BRAINDUO: DUAL-DOMAIN TOKENIZATION AND PRETRAINING FOR EEG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) foundation models use large-scale pretraining to learn a general-purpose EEG encoder that can be adapted to diverse brain-decoding tasks. Although current EEG foundation models benefit from large-scale self-supervised pretraining, learning transferable representations from high-dimensional, non-stationary EEG signals remains challenging. Existing approaches can struggle to jointly preserve time series and spectral information and capture dependencies at multiple spatiotemporal scales, due to inherent limitations in tailoring tokenizer design and pretraining strategies to the characteristics of EEG signals. To address these challenges, we introduce BrainDuo, an EEG foundation model that combines a dual-domain spectrotemporal codebook with discrete and continuous supervision to learn local and long-range dependencies among EEG tokens. Our BrainDuo features three novel components, including Duo Codebook, Duo Latent and DuoView Pretraining to learn effective, robust EEG representation. Specifically, Duo Codebook learns a shared spectrotemporal vocabulary, Duo Latent transfers both discrete codes and continuous embeddings into pretraining, and DuoView Pretraining applies cyclic and clustered attention to encode the temporal context along each channel and the topographic context among neighboring electrodes. We further incorporate brain-computer interface (BCI)-native skills into downstream tasks to help EEG foundation models exploit task-specific prior knowledge for various BCI decoding. Extensive experiments across 10 representative EEG tasks and 14 datasets demonstrate that BrainDuo outperforms strong EEG foundation baselines across general EEG benchmarks and moves closer to the models performance specialized in BCI tasks. These results highlight the potential of leveraging a hierarchical, flexible design to learn general-purpose EEG representations, demonstrating BrainDuo's versatility as a backbone for diverse EEG decoding and BCI applications.

1 INTRODUCTION

Electroencephalography (EEG) provides a non-invasive means of studying brain activity, with applications spanning clinical assessment and brain–computer interfaces (BCIs). EEG and BCI tasks such as sleep staging, motor imagery, emotion recognition, and visual decoding involve distinct neural states and decoding objectives, yet share common EEG characteristics, including canonical neural oscillations (e.g., δ, θ, α, β, and γ waves) and background noise. Building a separate model for each application limits the reuse of the knowledge across tasks. EEG foundation models offer a scalable alternative: large-scale self-supervised pretraining produces an encoder that can be adapted to a wide range of downstream problems (Yang et al., 2023a; Jiang et al., 2024; Wang et al., 2025).

Existing EEG foundation models (EFM) commonly divide recordings into channel–time patches and learn representations through signal reconstruction or latent-target prediction (Jiang et al., 2024; Wang et al., 2024; 2025). Codebook-based tokenization provides discrete targets for masked prediction. Recent architectures increasingly incorporate EEG-specific structure. CodeBrain combines separate temporal and frequency tokenization with global and local modeling (Ma et al., 2026). These advances motivate two connected design questions: what information should a patch representation obtain and transfer, and how should pretraining organize interactions across time and electrodes?

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 questions arise from the inherent challenges of EEG data. Scalp EEG is nonstationary, with a low signal-to-noise ratio due to neural signal attenuation through volume conduction and contamination by artifacts. These properties, together with substantial variability across tasks, subjects, and devices, make learning transferable representations challenging. To tackle these challenges, pretraining must preserve intrinsic temporal and spectral information within EEG patches while modeling diverse dependencies within the data. However, existing tokenization method, i.e. codebooks, often emphasize single domain information, either spectral information or temporal signal structure, leaving their complementary information less captured. During pretraining, models may also rely on uniform or fixed interaction patterns, limiting their flexibility in capturing abstract and diversified representation (Jiang et al., 2024; Zhou et al., 2025).

To address these issues, we introduce BrainDuo, a dual-domain EFM with tokenization and pretraining for EEG decoding and BCI applications. BrainDuo comprise three core components, **Duo Codebook**, **Duo Latent** and **DuoView Pretraining**, to synergistically learn effective representation during large-scale pretraining. Specifically, **Duo Codebook** learns a shared vocabulary through waveform and short-time Fourier transform (STFT) reconstruction to fully recognize temporal and spectral information and the underlying time–frequency interplay. Motivated by potential information loss during quantization, **Duo Latent** transfers both the *discrete code* and the pre-quantization *continuous embedding* through dual supervision, combining code prediction with feature alignment to more fully preserve the information in each EEG patch. These tokenization modules further provide supervision for **DuoView Pretraining**, which implements decomposed and clustered spatiotemporal modeling through the *multi-rate block* and *long-range block*, respectively. Together, these components provide substrate for self-supervised learning of hierarchical EEG representations that help overcome data variability and generalize to a wide range of downstream applications.

For BCI applications, can general-purpose EEG representations approach the performance of task-specialist BCI models through task-specific protocols? We investigate this question by incorporating **BCI training skills**, including a range of BCI processing protocols such as filter banks (Ang et al., 2008), channel selection based on task priors, denoising, and grand averaging, into downstream pipelines. These BCI training skills incorporate task-specific signal structure and prior knowledge into downstream adaptation, helping translate EFM representations into effective BCI decoders.

Our contributions are summarized as follows:

- **Dual-Domain Tokenizer for Unified EEG Patch Representations.** We learn a single codebook under joint spatiotemporal reconstruction, then apply both discrete and continuous tokenizer embeddings as pretraining targets. Duo Codebook and Duo Latent integrate EEG patch representation learning with code prediction and feature alignment, transferring information beyond discrete codes alone, as used in current codebook-based methods.
- **DuoView Pretraining for Local and Long-Range EEG Dependencies.** We introduce channel topological (CTopo) distance to guide neighborhood-based attention and cluster masking. We propose a multi-rate block to capture multi-cyclic temporal dependencies and modeling spatial coherence separately. A long-range block connects tokens within CTopo-guided patch neighborhoods to capture spatiotemporal dependencies through attention. Interlace stacked, the multi-rate and long-range blocks form the DuoView pretraining backbone, aggregating multi-scale context for hierarchical EEG representation learning.
- **BCI-Native Adaptation and Generalist–Specialist Evaluation.** We establish two evaluation tracks covering general EFM benchmarks and task-specific BCI skills across 10 tasks and 14 datasets. For general EFM benchmarks, we use evaluation pipelines for comparisons with other EFMs. The BCI track incorporates task-native techniques into foundation-model adaptation. Comparisons with specialist models further identify where pretrained EEG representations approach the performance of specialized systems and where performance gaps remain.

2 Related Work

General EEG foundation models. EEG foundation modeling seeks to learn reusable neural representations from large unlabeled corpora and transfer them across downstream tasks. Early work such as BENDR explored Transformer-based EEG representation learning (Kostas et al., 2021).

BIOT addressed cross-data learning from heterogeneous biosignal recordings (Yang et al., 2023a). LaBraM, EEGPT, and CBraMod subsequently advanced masked pretraining for transfer across EEG decoding tasks (Jiang et al., 2024; Wang et al., 2024; 2025). More recent models address variation in recording configurations: REVE uses electrode coordinates to accommodate different setups (El Ouahidi et al., 2025), and LUNA maps variable channel configurations into a fixed-size latent channel representation at each temporal patch (Döner et al., 2025).

EEG tokenization and pretraining supervision. Tokenizer design and pretraining supervision involve two distinct choices: the information retained by a tokenizer and the targets used to train the backbone. LaBraM learns spectrum-supervised vector-quantized codes (Jiang et al., 2024), while NeuroLM learns a shared temporal–frequency vocabulary with additional alignment to the text embedding space (Jiang et al., 2025). CodeBrain instead applies separate temporal and frequency codebooks to a shared encoding (Ma et al., 2026), and TFM-Tokenizer learns reusable time–frequency motifs from single-channel EEG (Pradeepkumar et al., 2026). These tokenization approaches are distinct from the choice of pretraining target: LaBraM predicts discrete code identities, whereas EEGPT combines continuous representation alignment with masked reconstruction (Jiang et al., 2024; Wang et al., 2024).

Spatiotemporal modeling and masking. EEG architectures organize interactions through axis-specific, multi-scale, and geometry-aware designs. CBraMod separates spatial and temporal attention into parallel head groups (Wang et al., 2025). CSBrain aggregates features within temporal windows and brain regions before modeling cross-window and cross-region dependencies with structured sparse attention (Zhou et al., 2025). CodeBrain combines structured global convolution and sliding-window attention to model long-range dependencies and local neural events (Ma et al., 2026). Geometry also informs self-supervised learning: REVE uses electrode coordinates in spatiotemporal positional encoding and constructs joint spatial–temporal block masks (El Ouahidi et al., 2025), while EEG-DisGCMAE combines graph contrastive and masked pretraining with topology distillation for high-to-low-density transfer (Wei et al., 2025).

BCI-native adaptation and evaluation. Task-specialist BCI models combine data preprocessing and architectures tailored to their decoding objectives. SSVEP systems exploit harmonic sub-bands aggregation (Guney et al., 2022; Lan et al., 2025). For visual EEG decoding, averaging responses to repeated presentations of the same image (*grand average*) is a common THINGS-EEG evaluation practice in NICE, ATM, and BrainHiVE (Song et al., 2024; Li et al., 2024; Zheng et al., 2026). PGCN and DAEST model spatial relationships and denoising techniques are applied for emotion recognition (Jin et al., 2024; Shen et al., 2025). These systems provide task-specific practices and reference models for evaluating general-purpose encoders. Recent benchmarking studies further show that the benefits of EEG pretraining depend on adaptation, data availability, and evaluation protocols (Lee et al., 2025; Yang et al., 2026; Xiong et al., 2026).

3 METHOD

3.1 OVERVIEW

BrainDuo unifies **Duo Codebook**, **Duo Latent**, and **DuoView Pretraining**. The Tokenizer first learns **Duo Codebook** through waveform and STFT reconstruction. It supplies the discrete code and continuous embedding for dual supervision of the backbone. DuoView Pretraining organizes its context through cyclic temporal groups and clustered spatio-temporal neighborhoods, implemented by the multi-rate block and long-range block, respectively. Masking, guided by CTopo distance, constructs the prediction inputs. Figure 1 presents an overview of BrainDuo and visualizes its performance across downstream tasks. Figure 2 details the training pipeline and its components. For an EEG segment $\boldsymbol{X} \in \mathbb{R}^{C\times L}$, we divide each of its C channels into A non-overlapping patches $\boldsymbol{x}_{c,a} \in \mathbb{R}^{P}$.

3.2 TOKENIZER AND DUO CODEBOOK

We first use Patch Embedding and the Tokenizer Encoder to produce the continuous embedding $\boldsymbol{h}^{\mathrm{tok}}_{c,a} \in \mathbb{R}^{D}$. We then map it to a normalized quantizer vector $\boldsymbol{z}_{c,a} \in \mathbb{R}^{d}$ through a learned projection.

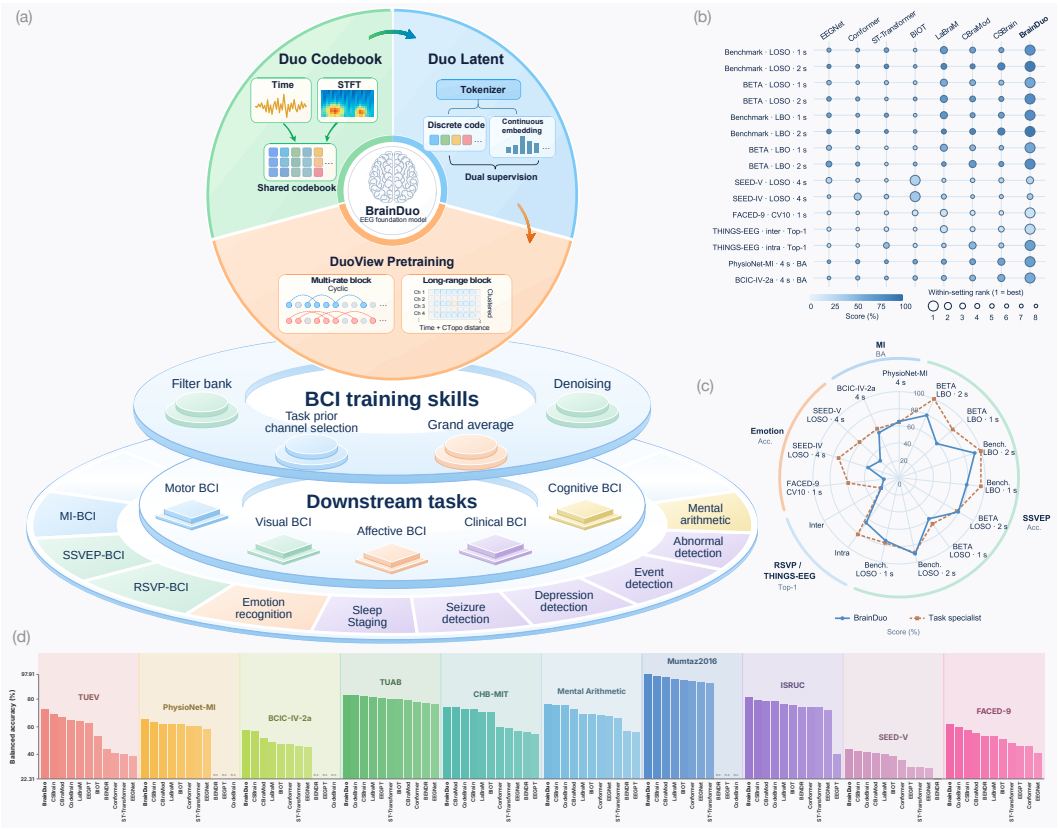

Figure 1: Overview of BrainDuo and its downstream evaluation. (a) Duo Codebook, Duo Latent, and DuoView Pretraining as its core components, together with BCI training skills and downstream task families. (b) BCI comparison with general EEG models: color encodes score and bubble size encodes within-setting rank. (c) Comparison with task specialists using the indicated task metrics. (d) General EEG evaluation using balanced accuracy; The y axis spans the global min–max range.

Given the normalized codebook $\mathcal{E} = \{\boldsymbol{e}_k\}_{k=1}^{K}$, the code assigned to a patch is

$$k_{c,a} = \arg\max_{k} \boldsymbol{z}_{c,a}^{\top} \boldsymbol{e}_k, \qquad \boldsymbol{q}_{c,a} = \boldsymbol{e}_{k_{c,a}}. \tag{1}$$

A commitment loss $\mathcal{L}_{\text{vq}}$ (van den Oord et al., 2017) aligns the projected features with their assigned codes, and exponential moving averages update the codebook.

Time Series and STFT reconstruction. The shared quantized representation is decoded through three reconstruction heads, i.e., **AHead**, **PHead**, and **THead**, to predict spectral *Amplitude*, *Phase*, and *Temporal waveform*, respectively. Amplitude denotes STFT magnitude. For each head, a separate loss is applied. We use *Cauchy loss* (Barron, 2019) for AHead, $\mathcal{L}_{\text{amp}} = \text{mean}\log(1 + (\widehat{m} - m)^2)$, where m_i is the i-th element of the STFT magnitude target and $\widehat{m}_i$ is the corresponding prediction. Its logarithmic growth limits the influence of large reconstruction errors, making amplitude reconstruction less sensitive to outliers. PHead uses *MSE loss*; Waveform reconstruction in THead includes *Corr loss* to encourage matching temporal waveform shape rather than absolute amplitude or baseline level. The full description is detailed in Appendix B.1. The complete objective is

$$\mathcal{L}_{\text{tok}} = \mathcal{L}_{\text{vq}} + \mathcal{L}_{\text{amp}} + \mathcal{L}_{\text{phase}} + \lambda_{\text{wav}}\mathcal{L}_{\text{wav}}. \tag{2}$$

Duo Codebook therefore learns one vocabulary under supervision from both time and frequency descriptions of each patch. More technical details such as STFT coefficient selection, target normalization, decoder settings, and waveform loss are given in Appendix B.1.

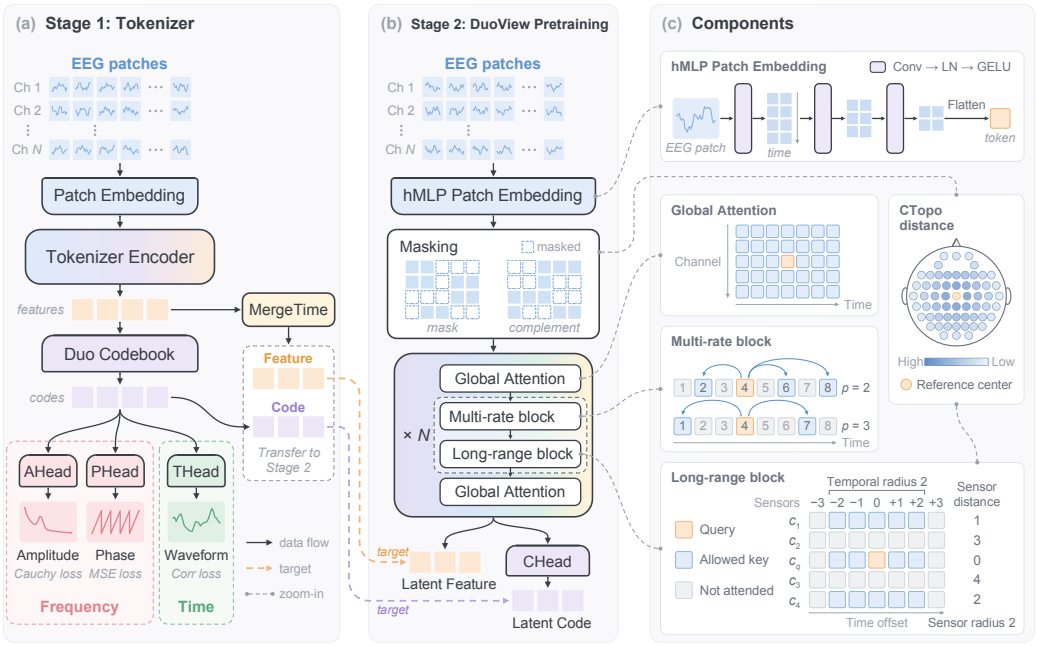

Figure 2: Training pipeline and components of BrainDuo. (a) The Tokenizer learns Duo Codebook through Amplitude, Phase, and Waveform reconstruction, providing the Continuous embedding and Discrete code. (b) DuoView Pretraining transfers both targets to a backbone with Masking and the mask/complement inputs, Global Attention, the Multi-rate block, and the Long-range block. (c) Component illustrations show hMLP Patch Embedding, attention patterns, and CTopo distance.

3.3 DuoView Pretraining: Cyclic and Clustered Context

hMLP Patch Embedding. To extract rich temporal features from EEG patches, we use a hierarchical MLP (hMLP) (Touvron et al., 2022) with three Conv $\rightarrow$ LN $\rightarrow$ GELU stages as the backbone.

CTopo distance. To guide neighborhood-based pre-training backbone and masking, we design channel topological (CTopo) distance as the shortest-path hop distance between electrodes on a canonical sensor layout. For each recording, we select the corresponding rows and columns, preserving canonical distances across channel configurations. Unlike conventional region-based distance, CTopo preserves physical sensor proximity across anatomical brain-region boundaries, defining neighborhoods for cross-region information flow in the long-range block and masking.

Cluster Masking To promote learning of local spatiotemporal interactions, we introduce cluster masking. A cluster comprises EEG patches within predefined CTopo and temporal distances from a sampled patch:

$$\mathcal{N}(c,a) = \{(c',a') : |a-a'| \le r_t,\ D_S[c,c'] \le r_c\}\,, \tag{3}$$

where (c,a) denotes a sampled patch, and r_c and r_t denote the channel and temporal radii, respectively. We randomly sample cluster centers and accumulate their neighborhoods until reaching the target mask count, with excess patches randomly removed. The DuoView pretraining backbone processes the masked patches and their complement in two shared-weight passes. Each pass predicts its hidden patches from the visible tokens. This design captures local interactions complementary to global EEG representations.

Multi-rate block. The multi-rate block (M) decouples temporal and spatial attention into separate heads within attention layers to capture multi-scale spatiotemporal interactions. Temporal heads use cyclic temporal groups defined by two periods: for a given period p, patches within the same channel interact when their time indices share a residue modulo p. Spatial heads connect channels at the same patch time. Different periods define cyclic temporal groups at multiple intervals, allowing

each patch to gather information from distant tokens without being restricted to a single cyclic pattern. The temporal- and spatial-head outputs are concatenated and jointly projected.

Long-range block. In parallel with the multi-rate block, the long-range block (L) attends within the clusters defined above to capture spatiotemporal dependencies. Learned query–key scores determine attention weights within these neighborhoods. This design integrates information beyond the decoupled temporal and spatial interactions of the multi-rate block.

3.4 DUO LATENT: DUAL SUPERVISION

To transfer tokenizer information beyond discrete codes, we propose continuous embedding supervision in Duo Latent (Fig. 2). The frozen tokenizer provides two types of latent code per EEG patch: a discrete code $k_{c,a}$ and a pre-quantization continuous embedding $\boldsymbol{h}^{\mathrm{tok}}_{c,a}$. For cluster-masked prediction, CHead predicts code probabilities for hidden patches. The code prediction loss sums the mean cross-entropies over the mask M and its complement $\overline{M}$:

$$\mathcal{L}_{\mathrm{code}} = -\sum_{v\in\{M,\overline{M}\}} \frac{1}{|v|} \sum_{(c,a)\in v} \log p^{v}_{c,a}(k_{c,a}). \tag{4}$$

MergeTime averages embeddings across temporal patches within each channel to reduce patch-level variability while preserving channel-specific information. Applied to backbone features and tokenizer embeddings, it yields $\overline{\boldsymbol{u}}_c$ and $\overline{\boldsymbol{h}}_c^{\,\mathrm{tok}}$, respectively, for continuous alignment:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{CD}\sum_{c=1}^{C} \left\|\overline{\boldsymbol{u}}_c - \mathrm{sg}(\overline{\boldsymbol{h}}_c^{\,\mathrm{tok}})\right\|_2^2, \qquad \mathcal{L}_{\mathrm{pre}} = \mathcal{L}_{\mathrm{code}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}, \tag{5}$$

where C and D denote the channel count and embedding dimension, sg denotes stop-gradient, and λ_{align} weights alignment in the pretraining loss. This dual supervision encourages EEG representations that capture both code-level structure and finer-grained feature information.

3.5 BCI SKILLS

We introduce a BCI-skills strategy that integrates task-specific prior knowledge into the transfer of pretrained EEG representations. This strategy incorporates four BCI training skills into downstream pipelines: 1) Filter Bank 2) Task prior channel selection 3) Denoising 4) Grand Average. Dataset-specific settembedding designs for downstream EEG decoding.ings are detailed in Appendix C.1.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Pretraining and downstream evaluation. We resample multichannel EEG to 200 Hz and divide it into 1-s patches, then perform Duo Codebook learning followed by DuoView pretraining with a 12-layer backbone. We evaluate downstream transfer across 10 tasks and 14 datasets through two tracks, assessing broad transfer on ten datasets in the general EEG track and performance in the BCI track. Table 1 summarizes both tracks' configuration.

Metrics and reporting. We report balanced accuracy (BA), Cohen's κ, and weighted F1 (F1-w) for multiclass EEG classification; BA, AUC-PR, and AUC for binary tasks; accuracy for SSVEP; and Top-k accuracy for visual retrieval. The specialist-facing emotion evaluations additionally report accuracy. Unless otherwise stated, model results are means and standard deviations over five fine-tuning seeds. Main comparison tables show representative metrics, while Appendix C gives every available metric and uncertainty. FACED-9 and SEED-V appear in both evaluation tracks with different protocols. Results are compared within these protocols.

Table 1: **Overview of downstream EEG datasets and evaluation contracts.** The upper block is the general EEG foundation-model benchmark and the lower block is the specialist-facing BCI evaluation. A dataset is listed twice when the two tracks use different windows or splits.

Task	Dataset / protocol	Rate (Hz)	#Channels	Duration	#Examples	#Subjects	Target
General EEG foundation-model benchmark							
Motor imagery classification	BCIC-IV-2a, fixed split	250	22	4 s	5,184	9	4-class
	PhysioNet-MI, fixed split	160	64	4 s	9,837	109	4-class
Emotion recognition	FACED-9, fixed subjects	250	30	10 s	10,332	123	9-class
	SEED-V, within subject	1000	62	1 s	117,744	16	5-class
Event type classification	TUEV, official split	256	23	5 s	112,237	370	6-class
Abnormality detection	TUAB, official evaluation	256	23	10 s	409,083	2,328	2-class
Sleep staging	ISRUC-S1, fixed subjects	200	6	20×30 s	4,462	100	5-class
Seizure detection	CHB-MIT, fixed cases	256	16	10 s	345,786	23	2-class
Depression classification	Mumtaz2016, fixed patients	256	19	5 s	7,143	63	2-class
Mental arithmetic classification	MentalArithmetic, fixed subjects	500	19	5 s	1,707	36	2-class
Specialist-facing BCI evaluation							
SSVEP classification	Benchmark, LOSO and LBO	250	9–21	1/2 s	8,400	35	40-class
	BETA, LOSO and LBO	250	9–21	1/2 s	11,200	70	40-class
Emotion recognition	SEED-V, subject LOSO	1000	62	4 s	29,168	16	5-class
	SEED-IV, subject LOSO	1000	62	4 s	6,878	15	4-class
	FACED-9, subject CV10	250	30	1 s	103,320	123	9-class
Visual retrieval	THINGS-EEG, inter/intra subject	250	17	1 s	16,740	10	200-way
Motor imagery classification	PhysioNet-MI, fixed split	160	64	4 s	9,837	109	4-class
	BCIC-IV-2a, fixed split	250	22	4 s	5,184	9	4-class

Table 2: **Controlled transfer across ten EEG tasks.** For compact reporting, we show two representative metrics per dataset. Bold and underline denote the best and second-best mean in each row, respectively; "–" denotes an unavailable result.

Dataset	Metric	EEGNet	Conformer	ST-Trans.	BENDR	BIOT	EEGPT	LaBraM	CBraMod	CSBrain	CodeBrain	BrainDuo
BCIC-IV-2a	BA	.4482	.4696	.4575	–	.4748	–	.4869	.5138	<u>.5657</u>	–	**.5715**
	F1-w	.4226	.4533	.4471	–	.4607	–	.4758	.4984	**.5637**	–	<u>.5582</u>
PhysioNet-MI	BA	.5814	.6049	.6035	–	.6153	–	.6173	.6174	<u>.6304</u>	–	**.6517**
	F1-w	.5796	.6062	.6053	–	.6158	–	.6177	.6179	<u>.6308</u>	–	**.6424**
SEED-V	BA	.2961	.3537	.3052	.2231	.3837	.3061	.3976	.4091	<u>.4197</u>	.4137	**.4349**
	F1-w	.2749	.3487	.2833	.2026	.3856	.3090	.3974	.4101	<u>.4280</u>	.4235	**.4447**
FACED-9	BA	.4090	.4559	.4810	.5320	.5118	.4607	.5273	.5509	.5752	<u>.5941</u>	**.6172**
	F1-w	.4124	.4514	.4795	.5340	.5136	.3924	.5288	.5618	.5796	<u>.5953</u>	**.6061**
TUEV	BA	.3876	.4074	.3984	.4363	.5281	.6232	.6409	.6671	<u>.6903</u>	.6428	**.7242**
	F1-w	.6539	.6983	.6823	.6755	.7492	.8187	.8312	.8342	.8333	<u>.8362</u>	**.8615**
TUAB	BA	.7642	.7758	.7966	.7714	.7959	.8038	.8140	.7891	.8172	<u>.8294</u>	**.8307**
	AUROC	.8412	.8445	.8707	.8426	.8815	.8811	.9022	.8606	.8957	<u>.9030</u>	**.9126**
MentalArithmetic	BA	.6770	.6805	.6631	.5681	.6875	.5597	.6909	.7256	<u>.7558</u>	.7514	**.7639**
	AUROC	.7321	.7424	.7132	.6248	.7536	.7162	.7721	.7905	.8478	<u>.8707</u>	**.9046**
CHB-MIT	BA	.5658	.5976	.5915	.5609	.7068	.5481	.7075	<u>.7398</u>	.7262	.7273	**.7419**
	AUROC	.8048	.8226	.8237	.8632	.8761	.8438	.8679	.8892	.8915	<u>.8961</u>	**.9806**
ISRUC	BA	.7154	.7400	.7381	.7401	.7527	.4012	.7633	.7865	<u>.7925</u>	.7835	**.8098**
	F1-w	.7513	.7634	.7681	.7569	.7790	.3111	.7810	.8011	.7990	<u>.8020</u>	**.8139**
Mumtaz2016	BA	.9232	.9308	.9135	–	.9358	–	.9409	.9560	<u>.9643</u>	–	**.9791**
	AUROC	.9639	.9702	.9594	–	.9758	–	.9782	.9921	**.9956**	–	<u>.9940</u>

4.2 Downstream performance

General EEG transfer. Table 2 compares BrainDuo with the reported baselines across the ten general EEG datasets. BrainDuo has the highest reported BA on every dataset. Most secondary metrics follow the same trend, with our model ranking second on two metrics.

BCI decoding. Table 3 compares EEG foundation models with task-specialist models under the BCI protocols. BrainDuo achieves the highest accuracy among the compared EFMs in most settings. At 2 s under LOSO, its accuracy exceeds that of MTSNet, a BCI specialist, on both Benchmark and BETA. These comparisons show that the performance gap between foundation and task-specialist models. Section 4.5 examines how BCI skill contributes to these results.

Table 3: **BCI transfer across specialist-facing protocols.** We show two representative rows per evaluation and means on the $[0, 1]$ scale; Bold and underline mark the best and second-best general EEG models; the specialist column is unranked. "–" is unreported.

Dataset / evaluation	Metric	EEGNet	Con-former	ST-Trans.	BIOT	LaBraM	CBraMod	CSBrain	BrainDuo	Task specialist
Benchmark (LOSO; MTSNet)	Acc (1 s)	.5837	.4945	.6158	.3210	.6916	.6298	.6274	**.7482**	.7737
	Acc (2 s)	.7869	.7608	.8006	.5406	.8276	.8243	.8477	**.9042**	.8890
BETA (LOSO; MTSNet)	Acc (1 s)	.4649	.2599	.3939	.3270	.5675	.5375	.4757	**.5905**	.6621
	Acc (2 s)	.6734	.5903	.4228	.5242	.7099	.6932	.6786	**.7949**	.7763
Benchmark (LBO; DNN)	Acc (1 s)	.6290	.5294	.6233	.3099	.7188	.6521	.6694	**.7905**	.9557
	Acc (2 s)	.8275	.7976	.7204	.5423	.8297	.8456	.8726	**.9245**	1.0000
BETA (LBO; DNN)	Acc (1 s)	.4714	.2462	.1525	.3037	.5527	.5188	.4648	**.5907**	.8378
	Acc (2 s)	.6847	.5824	.2641	.5067	.6776	.6909	.6812	**.7924**	1.0000
SEED-V (LOSO, 4 s; PGCN)	Acc	.2908	.2716	.2545	**.3196**	.2727	.2763	.2881	.2921	.6178
	BA	.2717	.2758	.2579	**.3190**	.2636	.2888	.2766	.2858	–
SEED-IV (LOSO, 4 s; PGCN)	Acc	.3429	.3846	.3556	**.3961**	.3691	.3500	.3529	.3777	.7369
	BA	.3350	.3808	.3392	**.3909**	.2524	.3389	.3486	.3714	–
FACED-9 (CV10, 1 s; DAEST)	Acc	.1198	.1173	.1211	.1537	.1741	.1505	.1321	**.1760**	.5930
	BA	.1148	.1118	.1205	.1395	.1687	.1463	.1279	**.1705**	–
THINGS-EEG (inter; BrainHiVE)	Top-1	.1337	.0580	.1515	.0595	.2055	.1682	.2037	**.2385**	.2410
	Top-5	.3680	.2183	.3735	.1860	.5200	.4237	.4788	**.5312**	.5740
THINGS-EEG (intra; BrainHiVE)	Top-1	.1722	.0157	.4515	.1955	.3487	.5077	.3803	**.6480**	.8140
	Top-5	.4262	.0708	.7875	.4765	.8112	**.8967**	.6690	.8938	.9530
PhysioNet-MI (4 s; AGTCNet)	BA	.5814	.6049	.6035	.6153	.6173	.6174	.6304	**.6517**	.6440
	F1-w	.5796	.6062	.6053	.6158	.6177	.6179	.6308	**.6424**	.6434
BCIC-IV-2a (4 s; AGTCNet)	BA	.4482	.4696	.4575	.4748	.4869	.5138	.5657	**.5715**	.6224
	F1-w	.4226	.4533	.4471	.4607	.4758	.4984	**.5637**	.5582	.6190

Table 4: **Complete Duo Codebook and Duo Latent ablations.** The upper panel changes how the Duo Codebook is learned while retaining the same backbone architecture, corpus, and downstream protocol; the lower panel fixes the tokenizer and changes Duo Latent in DuoView Pretraining.

	TUEV			TUAB		
Variant	**BA**	κ	**F1-w**	**BA**	**AUC-PR**	**AUROC**
Duo Codebook and reconstruction targets						
STFT only	.6701±.0044	.6593±.0059	.8198±.0045	.8154±.0075	.9047±.0013	.9042±.0043
Waveform only	.5854±.0057	.5590±.0049	.7772±.0078	.8046±.0058	.8909±.0095	.8928±.0087
FFT only	.6512±.0058	.6407±.0038	.8200±.0093	.8254±.0018	.9108±.0029	.9067±.0013
Shared STFT + waveform (MSE)	.7164±.0066	.7141±.0075	.8576±.0035	.8280±.0016	.9133±.0008	.9114±.0010
Separate STFT / waveform codebooks	.6918±.0073	.6677±.0087	.8333±.0027	.8297±.0024	.9138±.0017	.9117±.0025
Duo Codebook (ours)	**.7242±.0040**	**.7230±.0137**	**.8615±.0124**	**.8307±.0032**	**.9151±.0023**	**.9126±.0070**
Duo Latent supervision in DuoView Pretraining						
Discrete code only	.7074±.0097	.6803±.0078	.8250±.0063	.8202±.0091	.9018±.0038	.9003±.0097
Continuous embedding only	.6385±.0019	.6037±.0040	.7806±.0011	.7710±.0067	.8669±.0097	.8754±.0010
Duo Latent (Dual supervision; ours)	**.7242±.0040**	**.7230±.0137**	**.8615±.0124**	**.8307±.0032**	**.9151±.0023**	**.9126±.0070**

4.3 Duo Codebook and Duo Latent

Table 4 evaluates two design choices: what information an EEG patch representation preserves and how it transfers to pretraining. The upper panel compares reconstruction objectives for codebook learning. The result shows that Duo Codebook outperforms all other codebook designs, indicating that a unified codebook learned through spatiotemporal reconstruction preserves the most useful information for our model. The lower panel fixes the tokenizer and compares the information transferred to the backbone during pretraining. Duo Latent combines discrete code prediction with continuous embedding alignment and achieves the highest mean on all six metrics. Continuous embedding alignment alone performs worse than code prediction alone, but adding it to code prediction improves performance. These results show that using both the discrete codes and continuous embeddings of a shared spectrotemporal vocabulary to supervise pretraining gives the best performance.

4.4 DuoView Pretraining

To assess the contributions of the multi-rate (M) and long-range (L) blocks, we replace the seven middle blocks with only M or only L, retaining the two leading and three trailing global attention

Table 5: DuoView component switches and masking.

	TUEV			TUAB		
Variant	**BA**	κ	**F1-w**	**BA**	**AUC-PR**	**AUROC**
A. block ablation						
M only for middle blocks	0.6853 ± 0.0074	0.6866 ± 0.0042	0.8230 ± 0.0054	0.8160 ± 0.0038	0.9057 ± 0.0025	0.9033 ± 0.0032
L only for middle blocks	0.7194 ± 0.0031	0.7097 ± 0.0028	0.8451 ± 0.0034	0.8278 ± 0.0020	0.9111 ± 0.0033	0.9089 ± 0.0016
B. masking ablation						
Random masking	0.7091 ± 0.0487	0.6836 ± 0.0381	0.8373 ± 0.0201	0.8201 ± 0.0051	0.9044 ± 0.0052	0.9086 ± 0.0050
Montage-distance masking	0.6820 ± 0.0595	0.6417 ± 0.0463	0.8107 ± 0.0258	0.8148 ± 0.0048	0.8991 ± 0.0062	0.8975 ± 0.0051
C. other ablation						
Single convolution (no hMLP)	0.6909 ± 0.0069	0.6725 ± 0.0162	0.8362 ± 0.0075	0.8280 ± 0.0072	0.9079 ± 0.0076	0.9056 ± 0.0080
Without RoPE	0.7132 ± 0.0154	0.7035 ± 0.0158	0.8479 ± 0.0070	$\mathbf{0.8353 \pm 0.0056}$	0.9124 ± 0.0075	0.9098 ± 0.0040
D. Baseline	$\mathbf{0.7242 \pm 0.0040}$	$\mathbf{0.7230 \pm 0.0137}$	$\mathbf{0.8615 \pm 0.0124}$	$\mathbf{0.8307 \pm 0.0032}$	$\mathbf{0.9151 \pm 0.0023}$	$\mathbf{0.9126 \pm 0.0070}$

blocks (Table 5, panel A). We found that the L-only variant achieves higher means than the M-only variant on all six metrics, yet neither matches the combined baseline. For masking, preserving CTopo distance yields the strongest performance (Table 5, panel B). CTopo-based masking achieves the highest mean on all six metrics compared with standard random masking and distance-guided masking using distances recomputed from the electrode montage. In contrast, recomputing montage distances results in poorer performance than either alternative. The ablations in panel C further demonstrate the benefits of hMLP patch embedding and RoPE. Replacing hMLP with a single convolution lowers the means on all six metrics, highlighting the value of hierarchical patch embedding. Removing RoPE also lowers the means on all three TUEV metrics and both TUAB AUC metrics.

4.5 BCI TRAINING SKILLS

We evaluate task-specific BCI skills using BrainDuo. The following experiments separately examine filter bank, task prior channel selection, and denoising; they measure the benefit of these choices for BrainDuo.

- **Filter bank.** Table 26 panel A compares SSVEP decoding with and without filter bank on Benchmark and BETA under the 1-s LOSO and LBO protocols. Filter bank improves mean accuracy in all four settings, showcasing its effectiveness in boosting EFM representations.
- **Task prior channel selection.** Panel B of Table 26 compares performance using all available EEG channels with that using task-selected channels. Task-specific selection improves mean Top-1 accuracy in both inter- and intra-subject THINGS-EEG evaluations and increases SSVEP accuracy. These results suggest that channel priors facilitate more discriminative EFM representations for downstream decoding by reducing spatial redundancy.
- **Denoising.** Figure 5 evaluates filter denoising on FACED-9 dataset. With the lower cutoff at 0.5 Hz, the 15-Hz upper cutoff yields the highest mean on all four metrics, indicating that tailoring spectral bandwidth to BCI tasks during fine-tuning improves EFM representations.
- **Grand Average** Table 19 varies the number of averaged trials for inter- and intra-subject THINGS-EEG retrieval. Increasing the number of averaged trials improves Top-1 accuracy in both settings, with larger gains for intra-subject retrieval, which highlights its potential to improve EFM representations for event-related potential (ERP)-based BCI decoding.

5 CONCLUSION

We presented BrainDuo, a dual-domain EEG foundation model that learns shared spectrotemporal features through Duo Codebook, integrates discrete and continuous supervision through Duo Latent, and captures local and distant spatiotemporal interactions through DuoView Pretraining. Evaluation across 10 tasks and 14 datasets demonstrates strong transfer across general EEG tasks and competitive performance against task-specialist models in several BCI settings, supported by analyses of the representation and pretraining designs. Our downstream ablations further show the value of BCI-native skills in adapting pretrained representations to task-specific signal structure. Together, these findings help bridge the gap between general-purpose EEG representation learning and specialized BCI applications, advancing the development of broadly applicable EEG foundation models.

REFERENCES

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.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jonathan T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4331–4339, 2019.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. Technical Report 1-6, 2008.

Jianbo Chen, Yangsong Zhang, Yudong Pan, Peng Xu, and Cuntai Guan. A transformer-based deep neural network model for SSVEP classification. *Neural Networks*, 164:521–534, 2023a. doi: 10.1016/j.neunet.2023.04.045.

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, 2023b.

Xiaogang Chen, Yijun Wang, Shangkai Gao, Tzyy-Ping Jung, and Xiaorong Gao. Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface. *Journal of Neural Engineering*, 12(4):046008, 2015. doi: 10.1088/1741-2560/12/4/046008.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Wenlong Ding, Jianhua Shan, Bin Fang, Chengyin Wang, Fuchun Sun, and Xinyue Li. Filter bank convolutional neural network for short time-window steady-state visual evoked potential classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2615–2624, 2021. doi: 10.1109/TNSRE.2021.3132162.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/66969a9e6bd7a26dfeccea7227178ca7-Abstract-Conference.html.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/20a917f77773ac0fa8bea2bdd6606b66-Abstract-Conference.html.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Osman Berke Guney, Muhtasham Oblokulov, and Huseyin Ozkan. A deep neural network for SSVEP-based brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 69 (2):932–944, 2022. doi: 10.1109/TBME.2021.3110440. URL https://arxiv.org/abs/2011.08562.

Amir Harati, Meysam Golmohammadi, Silvia Lopez, Iyad Obeid, and Joseph Picone. Improved eeg event classification using differential energy. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4. IEEE, 2015.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47393e8594c82ce8fd83adc672cf9872-Abstract-Conference.html.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html.

Ming Jin, Changde Du, Huiguang He, Ting Cai, and Jinpeng Li. PGCN: Pyramidal graph convolutional network for EEG emotion recognition. *IEEE Transactions on Multimedia*, 26:9070–9082, 2024. doi: 10.1109/TMM.2024.3385676. URL https://doi.org/10.1109/TMM.2024.3385676.

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

Sirvan Khalighi, Teresa Sousa, José M. Santos, and Urbano Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659. URL https://doi.org/10.3389/fnhum.2021.653659.

Zhen Lan, Zixing Li, Chao Yan, Xiaojia Xiang, Dengqing Tang, Min Wu, and Zhenghua Chen. MTSNet: Convolution-based transformer network with multi-scale temporal-spectral feature fusion for SSVEP signal decoding. *IEEE Journal of Biomedical and Health Informatics*, 29(11): 8034–8047, 2025. doi: 10.1109/JBHI.2025.3573410. URL https://ieeexplore.ieee.org/document/11014538/.

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. doi: 10.1088/1741-2552/aace8c.

Na Lee, Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Are large brainwave foundation models capable yet? insights from fine-tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32878–32888. PMLR, 2025. URL https://proceedings.mlr.press/v267/lee25a.html.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ba5f1233efa77787ff9ec015877dbd1f-Abstract-Conference.html.

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery EEG classification algorithm based on CNN–LSTM feature fusion network. *Biomedical Signal Processing and Control*, 72:103342, 2022.

Galvin Brice S. Lim, Brian Godwin S. Lim, Argel A. Bandala, John Anthony C. Jose, Timothy Scott C. Chu, and Edwin Sybingco. AGTCNet: A graph-temporal approach for principled motor imagery EEG classification. *IEEE Access*, 13:187383–187409, 2025. doi: 10.1109/ACCESS.2025.3627419. URL https://arxiv.org/abs/2506.21338.

Bingchuan Liu, Xiaoshan Huang, Yijun Wang, Xiaogang Chen, and Xiaorong Gao. BETA: A large benchmark database toward SSVEP–BCI application. *Frontiers in Neuroscience*, 14:627, 2020.

Bingchuan Liu, Xiaogang Chen, Nanlin Shi, Yijun Wang, Shangkai Gao, and Xiaorong Gao. Improving the performance of individually calibrated SSVEP-BCI by task-discriminant component analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1998–2007, 2021. doi: 10.1109/TNSRE.2021.3114340.

Steven J. Luck. *An Introduction to the Event-Related Potential Technique*. MIT Press, 2 edition, 2014. ISBN 9780262525855.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/86d1ab582afb247ccaa84bec4a7e24f7-Abstract-Conference.html.

Wajid Mumtaz et al. A wavelet-based technique to predict treatment outcome for major depressive disorder. *PLOS ONE*, 2017.

Masaki Nakanishi, Yijun Wang, Xiaogang Chen, Yu-Te Wang, Xiaorong Gao, and Tzyy-Ping Jung. Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis. *IEEE Transactions on Biomedical Engineering*, 65(1):104–112, 2018. doi: 10.1109/TBME.2017.2694818.

Anthony M. Norcia, L. Gregory Appelbaum, Justin M. Ales, Benoit R. Cottereau, and Bruno Rossion. The steady-state visual evoked potential in vision research: A review. *Journal of Vision*, 15 (6):4, 2015. doi: 10.1167/15.6.4.

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

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. 2024:1862–1874, 2024.

Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Tokenizing single-channel EEG with time-frequency motif learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3dc85735f6e2fcf093e67b134fa00d21-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: A general-purpose brain–computer interface system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. URL https://arxiv.org/abs/1610.02391.

Xinke Shen, Lingyi Tao, Xuyang Chen, Sen Song, Quanying Liu, and Dan Zhang. Contrastive learning of shared spatiotemporal EEG representations across individuals for naturalistic neuroscience. *NeuroImage*, 301:120890, 2024. doi: 10.1016/j.neuroimage.2024.120890. URL https://arxiv.org/abs/2402.14213.

Xinke Shen, Runmin Gan, Kaixuan Wang, Shuyi Yang, Qingzhu Zhang, Quanying Liu, Dan Zhang, and Sen Song. Dynamic-attention-based EEG state transition modeling for emotion recognition. *IEEE Transactions on Affective Computing*, 16(4):3552–3568, 2025. doi: 10.1109/TAFFC.2025.3593630. URL https://arxiv.org/abs/2411.04568.

Ali Hossam Shoeb. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial–temporal feature learning for EEG decoding. *arXiv preprint arXiv:2106.11170*, 2021. URL https://arxiv.org/abs/2106.11170.

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, 2023. doi: 10.1109/TNSRE.2022.3230250.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d0865cbe51d35ec322f9af9db7806fc7-Abstract-Conference.html.

Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about Vision Transformers. In *Computer Vision – ECCV 2022*, volume 13684 of *Lecture Notes in Computer Science*, pp. 497–515. Springer, 2022. doi: 10.1007/978-3-031-20053-3_29. URL https://doi.org/10.1007/978-3-031-20053-3_29.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL https://jmlr.org/papers/v9/vandermaaten08a.html.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbbd6d915cb90be21c1254a82d45cedd-Abstract-Conference.html.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for SSVEP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2017.

Xinxu Wei, Kanhao Zhao, Yong Jiao, Hua Xie, Lifang He, and Yu Zhang. Pre-training graph contrastive masked autoencoders are strong distillers for EEG. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66358–66377. PMLR, 2025. URL https://proceedings.mlr.press/v267/wei25q.html.

Xun Wu, Wei-Long Zheng, and Bao-Liang Lu. Investigating EEG-based functional connectivity patterns for multimodal emotion recognition. *arXiv preprint arXiv:2004.01973*, 2020.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of EEG foundation models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/60932. Main conference; PMLR pagination pending at literature cutoff.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract.html.

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, 2023b.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f0f39b7686634fc81ca0112566b2c05f-Abstract-Conference.html.

Huiming Yao, Ke Liu, Xin Deng, Xianlun Tang, and Hong Yu. FB-EEGNet: A fusion neural network across multi-stimulus for SSVEP target detection. *Journal of Neuroscience Methods*, 379:109674, 2022. doi: 10.1016/j.jneumeth.2022.109674.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c5506f154e5f2203a7dca2bee2faeced-Abstract-Conference.html.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49 (3):1110–1122, 2019.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7e199ad8ae40eb19b2980f61f659cb07-Abstract-Conference.html.

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

A Data and Evaluation Protocols

This appendix gives shared data and evaluation settings (Appendix A), pretraining details and training dynamics (Appendix B), and task-specific protocols and complete results (Appendix C). Additional model ablations, pretraining and scaling studies, and qualitative analyses follow in Appendices D–F.

A.1 Pretraining corpus and segmentation

The pretraining corpus combines 9,000 hours of Temple University (TU) EEG with INRIA BCI, PhysioNet Motor Movement/Imagery (PhysioNet-MI), and SEED-IV data. Only training partitions

contribute to pretraining; downstream validation and test recordings are excluded. Recordings are resampled to 200 Hz and converted to 30-s examples. Each example is divided into thirty non-overlapping 1-s patches, so a recording with C retained electrodes produces a $C \times 30$ patch grid. The tokenizer uses a 20% validation partition of the pretraining corpus for checkpoint selection.

A.2 Downstream datasets and evaluation

Table 1 in the main paper summarizes both evaluation tracks. Its sampling rates describe the source recordings; all EEG foundation-model inputs are resampled to 200 Hz. Controlled comparisons share task-level input construction and evaluation protocols. Dataset-specific filtering, channel selection, window construction, splits, heads, and complete results are collected with each task in Appendix C. Published specialist references retain their documented native pipelines.

Unless noted otherwise, $\pm$ denotes the sample standard deviation over five fine-tuning seeds using one pretrained backbone. Published specialist results retain their source uncertainty definitions.

A.3 Metrics

Balanced accuracy (BA). Balanced accuracy averages recall across classes, giving each class equal weight regardless of its frequency in the test set:

$$\mathrm{BA} = \frac{1}{C} \sum_{c=1}^{C} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \tag{6}$$

where C is the number of classes, and TP_c and FN_c denote true positives and false negatives for class c. This prevents larger classes from dominating the metric.

Weighted F1 (F1-w). Weighted F1 combines classwise precision and recall through their harmonic mean and weights each class by its test-set support:

$$\mathrm{F1_w} = \sum_{c=1}^{C} \frac{n_c}{N} \mathrm{F1}_c, \qquad \mathrm{F1}_c = \frac{2\mathrm{TP}_c}{2\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{7}$$

where n_c is the number of test examples belonging to class c, $N = \sum_{c=1}^{C} n_c$, and FP_c denotes false positives for class c. Unlike balanced accuracy, F1-w gives greater weight to more frequent classes.

Cohen's κ. Cohen's κ measures agreement between predicted and true labels after correcting for the agreement expected by chance under their empirical class distributions:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_o = \frac{1}{N} \sum_{c=1}^{C} \mathrm{TP}_c, \qquad p_e = \sum_{c=1}^{C} \frac{n_c}{N} \frac{\widehat{n}_c}{N}, \tag{8}$$

where p_o is the observed agreement, p_e is the expected agreement, and $\widehat{n}_c$ is the number of examples predicted as class c. A value of 1 indicates perfect agreement, while 0 indicates agreement equal to the chance expectation. **Area under the precision–recall curve (AUC-PR).** AUC-PR summarizes the trade-off between precision and recall as the decision threshold varies over continuous positive-class scores:

$$\text{AUC-PR} = \int_0^1 P(r)\, \mathrm{d}r, \tag{9}$$

where $P(r)$ denotes precision along the precision–recall curve at recall r. At threshold t, precision and recall are $P(t) = \mathrm{TP}(t)/[\mathrm{TP}(t) + \mathrm{FP}(t)]$ and $R(t) = \mathrm{TP}(t)/[\mathrm{TP}(t) + \mathrm{FN}(t)]$. The metric summarizes positive-class detection across thresholds rather than at a single operating point.

Area under the receiver operating characteristic curve (AUROC). AUROC measures how well continuous positive-class scores distinguish positive from negative examples across decision thresholds:

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(u)\, \mathrm{d}u, \tag{10}$$

where u denotes the false positive rate and $\mathrm{TPR}(u)$ is the corresponding true positive rate along the ROC curve. At threshold t, these rates are $\mathrm{TPR}(t) = \mathrm{TP}(t)/[\mathrm{TP}(t) + \mathrm{FN}(t)]$ and $\mathrm{FPR}(t) =$

$\mathrm{FP}(t)/[\mathrm{FP}(t) + \mathrm{TN}(t)]$, with $\mathrm{TN}(t)$ denoting true negatives. **Classification accuracy (ACC).** Classification accuracy measures the proportion of test examples whose predicted labels match their true labels:

$$\mathrm{ACC} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\widehat{y}_i = y_i], \tag{11}$$

where y_i and $\widehat{y}_i$ are the true and predicted labels for example i, and $\mathbf{1}[\cdot]$ is the indicator function. Each example contributes equally to the metric.

Top-k retrieval accuracy. Top-k retrieval accuracy measures the proportion of queries for which the correct target appears among the k highest-ranked candidates:

$$\text{Top-}k = \frac{1}{Q} \sum_{i=1}^{Q} \mathbf{1}[r_i \leq k], \tag{12}$$

where Q is the number of test queries and r_i is the rank of the correct target for query i when candidates are ordered by their retrieval scores. The values of k and the candidate gallery follow the evaluation protocol specified for each dataset.

A.4 Baselines

Task-specific models:

- **EEGNet** (Lawhern et al., 2018) is a lightweight CNN for EEG decoding across different BCI paradigms. It uses temporal convolutions to learn frequency-selective filters, depthwise convolutions to extract spatial features, and separable convolutions to combine these features.
- **EEGConformer** (Song et al., 2023) combines CNN and Transformer modules for EEG decoding. The CNN extracts local spatiotemporal features, while the Transformer captures long-range temporal dependencies through self-attention.
- **ST-Transformer** (Song et al., 2021) applies spatial and temporal attention for EEG classification. Its feature-channel attention captures spatial relationships, while temporal self-attention models dependencies between signal segments.

General EEG models:

- **BENDR** (Kostas et al., 2021) combines a convolutional encoder with a Transformer to learn representations from unlabeled EEG. It uses masked contrastive learning to distinguish the latent representations of masked segments from distractors before adaptation to downstream tasks.
- **BIOT** (Yang et al., 2023a) is a biosignal Transformer designed for learning across datasets with different channel configurations, recording lengths, and missing segments. It converts channel-wise signal segments into tokens with spectral, channel, and temporal position information and processes them using linear attention.
- **EEGPT** (Wang et al., 2024) uses a hierarchical Transformer to learn EEG representations through separate spatial and temporal processing. Its pretraining objective combines masked EEG reconstruction with alignment to continuous representations produced by a momentum encoder.
- **LaBraM** (Jiang et al., 2024) learns a vector-quantized neural tokenizer that assigns discrete codes to EEG channel patches through spectrum prediction. It then pretrains a Transformer to predict the codes of masked patches and fine-tunes the learned representations for downstream EEG tasks.
- **CBraMod** (Wang et al., 2025) uses a criss-cross Transformer with parallel spatial and temporal attention for masked EEG reconstruction. It combines time- and frequency-domain patch features with asymmetric conditional positional encoding to accommodate different channel configurations and recording lengths.

Algorithm 1 BrainDuo pretraining

Input: EEG patches X, electrode identities S, CTopo distances $D^\star$.

Duo Codebook learning

for each minibatch X **do**
  $(h^{\mathrm{tok}}, k, q) \leftarrow \mathrm{Tokenizer}(X)$.
  Decode q through AHead, PHead, and THead.
  $\mathcal{L}_{\mathrm{tok}} \leftarrow \mathcal{L}_{\mathrm{vq}} + \mathcal{L}_{\mathrm{amp}} + \mathcal{L}_{\mathrm{phase}} + 0.1\mathcal{L}_{\mathrm{wav}}$.
  Update tokenizer with straight-through gradients; update codebook by EMA.
end for; select and freeze the tokenizer.

DuoView Pretraining with Duo Latent

for each minibatch (X, S) **do**
  Obtain frozen targets (k, h^{tok}) from uncorrupted X.
  $D_S \leftarrow D^\star[S, S]$; $M \leftarrow \mathrm{ClusterMask}(X, D_S)$; $\overline{M} \leftarrow \mathrm{complement}(M)$.
  $U^M \leftarrow F_\theta(X; M)$; $U^{\overline{M}} \leftarrow F_\theta(X; \overline{M})$.
  $\mathcal{L}_{\mathrm{code}} \leftarrow \mathrm{CE}_M(\mathrm{CHead}(U^M), k) + \mathrm{CE}_{\overline{M}}(\mathrm{CHead}(U^{\overline{M}}), k)$.
  $\mathcal{L}_{\mathrm{align}} \leftarrow \mathrm{MSE}(\mathrm{MergeTime}(U^M), \mathrm{MergeTime}(h^{\mathrm{tok}}))$.
  Update backbone and CHead using $\mathcal{L}_{\mathrm{code}} + 0.5\mathcal{L}_{\mathrm{align}}$.
end for; **return** pretrained backbone F_θ.

h^{tok} is the embedding before quantizer projection; $\mathcal{L}_{\mathrm{wav}}$ is Corr loss. F_θ applies each mask to fresh patch embeddings. CE_V averages over hidden positions; MergeTime averages patch times within each electrode.

- **CSBrain** (Zhou et al., 2025) learns cross-scale spatiotemporal representations by aggregating EEG features within temporal windows and anatomically defined brain regions. It combines this tokenization with structured sparse attention across windows and regions and pretrains the model through masked EEG reconstruction.
- **CodeBrain** (Ma et al., 2026) uses a TFDual-Tokenizer to learn separate temporal and frequency codebooks from EEG signals. Its EEGSSM backbone combines structured global convolution with sliding-window attention and is pretrained to predict discrete codes for masked EEG patches.

B Additional Pretraining and Implementation Details

B.1 Duo Codebook

Duo Codebook is learned by the Tokenizer through waveform and STFT reconstruction. For each waveform patch $x \in \mathbb{R}^{200}$, it constructs an STFT with a 64-sample Hann window, hop 8, and a one-sided spectrum. The magnitude and phase arrays are flattened in implementation order and cropped to the decoder width of 200. Each target is standardized per example over its sensor, patch, and feature axes. AHead reconstructs Amplitude (STFT magnitude) using Cauchy loss,

$$\mathcal{L}_{\mathrm{amp}} = \frac{1}{N} \sum_i \log\bigl(1 + (\widehat{m}_i - m_i)^2\bigr)\,, \tag{13}$$

while PHead reconstructs Phase using MSE loss. THead reconstructs Waveform with the Corr loss

$$\mathcal{L}_{\mathrm{wav}} = 1 - \mathrm{corr}(\widehat{x}, x). \tag{14}$$

The total tokenizer loss is $\mathcal{L}_{\mathrm{vq}} + \mathcal{L}_{\mathrm{amp}} + \mathcal{L}_{\mathrm{phase}} + 0.1\mathcal{L}_{\mathrm{wav}}$.

The Tokenizer Encoder and decoder operate on the same electrode–patch grid. Both use width 200; the encoder has 12 Transformer blocks and 10 heads, and the decoder has three blocks. A two-layer projection maps encoder tokens to a 64-dimensional normalized latent. Nearest-code assignment is by cosine similarity to an 8,192-entry codebook. The normalized EMA quantizer uses decay 0.99, commitment weight 1, and no k-means initialization in the selected run. Thus waveform and STFT targets update one bottleneck rather than two separately indexed vocabularies.

Table 6: Duo Codebook architecture and optimization recipe. The effective peak learning rate accounts for linear scaling from the configured reference batch size.

Component	Setting
Input / patch	30 s at 200 Hz / 1 s (200 samples)
Encoder / decoder	12 / 3 blocks; width 200; 10 heads
Vocabulary	8,192 codes; 64 dimensions; cosine assignment
EMA quantizer	decay 0.99; commitment 1; no k-means initialization
STFT	$n_{\mathrm{fft}} = 64$, Hann 64, hop 8, centered
Targets	standardized linear magnitude, phase, and waveform
Loss weights	VQ 1; magnitude 1; phase 1; waveform 0.1
Configured optimization	AdamW; batch 48; maximum 100 epochs; weight decay 10^{-4}
Learning rate	effective peak 3.75×10^{-5}; minimum 10^{-5}
Validation	20% tokenizer partition

B.2 hMLP Patch Embedding and positional information

The backbone receives raw 200-sample patches, not the decoded STFT arrays. The hMLP Patch Embedding adapts the hierarchical MLP stem of Touvron et al. (2022) to one-dimensional EEG patches. The temporal stem has three convolutional layers with feature-channel widths $1 \to 8 \to 8 \to 8$, kernel $(1, 3)$, stride $(1, 2)$, and padding $(0, 1)$, each followed by LayerNorm over the remaining temporal dimension and GELU. The same stem is applied independently to every electrode–time patch, without mixing different patches during embedding. The temporal extent contracts $200 \to 100 \to 50 \to 25$ and is flattened to $8 \times 25 = 200$ dimensions. Electrode identity and patch time are added as separate positional embeddings; RoPE with base 10,000 is also applied inside attention.

B.3 CTopo distance

The base of this distance is an undirected international 10–20/10–10 sensor layout. Channel labels are case-normalized and legacy aliases are retained in the registry. We symmetrize every listed edge and run breadth-first search on the master graph to obtain all-pairs CTopo distance $D^\star$, measured in hops. For an observed ordered channel list $S = (s_1, \ldots, s_C)$, the model gathers $D_S[i, j] = D^\star[s_i, s_j]$. Paths are computed on the master layout before restriction, so removing an unobserved intermediate electrode does not change the distance between two retained electrodes. Unsupported labels receive a distance larger than the maximum permitted radius and are excluded by the pretraining channel filter. CTopo distance is a channel-layout metric, not a functional-connectivity graph. It determines which token pairs are eligible in the Long-range block and which patches can enter the same corruption neighborhood; learned query–key scores still determine the relative weight of eligible pairs.

B.4 DuoView Pretraining and block composition

The 12-block DuoView Pretraining backbone has width 200 and 10 heads. We denote Global Attention, the Multi-rate block, and the Long-range block by G, M, and L, respectively. Blocks follow the sequence

$$\mathtt{G\ G\ M\ L\ M\ L\ M\ L\ M\ G\ G\ G}. \tag{15}$$

In Global Attention, all tokens are eligible to interact. Each Multi-rate block allocates five heads to within-electrode temporal interaction and five heads to same-time cross-electrode interaction. Of the temporal heads, three use time-index residue classes modulo 2 and two use residue classes modulo 3, implementing the Cyclic temporal grouping. Each Long-range block assigns all 10 heads to a joint neighborhood with temporal radius 7 patches and CTopo-distance radius 2 hops. In the Multi-rate and Long-range blocks, the class token only self-attends; the surrounding Global Attention blocks exchange information between the class token and patch tokens.

B.5 Cluster Masking

Masking is guided by temporal and CTopo distances. The selected mask uses temporal radius one and CTopo-distance radius two, which need not equal the neighborhood radii of the Long-range block. The mask ratio is 0.5. The implementation is:

Table 7: DuoView Pretraining architecture and optimization recipe.

Component	Setting
Encoder	12 blocks; width 200; 10 heads
hMLP Patch Embedding	three stride-2 temporal convolutions; output 200
Program	GGMLMLMLMGGG
Multi-rate block heads	5 temporal + 5 channel; temporal modulo 2/3
Long-range block support	temporal radius 7; CTopo-distance radius 2
Mask support	temporal radius 1; CTopo-distance radius 2
Mask ratio	0.5
Targets	Discrete code + Continuous embedding
Alignment	channelwise temporal mean; MSE; weight 0.5
Optimization recipe	AdamW; batch 80; 30 epochs; peak LR 5×10^{-4}
Weight decay	0.05

1. randomly permute all electrode–time anchor locations;
2. visit an anchor if its center has not already been masked;
3. add every token within one patch in time and at most two hops in CTopo distance of that anchor;
4. stop once at least half of the tokens are masked; and
5. if the last neighborhood overshoots the target count, randomly unmask excess positions to obtain the exact ratio.

The resulting mask is a union of local neighborhoods and need not be one connected component. The montage-recomputed ablation changes only the distance used by this procedure; it is not a learned-graph experiment.

B.6 Duo Latent and Dual supervision

Duo Latent transfers Discrete code and Continuous embedding from the frozen Tokenizer for Dual supervision. Let M denote a sampled mask. The student is evaluated on M and its complement. CHead produces Latent Code predictions, and cross-entropy is accumulated against the tokenizer's Discrete code at every position hidden in the corresponding view. The two views therefore supervise all positions while preventing a target patch from being visible in the view that predicts it. For continuous transfer, the student's Latent Feature from the first masked pass is mean-pooled over time for each electrode and matched by MSE to the Continuous embedding from the uncorrupted Tokenizer. The teacher embedding is taken before its 64-dimensional quantizer projection; MergeTime averages it over time per electrode. Continuous embedding alignment requires no additional projection. The complete objective is

$$\mathcal{L}_{\text{pre}} = \text{CE}_M + \text{CE}_{\overline{M}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \qquad \lambda_{\text{align}} = 0.5. \tag{16}$$

B.7 Training dynamics

Duo Codebook. Figure 3 presents reconstruction losses and codebook usage over tokenizer epochs. Waveform reconstruction loss decreases from 0.3681 to 0.2597, while amplitude and phase reconstruction losses decrease from 0.0830 to 0.0631 and from 0.8272 to 0.6897, respectively. The reconstruction objectives therefore improve jointly during optimization of the shared vocabulary. Their magnitudes reflect different objectives and should not be compared directly across panels. The number of unused entries during training falls from 4,852 to zero at epoch and remains zero over the subsequent recorded epochs. Validation unused-code counts also fall from 6,264 to zero by the end of the trace. These counts describe codebook coverage on the respective passes. Reconstruction panels use training components only.

DuoView Pretraining and Duo Latent. Figure 4 separates the recorded backbone objective into Discrete code prediction and Continuous embedding alignment. The total and prediction losses decline throughout the trace; alignment decreases sharply at first and then varies within a narrow range before reaching its final value. The prediction loss is the sum over the two masking views, and the total equals the sum of the prediction and alignment components.

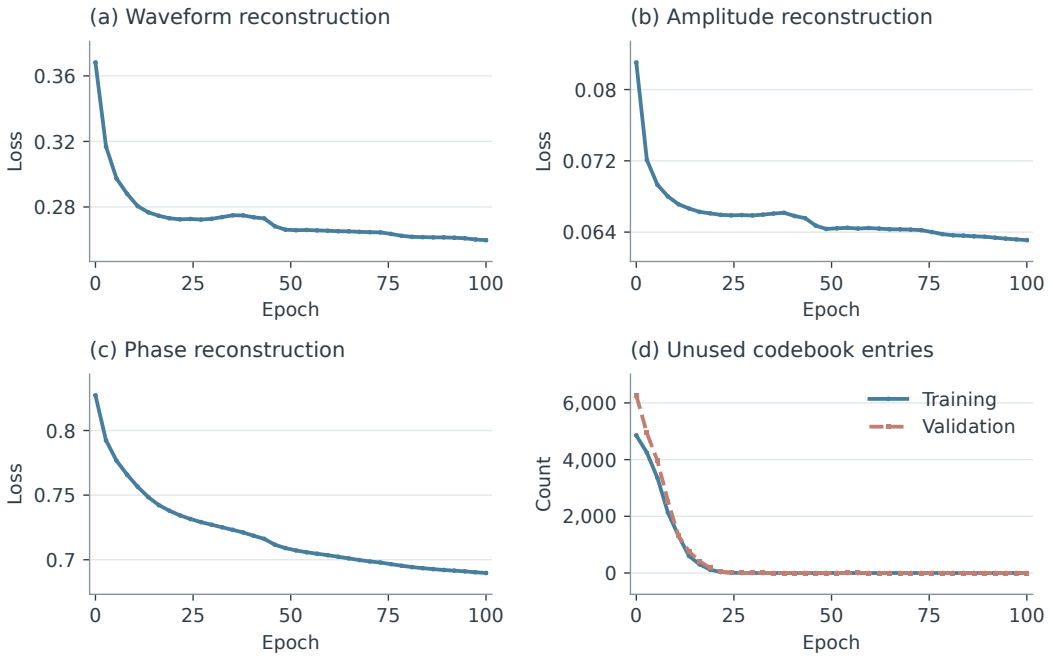

Figure 3: Duo Codebook training dynamics. Panels (a–c) show the recorded training losses for waveform, amplitude, and phase reconstruction, each on its own vertical scale. Waveform loss is shown before its objective weight of 0.1 is applied. Panel (d) shows the logged numbers of unused codebook entries during training and validation.

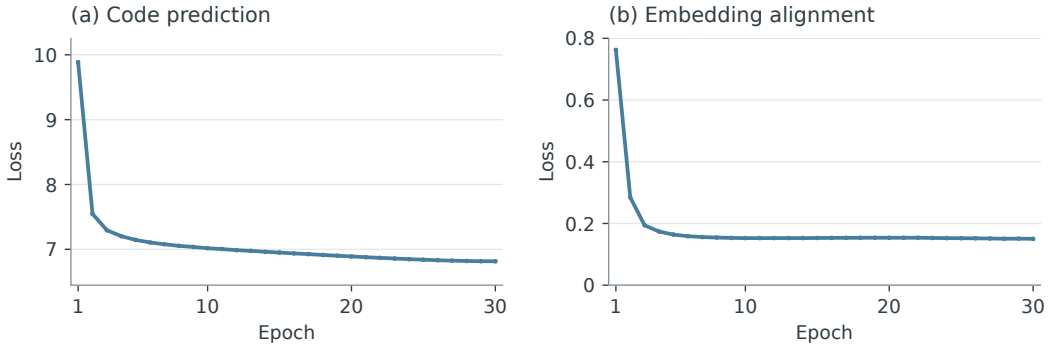

Figure 4: DuoView Pretraining with Duo Latent's Dual supervision. Panels show the Discrete code prediction loss, and Continuous embedding alignment loss. Each panel uses its own vertical scale.

C Downstream Experimental Details and Results

This section provides the complete downstream results underlying the condensed comparisons in the main paper. Following the task-by-task organization of prior EEG foundation-model evaluations, each dataset description specifies the signal construction, evaluation split, and task-native metrics before interpreting the corresponding table. Results from the broad-transfer benchmark and protocol-specific specialist comparisons are identified separately whenever their inputs or evaluation units differ. The task descriptions cover the Motor BCI, Visual BCI, Affective BCI, Clinical BCI, and Cognitive BCI categories in the overview figure while retaining each task's specific input construction and evaluation protocol.

The encoder produces a channel–time grid of patch representations $H \in \mathbb{R}^{C\times A\times D}$ with $D = 200$. Classification inputs sampled at 200 Hz are partitioned into 200-sample patches, so a T-second window normally gives A patches per channel. The task head specifies how this grid is pooled or flattened and maps it to K class logits, or a single logit for binary classification. Binary tasks use binary cross-entropy with logits; multiclass tasks use cross-entropy with the configured label smoothing. The backbone and head are fine-tuned jointly from one selected pretrained backbone. Each dataset pipeline supplies its channel identities and temporal input construction, and only the task head is replaced. Model selection uses validation data, followed by evaluation on the held-out test partition. The descriptions below restore the supplied task-specific adaptation recipes, supplemented by head definitions in the available implementation. Table 8 records their optimization settings.

Table 8 restores the per-task optimization settings from the supplied adaptation specification. Its shared settings are AdamW, cosine learning-rate decay, weight decay 0.05, and stochastic depth 0.1. Multiclass recipes use label smoothing 0.1 unless disabled in the run configuration; binary tasks use binary cross-entropy without multiclass label smoothing.

Table 8: Recorded task-specific fine-tuning configurations.

Dataset / protocol	Learning rate	Batch	Epochs	Warm-up	Layer decay	Dropout	EMA decay	Validation selection
PhysioNet-MI	5×10^{-4}	32	50	5	0.65	0.1	–	Cohen's κ
BCIC-IV-2a	1×10^{-4}	64	50	2	0.65	0.1	–	Balanced accuracy
SEED-V, 1-s broad	2×10^{-4}	64	50	5	0.65	0.1	–	Cohen's κ
SEED-V, 4-s LOSO	5×10^{-4}	64	30	5	0.65	0.1	0.999	Cohen's κ
SEED-IV, 4-s LOSO	2×10^{-4}	64	20	5	0.65	0.1	0.999	Balanced accuracy
FACED-9, 10-s broad	2×10^{-3}	64	80	5	0.9	0.0	0.9999	Balanced accuracy
FACED-9, 1-s CV10	5×10^{-4}	128	20	4	0.9	0.0	0.9999	Balanced accuracy
Benchmark / BETA, 1 s	5×10^{-4}	64	60	5	0.65	0.1	0.999	Accuracy
Benchmark / BETA, 2 s	5×10^{-4}	64	30	5	0.65	0.1	0.999	Accuracy
THINGS-EEG retrieval	5×10^{-4}	256	25	0	*n.s.*	0.1	–	Top-1 accuracy
CHB-MIT	2×10^{-4}	64	50	5	0.65	0	–	Balanced accuracy
ISRUC-S1	10^{-3}	16	20	2	0.8	0.0	–	Balanced accuracy
Mental Arithmetic	2×10^{-4}	16	10	5	0.9	0.0	–	Balanced accuracy
Mumtaz2016	10^{-4}	64	40	3	0.65	0.1	–	AUROC
TUEV	3×10^{-4}	144	8	0	0.9	0.0	–	Cohen's κ
TUAB	10^{-4}	144	1	0	0.9	0.0	–	Balanced accuracy

C.1 BCI Skills

To adapt general-purpose EEG representations to task-specific signal structure, we incorporate four BCI training skills into downstream pipelines: Denoising, Grand average, Filter bank, and Task prior channel selection. Each task uses the applicable techniques while retaining the pretrained backbone architecture. The backbone and task head are fine-tuned jointly.

Denoising. To attenuate out-of-band activity while retaining task-relevant spectral components, we apply task-dependent band-pass filtering before encoding. For emotion recognition, the retained frequency range follows each dataset's adaptation protocol. For example, the FACED-9 BCI setting uses a 0.5–15 Hz passband to emphasize low-frequency components. This constructs a band-limited waveform input for the pretrained encoder.

Grand average. For visual retrieval, we aggregate repeated EEG responses to obtain a more reliable estimate of the stimulus-locked signal. Within each subject, epochs corresponding to the same image are aligned to stimulus onset and averaged at corresponding channels and time points:

$$\overline{X} = \frac{1}{R} \sum_{r=1}^{R} X^{(r)}, \tag{17}$$

where $X^{(r)}$ is the EEG response to the r-th presentation and R is the number of averaged repetitions. Averaging retains the consistent evoked response while attenuating background activity and recording noise that vary across repetitions. The averaged waveform is then encoded as a single retrieval query.

Filter bank. To exploit responses at the stimulation frequency and its harmonics, we construct multiple frequency-selective inputs for SSVEP decoding. Three overlapping Butterworth filters with passbands of 8–90, 16–90, and 24–90 Hz produce sub-band signals that are processed using shared encoder weights. Temporal aggregation produces one representation per channel and band. These representations are concatenated and passed to the task head, allowing the classifier to combine information from different harmonic components rather than relying on a single broadband input.

Task prior channel selection. To focus the input on sensors relevant to the decoding task, we retain a task-specific electrode subset before patch embedding. SSVEP decoding on Benchmark and BETA uses 21 parietal, parieto-occipital, and occipital electrodes, while THINGS-EEG retrieval uses 17 posterior electrodes. The selected signals and their electrode identities are supplied to the backbone, and the task head aggregates representations from the retained channels.

C.2 Motor Imagery Classification

We evaluate four-class motor imagery on **PhysioNet-MI** and **BCIC-IV-2a**, using fixed subject partitions for both evaluation tracks. The BCI specialist for these is AGTCNet.

PhysioNet-MI (Schalk et al., 2004) contains recordings from 109 subjects acquired with 64 electrodes at 160 Hz. We extract four-second trials and apply common-average referencing, 0.3-Hz high-pass filtering, and 60-Hz notch filtering before resampling each trial to 200 Hz. Subjects 1–70, 71–89, and 90–109 are used for training, validation, and testing, respectively, yielding 6,300, 1,734, and 1,803 trials. Both evaluation tracks and the matched AGTCNet rerun use this partition. The input before any channel restriction is $64 \times 4 \times 200$. The task head averages the four time tokens within each channel and concatenates the resulting channel features. For C_{active} retained channels, its classifier has dimensions $200C_{\mathrm{active}} \rightarrow 4096 \rightarrow 200 \rightarrow 4$, with GELU and dropout after the first two linear layers. Table 9 reports the results. BrainDuo achieves the highest balanced accuracy of 0.6517 and Cohen's κ of 0.5289. Relative to the strongest foundation-model baseline, CSBrain, these scores improve balanced accuracy and Cohen's κ by 2.13 and 2.18 percentage points. The matched AGTCNet rerun obtains the highest weighted F1 of 0.6434, 0.10 percentage points above BrainDuo.

Table 9: Results on Motor Imagery Classification (PhysioNet-MI, 4-class).

	Method	Balanced Accuracy	Cohen's κ	Weighted F1
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.5814 ± 0.0125	0.4468 ± 0.0199	0.5796 ± 0.0115
	EEGConformer (Song et al., 2023)	0.6049 ± 0.0104	0.4736 ± 0.0171	0.6062 ± 0.0095
	SPaRCNet (Jing et al., 2023)	0.5932 ± 0.0152	0.4564 ± 0.0234	0.5937 ± 0.0147
	ContraWR (Yang et al., 2023b)	0.5892 ± 0.0133	0.4527 ± 0.0248	0.5918 ± 0.0116
	CNN-Transformer (Peh et al., 2022)	0.6053 ± 0.0118	0.4725 ± 0.0223	0.6041 ± 0.0105
	FFCL (Li et al., 2022)	0.5726 ± 0.0092	0.4323 ± 0.0182	0.5701 ± 0.0079
	ST-Transformer (Song et al., 2021)	0.6035 ± 0.0081	0.4712 ± 0.0199	0.6053 ± 0.0075
	AGTCNet (Lim et al., 2025)	$\underline{0.6440 \pm 0.0091}$	$\underline{0.5254 \pm 0.0121}$	$\mathbf{0.6434 \pm 0.0085}$
Foundation Models	BENDR (Kostas et al., 2021)	*n.r.*	*n.r.*	*n.r.*
	BIOT (Yang et al., 2023a)	0.6153 ± 0.0154	0.4875 ± 0.0272	0.6158 ± 0.0197
	EEGPT (Wang et al., 2024)	*n.r.*	*n.r.*	*n.r.*
	LaBraM-Base (Jiang et al., 2024)	0.6173 ± 0.0122	0.4912 ± 0.0192	0.6177 ± 0.0141
	CBraMod (Wang et al., 2025)	0.6174 ± 0.0036	0.4898 ± 0.0048	0.6179 ± 0.0035
	CSBrain (Zhou et al., 2025)	0.6304 ± 0.0090	0.5071 ± 0.0120	0.6308 ± 0.0095
	CodeBrain (Ma et al., 2026)	*n.r.*	*n.r.*	*n.r.*
	BrainDuo (Ours)	$\mathbf{0.6517 \pm 0.0052}$	$\mathbf{0.5289 \pm 0.0096}$	$\underline{0.6424 \pm 0.0128}$

BCIC-IV-2a (Brunner et al., 2008) contains two 250-Hz sessions (T and E) from nine subjects, with 22 EEG channels and 288 trials per session covering four imagined movements. We retain the first 22 EEG channels, crop the interval $[2, 6)$ s relative to trial onset, and resample it to 200 Hz. The optional band-pass filter is unset in the evaluated preprocessing configuration. Subjects 1–5, 6–7, and 8–9 form the training, validation, and test sets, respectively, containing 2,880, 1,152, and 1,152 trials. Both evaluation tracks use these subject partitions, with both sessions retained for BrainDuo and the matched AGTCNet rerun. The input has shape $22 \times 4 \times 200$. The task specific head pools the representation grid to 22×4 tokens, flattens it, and applies a $17{,}600 \to 800 \to 200 \to 4$ classifier with ELU and dropout after the first two linear layers. As shown in Table 10, AGTCNet is strongest among the two 5,184-trial evaluations, with balanced accuracy of 0.6224, Cohen's κ of 0.4965, and weighted F1 of 0.6190. Among the foundation models, BrainDuo attains the highest in BA and κ.

Table 10: Results on Motor Imagery Classification (BCIC-IV-2a, 4-class).

	Method	Balanced Accuracy	Cohen's κ	Weighted F1
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.4482 ± 0.0094	0.2693 ± 0.0121	0.4226 ± 0.0108
	EEGConformer (Song et al., 2023)	0.4696 ± 0.0106	0.2924 ± 0.0141	0.4533 ± 0.0128
	SPaRCNet (Jing et al., 2023)	0.4635 ± 0.0117	0.2847 ± 0.0147	0.4432 ± 0.0126
	ContraWR (Yang et al., 2023b)	0.4678 ± 0.0125	0.2905 ± 0.0160	0.4413 ± 0.0142
	CNN-Transformer (Peh et al., 2022)	0.4600 ± 0.0108	0.2800 ± 0.0148	0.4460 ± 0.0114
	FFCL (Li et al., 2022)	0.4470 ± 0.0143	0.2627 ± 0.0176	0.4238 ± 0.0139
	ST-Transformer (Song et al., 2021)	0.4575 ± 0.0145	0.2733 ± 0.0198	0.4471 ± 0.0142
	AGTCNet (Lim et al., 2025)	$\mathbf{0.6224 \pm 0.0384}$	$\mathbf{0.4965 \pm 0.0512}$	$\mathbf{0.6190 \pm 0.0384}$
Foundation Models	BENDR (Kostas et al., 2021)	*n.r.*	*n.r.*	*n.r.*
	BIOT (Yang et al., 2023a)	0.4748 ± 0.0093	0.2997 ± 0.0139	0.4607 ± 0.0125
	EEGPT (Wang et al., 2024)	*n.r.*	*n.r.*	*n.r.*
	LaBraM-Base (Jiang et al., 2024)	0.4869 ± 0.0085	0.3159 ± 0.0154	0.4758 ± 0.0103
	CBraMod (Wang et al., 2025)	0.5138 ± 0.0066	0.3518 ± 0.0094	0.4984 ± 0.0085
	CSBrain (Zhou et al., 2025)	0.5657 ± 0.0071	0.4209 ± 0.0093	$\underline{0.5637 \pm 0.0087}$
	CodeBrain (Ma et al., 2026)	*n.r.*	*n.r.*	*n.r.*
	BrainDuo (Ours)	$\underline{0.5715 \pm 0.0050}$	$\underline{0.4287 \pm 0.0067}$	0.5582 ± 0.0049

C.3 Emotion Recognition

We evaluate emotion recognition on **SEED-V** and **FACED-9** in the general EEG track and on SEED-V, SEED-IV, and FACED-9 in the BCI-native track. The tracks use different windows and splits,

so their results are reported separately. DAEST is the BCI specialist for FACED and PGCN is the specialist for SEED-IV and SEED-V.

BCI Skills: Frequency-selective emotion representations. DAEST reports informative learned components around 4–12 Hz on FACED for several emotions, while other emotions recruit higher frequencies (Shen et al., 2025). CL-SSTER also identifies low-frequency components associated across subjects (Shen et al., 2024). These observations motivate task-dependent frequency selection. Our FACED setting is evaluated in Figure 5; the SEED-V and SEED-IV bands pases are specified below.

SEED-V (General Setting) (Wu et al., 2020) contains three recording sessions from 16 subjects, with five emotion classes and 62 channels sampled at 1,000 Hz. We band-limit the signals to 0.3–75 Hz, form one-second windows, and resample them to 200 Hz. Within each session, trials 0–4, 5–9, and 10–14 define the training, validation, and test sets, producing 34,432, 42,960, and 40,352 windows. Each input has shape $62 \times 1 \times 200$. The task head uses one representation per channel and applies a $12{,}400 \to 800 \to 200 \to 5$ classifier, with ELU and dropout after the first two linear layers. This adapter assigns consecutive channel-position slots to the 62 electrodes. The supplied recipe disables exponential moving averaging (EMA). As shown in Table 11, BrainDuo obtains the best balanced accuracy of 0.4349, Cohen's κ of 0.2953, and weighted F1 of 0.4447. These results exceed CSBrain, the strongest baseline in all three metrics, by 1.52, 1.68, and 1.67 percentage points, respectively.

Table 11: Results on Emotion Recognition (SEED-V, 5-class).

	Method	Balanced Accuracy	Cohen's κ	Weighted F1
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.2961 ± 0.0102	0.1006 ± 0.0143	0.2749 ± 0.0098
	EEGConformer (Song et al., 2023)	0.3537 ± 0.0112	0.1772 ± 0.0174	0.3487 ± 0.0136
	SPaRCNet (Jing et al., 2023)	0.2949 ± 0.0078	0.1121 ± 0.0139	0.2979 ± 0.0083
	ContraWR (Yang et al., 2023b)	0.3546 ± 0.0105	0.1905 ± 0.0188	0.3544 ± 0.0121
	CNN-Transformer (Peh et al., 2022)	0.3678 ± 0.0078	0.2072 ± 0.0183	0.3642 ± 0.0088
	FFCL (Li et al., 2022)	0.3641 ± 0.0092	0.2078 ± 0.0201	0.3645 ± 0.0132
	ST-Transformer (Song et al., 2021)	0.3052 ± 0.0072	0.1083 ± 0.0121	0.2833 ± 0.0105
Foundation Models	BENDR (Kostas et al., 2021)	0.2231 ± 0.0059	0.0335 ± 0.0062	0.2026 ± 0.0330
	BIOT (Yang et al., 2023a)	0.3837 ± 0.0187	0.2261 ± 0.0262	0.3856 ± 0.0203
	EEGPT (Wang et al., 2024)	0.3061 ± 0.0044	0.1323 ± 0.0062	0.3090 ± 0.0052
	LaBraM-Base (Jiang et al., 2024)	0.3976 ± 0.0138	0.2386 ± 0.0209	0.3974 ± 0.0111
	CBraMod (Wang et al., 2025)	0.4091 ± 0.0097	0.2569 ± 0.0143	0.4101 ± 0.0108
	CSBrain (Zhou et al., 2025)	$\underline{0.4197 \pm 0.0033}$	$\underline{0.2785 \pm 0.0034}$	$\underline{0.4280 \pm 0.0023}$
	CodeBrain (Ma et al., 2026)	0.4137 ± 0.0023	0.2735 ± 0.0032	0.4235 ± 0.0022
	BrainDuo (Ours)	$\mathbf{0.4349 \pm 0.0085}$	$\mathbf{0.2953 \pm 0.0149}$	$\mathbf{0.4447 \pm 0.0098}$

FACED-9 (General Setting) (Chen et al., 2023b) contains recordings from 123 subjects across nine emotion classes. The broad-transfer protocol retains 30 channels, uses ten-second records, and resamples the signals from 250 to 200 Hz. The stored inputs are standardized and clipped; the upstream filter configuration that produced the evaluated broad-transfer data is not uniquely identifiable from the available preprocessing record. Subjects 0–79, 80–99, and 100–122 define the training, validation, and test sets, respectively, yielding 6,720, 1,680, and 1,932 records. The $30 \times 10 \times 200$ input produces a 30×10 representation grid. The task head flattens this grid and uses linear dimensions $60{,}000 \to 2000 \to 200 \to 9$. Layer normalization precedes each of the first two linear layers, and GELU and dropout follow them. Table 12 summarizes the results. BrainDuo achieves the highest balanced accuracy of 0.6172, Cohen's κ of 0.5531, and weighted F1 of 0.6061, improving on CodeBrain by 2.31, 1.25, and 1.08 percentage points.

Table 12: Results on Emotion Recognition (FACED-9, 9-class).

	Method	Balanced Accuracy	Cohen's κ	Weighted F1
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.4090 ± 0.0122	0.3342 ± 0.0251	0.4124 ± 0.0141
	EEGConformer (Song et al., 2023)	0.4559 ± 0.0125	0.3858 ± 0.0186	0.4514 ± 0.0107
	SPaRCNet (Jing et al., 2023)	0.4673 ± 0.0155	0.3978 ± 0.0289	0.4729 ± 0.0133
	ContraWR (Yang et al., 2023b)	0.4887 ± 0.0078	0.4231 ± 0.0151	0.4884 ± 0.0074
	CNN-Transformer (Peh et al., 2022)	0.4697 ± 0.0132	0.4017 ± 0.0166	0.4702 ± 0.0125
	FFCL (Li et al., 2022)	0.4673 ± 0.0158	0.3987 ± 0.0338	0.4699 ± 0.0145
	ST-Transformer (Song et al., 2021)	0.4810 ± 0.0079	0.4137 ± 0.0133	0.4795 ± 0.0096
Foundation Models	BENDR (Kostas et al., 2021)	0.5320 ± 0.0083	0.4716 ± 0.0095	0.5340 ± 0.0086
	BIOT (Yang et al., 2023a)	0.5118 ± 0.0118	0.4476 ± 0.0254	0.5136 ± 0.0112
	EEGPT (Wang et al., 2024)	0.4607 ± 0.0014	0.4639 ± 0.0023	0.3924 ± 0.0017
	LaBraM-Base (Jiang et al., 2024)	0.5273 ± 0.0107	0.4698 ± 0.0188	0.5288 ± 0.0102
	CBraMod (Wang et al., 2025)	0.5509 ± 0.0089	0.5041 ± 0.0122	0.5618 ± 0.0093
	CSBrain (Zhou et al., 2025)	0.5752 ± 0.0042	0.5204 ± 0.0036	0.5796 ± 0.0031
	CodeBrain (Ma et al., 2026)	$\underline{0.5941 \pm 0.0098}$	$\underline{0.5406 \pm 0.0084}$	$\underline{0.5953 \pm 0.0113}$
	BrainDuo (Ours)	$\mathbf{0.6172 \pm 0.0039}$	$\mathbf{0.5531 \pm 0.0024}$	$\mathbf{0.6061 \pm 0.0082}$

SEED-V (BCI setting) uses a 0.5–50-Hz passband, four-second windows, and 16 leave-one-subject-out folds, holding out one subject for testing in each fold. Validation windows are drawn from the pooled training subjects rather than from an additional held-out subject. The available split record contains 24,611 training, 2,734 validation, and 1,823 test windows. These counts describe that recorded split, not every fold. The encoder input is $62 \times 4 \times 200$. In the task head, the time tokens are aggregated within each channel before a learned query attends to the 62 channel representations. Bias-free query, key, and value projections have width 200, and attention scores are scaled by $1/\sqrt{200}$. Layer normalization, dropout, and a $200 \to 5$ linear classifier follow the attention pooling. The specification also describes learned attention over the four time tokens as an alternative to averaging. The recorded recipe uses EMA with decay 0.999 and validation Cohen's κ for model selection. Table 13 reports accuracy, balanced accuracy, Cohen's κ, and weighted F1 under this protocol. PGCN (Jin et al., 2024) obtains the highest reported accuracy of 0.6178 using five-band differential-entropy features. Its SEED-V and SEED-IV references retain the published accuracy and source-reported uncertainty; other metrics were not provided. Among the foundation models, BIOT has the strongest reported results, including an accuracy of 0.3196 and a weighted F1 of 0.2858. Since PGCN operates on engineered spectral features rather than the shared raw-waveform input, its result represents the task-specific frontier rather than a controlled backbone comparison.

Table 13: Results on Emotion Recognition (SEED-V, 5-class; subject LOSO, 4 s).

	Method	Accuracy	Balanced Accuracy	Cohen's κ	Weighted F1
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.2908 ± 0.0090	0.2717 ± 0.0126	0.0903 ± 0.0151	0.2130 ± 0.0171
	EEGConformer (Song et al., 2023)	0.2716 ± 0.0032	0.2758 ± 0.0022	0.0890 ± 0.0030	0.2072 ± 0.0065
	ST-Transformer (Song et al., 2021)	0.2545 ± 0.0413	0.2579 ± 0.0432	0.0652 ± 0.0460	0.1947 ± 0.0504
	PGCN (Jin et al., 2024)	$\mathbf{0.6178 \pm 0.0859}$	*n.r.*	*n.r.*	*n.r.*
Foundation Models	BIOT (Yang et al., 2023a)	$\underline{0.3196 \pm 0.0742}$	$\mathbf{0.3190 \pm 0.0774}$	$\mathbf{0.1412 \pm 0.0937}$	$\mathbf{0.2858 \pm 0.0829}$
	LaBraM-Base (Jiang et al., 2024)	0.2727 ± 0.0036	0.2636 ± 0.0589	0.0624 ± 0.0060	0.1930 ± 0.0114
	CBraMod (Wang et al., 2025)	0.2763 ± 0.0077	$\underline{0.2888 \pm 0.0275}$	0.1018 ± 0.0084	0.2315 ± 0.0088
	CSBrain (Zhou et al., 2025)	0.2881 ± 0.0029	0.2766 ± 0.0239	0.0919 ± 0.0040	$\underline{0.2430 \pm 0.0049}$
	BrainDuo (Ours)	0.2921 ± 0.0066	0.2858 ± 0.0342	$\underline{0.1190 \pm 0.0066}$	0.2406 ± 0.0080

SEED-IV (subject LOSO) (Zheng et al., 2019) is evaluated with a 0.5–70-Hz passband and four-second windows under 15-fold leave-one-subject-out evaluation. Validation is sampled at the pooled-window level from training data. The processed dataset contains 6,878 windows in total; training, validation, and test counts depend on the held-out subject. Each encoder input has shape $62 \times 4 \times 200$. The task head uses hidden widths 800 and 200, ELU and dropout after both hidden layers, and four output logits. The available implementation retains the full 62×4 token grid, giving a flattened input width of 49,600; the supplied head inventory instead describes temporal pooling to one token per channel. Both use consecutive channel-position slots. As shown in Table 14, PGCN

achieves the highest reported accuracy of 0.7369 under its five-band differential-entropy pipeline. Among the foundation models, BIOT obtains the highest accuracy, balanced accuracy, and Cohen's κ, whereas BrainDuo achieves the highest weighted F1 of 0.3714. The PGCN row is again reported as a task-specific reference because its input representation differs from the raw-waveform evaluation.

Table 14: Results on Emotion Recognition (SEED-IV, 4-class; subject LOSO, 4 s).

	Method	Accuracy	Balanced Accuracy	Cohen's κ	Weighted F1
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.3429 ± 0.0041	0.3350 ± 0.0014	0.1157 ± 0.0020	0.3264 ± 0.0016
	EEGConformer (Song et al., 2023)	0.3846 ± 0.0105	0.3808 ± 0.0097	0.1752 ± 0.0125	0.3495 ± 0.0175
	ST-Transformer (Song et al., 2021)	0.3556 ± 0.0393	0.3392 ± 0.0328	0.1272 ± 0.0468	0.2957 ± 0.0421
	PGCN (Jin et al., 2024)	$\mathbf{0.7369 \pm 0.0716}$	*n.r.*	*n.r.*	*n.r.*
Foundation Models	BIOT (Yang et al., 2023a)	0.3961 ± 0.0629	$\mathbf{0.3909 \pm 0.0595}$	$\mathbf{0.1890 \pm 0.0817}$	0.3669 ± 0.0792
	LaBraM-Base (Jiang et al., 2024)	0.3691 ± 0.0033	0.2524 ± 0.0185	0.1562 ± 0.0034	0.3510 ± 0.0038
	CBraMod (Wang et al., 2025)	0.3500 ± 0.0086	0.3389 ± 0.0241	0.1225 ± 0.0121	0.3082 ± 0.0126
	CSBrain (Zhou et al., 2025)	0.3529 ± 0.0038	0.3486 ± 0.0041	0.1313 ± 0.0058	0.3359 ± 0.0010
	BrainDuo (Ours)	0.3777 ± 0.0052	0.3714 ± 0.0124	0.1656 ± 0.0070	$\mathbf{0.3714 \pm 0.0094}$

FACED-9 (subject CV10) uses the declared one-second, ten-fold subject-level protocol: fold k is used for testing and the next fold cyclically for validation. Its supplied recipe uses a 0.5–15-Hz band-pass filter and a $30 \times 1 \times 200$ input. The task head uses the same normalization and activation ordering as the ten-second protocol, with linear dimensions $6000 \to 2000 \to 200 \to 9$. The one-second construction requires the temporal pooling target to be one patch, while the hidden width remains 2,000. The recorded recipe selects by validation balanced accuracy and uses EMA with decay 0.9999. Table 15 reports the results. We retain DAEST's published FACED-9 accuracy of 0.5930 ± 0.0770 (Shen et al., 2025) as the task-specialist reference. DAEST extracts classification features with a one-second window and stride; its five-second contrastive crops belong to pretraining. Its native classifier uses running normalization and within-trial linear-dynamical-system smoothing. Among the general EEG models evaluated under this construction, BrainDuo obtains the highest accuracy of 0.1760, balanced accuracy of 0.1705, Cohen's κ of 0.0682, and weighted F1 of 0.1586.

Table 15: Results on Emotion Recognition (FACED, 9-class; subject CV10, 1 s).

	Method	Accuracy	Balanced Accuracy	Cohen's κ	Weighted F1
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.1198 ± 0.0063	0.1148 ± 0.0038	0.0042 ± 0.0044	0.1084 ± 0.0040
	EEGConformer (Song et al., 2023)	0.1173 ± 0.0048	0.1118 ± 0.0026	0.0014 ± 0.0034	0.0663 ± 0.0032
	ST-Transformer (Song et al., 2021)	0.1211 ± 0.0100	0.1205 ± 0.0084	0.0100 ± 0.0090	0.0650 ± 0.0180
	DAEST (Shen et al., 2025)	$\mathbf{0.5930 \pm 0.0770}$	*n.r.*	*n.r.*	*n.r.*
Foundation Models	BIOT (Yang et al., 2023a)	0.1537 ± 0.0170	0.1395 ± 0.0088	0.0343 ± 0.0210	0.1079 ± 0.0250
	LaBraM-Base (Jiang et al., 2024)	0.1741 ± 0.0010	0.1687 ± 0.0125	0.0649 ± 0.0019	0.1555 ± 0.0012
	CBraMod (Wang et al., 2025)	0.1505 ± 0.0076	0.1463 ± 0.0101	0.0432 ± 0.0054	0.1264 ± 0.0027
	CSBrain (Zhou et al., 2025)	0.1321 ± 0.0111	0.1279 ± 0.0034	0.0189 ± 0.0034	0.0686 ± 0.0050
	BrainDuo (Ours)	0.1760 ± 0.0020	$\mathbf{0.1705 \pm 0.0061}$	$\mathbf{0.0682 \pm 0.0015}$	$\mathbf{0.1586 \pm 0.0024}$

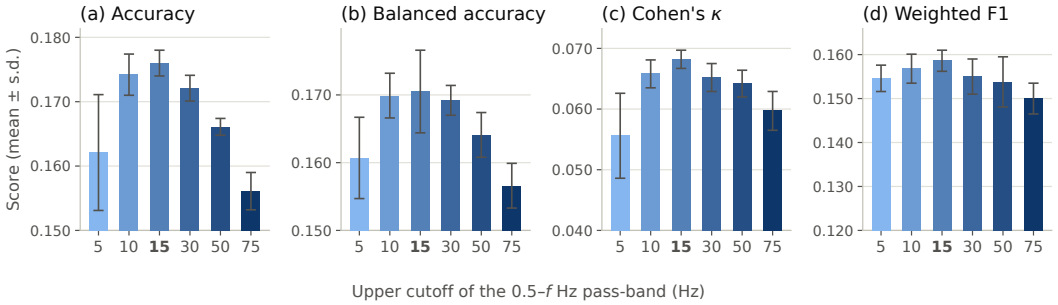

Figure 5: Denoising ablation on FACED-9 (1-s windows, subject CV10)

C.4 SSVEP CLASSIFICATION

We evaluate 40-target SSVEP classification on **Benchmark** and **BETA**, using one- and two-second windows under leave-one-subject-out (LOSO) and leave-one-block-out (LBO) protocols. We use DNN as the BCI specialist for LBO and MTSNET as LOSO specialist.

Channel selection and validation. Both datasets use the same 21 task-selected electrodes, in the following input order: P7, P5, P3, P1, PZ, P2, P4, P6, P8 (nine parietal); PO7, PO5, PO3, POZ, PO4, PO6, PO8 (seven parieto-occipital); and CB1, O1, OZ, O2, CB2 (five occipital). For each LOSO or LBO fold, 5% of the training partition is held out for validation and the other 95% is used for training.

BCI skill: Filter Bank. Filter-bank processing is widely used in SSVEP decoding, including FBCCA (Chen et al., 2015), TRCA (Nakanishi et al., 2018), TDCA (Liu et al., 2021), and neural models such as FB-tCNN (Ding et al., 2021), the DNN of Guney et al. (2022), FB-EEGNet (Yao et al., 2022), and FB-SSVEPformer (Chen et al., 2023a). Building on classical sub-band signal processing, FBCCA applies multiple band-pass filters to exploit information at the stimulation frequency and its harmonics (Chen et al., 2015). Its physiological motivation is that periodic visual stimulation evokes stimulus-locked responses at these frequencies, with harmonics arising in part from nonlinear processing in the visual system (Norcia et al., 2015). The resulting, often overlapping sub-bands emphasize different harmonic components, allowing their contributions to be combined rather than treated as a single broadband input. Neural implementations extend this approach through learned sub-band weighting or feature and prediction fusion (Guney et al., 2022; Yao et al., 2022; Chen et al., 2023a).

Benchmark (Wang et al., 2017) contains 35 subjects and six blocks per subject, giving 8,400 trials acquired at 250 Hz. One- and two-second windows begin 140 ms after stimulus onset, and inputs are resampled to 200 Hz. The BrainDuo recipe applies three Butterworth filter bands (8–90, 16–90, and 24–90 Hz) and processes each band with the same encoder weights by folding the band axis into the batch. For the ordered 21-channel layout above, the two-second model input is $3 \times 21 \times 2 \times 200$; the one-second counterpart has one time patch. Temporal aggregation leaves one feature per channel and band, which the head concatenates into a 12,600-dimensional vector. The task head then maps $12{,}600 \to 800 \to 200 \to 40$, with ELU and dropout between linear layers. The supplied schedule uses 60 epochs for one-second windows and 30 for two-second windows. Table 16 reports LOSO and LBO accuracy for one- and two-second observation windows. Under LOSO, MTSNet obtains the highest one-second accuracy of 0.7737, while BrainDuo achieves the highest two-second accuracy of 0.9042, numerically exceeding MTSNet by 1.52 percentage points. Under LBO, the DNN results are strongest at their respective durations; BrainDuo ranks second with accuracies of 0.7905 and 0.9245.

Table 16: Results on SSVEP Classification (Benchmark, 40-class).

	Method	LOSO Accuracy		LBO Accuracy	
		1 s	2 s	1 s	2 s
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.5837 ± 0.0061	0.7869 ± 0.0079	0.6290 ± 0.0046	0.8275 ± 0.0077
	EEGConformer (Song et al., 2023)	0.4945 ± 0.0125	0.7608 ± 0.0042	0.5294 ± 0.0082	0.7976 ± 0.0051
	ST-Transformer (Song et al., 2021)	0.6158 ± 0.0158	0.8006 ± 0.0068	0.6233 ± 0.0055	0.7204 ± 0.0066
	MTSNet (Lan et al., 2025)	$\mathbf{0.7737 \pm 0.0049}$	$\underline{0.8890 \pm 0.0015}$	*n.r.*	*n.r.*
	DNN (Guney et al., 2022)	*n.r.*	*n.r.*	$\mathbf{0.9557 \pm 0.0117}$	$\mathbf{1.0000}$
Foundation Models	BIOT (Yang et al., 2023a)	0.3210 ± 0.0327	0.5406 ± 0.0129	0.3099 ± 0.0043	0.5423 ± 0.0129
	LaBraM-Base (Jiang et al., 2024)	0.6916 ± 0.0084	0.8276 ± 0.0096	0.7188 ± 0.0029	0.8297 ± 0.0040
	CBraMod (Wang et al., 2025)	0.6298 ± 0.0085	0.8243 ± 0.0023	0.6521 ± 0.0086	0.8456 ± 0.0073
	CSBrain (Zhou et al., 2025)	0.6274 ± 0.0047	0.8477 ± 0.0056	0.6694 ± 0.0068	0.8726 ± 0.0084
	BrainDuo (Ours)	$\underline{0.7482 \pm 0.0193}$	$\mathbf{0.9042 \pm 0.0095}$	$\underline{0.7905 \pm 0.0062}$	$\underline{0.9245 \pm 0.0044}$

BETA (Liu et al., 2020) contains 70 subjects and four blocks per subject, for 11,200 trials acquired at 250 Hz. One- and two-second windows begin 132 ms after stimulus onset, and inputs are resampled to 200 Hz. The adaptation uses the same 21-channel selection, three-band shared-encoder filter bank, and 40-class MLP head as Benchmark, with the same duration-dependent training budget. BETA

has four rather than six blocks, so LBO evaluation uses four held-out-block folds per subject; LOSO evaluation holds out each of its 70 subjects in turn. Table 17 follows the same LOSO/LBO and one-/two-second evaluation structure. MTSNet leads one-second LOSO with an accuracy of 0.6621, whereas BrainDuo leads two-second LOSO with 0.7949, a numerical difference of 1.86 percentage points over MTSNet. Under LBO, DNN results are strongest at their respective durations, and BrainDuo is second with accuracies of 0.5907 and 0.7924.

Table 17: Results on SSVEP Classification (BETA, 40-class).

	Method	LOSO Accuracy		LBO Accuracy	
		1 s	2 s	1 s	2 s
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.4649 ± 0.0110	0.6734 ± 0.0176	0.4714 ± 0.0041	0.6847 ± 0.0045
	EEGConformer (Song et al., 2023)	0.2599 ± 0.0090	0.5903 ± 0.0022	0.2462 ± 0.0122	0.5824 ± 0.0050
	ST-Transformer (Song et al., 2021)	0.3939 ± 0.0064	0.4228 ± 0.0028	0.1525 ± 0.0053	0.2641 ± 0.0023
	MTSNet (Lan et al., 2025)	$\mathbf{0.6621 \pm 0.0194}$	0.7763 ± 0.0062	*n.r.*	*n.r.*
	DNN (Guney et al., 2022)	*n.r.*	*n.r.*	$\mathbf{0.8378 \pm 0.0160}$	$\mathbf{1.0000}$
Foundation Models	BIOT (Yang et al., 2023a)	0.3270 ± 0.0198	0.5242 ± 0.0024	0.3037 ± 0.0070	0.5067 ± 0.0072
	LaBraM-Base (Jiang et al., 2024)	0.5675 ± 0.0043	0.7099 ± 0.0020	0.5527 ± 0.0018	0.6776 ± 0.0058
	CBraMod (Wang et al., 2025)	0.5375 ± 0.0060	0.6932 ± 0.0025	0.5188 ± 0.0072	0.6909 ± 0.0045
	CSBrain (Zhou et al., 2025)	0.4757 ± 0.0051	0.6786 ± 0.0018	0.4648 ± 0.0040	0.6812 ± 0.0062
	BrainDuo (Ours)	0.5905 ± 0.0093	$\mathbf{0.7949 \pm 0.0022}$	0.5907 ± 0.0065	0.7924 ± 0.0051

C.5 Visual Retrieval

We evaluate inter- and intra-subject visual retrieval on **THINGS-EEG**. The dataset was collected using a rapid serial visual presentation (RSVP) protocol. Each EEG query is matched to an image from a 200-candidate test set. BrainHive is the BCI specialist.

BCI Skill: Grand Average Grand averaging is a standard procedure in event-related potential (ERP) analysis and is widely used in THINGS-EEG visual decoding (Luck, 2014; **?**; **?**). For each subject, EEG epochs from repeated presentations of the same image are aligned to stimulus onset and averaged at corresponding channels and time points. This preserves the consistent stimulus-locked response while attenuating background activity and recording noise that vary across repetitions, improving the signal-to-noise ratio. NICE and ATM follow this approach by averaging the 80 repetitions of each test image into a single EEG query (Song et al., 2024; Li et al., 2024). The resulting average provides a more reliable estimate of the visual evoked response for image retrieval.

THINGS-EEG (Gifford et al., 2022) uses 17 posterior channels and one-second windows resampled from 250 to 200 Hz, with repeated trials averaged for each image condition. Each model input contains 200 samples per channel. Intra-subject evaluation uses 16,540 training conditions and 200 test concepts for each subject. Inter-subject evaluation holds out each of the ten subjects in turn, producing 148,860 training pairs from the other nine subjects and 200 held-out-subject test pairs evaluated against 200 candidate images in each fold. The retrieval head takes one representation per electrode, concatenates the 17 vectors, and applies $\mathrm{LN}(3400) \to \mathrm{Linear}(3400, 2048) \to \mathrm{GELU} \to \mathrm{Dropout} \to \mathrm{Linear}(2048, 1024)$, followed by ℓ_2 normalization. It is trained with symmetric InfoNCE against a sum-fused, ℓ_2-normalized visual target from four frozen image encoders (**CLIP ViT-B/32** (Radford et al., 2021) , **SDXL-VAE** (Podell et al., 2024) , **I-JEPA ViT-H/14** (Assran et al., 2023) and **ViT-H/14 (layer 16)** (Cherti et al., 2023) provides 1280-dimensional intermediate features from layer 16 of the contrastively pretrained image encoder rather than its final image embedding..). We reserves 3% of training concepts for validation, selecting by validation Top-1 accuracy. Table 18 reports Top-1, Top-5, and Top-10 retrieval accuracy. The BrainHiVE evaluation uses the same four-encoder fused visual target and attains the highest reported Top-1 scores. Among the general EEG models, BrainDuo has the highest inter-subject Top-1 and Top-5 scores, improving over LaBraM-Base by 3.30 and 1.12 percentage points.

Table 18: Results on Visual Retrieval (THINGS-EEG). Bold and underline indicate the best and second-best mean in each column, respectively, across all methods.

	Method	Inter-subject Accuracy			Intra-subject Accuracy		
		Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.1337 ± 0.0013	0.3680 ± 0.0092	0.5190 ± 0.0055	0.1722 ± 0.0190	0.4262 ± 0.0240	0.5860 ± 0.0260
	EEGConformer (Song et al., 2023)	0.0580 ± 0.0023	0.2183 ± 0.0051	0.3573 ± 0.0068	0.0157 ± 0.0050	0.0708 ± 0.0160	0.1262 ± 0.0250
	ST-Transformer (Song et al., 2021)	0.1515 ± 0.0142	0.3735 ± 0.0420	0.5260 ± 0.0780	0.4515 ± 0.0659	0.7875 ± 0.0610	0.8920 ± 0.0400
	BrainHiVE (Zheng et al., 2026)	$\mathbf{0.2410 \pm 0.057}$	$\mathbf{0.5740 \pm 0.093}$	$\mathbf{0.6905 \pm 0.021}$	$\mathbf{0.8140 \pm 0.0043}$	$\mathbf{0.9530 \pm 0.0270}$	$\mathbf{1 \pm 0.0000}$
Foundation Models	BIOT (Yang et al., 2023a)	0.0595 ± 0.0210	0.1860 ± 0.0296	0.3080 ± 0.0940	0.1955 ± 0.0680	0.4765 ± 0.1120	0.6325 ± 0.1010
	LaBraM-Base (Jiang et al., 2024)	0.2055 ± 0.0034	0.5200 ± 0.0109	$\underline{0.6607 \pm 0.0074}$	0.3487 ± 0.0035	0.8112 ± 0.0140	0.8955 ± 0.0030
	CBraMod (Wang et al., 2025)	0.1682 ± 0.0078	0.4237 ± 0.0060	0.5737 ± 0.0141	0.5077 ± 0.0036	$\underline{0.8967 \pm 0.0040}$	0.9543 ± 0.0050
	CSBrain (Zhou et al., 2025)	0.2037 ± 0.0075	0.4788 ± 0.0043	0.6287 ± 0.0044	0.3803 ± 0.0300	0.6690 ± 0.0320	0.7755 ± 0.0320
	BrainDuo (Ours)	$\underline{0.2385 \pm 0.0024}$	$\underline{0.5312 \pm 0.0072}$	0.6500 ± 0.0040	$\underline{0.6480 \pm 0.0090}$	0.8938 ± 0.0090	$\underline{0.9640 \pm 0.0251}$

C.5.1 GRAND AVERAGE

Grand average combines repeated EEG trials for the same image condition before retrieval. We vary the number of averaged trials $k \in \{1, 5, 10, 20, 40, 80\}$ for inter- and intra-subject THINGS-EEG, retaining four fused image encoders and task-selected EEG channels. Table 19 reports all Top-1 results. Mean Top-1 accuracy increases across the k , from 6.05% to 23.85% inter-subject and from 12.01% to 64.80% intra-subject. Relative to a single trial, the 80-trial baseline gains 17.80 and 52.79 percentage points, respectively. These results support repeated-trial aggregation in this BCI-native retrieval pipeline and quantify the trade-off between accuracy and the number of EEG repetitions used per prediction.

Table 19: Grand average on THINGS-EEG. The $k = 80$ column is the existing headline baseline.

Averaged trials (k)	**1**	**5**	**10**	**20**	**40**	**80**
Inter-subject	0.0605 ± 0.0012	0.1197 ± 0.0039	0.1461 ± 0.0019	0.1851 ± 0.0031	0.2150 ± 0.0029	0.2385 ± 0.0024
Intra-subject	0.1201 ± 0.0094	0.4003 ± 0.0102	0.5057 ± 0.0079	0.5721 ± 0.0060	0.6091 ± 0.0083	0.6480 ± 0.0090

C.6 Seizure Detection

CHB-MIT (Shoeb, 2009) contains long-term pediatric scalp EEG from 23 cases. We retain 16 shared channels, resample the recordings from 256 to 200 Hz, and construct ten-second segments. The available preprocessing record does not establish an additional band-pass filter. Cases 01–19, 20–21, and 22–23 define the training, validation, and test sets, containing 303,893, 21,041, and 20,852 segments. Non-seizure (baseline) windows are non-overlapping; a five-second stride is used only for seizure-centered positive windows in the training set. The implemented adapter accepts $16 \times 10 \times 200$ inputs, pools the representation grid to 16×10 tokens, and flattens it into 32,000 features. Its binary classifier has dimensions $32{,}000 \to 2000 \to 200 \to 1$, with ELU and dropout after the first two linear layers. Binary cross-entropy with logits trains the single output. Table 20 reports the results. BrainDuo achieves the highest balanced accuracy of 0.7419, AUC-PR of 0.7671, and AUROC of 0.9806. Relative to the strongest baseline in each column, these correspond to gains of 0.21, 25.07, and 8.45 percentage points, with the largest separation on the imbalance-sensitive AUC-PR metric.

Table 20: Results on Seizure Detection (CHB-MIT, 2-class).

	Method	**Balanced Accuracy**	**AUC-PR**	**AUROC**
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.5658 ± 0.0106	0.1914 ± 0.0182	0.8048 ± 0.0136
	EEGConformer (Song et al., 2023)	0.5976 ± 0.0141	0.2209 ± 0.0215	0.8226 ± 0.0170
	SPaRCNet (Jing et al., 2023)	0.5876 ± 0.0191	0.1247 ± 0.0119	0.8143 ± 0.0148
	ContraWR (Yang et al., 2023b)	0.6344 ± 0.0002	0.2264 ± 0.0174	0.8097 ± 0.0114
	CNN-Transformer (Peh et al., 2022)	0.6389 ± 0.0067	0.2479 ± 0.0227	0.8662 ± 0.0082
	FFCL (Li et al., 2022)	0.6262 ± 0.0104	0.2049 ± 0.0346	0.8271 ± 0.0051
	ST-Transformer (Song et al., 2021)	0.5915 ± 0.0195	0.1422 ± 0.0094	0.8237 ± 0.0491
Foundation Models	BENDR (Kostas et al., 2021)	0.5609 ± 0.0432	0.3071 ± 0.1240	0.8632 ± 0.0526
	BIOT (Yang et al., 2023a)	0.7068 ± 0.0457	0.3277 ± 0.0460	0.8761 ± 0.0284
	EEGPT (Wang et al., 2024)	0.5481 ± 0.0151	0.3073 ± 0.0641	0.8438 ± 0.0066
	LaBraM-Base (Jiang et al., 2024)	0.7075 ± 0.0358	0.3287 ± 0.0402	0.8679 ± 0.0199
	CBraMod (Wang et al., 2025)	$\underline{0.7398 \pm 0.0284}$	0.3689 ± 0.0382	0.8892 ± 0.0154
	CSBrain (Zhou et al., 2025)	0.7262 ± 0.0115	$\underline{0.5164 \pm 0.0449}$	0.8915 ± 0.0321
	CodeBrain (Ma et al., 2026)	0.7273 ± 0.0240	0.4377 ± 0.0288	$\underline{0.8961 \pm 0.0174}$
	BrainDuo (Ours)	$\mathbf{0.7419 \pm 0.0014}$	$\mathbf{0.7671 \pm 0.0118}$	$\mathbf{0.9806 \pm 0.0143}$

C.7 Sleep Staging

ISRUC-S1 (Khalighi et al., 2016) contains overnight polysomnographic recordings from 100 subjects. We use six EEG derivations (F3-A2, C3-A2, O1-A2, F4-A1, C4-A1, and O2-A1) sampled at 200 Hz. The recordings comprise 89,240 labeled 30-second epochs across five stages; consecutive epochs are grouped into sequences of 20 and the model predicts one stage for each epoch. Subjects 1–80, 81–90, and 91–100 are used for training, validation, and testing. We apply 0.3–35-Hz band-pass filtering, a 50-Hz notch filter, channelwise standardization, and clipping; this split yields 3,559, 468, and 435 sequences. The task head accepts $6 \times 30 \times 200$ and flattens its 180 patch representations before a $36{,}000 \to 512$ linear projection and GELU. A single pre-norm Transformer layer then mixes the 20 epoch embeddings, using width 512, four attention heads, and an FFN of width 2,048 with GELU. Dropout and a $512 \to 5$ linear layer predict the result. Table 21 reports the results. BrainDuo achieves the highest balanced accuracy of 0.8098, Cohen's κ of 0.7748, and weighted F1 of 0.8139.

Table 21: Results on Sleep Staging (ISRUC, 5-class).

	Method	Balanced Accuracy	Cohen's κ	Weighted F1
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.7154 ± 0.0121	0.7040 ± 0.0173	0.7513 ± 0.0124
	EEGConformer (Song et al., 2023)	0.7400 ± 0.0133	0.7143 ± 0.0162	0.7634 ± 0.0151
	SPaRCNet (Jing et al., 2023)	0.7487 ± 0.0075	0.7097 ± 0.0132	0.7624 ± 0.0092
	ContraWR (Yang et al., 2023b)	0.7402 ± 0.0126	0.7178 ± 0.0156	0.7610 ± 0.0137
	CNN-Transformer (Peh et al., 2022)	0.7363 ± 0.0087	0.7129 ± 0.0121	0.7719 ± 0.0153
	FFCL (Li et al., 2022)	0.7277 ± 0.0182	0.7016 ± 0.0292	0.7614 ± 0.0197
	ST-Transformer (Song et al., 2021)	0.7381 ± 0.0205	0.7013 ± 0.0352	0.7681 ± 0.0175
Foundation Models	BENDR (Kostas et al., 2021)	0.7401 ± 0.0056	0.6956 ± 0.0053	0.7569 ± 0.0049
	BIOT (Yang et al., 2023a)	0.7527 ± 0.0121	0.7291 ± 0.0230	0.7790 ± 0.0146
	EEGPT (Wang et al., 2024)	0.4012 ± 0.0177	0.2223 ± 0.0227	0.3111 ± 0.0110
	LaBraM-Base (Jiang et al., 2024)	0.7633 ± 0.0102	0.7231 ± 0.0182	0.7810 ± 0.0133
	CBraMod (Wang et al., 2025)	0.7865 ± 0.0110	0.7442 ± 0.0152	0.8011 ± 0.0099
	CSBrain (Zhou et al., 2025)	$\underline{0.7925 \pm 0.0030}$	0.7406 ± 0.0102	0.7990 ± 0.0091
	CodeBrain (Ma et al., 2026)	0.7835 ± 0.0033	$\underline{0.7476 \pm 0.0040}$	$\underline{0.8020 \pm 0.0018}$
	BrainDuo (Ours)	$\mathbf{0.8098 \pm 0.0069}$	$\mathbf{0.7748 \pm 0.0043}$	$\mathbf{0.8139 \pm 0.0156}$

C.8 Mental Stress Detection

Mental Arithmetic (Zyma et al., 2019) contains EEG recorded before and during a cognitive arithmetic task for 36 subjects. Removing the stored reference channel leaves 19 channels, which are resampled from 500 to 200 Hz and divided into five-second windows. Subjects 00–27, 28–31, and 32–35 form the training, validation, and test sets with 1,343, 172, and 192 examples. The task head receives $19 \times 5 \times 200$ input, pools to a 19×5 token grid, and flattens it into 19,000 features. Its classifier maps $19{,}000 \to 1000 \to 200 \to 1$, with GELU and dropout after the first two linear layers, and is trained using binary cross-entropy with logits. Table 22 reports the results. BrainDuo obtains the highest balanced accuracy of 0.7639, AUC-PR of 0.7491, and AUROC of 0.9046. Compared with the strongest baseline in each metric, these results improve by 0.81, 3.14, and 3.39 percentage points, respectively.

Table 22: Results on Mental Stress Detection (Mental Arithmetic, 2-class).

	Method	Balanced Accuracy	AUC-PR	AUROC
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.6770 ± 0.0116	0.5763 ± 0.0102	0.7321 ± 0.0108
	EEGConformer (Song et al., 2023)	0.6805 ± 0.0123	0.5829 ± 0.0134	0.7424 ± 0.0128
	SPaRCNet (Jing et al., 2023)	0.6879 ± 0.0107	0.5825 ± 0.0193	0.7418 ± 0.0132
	ContraWR (Yang et al., 2023b)	0.6631 ± 0.0097	0.5787 ± 0.0164	0.7332 ± 0.0082
	CNN-Transformer (Peh et al., 2022)	0.6779 ± 0.0268	0.5777 ± 0.0285	0.7258 ± 0.0336
	FFCL (Li et al., 2022)	0.6798 ± 0.0142	0.5786 ± 0.0266	0.7330 ± 0.0198
	ST-Transformer (Song et al., 2021)	0.6631 ± 0.0173	0.5672 ± 0.0259	0.7132 ± 0.0174
Foundation Models	BENDR (Kostas et al., 2021)	0.5681 ± 0.0448	0.3661 ± 0.0672	0.6248 ± 0.0765
	BIOT (Yang et al., 2023a)	0.6875 ± 0.0186	0.6004 ± 0.0195	0.7536 ± 0.0144
	EEGPT (Wang et al., 2024)	0.5597 ± 0.0171	0.5081 ± 0.0275	0.7162 ± 0.0171
	LaBraM-Base (Jiang et al., 2024)	0.6909 ± 0.0125	0.5999 ± 0.0155	0.7721 ± 0.0093
	CBraMod (Wang et al., 2025)	0.7256 ± 0.0132	0.6267 ± 0.0099	0.7905 ± 0.0073
	CSBrain (Zhou et al., 2025)	$\underline{0.7558 \pm 0.0106}$	0.6696 ± 0.0221	0.8478 ± 0.0297
	CodeBrain (Ma et al., 2026)	0.7514 ± 0.0203	$\underline{0.7177 \pm 0.0421}$	$\underline{0.8707 \pm 0.0209}$
	BrainDuo (Ours)	$\mathbf{0.7639 \pm 0.0038}$	$\mathbf{0.7491 \pm 0.0072}$	$\mathbf{0.9046 \pm 0.0092}$

C.9 Mental Disorder Diagnosis

Mumtaz2016 (Mumtaz et al., 2017) contains 63 participants in the evaluation. We retain 19 channels, apply a 0.3–75-Hz band-pass filter and a 50-Hz notch filter, resample the signals from 256 to 200 Hz, clip them at $\pm 100\,\mu$V before input scaling, and construct five-second windows. The 43/9/11-participant partition yields 4,891, 1,041, and 1,211 training, validation, and test examples.

The input is $19 \times 5 \times 200$. Then it averages patch representations and uses the backbone's normalized linear readout, $\mathrm{LN}(200) \to \mathrm{Linear}(200, 1)$, trained with binary cross-entropy. The specification additionally defines a patient-level prediction by averaging window probabilities within each patient; this aggregation is distinct from window-level scoring. Table 23 summarizes the results. BrainDuo achieves the highest balanced accuracy of 0.9791 and AUC-PR of 0.9942, improving over CSBrain by 1.48 and 0.06 percentage points. CSBrain obtains the highest AUROC of 0.9956, 0.16 percentage points above BrainDuo.

Table 23: Results on Depression Detection (Mumtaz2016, 2-class).

	Method	Balanced Accuracy	AUC-PR	AUROC
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.9232 ± 0.0104	0.9626 ± 0.0095	0.9639 ± 0.0093
	EEGConformer (Song et al., 2023)	0.9308 ± 0.0117	0.9684 ± 0.0105	0.9702 ± 0.0101
	SPaRCNet (Jing et al., 2023)	0.9316 ± 0.0095	0.9754 ± 0.0065	0.9781 ± 0.0063
	ContraWR (Yang et al., 2023b)	0.9195 ± 0.0115	0.9589 ± 0.0102	0.9621 ± 0.0092
	CNN-Transformer (Peh et al., 2022)	0.9305 ± 0.0068	0.9757 ± 0.0074	0.9742 ± 0.0059
	FFCL (Li et al., 2022)	0.9314 ± 0.0083	0.9717 ± 0.0021	0.9753 ± 0.0033
	ST-Transformer (Song et al., 2021)	0.9135 ± 0.0103	0.9578 ± 0.0086	0.9594 ± 0.0059
Foundation Models	BENDR (Kostas et al., 2021)	*n.r.*	*n.r.*	*n.r.*
	BIOT (Yang et al., 2023a)	0.9358 ± 0.0052	0.9736 ± 0.0034	0.9758 ± 0.0042
	EEGPT (Wang et al., 2024)	*n.r.*	*n.r.*	*n.r.*
	LaBraM-Base (Jiang et al., 2024)	0.9409 ± 0.0079	0.9798 ± 0.0093	0.9782 ± 0.0057
	CBraMod (Wang et al., 2025)	0.9560 ± 0.0056	0.9923 ± 0.0032	0.9921 ± 0.0025
	CSBrain (Zhou et al., 2025)	$\underline{0.9643 \pm 0.0155}$	$\underline{0.9936 \pm 0.0034}$	$\mathbf{0.9956 \pm 0.0020}$
	CodeBrain (Ma et al., 2026)	*n.r.*	*n.r.*	*n.r.*
	BrainDuo (Ours)	$\mathbf{0.9791 \pm 0.0079}$	$\mathbf{0.9942 \pm 0.0053}$	$\underline{0.9940 \pm 0.0016}$

C.10 Event Type Classification

TUEV (Harati et al., 2015) assigns one of six event labels to each five-second EEG segment. We use 23 channels, a 0.1–75-Hz band-pass filter, a 50-Hz notch filter, and resampling from 256 to 200 Hz. The training, validation, and test sets contain 62,630, 21,302, and 28,305 examples; training and validation use 232 and 58 subjects, and testing follows the official partition. The $23 \times 5 \times 200$ input produces 115 patch representations. The task head averages them, applies layer normalization, and uses a $200 \to 6$ linear classifier. Table 24 reports the results. BrainDuo achieves the highest balanced accuracy of 0.7242, Cohen's κ of 0.7230, and weighted F1 of 0.8615. Relative to the strongest baseline in each column, the gains are 3.39, 3.18, and 2.53 percentage points.

Table 24: Results on Event Type Classification (TUEV, 6-class).

	Method	Balanced Accuracy	Cohen's κ	Weighted F1
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.3876 ± 0.0143	0.3577 ± 0.0155	0.6539 ± 0.0120
	EEGConformer (Song et al., 2023)	0.4074 ± 0.0164	0.3967 ± 0.0195	0.6983 ± 0.0152
	SPaRCNet (Jing et al., 2023)	0.4161 ± 0.0262	0.4233 ± 0.0181	0.7024 ± 0.0104
	ContraWR (Yang et al., 2023b)	0.4384 ± 0.0349	0.3912 ± 0.0237	0.6893 ± 0.0136
	CNN-Transformer (Peh et al., 2022)	0.4087 ± 0.0161	0.3815 ± 0.0134	0.6854 ± 0.0293
	FFCL (Li et al., 2022)	0.3979 ± 0.0104	0.3732 ± 0.0188	0.6783 ± 0.0120
	ST-Transformer (Song et al., 2021)	0.3984 ± 0.0228	0.3765 ± 0.0306	0.6823 ± 0.0190
Foundation Models	BENDR (Kostas et al., 2021)	0.4363 ± 0.0245	0.4271 ± 0.0238	0.6755 ± 0.0216
	BIOT (Yang et al., 2023a)	0.5281 ± 0.0225	0.5273 ± 0.0249	0.7492 ± 0.0082
	EEGPT (Wang et al., 2024)	0.6232 ± 0.0114	0.6351 ± 0.0134	0.8187 ± 0.0063
	LaBraM-Base (Jiang et al., 2024)	0.6409 ± 0.0065	0.6637 ± 0.0093	0.8312 ± 0.0052
	CBraMod (Wang et al., 2025)	0.6671 ± 0.0107	0.6772 ± 0.0096	0.8342 ± 0.0064
	CSBrain (Zhou et al., 2025)	$\underline{0.6903 \pm 0.0059}$	0.6833 ± 0.0047	0.8333 ± 0.0057
	CodeBrain (Ma et al., 2026)	0.6428 ± 0.0062	$\underline{0.6912 \pm 0.0101}$	$\underline{0.8362 \pm 0.0048}$
	BrainDuo (Ours)	$\mathbf{0.7242 \pm 0.0040}$	$\mathbf{0.7230 \pm 0.0137}$	$\mathbf{0.8615 \pm 0.0124}$

C.11 ABNORMALITY DETECTION

TUAB (Obeid & Picone, 2016) is a binary normal-versus-abnormal EEG benchmark. We use 23 channels, ten-second non-overlapping windows, a 0.1–75-Hz band-pass filter, a 50-Hz notch filter, and resampling from 256 to 200 Hz. The training, validation, and test sets contain 301,464, 70,674, and 36,945 windows. The ten-second input has shape $23 \times 10 \times 200$. The task head averages the 230 patch representations, applies layer normalization, and maps $200 \to 1$ for binary cross-entropy with logits. Table 25 reports the results. BrainDuo achieves the highest balanced accuracy of 0.8307, AUC-PR of 0.9151, and AUROC of 0.9126, exceeding CodeBrain by 0.13, 0.51, and 0.96 percentage points.

Table 25: Results on Abnormal EEG Detection (TUAB, 2-class).

	Method	Balanced Accuracy	AUC-PR	AUROC
Task-specific Models	EEGNet (Lawhern et al., 2018)	0.7642 ± 0.0036	0.8299 ± 0.0043	0.8412 ± 0.0031
	EEGConformer (Song et al., 2023)	0.7758 ± 0.0049	0.8427 ± 0.0054	0.8445 ± 0.0038
	SPaRCNet (Jing et al., 2023)	0.7896 ± 0.0018	0.8414 ± 0.0018	0.8676 ± 0.0012
	ContraWR (Yang et al., 2023b)	0.7746 ± 0.0041	0.8421 ± 0.0104	0.8456 ± 0.0074
	CNN-Transformer (Peh et al., 2022)	0.7777 ± 0.0022	0.8433 ± 0.0039	0.8461 ± 0.0013
	FFCL (Li et al., 2022)	0.7848 ± 0.0038	0.8448 ± 0.0065	0.8569 ± 0.0051
	ST-Transformer (Song et al., 2021)	0.7966 ± 0.0023	0.8521 ± 0.0026	0.8707 ± 0.0019
Foundation Models	BENDR (Kostas et al., 2021)	0.7714 ± 0.0248	0.8412 ± 0.0215	0.8426 ± 0.0237
	BIOT (Yang et al., 2023a)	0.7959 ± 0.0057	0.8792 ± 0.0023	0.8815 ± 0.0043
	EEGPT (Wang et al., 2024)	0.8038 ± 0.0040	0.8891 ± 0.0018	0.8811 ± 0.0015
	LaBraM-Base (Jiang et al., 2024)	0.8140 ± 0.0019	0.8965 ± 0.0016	0.9022 ± 0.0009
	CBraMod (Wang et al., 2025)	0.7891 ± 0.0030	0.8636 ± 0.0063	0.8606 ± 0.0057
	CSBrain (Zhou et al., 2025)	0.8172 ± 0.0043	0.9005 ± 0.0066	0.8957 ± 0.0046
	CodeBrain (Ma et al., 2026)	$\underline{0.8294 \pm 0.0013}$	$\underline{0.9100 \pm 0.0006}$	$\underline{0.9030 \pm 0.0009}$
	BrainDuo (Ours)	$\mathbf{0.8307 \pm 0.0032}$	$\mathbf{0.9151 \pm 0.0023}$	$\mathbf{0.9126 \pm 0.0070}$

D Additional Model Ablations

D.1 Duo Latent alignment-weight sensitivity

Table 27 reports the supplied sweep of λ_{align} in DuoView Pretraining. This coefficient weights Continuous embedding alignment relative to Discrete code prediction; it is not a reconstruction weight for Duo Codebook learning. At zero, pretraining uses only the Discrete code objective. The $\lambda_{\text{align}} = 0.5$ condition is the baseline used for the headline results. The baseline has the highest reported mean on all six metrics. The response u-shaped: weight 1.0 attains seconds comparing to all 4 epxeriments all the three TUEV metrics and both TUAB AUC metrics and weight 0.1 and 0 are lower. The sweep favors supports balancing continuous alignment with code prediction.

D.2 Multi-rate temporal-period sensitivity

Table 28 examines the temporal periods used for Cyclic temporal grouping in the Multi-rate block. The baseline assigns three temporal heads to period two and two temporal heads to period three, alongside five spatial heads. The supplied sweep compares the repeated-period configurations $(2, 2)$ and $(3, 3)$ with $(2, 3)$, $(5, 7)$, and $(2, 7)$.

The $(2, 3)$ baseline achieves the highest reported mean across all three TUEV metrics and TUAB BA. Compared with $(3, 3)$, it improves TUEV BA from 0.7097 to 0.7242, κ from 0.6975 to 0.7230, and F1-w from 0.8276 to 0.8615, while increasing TUAB BA from 0.8270 to 0.8307. The results also reveal a trade-off: $(2, 7)$ achieves slightly higher TUAB AUC-PR and AUROC, but lower TUAB BA (0.8207 versus 0.8307). These observations favor the selected pair over the tested repeated-period controls and suggest a benefit from combining two distinct periods rather than repeating a single period. The complementary cyclic groups provide broad temporal coverage while decomposing temporal interactions into structured subsets that may help the encoder extract useful information.

D.3 DuoView Pretraining block composition

All configurations in Table 29 contain five Global Attention blocks (G), four Multi-rate blocks (M), and three Long-range blocks (L). The first four controls fix Global Attention blocks at depths $\{0, 1, 9, 10, 11\}$ and vary only the order of Multi-rate and Long-range blocks. The next three preserve the relative subsequence `MLMLMLM` while repositioning the five Global Attention blocks. These controls test whether alternating Multi-rate and Long-range blocks improves performance and where Global Attention should be placed. The proposed `GGMLMLMLMGGG` sequence achieves the highest mean on all three TUEV metrics and the second-highest mean on all three TUAB metrics. With Global Attention positions fixed, the alternating sequence outperforms both grouped orders and the tested permutations on all six metrics. With the relative M/L order fixed, placing two Global Attention blocks before and three after the alternating sequence outperforms periodic and center-loaded placements on all six metrics. Reversing this boundary allocation yields higher TUAB means but lower TUEV means, indicating a dataset-dependent preference for the allocation of Global Attention blocks.

Table 26: Effects of Filter bank and Task prior channel selection. Panels A and B are separate ablations; bold marks the higher mean within each evaluation setting in each panel.

A. Filter bank: SSVEP accuracy (1-s windows)

Filter bank	Benchmark		BETA	
	LOSO	LBO	LOSO	LBO
Without	0.6950 ± 0.0073	0.7150 ± 0.0049	0.5391 ± 0.0023	0.5061 ± 0.0063
With	$\mathbf{0.7482 \pm 0.0193}$	$\mathbf{0.7905 \pm 0.0062}$	$\mathbf{0.5905 \pm 0.0093}$	$\mathbf{0.5907 \pm 0.0065}$

B. Task prior channel selection

EEG channels	THINGS-EEG (Top-1)		Benchmark, 1 s (accuracy)		BETA, 1 s (accuracy)	
	Inter-subject	Intra-subject	LOSO	LBO	LOSO	LBO
All available	0.2135 ± 0.00538	0.5910 ± 0.01194	0.6738 ± 0.0045	0.7218 ± 0.0037	0.5069 ± 0.0062	0.5307 ± 0.0092
Task-selected	$\mathbf{0.2385 \pm 0.0024}$	$\mathbf{0.6480 \pm 0.0090}$	$\mathbf{0.7482 \pm 0.0193}$	$\mathbf{0.7905 \pm 0.0062}$	$\mathbf{0.5905 \pm 0.0093}$	$\mathbf{0.5907 \pm 0.0065}$

Table 27: Sensitivity to Duo Latent's alignment weight. The pretraining objective is $\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{code}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$.

	TUEV			TUAB		
λ_{align}	BA	κ	F1-w	BA	AUC-PR	AUROC
0	$\underline{0.7074 \pm 0.0097}$	0.6803 ± 0.0078	0.8250 ± 0.0063	0.8202 ± 0.0091	0.9018 ± 0.0038	0.9003 ± 0.0097
0.1	0.7000 ± 0.0088	0.6764 ± 0.0142	0.8256 ± 0.0085	0.8250 ± 0.0018	0.9048 ± 0.0018	0.9031 ± 0.0015
0.5 (baseline)	$\mathbf{0.7242 \pm 0.0040}$	$\mathbf{0.7230 \pm 0.0137}$	$\mathbf{0.8615 \pm 0.0124}$	$\mathbf{0.8307 \pm 0.0032}$	$\mathbf{0.9151 \pm 0.0023}$	$\mathbf{0.9126 \pm 0.0070}$
1	0.7135 ± 0.0021	$\underline{0.7081 \pm 0.0044}$	$\underline{0.8461 \pm 0.0033}$	$\underline{0.8274 \pm 0.0021}$	$\underline{0.9124 \pm 0.0011}$	$\underline{0.9101 \pm 0.0016}$

Table 28: Temporal-period sensitivity of the Multi-rate block.

	TUEV			TUAB		
Temporal periods	BA	κ	F1-w	BA	AUC-PR	AUROC
$(2, 2)$	0.6735 ± 0.0082	0.6430 ± 0.0120	0.8197 ± 0.0075	0.8230 ± 0.0034	0.9006 ± 0.0042	0.8997 ± 0.0051
$(3, 3)$	0.7097 ± 0.0054	0.6975 ± 0.0154	0.8276 ± 0.0079	0.8270 ± 0.0056	0.9074 ± 0.0064	0.9093 ± 0.0039
$(2, 3)$ **(baseline)**	$\mathbf{0.7242 \pm 0.0040}$	$\mathbf{0.7230 \pm 0.0137}$	$\mathbf{0.8615 \pm 0.0124}$	$\mathbf{0.8307 \pm 0.0032}$	0.9151 ± 0.0023	0.9126 ± 0.0070
$(5, 7)$	0.7138 ± 0.0118	0.7077 ± 0.0076	0.8425 ± 0.0067	0.8234 ± 0.0059	0.9091 ± 0.0074	0.9088 ± 0.0116
$(2, 7)$	0.7107 ± 0.0222	0.7183 ± 0.0101	0.8504 ± 0.0043	0.8207 ± 0.0083	$\mathbf{0.9153 \pm 0.0077}$	$\mathbf{0.9137 \pm 0.0057}$

Table 29: **Complete controlled depth-program ablation.** All rows contain exactly five Global Attention (G), four Multi-rate (M), and three Long-range (L) blocks. The first panel fixes the five Global Attention layers and changes only the order of M/L. The second panel fixes the relative MLMLMLM subsequence and changes only the placement of G.

		TUEV			TUAB		
Variant	12-layer program	BA ↑	κ ↑	F1-w ↑	BA ↑	AUC-PR ↑	AUROC ↑
M/L order; Global Attention layers fixed							
Grouped M→L	GGMMMMLLLGGG	0.6980 ± 0.0045	0.6830 ± 0.0037	0.8302 ± 0.0023	0.8232 ± 0.0073	0.9008 ± 0.0072	0.9042 ± 0.0084
Grouped L→M	GGLLLMMMMGGG	0.6982 ± 0.0072	0.6780 ± 0.0051	0.8327 ± 0.0020	0.8287 ± 0.0048	0.9092 ± 0.0040	0.9071 ± 0.0064
Fixed permutation P_2	GGMMLLLMMGGG	0.7017 ± 0.0059	0.6982 ± 0.0026	0.8391 ± 0.0016	0.8225 ± 0.0023	0.9065 ± 0.0030	0.9045 ± 0.0032
Fixed permutation P_3	GGMMMLMLLGGG	0.7046 ± 0.0039	0.6995 ± 0.0040	0.8474 ± 0.0023	0.8292 ± 0.0085	0.9052 ± 0.0114	0.9073 ± 0.0081
Global Attention placement; relative M/L order fixed							
Reversed boundary (3/2)	GGGMLMLMLMGG	0.7129 ± 0.0061	$\underline{0.7187 \pm 0.0057}$	$\underline{0.8582 \pm 0.0046}$	$\mathbf{0.8382 \pm 0.0034}$	$\mathbf{0.9190 \pm 0.0038}$	$\mathbf{0.9172 \pm 0.0035}$
Periodic Global Attention	GMLGMLGMLGMG	$\underline{0.7158 \pm 0.0035}$	0.6980 ± 0.0036	0.8339 ± 0.0018	0.8281 ± 0.0056	0.9095 ± 0.0062	0.9072 ± 0.0066
Center-loaded Global Attention	MLMGGGGGLMLM	0.6959 ± 0.0053	0.6882 ± 0.0031	0.8214 ± 0.0079	0.8298 ± 0.0058	0.9013 ± 0.0055	0.9011 ± 0.0073
Interlaced with Global Attention boundaries (ours)	GGMLMLMLMGGG	$\mathbf{0.7242 \pm 0.0040}$	$\mathbf{0.7230 \pm 0.0137}$	$\mathbf{0.8615 \pm 0.0124}$	$\underline{0.8307 \pm 0.0032}$	$\underline{0.9151 \pm 0.0023}$	$\underline{0.9126 \pm 0.0070}$

E PRETRAINING BENEFITS AND SCALING

E.1 PRETRAINING VERSUS TRAINING FROM SCRATCH

Table 30 compares pretrained BrainDuo with the supplied results for random initialization followed directly by supervised downstream training. The comparison covers TUEV event detection, TUAB abnormality detection, and 1-s LBO SSVEP classification on Benchmark and BETA. The scratch entries use the same architecture, data splits, downstream training settings, and BCI training skills. Only initialization differs: pretrained weights versus random weights.

Pretrained means are higher on every reported metric. TUEV BA increases from 0.5528 to 0.7242, κ from 0.4570 to 0.7230, and F1-w from 0.7258 to 0.8615. TUAB BA increases from 0.7523 to 0.8307, AUROC from 0.8352 to 0.9126, and AUPRC from 0.8285 to 0.9151. For SSVEP, accuracy increases from 0.4850 to 0.7905 on Benchmark and from 0.3295 to 0.5907 on BETA, differences of 30.55 and 26.12 percentage points, respectively. The comparison therefore records substantial higher pretrained performance for both general EEG classification and the BCI-native SSVEP settings.

Table 30: Pretraining versus training from scratch.

Dataset / protocol	Metric	Scratch	Pretrained
TUEV	BA	0.5528 ± 0.0057	$\mathbf{0.7242 \pm 0.0040}$
	κ	0.4570 ± 0.0079	$\mathbf{0.7230 \pm 0.0137}$
	F1-w	0.7258 ± 0.0140	$\mathbf{0.8615 \pm 0.0124}$
TUAB	BA	0.7523 ± 0.0073	$\mathbf{0.8307 \pm 0.0032}$
	AUROC	0.8352 ± 0.0028	$\mathbf{0.9126 \pm 0.0070}$
	AUPRC	0.8285 ± 0.0053	$\mathbf{0.9151 \pm 0.0023}$
Benchmark (1-s LBO)	Accuracy	0.4850 ± 0.0094	$\mathbf{0.7905 \pm 0.0062}$
BETA (1-s LBO)	Accuracy	0.3295 ± 0.0052	$\mathbf{0.5907 \pm 0.0065}$

E.2 Pretraining-data volume

Figure 6 shows seven TU EEG data-volume settings, from 100 to 9,000 hours. The INRIA BCI, PhysioNet-MI, and SEED-IV data are unchanged across settings, and every data-size run uses the same number of pretraining epochs. The reported hours measure TU EEG data volume, not wall-clock training time. TUEV BA increases from 0.5872 to 0.7242 and TUAB BA from 0.7933 to 0.8307. Gains become smaller beyond 1,000 hours: TUEV BA rises from 0.7004 to 0.7242 over the remaining range, while TUAB BA rises from 0.8253 to 0.8307. The 9,000-hour reference uses the headline results, including their standard deviations. This setup includes 9,000 hours of TU EEG together with the additional sources described in Appendix A.1.

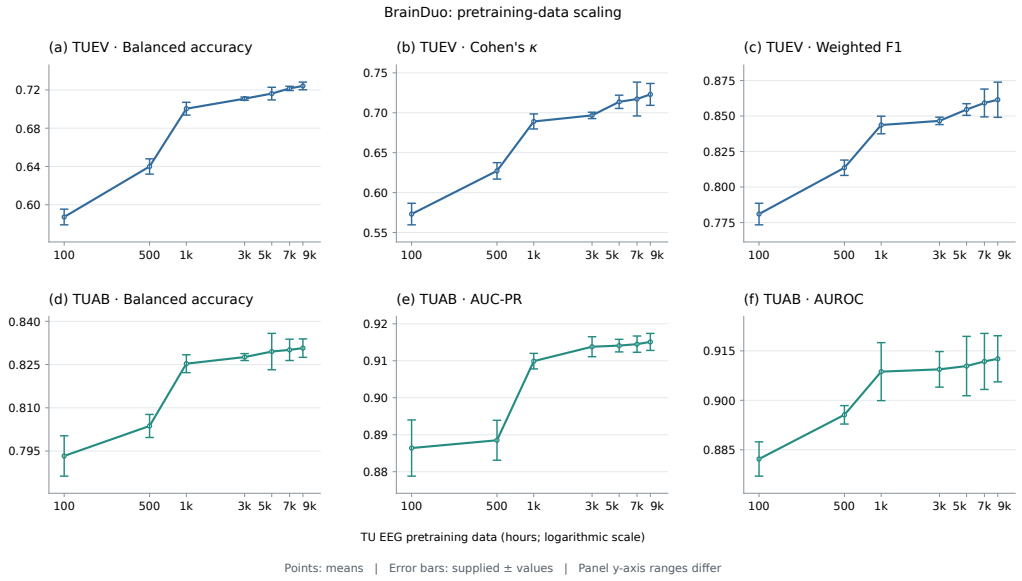

Figure 6: Pretraining-data scaling on TUEV and TUAB from 100 to 9,000 hours of TU EEG, shown on a logarithmic horizontal axis.

E.3 Model-depth scaling

Figure 7 shows two-, four-, six-, eight-, ten-, and twelve-layer settings. Reduced-depth backbones use condensed versions of the same block design, distributing Global Attention, Multi-rate blocks, and Long-range blocks through the available depth where it permits. The two-layer case contains one Multi-rate block and one Global block. Model-depth scaling therefore changes both capacity and the available block composition. TUEV BA increases from 0.5567 to 0.7242 and TUAB BA from 0.7773 to 0.8307.

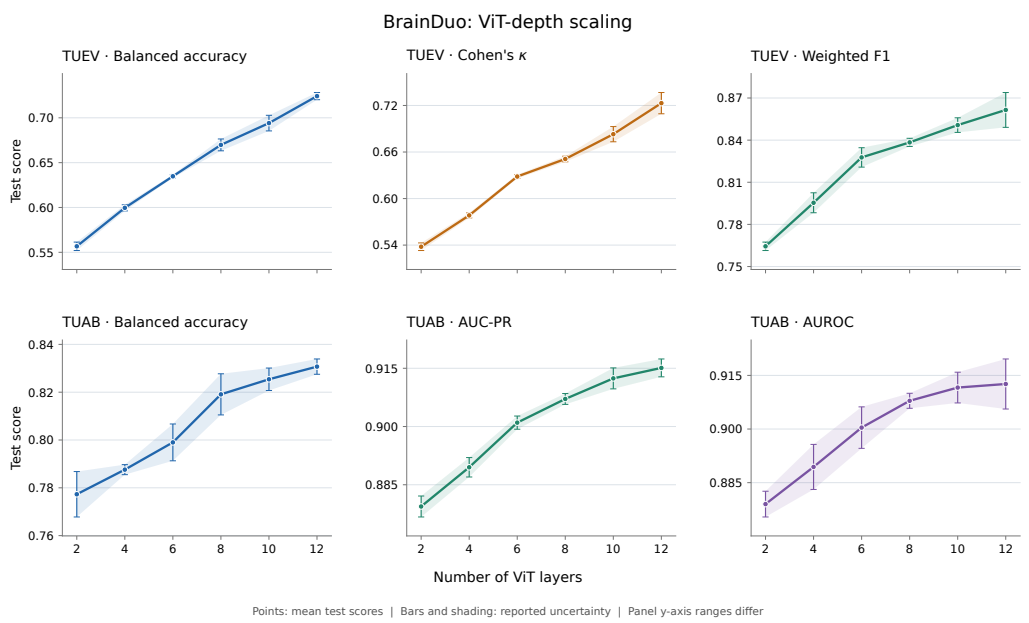

Figure 7: ViT-depth scaling on TUEV and TUAB from 2 to 12 layers. Points show mean test scores; bars and shading show reported uncertainty. Vertical-axis ranges differ across panels. Reduced-depth backbones use condensed versions of the three-block design described in Appendix E.

F Qualitative Analysis

F.1 Scalp topography

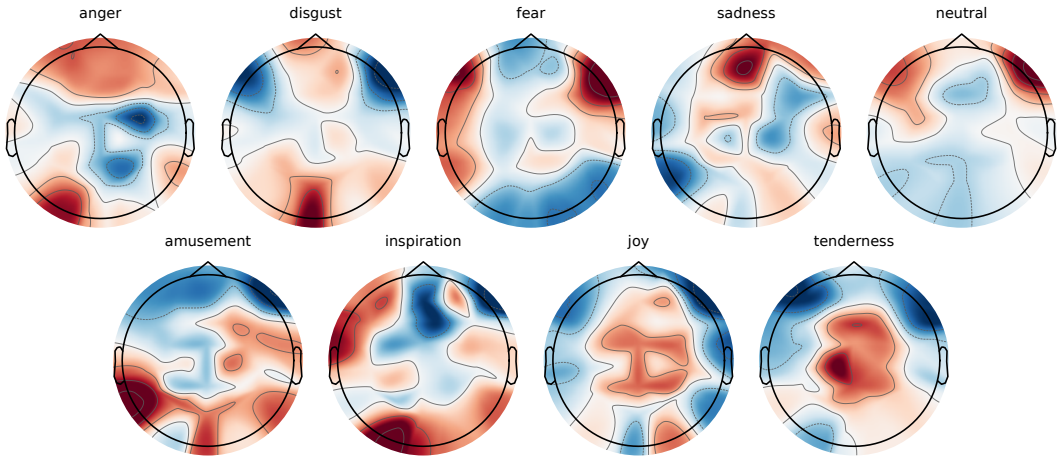

Figure 8: Class-contrast Grad-CAM scalp maps of fine-tuned BrainDuo for FACED-9 emotion recognition. Red and blue indicate higher and lower relative attribution after subtracting the electrode-wise across-class mean and standardizing each map across electrodes. Colors are not directly comparable across maps.

We apply Grad-CAM (Selvaraju et al., 2017) to the final transformer-block tokens of fine-tuned BrainDuo classifiers. Figure 8 shows FACED 9 emotion recognition , included as an additional qualitative case study. The maps exhibit class-dependent spatial patterns: for example, FACED anger and joy emphasize different anterior and central scalp locations. These patterns describe

relative model attribution. Figure 8 30 electrodes with 207 examples per emotion and 276 neutral examples for FACED-9.

F.2 Representation visualization

Figure 9 compares t-SNE projections (van der Maaten & Hinton, 2008) of raw EEG and fine-tuned representation on BCIC-IV-2a and SEED-V. For each dataset, we compare the feature distributions of raw EEG signals and those produced by BrainDuo after encoding. SEED-V shows several class-enriched regions in the representation, whereas BCIC-IV-2a retains more class overlap. This analysis illustrates the separability of pretrained hidden features.

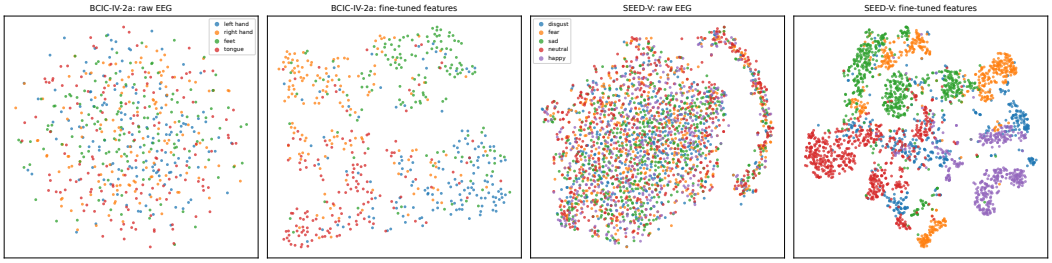

Figure 9: Independent t-SNE projections of raw EEG and fine-tuned for BCIC-IV-2a and SEED-V. Colors denote ground-truth classes

G Limitations

First, BCI-native skills improve BrainDuo's downstream performance, while gaps to task-specialist systems remain in some settings. Our evaluation is currently limited to offline datasets. Future work could extend the evaluation to additional BCI datasets and explore online, closed-loop tasks to assess real-time decoding, calibration requirements, and robustness across sessions and devices for practical BCI use.

Second, our pretraining and evaluation are restricted to scalp EEG, and CTopo defines spatial neighborhoods through a canonical scalp-electrode layout. The applicability of BrainDuo to intracranial recordings such as electrocorticography (ECoG) and stereo-EEG (sEEG) remains untested. Extending the framework to these settings would require examining whether the shared spectrotemporal vocabulary remains suitable and how spatial context should be defined for different recording configurations. Future work could adapt the temporal and spectral resolution of tokenization and construct recording-specific spatial neighborhoods to study which representations and contextual relationships transfer across recording modalities.

H AI use Statement

We used large language models (LLMs) to support manuscript preparation and revision through grammar checks, language refinement, and suggestions to improve clarity and paragraph structure. The authors are responsible for critically reviewing AI-assisted suggestions and checking the manuscript's text, references, and numerical claims against original sources and experimental records. Full responsibility for the final manuscript, research decisions, and reported results rests with the authors.

I Ethics Statement

This work uses previously released EEG datasets to study general-purpose representation learning and downstream BCI decoding. Dataset access, consent conditions, and redistribution restrictions remain governed by the original providers. BrainDuo is a research model and has not been validated for clinical diagnosis, autonomous treatment decisions, subject identification, or real-time safety-critical use.