining datasets reveal metric-dependent trade-offs. LaBraM achieves higher B-Acc on CHB-MIT, while DualBrain achieves higher AUPRC and AUROC. The two encoders obtain identical rounded B-Acc and kappa on P-MI. DualBrain leads weighted F1 on ISRUC, whereas LaBraM leads its B-Acc and kappa. LaBraM also leads all three metrics on HMC. Despite these trade-offs, DualBrain achieves the highest mean B-Acc across the evaluated datasets, supporting its broad applicability across EEG tasks without implying superiority on every dataset or metric. Its explicit pathways also enable targeted ablations of processing order and analysis of stage-wise fusion.

Table 22: Positional encoding comparison across 11 downstream datasets. Results are mean ± SD over 5 seeds. Bold and underline indicate the highest and second-highest means, respectively. The final row averages B-Acc across datasets.

Dataset	Metric	Ch-ID	ACPE	4D PE	SHPE (Ours)
TUAB	B-Acc	0.8021 ± 0.0075	0.8019 ± 0.0083	$\mathbf{0.8103 \pm 0.0095}$	$\underline{0.8100 \pm 0.0073}$
	AUPRC	0.8859 ± 0.0082	0.8920 ± 0.0093	$\underline{0.8966 \pm 0.0084}$	$\mathbf{0.8977 \pm 0.0084}$
	AUROC	0.8769 ± 0.0108	0.8846 ± 0.0087	$\underline{0.8879 \pm 0.0092}$	$\mathbf{0.8905 \pm 0.0094}$
TUEV	B-Acc	0.6714 ± 0.0151	0.6841 ± 0.0146	$\mathbf{0.7015 \pm 0.0129}$	$\underline{0.6978 \pm 0.0104}$
	Kappa	0.6609 ± 0.0355	0.6899 ± 0.0242	$\mathbf{0.7043 \pm 0.0132}$	$\underline{0.6979 \pm 0.0100}$
	F1-W	0.8204 ± 0.0189	0.8385 ± 0.0146	$\mathbf{0.8457 \pm 0.0070}$	$\underline{0.8430 \pm 0.0045}$
TUSZ	B-Acc	0.9053 ± 0.0245	0.9080 ± 0.0144	$\underline{0.9103 \pm 0.0113}$	$\mathbf{0.9279 \pm 0.0187}$
	AUPRC	0.9939 ± 0.0021	$\underline{0.9950 \pm 0.0017}$	0.9922 ± 0.0033	$\mathbf{0.9960 \pm 0.0013}$
	AUROC	0.9778 ± 0.0073	$\underline{0.9814 \pm 0.0062}$	0.9709 ± 0.0124	$\mathbf{0.9852 \pm 0.0047}$
CHB-MIT	B-Acc	0.8102 ± 0.0370	0.7988 ± 0.0451	$\underline{0.8244 \pm 0.0308}$	$\mathbf{0.8261 \pm 0.0348}$
	AUPRC	0.3651 ± 0.0935	0.3141 ± 0.0539	$\underline{0.3832 \pm 0.0849}$	$\mathbf{0.4213 \pm 0.0172}$
	AUROC	$\underline{0.8976 \pm 0.0186}$	0.8965 ± 0.0302	0.8969 ± 0.0244	$\mathbf{0.9180 \pm 0.0064}$
Siena	B-Acc	$\underline{0.8499 \pm 0.0181}$	0.8254 ± 0.0378	$\mathbf{0.8509 \pm 0.0200}$	0.8410 ± 0.0114
	AUPRC	0.3338 ± 0.0648	0.4023 ± 0.0488	$\underline{0.4502 \pm 0.0170}$	$\mathbf{0.4557 \pm 0.0637}$
	AUROC	$\underline{0.9347 \pm 0.0087}$	0.9310 ± 0.0061	0.9319 ± 0.0043	$\mathbf{0.9354 \pm 0.0025}$
SEED-V	B-Acc	$\underline{0.4094 \pm 0.0053}$	0.4060 ± 0.0045	0.4079 ± 0.0084	$\mathbf{0.4199 \pm 0.0034}$
	Kappa	$\underline{0.2646 \pm 0.0083}$	0.2600 ± 0.0060	0.2604 ± 0.0130	$\mathbf{0.2781 \pm 0.0045}$
	F1-W	$\underline{0.4153 \pm 0.0076}$	0.4115 ± 0.0041	0.4127 ± 0.0125	$\mathbf{0.4257 \pm 0.0040}$
FACED	B-Acc	0.6010 ± 0.0066	$\mathbf{0.6108 \pm 0.0066}$	0.5906 ± 0.0130	$\underline{0.6082 \pm 0.0052}$
	Kappa	0.5500 ± 0.0072	$\mathbf{0.5605 \pm 0.0077}$	0.5380 ± 0.0141	$\underline{0.5578 \pm 0.0061}$
	F1-W	0.6055 ± 0.0064	$\mathbf{0.6139 \pm 0.0068}$	0.5942 ± 0.0109	$\underline{0.6124 \pm 0.0058}$
M-AT	B-Acc	$\mathbf{0.8056 \pm 0.0326}$	0.7688 ± 0.0750	0.7125 ± 0.0083	$\underline{0.7938 \pm 0.0393}$
	AUPRC	$\mathbf{0.8376 \pm 0.0368}$	0.8102 ± 0.0566	0.6660 ± 0.0768	$\underline{0.8130 \pm 0.0365}$
	AUROC	$\mathbf{0.9230 \pm 0.0163}$	$\underline{0.9026 \pm 0.0181}$	0.8063 ± 0.0453	0.8769 ± 0.0152
P-MI	B-Acc	0.6295 ± 0.0039	0.6422 ± 0.0099	$\mathbf{0.6519 \pm 0.0043}$	$\underline{0.6503 \pm 0.0034}$
	Kappa	0.5059 ± 0.0052	0.5229 ± 0.0132	$\mathbf{0.5358 \pm 0.0057}$	$\underline{0.5337 \pm 0.0046}$
	F1-W	0.6303 ± 0.0040	0.6432 ± 0.0098	$\mathbf{0.6520 \pm 0.0047}$	$\underline{0.6510 \pm 0.0036}$
ISRUC	B-Acc	0.7806 ± 0.0084	$\mathbf{0.7997 \pm 0.0076}$	0.7835 ± 0.0082	$\underline{0.7883 \pm 0.0033}$
	Kappa	0.7282 ± 0.0172	$\mathbf{0.7584 \pm 0.0106}$	0.7439 ± 0.0160	$\underline{0.7480 \pm 0.0093}$
	F1-W	0.7911 ± 0.0129	$\mathbf{0.8143 \pm 0.0071}$	0.8027 ± 0.0125	$\underline{0.8068 \pm 0.0067}$
HMC	B-Acc	0.7315 ± 0.0045	$\underline{0.7409 \pm 0.0026}$	0.7368 ± 0.0037	$\mathbf{0.7414 \pm 0.0042}$
	Kappa	0.6842 ± 0.0030	$\underline{0.6912 \pm 0.0027}$	0.6881 ± 0.0046	$\mathbf{0.6920 \pm 0.0052}$
	F1-W	0.7539 ± 0.0026	$\underline{0.7598 \pm 0.0023}$	0.7565 ± 0.0038	$\mathbf{0.7604 \pm 0.0041}$
Average	B-Acc	$\underline{0.7270}$	0.7261	0.7255	$\mathbf{0.7368}$

D.2 Positional Encoding Comparison

We compare SHPE with channel-ID embeddings, ACPE, and 4D PE to examine how electrode positions are represented within the same dual-path architecture. The alternatives provide discrete electrode identity, convolutional neighborhood information, and continuous spatial-temporal coordinates. We retain the time-frequency tokenizer, 3-stage MJDE, patch-feature gates, masking strategy, decoder architecture, and reconstruction objective. The positional encoding is replaced in both the encoder and decoder. Each variant is pretrained with its assigned encoding, rather than changing the encoding only during downstream adaptation. This experiment therefore evaluates positional encoding throughout masked pretraining, not solely its effect inside the encoder.

The channel-ID alternative replaces the spherical harmonic features with a learned lookup over 364 canonical channel identifiers. It retains the same temporal encoding and post-processing. The identifiers are shared across datasets and refer to channel names rather than their positions in an input array. ACPE uses the depthwise 19×7 convolution from CBraMod over the channel-time grid. Masked token content is zeroed before the convolution to prevent access to masked signal information. The 4D PE alternative uses REVE's Fourier and learned coordinate embeddings for (x, y, z, t), followed by LayerNorm. We use 4 Fourier frequencies, restore spatial coordinates to the metric scale expected by the implementation, and disable coordinate-noise augmentation. Its output widths are adapted to 200 for the encoder and 100 for the decoder.

The encoder-side positional modules contain 2,700 parameters for SHPE, 36,600 for channel-ID embeddings, 26,800 for ACPE, and 1,600 for 4D PE. As shown in Table 22, SHPE achieves the highest mean B-Acc of 0.7368. Channel-ID embeddings, ACPE, and 4D PE obtain 0.7270, 0.7261, and 0.7255, respectively. SHPE remains within 1.18 percentage points of the highest B-Acc on every dataset. It leads all 3 metrics on TUSZ, CHB-MIT, SEED-V, and HMC. On CHB-MIT, the B-Acc advantage over the strongest alternative is only 0.17 points, but the AUPRC and AUROC advantages are 3.81 and 2.04 points. On TUAB and Siena, SHPE obtains the highest AUPRC and AUROC despite not achieving the highest B-Acc.

The alternatives show distinct task-specific strengths. ACPE leads all 3 metrics on FACED and ISRUC, while 4D PE leads all 3 on TUEV and P-MI. Channel-ID embeddings lead all 3 metrics on M-AT. SHPE's advantage is therefore its overall performance across heterogeneous datasets rather than dominance on every task. Its small positional-module parameter count also shows that this performance does not require a larger positional embedding module than channel-ID embeddings or ACPE. These results are consistent with the usefulness of a shared spherical-harmonic representation across electrode configurations.

Table 23: Encoder ablation across 11 downstream datasets. Both single-path variants use **3 stages**. Dual fixed averages the two pathway outputs at each stage, whereas DualBrain uses learned gated fusion. Results are mean ± SD over 5 seeds. Bold and underline indicate the highest and second-highest distinct means, respectively, at the reported precision.

Dataset	Metric	$\mathcal{TM} \to \mathcal{TP}$	$\mathcal{TP} \to \mathcal{TM}$	Dual fixed	DualBrain
TUAB	B-Acc	0.8059 ± 0.0109	$\mathbf{0.8164 \pm 0.0050}$	0.8101 ± 0.0081	0.8100 ± 0.0073
	AUPRC	0.8934 ± 0.0131	$\mathbf{0.9010 \pm 0.0087}$	0.8950 ± 0.0091	0.8977 ± 0.0084
	AUROC	0.8857 ± 0.0143	$\mathbf{0.8936 \pm 0.0103}$	0.8865 ± 0.0098	0.8905 ± 0.0094
TUEV	B-Acc	0.6879 ± 0.0116	0.6835 ± 0.0119	0.6865 ± 0.0089	$\mathbf{0.6978 \pm 0.0104}$
	Kappa	0.6906 ± 0.0297	0.6694 ± 0.0279	0.6969 ± 0.0129	$\mathbf{0.6979 \pm 0.0100}$
	F1-W	0.8367 ± 0.0161	0.8259 ± 0.0146	$\mathbf{0.8435 \pm 0.0066}$	0.8430 ± 0.0045
TUSZ	B-Acc	0.9202 ± 0.0073	0.9085 ± 0.0107	0.9102 ± 0.0118	$\mathbf{0.9279 \pm 0.0187}$
	AUPRC	0.9921 ± 0.0016	0.9920 ± 0.0025	0.9954 ± 0.0028	$\mathbf{0.9960 \pm 0.0013}$
	AUROC	0.9713 ± 0.0057	0.9708 ± 0.0092	0.9824 ± 0.0103	$\mathbf{0.9852 \pm 0.0047}$
CHB-MIT	B-Acc	0.8248 ± 0.0264	0.7887 ± 0.0347	0.8074 ± 0.0208	$\mathbf{0.8261 \pm 0.0348}$
	AUPRC	0.3987 ± 0.0492	0.3736 ± 0.0540	0.3713 ± 0.0378	$\mathbf{0.4213 \pm 0.0172}$
	AUROC	0.8876 ± 0.0222	0.9098 ± 0.0116	0.9106 ± 0.0136	$\mathbf{0.9180 \pm 0.0064}$
Siena	B-Acc	0.8390 ± 0.0249	$\mathbf{0.8694 \pm 0.0099}$	0.8182 ± 0.0200	0.8410 ± 0.0114
	AUPRC	0.4329 ± 0.0440	0.4159 ± 0.0495	0.4465 ± 0.0289	$\mathbf{0.4557 \pm 0.0637}$
	AUROC	0.9344 ± 0.0045	$\mathbf{0.9373 \pm 0.0027}$	0.9354 ± 0.0039	0.9354 ± 0.0025
SEED-V	B-Acc	0.4050 ± 0.0078	0.4124 ± 0.0052	0.4116 ± 0.0088	$\mathbf{0.4199 \pm 0.0034}$
	Kappa	0.2582 ± 0.0112	0.2681 ± 0.0074	0.2672 ± 0.0119	$\mathbf{0.2781 \pm 0.0045}$
	F1-W	0.4110 ± 0.0084	0.4192 ± 0.0069	0.4185 ± 0.0107	$\mathbf{0.4257 \pm 0.0040}$
FACED	B-Acc	0.5851 ± 0.0049	$\mathbf{0.6171 \pm 0.0098}$	0.5859 ± 0.0136	0.6082 ± 0.0052
	Kappa	0.5311 ± 0.0038	$\mathbf{0.5667 \pm 0.0122}$	0.5330 ± 0.0149	0.5578 ± 0.0061
	F1-W	0.5862 ± 0.0029	$\mathbf{0.6197 \pm 0.0113}$	0.5905 ± 0.0130	0.6124 ± 0.0058
M-AT	B-Acc	0.7389 ± 0.0704	0.6938 ± 0.0691	0.7382 ± 0.0446	$\mathbf{0.7938 \pm 0.0393}$
	AUPRC	$\mathbf{0.8566 \pm 0.0440}$	0.7462 ± 0.0559	0.8178 ± 0.0296	0.8130 ± 0.0365
	AUROC	$\mathbf{0.9319 \pm 0.0160}$	0.8309 ± 0.0266	0.9010 ± 0.0286	0.8769 ± 0.0152
P-MI	B-Acc	0.6363 ± 0.0041	0.6285 ± 0.0069	0.6483 ± 0.0034	$\mathbf{0.6503 \pm 0.0034}$
	Kappa	0.5151 ± 0.0055	0.5046 ± 0.0092	0.5310 ± 0.0045	$\mathbf{0.5337 \pm 0.0046}$
	F1-W	0.6370 ± 0.0046	0.6299 ± 0.0073	0.6495 ± 0.0035	$\mathbf{0.6510 \pm 0.0036}$
ISRUC	B-Acc	0.7826 ± 0.0121	$\mathbf{0.7917 \pm 0.0029}$	0.7831 ± 0.0117	0.7883 ± 0.0033
	Kappa	0.7292 ± 0.0230	0.7434 ± 0.0097	0.7333 ± 0.0224	$\mathbf{0.7480 \pm 0.0093}$
	F1-W	0.7872 ± 0.0207	0.8019 ± 0.0076	0.7951 ± 0.0190	$\mathbf{0.8068 \pm 0.0067}$
HMC	B-Acc	0.7368 ± 0.0015	0.7351 ± 0.0011	$\mathbf{0.7439 \pm 0.0010}$	0.7414 ± 0.0042
	Kappa	0.6893 ± 0.0038	0.6873 ± 0.0026	$\mathbf{0.6959 \pm 0.0013}$	0.6920 ± 0.0052
	F1-W	0.7559 ± 0.0022	0.7539 ± 0.0020	$\mathbf{0.7619 \pm 0.0011}$	0.7604 ± 0.0041
Average	B-Acc	0.7239	0.7223	0.7221	**0.7368**

D.3 Encoder Ablation

We examine whether both processing orders and learned stage-wise fusion are necessary by comparing DualBrain with 2 single-path variants and a fixed-fusion variant. The temporal-to-topographic variant applies temporal attention followed by spatial attention in each stage. The topographic-to-temporal variant reverses this order. Both contain 3 stages with 2 attention-FFN blocks per stage, giving 6 blocks in total. They retain the same embedding dimension of 200, feed-forward dimension of 800, 4 attention heads per axis, pre-norm RMSNorm, GELU activation, and final RMSNorm. Each block has independent parameters, and neither single-path variant contains a fusion gate.

Dual fixed retains both pathways and all 12 blocks of the 3-stage dual-path encoder. At every stage, it averages the two outputs element-wise with fixed weights of 0.5 and 0.5. The resulting representation is passed to both pathways in the next stage. DualBrain follows the same stage-wise arrangement but replaces the fixed average with an input-dependent patch-feature gate. Each gate maps the concatenated 400-dimensional pathway representations to 200-dimensional through a linear projection and sigmoid.

Each 3-stage single-path encoder contains 2,893,400 parameters. Dual fixed contains 5,786,600, while DualBrain contains 6,027,200. These counts exclude the tokenizer, positional encoding, decoder, and downstream head. Removing the gates reduces the encoder by 240,600 parameters. We use the 3-stage variants as our primary ablation, prioritizing matched stage count over parameter

count under the hypothesis that repeated sequential transformations are a key factor in this comparison. Extending a single pathway to six stages approximately matches parameter count but also increases sequential depth, potentially changing optimization and training stability. We therefore report the 6-stage variants as supplementary capacity controls. The primary comparison does not isolate processing order from capacity, and we interpret it alongside these supplementary results. Dual fixed preserves both pathways and their stage-wise interaction, providing a closer comparison of fusion mechanisms.

Table 24: Encoder ablation across 11 downstream datasets. Both single-path variants use **6 stages**. Results are mean ± SD over 5 seeds. Bold and underline indicate the highest and second-highest distinct means, respectively, at the reported precision.

Dataset	Metric	$\mathcal{TM} \to \mathcal{TP}$	$\mathcal{TP} \to \mathcal{TM}$	DualBrain
TUAB	B-Acc	0.8001 ± 0.0058	$\underline{0.8084 \pm 0.0086}$	$\mathbf{0.8100 \pm 0.0073}$
	AUPRC	0.8913 ± 0.0069	$\underline{0.8949 \pm 0.0101}$	$\mathbf{0.8977 \pm 0.0084}$
	AUROC	0.8821 ± 0.0072	$\underline{0.8865 \pm 0.0108}$	$\mathbf{0.8905 \pm 0.0094}$
TUEV	B-Acc	0.6555 ± 0.0053	$\mathbf{0.6985 \pm 0.0030}$	$\underline{0.6978 \pm 0.0104}$
	Kappa	0.6375 ± 0.0457	$\underline{0.6825 \pm 0.0202}$	$\mathbf{0.6979 \pm 0.0100}$
	F1-W	0.8096 ± 0.0266	$\underline{0.8328 \pm 0.0115}$	$\mathbf{0.8430 \pm 0.0045}$
TUSZ	B-Acc	0.9079 ± 0.0138	$\mathbf{0.9311 \pm 0.0151}$	$\underline{0.9279 \pm 0.0187}$
	AUPRC	0.9915 ± 0.0035	$\underline{0.9956 \pm 0.0024}$	$\mathbf{0.9960 \pm 0.0013}$
	AUROC	0.9691 ± 0.0128	$\underline{0.9835 \pm 0.0086}$	$\mathbf{0.9852 \pm 0.0047}$
CHB-MIT	B-Acc	$\underline{0.8220 \pm 0.0267}$	0.7865 ± 0.0260	$\mathbf{0.8261 \pm 0.0348}$
	AUPRC	$\underline{0.3495 \pm 0.0364}$	0.3384 ± 0.0369	$\mathbf{0.4213 \pm 0.0172}$
	AUROC	0.8977 ± 0.0186	$\underline{0.8987 \pm 0.0119}$	$\mathbf{0.9180 \pm 0.0064}$
Siena	B-Acc	0.8177 ± 0.0195	$\underline{0.8392 \pm 0.0122}$	$\mathbf{0.8410 \pm 0.0114}$
	AUPRC	$\underline{0.4525 \pm 0.0368}$	0.4412 ± 0.0181	$\mathbf{0.4557 \pm 0.0637}$
	AUROC	$\underline{0.9363 \pm 0.0089}$	$\mathbf{0.9397 \pm 0.0028}$	0.9354 ± 0.0025
SEED-V	B-Acc	0.4150 ± 0.0030	$\underline{0.4189 \pm 0.0036}$	$\mathbf{0.4199 \pm 0.0034}$
	Kappa	0.2716 ± 0.0042	$\underline{0.2766 \pm 0.0048}$	$\mathbf{0.2781 \pm 0.0045}$
	F1-W	0.4225 ± 0.0039	$\mathbf{0.4265 \pm 0.0037}$	$\underline{0.4257 \pm 0.0040}$
FACED	B-Acc	$\underline{0.5951 \pm 0.0045}$	0.5889 ± 0.0094	$\mathbf{0.6082 \pm 0.0052}$
	Kappa	$\underline{0.5438 \pm 0.0048}$	0.5376 ± 0.0103	$\mathbf{0.5578 \pm 0.0061}$
	F1-W	$\underline{0.5989 \pm 0.0035}$	0.5948 ± 0.0093	$\mathbf{0.6124 \pm 0.0058}$
M-AT	B-Acc	$\underline{0.7465 \pm 0.0898}$	0.7188 ± 0.0442	$\mathbf{0.7938 \pm 0.0393}$
	AUPRC	0.7714 ± 0.0821	$\underline{0.7772 \pm 0.0382}$	$\mathbf{0.8130 \pm 0.0365}$
	AUROC	$\mathbf{0.8925 \pm 0.0579}$	0.8759 ± 0.0159	$\underline{0.8769 \pm 0.0152}$
P-MI	B-Acc	$\mathbf{0.6507 \pm 0.0029}$	0.6451 ± 0.0035	$\underline{0.6503 \pm 0.0034}$
	Kappa	$\mathbf{0.5343 \pm 0.0039}$	0.5267 ± 0.0046	$\underline{0.5337 \pm 0.0046}$
	F1-W	$\mathbf{0.6511 \pm 0.0029}$	0.6457 ± 0.0038	$\underline{0.6510 \pm 0.0036}$
ISRUC	B-Acc	0.7762 ± 0.0051	$\underline{0.7880 \pm 0.0072}$	$\mathbf{0.7883 \pm 0.0033}$
	Kappa	0.7291 ± 0.0062	$\underline{0.7401 \pm 0.0140}$	$\mathbf{0.7480 \pm 0.0093}$
	F1-W	0.7932 ± 0.0058	$\underline{0.7994 \pm 0.0112}$	$\mathbf{0.8068 \pm 0.0067}$
HMC	B-Acc	0.7348 ± 0.0009	$\underline{0.7363 \pm 0.0031}$	$\mathbf{0.7414 \pm 0.0042}$
	Kappa	0.6875 ± 0.0019	$\mathbf{0.6926 \pm 0.0025}$	$\underline{0.6920 \pm 0.0052}$
	F1-W	0.7544 ± 0.0014	$\underline{0.7583 \pm 0.0024}$	$\mathbf{0.7604 \pm 0.0041}$
Average	B-Acc	0.7201	$\underline{0.7236}$	**0.7368**

To complement the primary three-stage ablation, we also evaluate six-stage single-path encoders with 5,786,600 parameters each, matching Dual fixed. DualBrain contains 6,027,200 encoder parameters, with the additional 240,600 parameters attributable entirely to its gating modules. As shown in Table 24, across the 11 datasets with complete results, DualBrain achieves a mean B-Acc of 0.7368, compared with 0.7201 for temporal-to-topographic and 0.7236 for topographic-to-temporal processing. Increasing single-path capacity therefore does not close the average performance gap under the shared training recipe. However, extending the single-path encoders to six stages also increases sequential depth, potentially affecting optimization and training stability. We prioritize preserving the number of sequential stages in our primary ablation and report the six-stage variants as supplementary capacity controls; neither comparison alone fully isolates the effect of processing order.

The preferred single-path order depends on both the dataset and metric. The topographic-to-temporal variant leads all 3 metrics on TUAB and FACED and obtains the highest B-Acc and AU-

ROC on Siena. The temporal-to-topographic variant leads AUPRC and AUROC on M-AT, although DualBrain achieves higher B-Acc. On CHB-MIT, temporal-to-topographic processing has higher B-Acc and AUPRC than the reverse order but lower AUROC. These differences show why a single metric is insufficient to assign an entire task to one preferred processing order.

Learned fusion is also not uniformly superior to fixed averaging. Dual fixed leads all 3 metrics on HMC and slightly exceeds DualBrain in weighted F1 on TUEV. Nevertheless, DualBrain's higher average B-Acc and improvements across multiple metrics support the usefulness of learned stage-wise fusion.

Table 24 shows that DualBrain obtains a mean B-Acc of 0.7368, compared with 0.7201 for the six-stage $\mathcal{TM} \rightarrow \mathcal{TP}$ model and 0.7236 for $\mathcal{TP} \rightarrow \mathcal{TM}$, over the same 11 datasets. Its mean B-Acc is higher on 10 and 9 datasets, respectively. Dual fixed reaches 0.7221. It falls between the two single-path controls, so retaining both paths with fixed averaging does not provide a consistent advantage.

The differences also depend on the metric: $\mathcal{TP} \rightarrow \mathcal{TM}$ has slightly higher B-Acc on TUEV but lower weighted F1, and both six-stage single-path models have higher AUROC on Siena. Matching the total block count therefore supports a comparison at similar encoder capacity, but leaves a difference in sequential depth. The results favor the full gated design on average under the shared training recipe; they do not establish that both processing orders are necessary or isolate the effect of repeated fusion.

E Additional Visualization

E.1 Gate Visualization

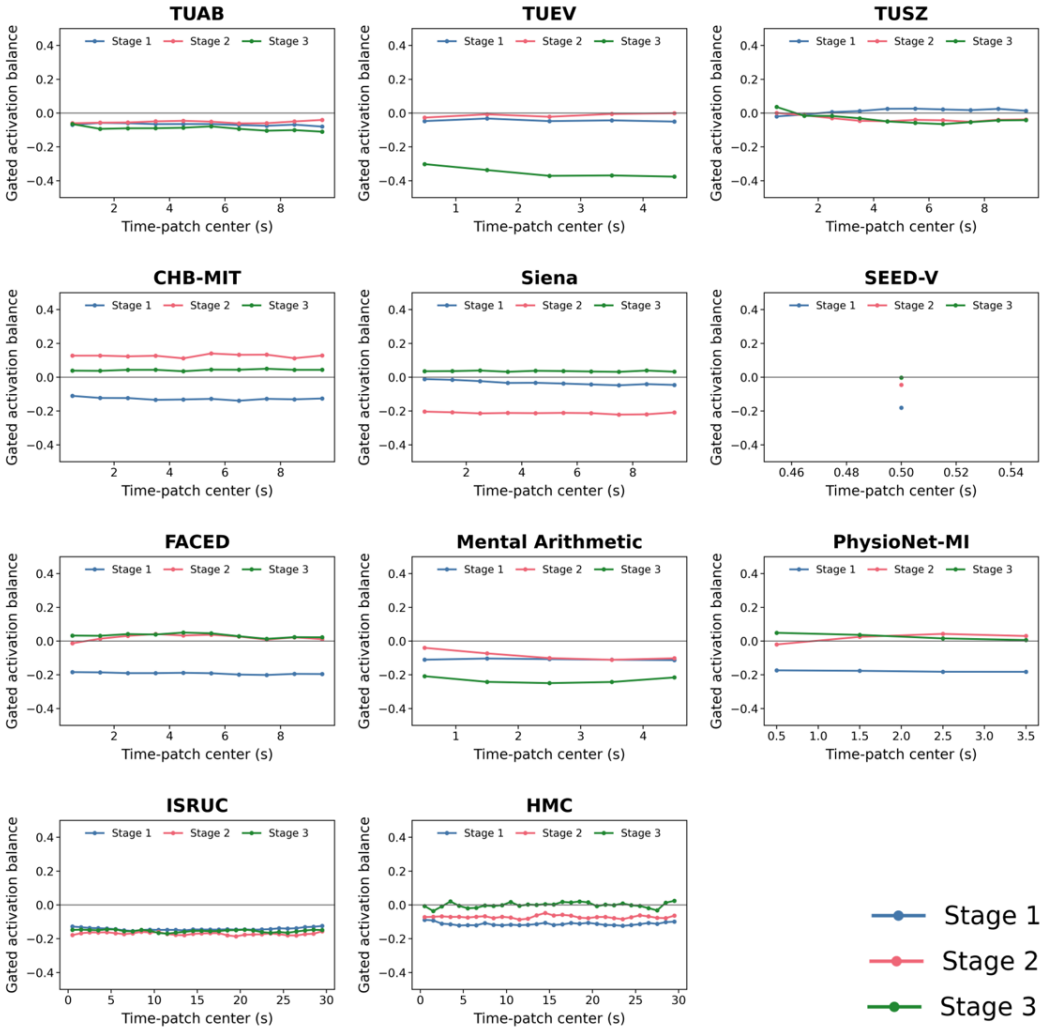

Figure 3: Activation-aware branch contribution balance across the 11 downstream EEG datasets. At each encoder stage, we measure the RMS magnitude of the two gated branch outputs, over all 200 latent dimensions. Their normalized difference is averaged over valid EEG channels and test observations. Positive values indicate a larger contribution magnitude from the $\mathcal{TP} \rightarrow \mathcal{TM}$ branch, whereas negative values indicate a larger contribution from the $\mathcal{TM} \rightarrow \mathcal{TP}$ branch. SEED-V contains a single temporal patch and therefore does not provide a temporal trajectory.

To examine how the learned gate combines the two processing-order branches, we analyze their gated activation magnitudes at each of the 3 fusion stages. Because the mean gate value alone ignores the magnitude of the corresponding branch representations, we compute the RMS magnitude of each gated branch output over the $D = 200$ latent dimensions. We then define the normalized activation balance as

$$R_{b,c,t} = \frac{\mathrm{RMS}_d\big(G \odot H^{(\mathrm{Topo}\rightarrow\mathrm{Temp})}\big) - \mathrm{RMS}_d\big((1-G) \odot H^{(\mathrm{Temp}\rightarrow\mathrm{Topo})}\big)}{\mathrm{RMS}_d\big(G \odot H^{(\mathrm{Topo}\rightarrow\mathrm{Temp})}\big) + \mathrm{RMS}_d\big((1-G) \odot H^{(\mathrm{Temp}\rightarrow\mathrm{Topo})}\big) + \epsilon}. \tag{14}$$

Positive values indicate a larger contribution from the $\mathcal{TP} \rightarrow \mathcal{TM}$ branch, whereas negative values indicate a larger contribution from the $\mathcal{TM} \rightarrow \mathcal{TP}$ branch. Values near zero indicate comparable

magnitudes. We compute $R_{b,c,t}$ for every test sample, channel, and temporal patch and then average it across test samples and valid EEG channels.

As shown in Fig. 3, the learned balance varies across tasks and stages but remains stable across temporal patches within each task. The temporal standard deviation of the stage-wise mean balance ranges from 0.003 to 0.028. CHB-MIT places greater emphasis on the $\mathcal{TP} \rightarrow \mathcal{TM}$ branch in Stage 2, whereas TUEV Stage 3 and Siena Stage 2 favor the $\mathcal{TM} \rightarrow \mathcal{TP}$ branch. Thus, the gate learns task- and stage-specific fusion patterns without irregular temporal fluctuations.

The branch imbalance is relatively large in Stage 1 but generally moves toward zero in later stages, except for ISRUC. Together with the improvements over the two single-path variants in Table 4, this convergence suggests that DualBrain benefits not only from the individual branches but also from their progressive integration into a balanced joint representation.

E.2 SHPE VISUALIZATION

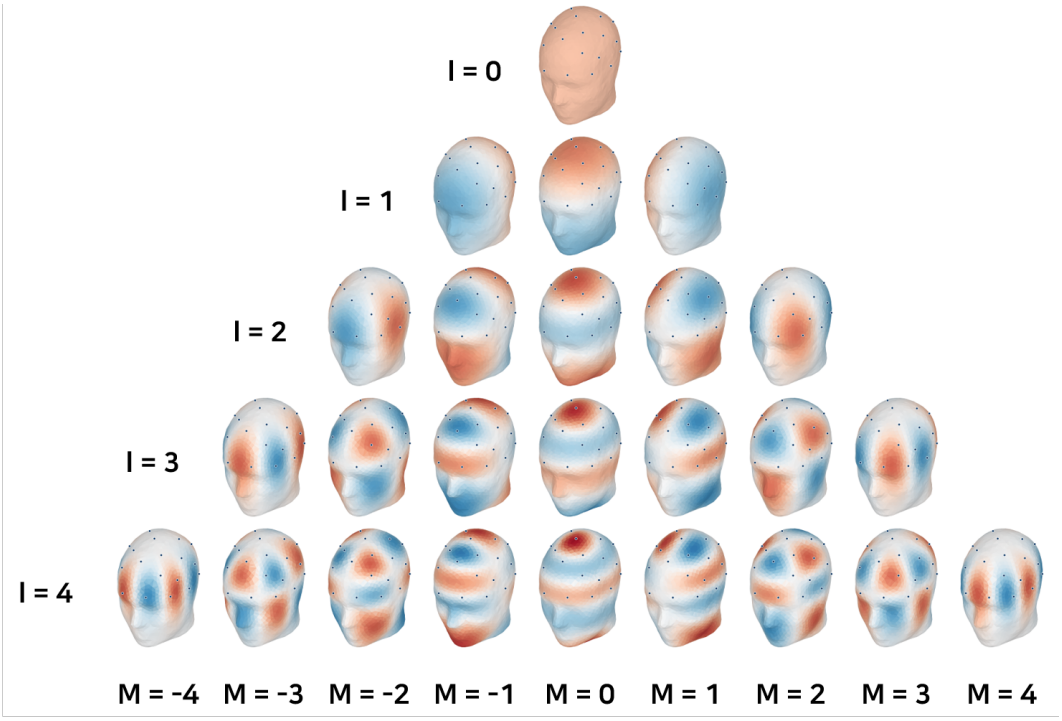

Figure 4: Visualization of spherical harmonic positional encoding (SHPE) over the scalp surface. We use real spherical harmonics up to degree $L = 4$, yielding $(L+1)^2 = 25$ basis functions in total.

Attention-based sequence models require positional information because self-attention alone cannot encode the order or relative distances between observations. Sinusoidal positional encoding addresses this limitation by representing the position at each time step using sine and cosine functions of varying frequencies. Low-frequency components change gradually and capture long-term temporal structures, whereas high-frequency components change rapidly and distinguish subtle differences between adjacent time steps. This encoding provides a structured representation of temporal relationships without requiring independent learning for every time index.

We extend this frequency-based principle from the temporal axis to the spatial geometry of EEG electrodes. Sine and cosine functions form a natural frequency basis on a one-dimensional temporal domain. For electrode locations distributed over the curved surface of the head, the corresponding frequency basis is given by spherical harmonics. Specifically, a real spherical-harmonic basis function combines sinusoidal variation along the azimuthal direction through $\sin(m\phi)$ and $\cos(m\phi)$ with elevation-dependent associated Legendre functions. Spherical harmonics can therefore be un-

derstood as a generalization of sinusoidal encoding from a one-dimensional axis to a spherical domain.

For each electrode, SHPE first normalizes its Cartesian coordinate onto the unit sphere and then evaluates the real spherical-harmonic basis functions $Y_{\ell,m}(\theta, \phi)$ for degrees $\ell = 0, \ldots, 4$ and orders $m = -\ell, \ldots, \ell$. This produces

$$\sum_{\ell=0}^{4}(2\ell + 1) = 1 + 3 + 5 + 7 + 9 = 25$$

deterministic spatial features per electrode.

The spherical-harmonic degree ℓ can be interpreted as spatial frequency. Low-degree harmonics vary smoothly over the scalp and represent coarse, global spatial organization, including broad anterior–posterior or inter-hemispheric gradients. As ℓ increases, the basis functions vary more rapidly over the surface and represent progressively finer spatial patterns. SHPE therefore describes electrode locations at multiple spatial scales, analogous to how temporal sinusoidal encoding represents time using multiple temporal frequencies.

Figure 4 visualizes representative basis functions from degrees 0 through 4 on the `fsaverage` scalp surface. The markers denote the EEG electrode locations. The surface colors show the values of the fixed spherical-harmonic basis functions rather than activations learned by the model.

This representation provides an inductive bias that is well matched to EEG geometry: electrodes are located on an approximately spherical surface, while EEG signals may exhibit both broad inter-regional structure and localized spatial variation. Instead of treating channel identities as unrelated categorical embeddings, SHPE derives their representations from physical coordinates and expresses them through a shared multiscale basis. The resulting 25-dimensional vector is mapped to the model's spatial embedding dimension by a learned linear projection. The model can consequently learn task-relevant combinations of temporal- and spatial-frequency components while retaining an explicit representation of electrode geometry.

F Discussion

F.1 Implication

Geometry-aware positional encoding for heterogeneous EEG. SHPE uses spherical harmonic basis functions to express electrode locations in a continuous, multiscale coordinate system. Table 5 shows an average performance advantage over the compared positional encodings, although ACPE performs better on FACED and ISRUC. These comparisons support the usefulness of physical electrode coordinates under dataset-specific supervised fine-tuning; they do not establish robustness to unseen montages, electrode removal, or zero-shot transfer. Whether SHPE's finer geometric structure improves robustness under these conditions remains to be tested.

Simple but General Encoder. DualBrain combines two processing orders within a shared encoder design. The Topo→Temp pathway models the temporal evolution of spatial features, while the Temp→Topo pathway models spatial relationships among temporal features. This design accommodates both perspectives without requiring a different encoder architecture for each downstream task. Its generality therefore lies in architectural reuse across the evaluated tasks, with task-specific supervised adaptation, rather than an assumption that either processing order is universally preferable.

F.2 Limitation

Although DualBrain was evaluated on 11 downstream datasets, the task coverage remains comparable to existing EEG foundation-model benchmarks rather than substantially broader. Moreover, our gate analysis measures relative branch activation magnitudes but does not establish the causal contribution of each pathway to individual predictions. Future work should evaluate additional EEG applications and employ causal attribution or pathway-intervention analyses to clarify how each processing order influences model decisions.