# SHORTEEG-BENCH: A COMPREHENSIVE BENCHMARK FOR EEG FOUNDATION MODELS UNDER SHORT-WINDOW PROTOCOLS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

EEG Foundation Models (EEG-FMs) have demonstrated promising neural decoding capabilities across diverse tasks and subjects. However, current evaluations mainly rely on long temporal windows (4–30s), which may obscure fine-grained temporal dynamics and limit the applicability of these models to low-latency brain–computer interfaces. Therefore, we introduce **ShortEEG-Bench**, a unified benchmark for systematically evaluating EEG-FMs under short-window protocols (1/8-1/2 of the existing long windows). We evaluate 10 models on 11 datasets, covering cross-model comparisons, temporal analysis and multi-task training. Our results reveal several key findings. **First**, EEG-FMs maintain competitive performance with shortened inputs, and task-specific models do not consistently dominate, where $36\%$ short-window majority voting outperforming long-window predictions. **Second**, short-window analysis reveals substantial temporal fluctuations hidden by long-window aggregation, including higher accuracy in early post-cue intervals followed by a gradual decline in specific tasks. **Third**, in-domain multi-task learning yields positive transfer for $57\%$ of the evaluated EEG-FMs, whereas cross-domain training more often induces negative transfer. Overall, ShortEEG-Bench reveals that short-window evaluation preserves competitive decoding performance while exposing temporal dynamics and transfer behaviors masked by long-window aggregation, enabling more temporally resolved and diagnostically meaningful EEG-FM evaluation.

## 1 INTRODUCTION

Electroencephalography (EEG) foundation models (EEG-FMs) aim to learn transferable representations from large-scale neural recording, and recent models have explored several architectures, tokenization schemes, and pretraining objectives to accommodate heterogeneous EEG signals (Jiang et al., 2024b; Wang et al., 2024; Yang et al., 2023; El Ouahidi et al., 2026). This diversity makes standardized evaluation essential for assessing the practical value of pretrained representations. Benchmarks such as EEG-FM-Bench (Xiong et al., 2026) provide a shared framework for comparing released checkpoints across datasets and adaptation settings. However, input duration remains an underexamined dimension of EEG-FM evaluation.

Existing evaluation settings typically use EEG windows lasting several seconds to tens of seconds, including 4-s inputs for motor imagery and 30-s epochs for sleep staging (Table 1). Such windows provide extended temporal context, but also impose a longer observation period before a prediction can be made. Moreover, a single prediction over an extended segment does not reveal how decoding performance varies within that segment. In addition, as shown in Fig. 1(a), a shorter input can support earlier decisions and aggregation of several short-window decisions can use the full observation period. Both are useful reference points for understanding temporal information.

Therefore, we introduce **ShortEEG-Bench**, a unified benchmark covering 11 datasets across five task families. As shown in Fig. 1(b), the benchmark evaluates windows from 0.5 to 15 s, corresponding to one-eighth to one-half of the respective long-window inputs. In addition, our baselines include three conventional deep neural networks (EEGNet, EEG Conformer, and ADFCNN) and seven EEG-FMs (LaBraM, EEGPT, BENDR, BIOT, CBraMod, REVE, and CSBrain). Each



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.

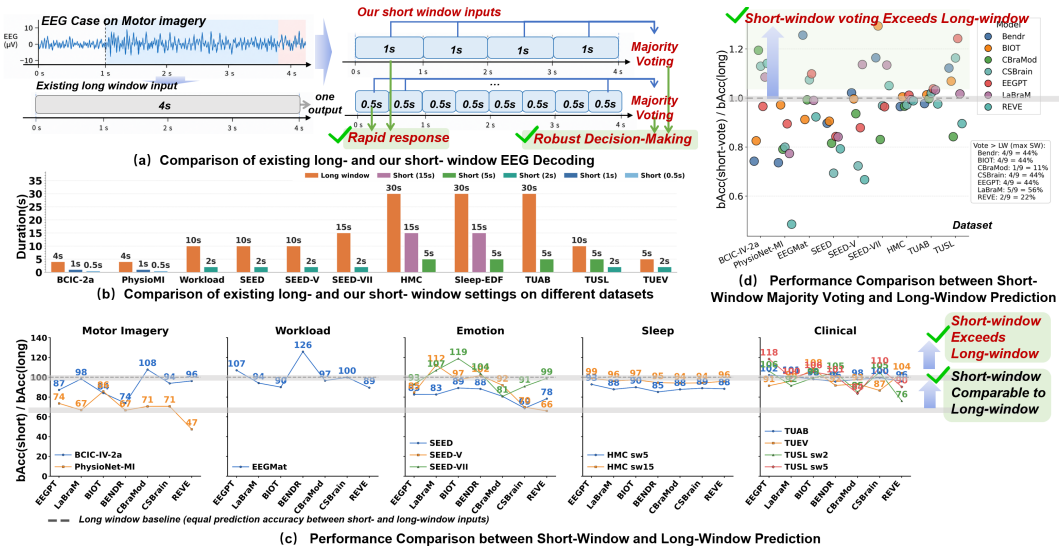


Figure 1: Illustration of the settings and effectiveness of short window-based EEG decoding. (a) Comparison of existing long-window EEG decoding with ours short ones on motor imagery (MI), where our fine-grained 1-s (0.5-s) windows enable rapid responses and more robust predictions through majority voting across multiple windows. (b) Comparison of long- and our short- window settings for EEG decoding benchmarks, with short-window durations ranging from 12.5% to 50% of their corresponding long-window inputs. (c-d) Short(Short voting)-to-long window accuracy ratio (y-axis) across different models (x-axis). Despite substantially shorter input durations, short windows achieve comparable or even superior performance to long windows on certain tasks.

EEG-FM is systematically evaluated under a comprehensive set of training and inference configurations along three dimensions: 1) adaptation strategies, including full fine-tuning, linear probing, and LoRA; 2) training paradigms, including single-dataset training, within-domain multi-dataset training, and cross-domain multi-task training; and 3) inference strategies, including window-level prediction and majority voting across windows within the same parent segment.

Our results yield four main findings. **First**, as shown in Fig. 1(c), short-window EEG decoding remains viable. Under short-window settings, EEG-FMs generally achieve stronger balanced accuracy than conventional deep neural network-based models, obtaining the best performance in most datasets. **Second**, as shown in Fig. 1(d), majority voting often improves local predictions, bringing short-window performance closer to or even surpassing that of long-window predictions. In addition, leading short-window EEG-FMs also improve performance for some subjects with weak long-window conventional baselines, although these gains are not universal. **Third**, for explicitly cued tasks, such as motor imagery (MI), early post-cue windows generally achieve higher decoding accuracy than later windows across both MI datasets. Moreover, finer temporal resolution (0.5-s windows) reveals more pronounced position-dependent fluctuations in decoding performance. **Fourth**, multi-dataset training offers selective benefits: in-domain training is generally more favorable, whereas cross-domain training can introduce substantial negative transfer. These effects depend jointly on dataset composition, backbone, and adaptation strategy, and majority voting cannot reliably offset the adverse effects of joint training[1].

Our contributions are summarized as follows:

- We introduce ShortEEG-Bench, a unified short-window evaluation benchmark spanning 11 datasets, five task families, seven EEG foundation models, and three conventional deep learning baselines.
- We establish empirical baselines across input durations, adaptation strategies, training configurations, and prediction granularities, characterizing the performance cost of reduced temporal context and the conditions under which voting or multi-dataset training helps.

[1]The source code and evaluation framework will be released upon completion of the review process.

Table 1: Overview of input-window configurations in existing benchmarks (primarily adopt long windows of 4–30 s) and our ShortEEG-Bench (diverse short-window settings ranging from 0.5-5 s).

| Task family | Dataset | Classes | Original duration(s) | Existing window(s) | Our short window(s) |
|---|---|---|---|---|---|
| Motor imagery | BCIC-IV-2a | 4 | 4 | 4 | 1, 0.5 |
| | PhysioNet-MI | 4 | 4 | 4 | 1, 0.5 |
| Mental Stress | EEGMat | 2 | 60 | 10 | 2 |
| Emotion | SEED | 3 | ∼226 | 10 | 2 |
| | SEED-V | 5 | ∼163 | 10 | 2 |
| | SEED-VII | 7 | ∼174 | 15 | 2 |
| Sleep staging | HMC | 5 | 30 | 30 | 5, 15 |
| | Sleep-EDF | 5 | 30 | 30 | 5, 15 |
| Clinical | TUAB | 2 | ∼1300 | 30 | 5 |
| | TUEV | 6 | 1 | 5 | 2 |
| | TUSL | 3 | 10 | 10 | 2, 5 |

- We provide temporal-position and subject-level analyses that reveal variation hidden by aggregate scores, informing the evaluation of EEG models for earlier decisions and temporally resolved decoding.

## 2 Related Work

Electroencephalography (EEG) decoding has been widely studied for brain–computer interfaces (BCIs) and neuroscience applications, including motor imagery, emotion recognition, sleep staging, and clinical diagnosis. Early approaches mainly relied on handcrafted features and task-specific feature engineering Ramoser et al. (2000); Ang et al. (2012). With the development of deep learning, models such as DeepConvNet and EEGNet demonstrated the ability to automatically capture hierarchical temporal and spatial EEG representations Schirrmeister et al. (2017); Lawhern et al. (2018). Inspired by foundation models in natural language processing and computer vision, recent EEG studies have explored large-scale pretraining to learn transferable representations from heterogeneous datasets. LaBraM and EEGPT leverage token-based and self-supervised learning strategies for general EEG representation learning Jiang et al. (2024b); Wang et al. (2024), while BIOT and REVE further investigate transformer-based architectures and cross-dataset pretraining for robust EEG decoding Yang et al. (2023); El Ouahidi et al. (2026). Benchmarking is essential for understanding the generalization ability of EEG models Kommineni et al. (2026); Kontras et al. (2026). Recently, EEG-FM-Bench Xiong et al. (2026) introduced a comprehensive framework for evaluating EEG foundation models, while EEG-FM-Compass Liu et al. (2026) and Yang et al. Yang et al. (2026) further investigated adaptation strategies of EEG foundation models. However, existing benchmarks predominantly focus on conventional long-window settings, leaving the performance and temporal characteristics of EEG-FMs under short-window conditions largely unexplored.

## 3 EEG Decoding: Task Formulation and Baseline Approaches

### 3.1 Task Formulation

All downstream tasks in ShortEEG-Bench are formulated as supervised, single-label classification problems with dataset-specific label spaces. For each dataset $d$, we denote its labeled data as $\mathcal{D}_d = \{(X_i^{(d)}, y_i^{(d)})\}_{i=1}^{N_d}$, where $X_i^{(d)} \in \mathbb{R}^{C_d \times T_i}$ represents an EEG recording with $C_d$ channels and $T_i$ temporal samples. The corresponding label $y_i^{(d)} \in \mathcal{Y}_d = \{1, \ldots, K_d\}$ indicates one of $K_d$ mutually exclusive classes.

To investigate short-window EEG decoding, each original recording $X_i^{(d)}$ is segmented into $M^{(d)}$ short windows, $\{x_{i,j}^{(d)}\}_{j=1}^{M^{(d)}}$. Each short window inherits the label of its parent recording, i.e., $y_{i,j}^{(d)} = y_i^{(d)}$. Table 1 summarizes the short-window durations for each dataset; appendix provides dataset-specific segmentation details.

### 3.2 BENCHMARK MODELS

**Foundation Models** We group the selected EEG foundation models into three categories based on their representation learning approaches. (1) Masked token and representation prediction: LaBraM (Jiang et al., 2024b) learns discrete neural tokens through vector-quantized spectral modeling and pretrains a transformer to predict masked tokens. EEGPT (Wang et al., 2024) combines masked reconstruction with spatiotemporal representation alignment. (2) Contrastive and heterogeneous-signal representation learning: BENDR (Kostas et al., 2021) uses a convolutional encoder and a transformer contextualizer with masked contrastive learning. BIOT (Yang et al., 2023) tokenizes channel-wise time–frequency segments and incorporates channel and temporal information. (3) Structured spatiotemporal reconstruction: CBraMod (Wang et al., 2025) combines masked reconstruction with criss-cross attention that models spatial and temporal interactions separately within transformer layers. CSBrain (Zhou et al., 2026) uses cross-scale spatiotemporal tokenization and structured sparse attention to capture EEG structure at multiple resolutions. REVE (El Ouahidi et al., 2026) combines large-scale masked autoencoding with four-dimensional positional encoding of electrode location and time to accommodate different recording setups.

**Classic Deep Neural Network Models** EEGNet (Lawhern et al., 2018) combines temporal convolutions, depthwise spatial filtering, and separable convolutions to learn compact spectral–spatial features. EEG Conformer Song et al. (2022) combines convolutional feature extraction with self-attention, coupling local EEG patterns with interactions across temporal tokens. ADFCNN (Tao et al., 2023) uses attention-based dual-scale convolutional feature fusion for motor imagery, providing a task-specific comparator where evaluated. These models are trained from scratch to optimize the downstream classification objective, with all network parameters updated during training.

### 3.3 TASK SETUPS.

**Dataset.** We construct ShortEEG-Bench using 11 publicly available EEG datasets spanning five representative task families. For motor imagery decoding, we include BCIC-IV-2a (Tangermann et al., 2012) and PhysioNet-MI Schalk (2009); Goldberger et al. (2000), which are widely used benchmarks for multi-class motor intention recognition. For cognitive workload assessment, we adopt EEGMat (Zyma et al., 2019), a mental arithmetic EEG dataset containing recordings collected during cognitive task performance. For emotion recognition, we evaluate three widely used affective EEG datasets, including SEED (Zheng & Lu, 2015; Duan et al., 2013), SEED-V Liu et al. (2021), and SEED-VII Jiang et al. (2024a), covering different emotion categories and recording durations. For sleep staging, we use HMC Alvarez-Estevez & Rijsman (2021) and Sleep-EDF (Kemp et al., 2000), both consisting of sleep EEG recordings annotated with standard sleep stages. For clinical EEG analysis, we include three datasets from Temple University Hospital EEG resources Obeid & Picone (2016): TUAB Lopez et al. (2015) for abnormal EEG detection, TUEV Harati et al. (2015) for EEG event classification, and TUSL (von Weltin et al., 2017) for EEG slowing detection. Overall, these datasets cover diverse EEG characteristics, including different task paradigms, recording durations, subject populations, and classification granularities, enabling a comprehensive evaluation of EEG foundation models under short-window decoding scenarios.

**Training configurations.** We consider three training configurations for short-window models. Single-dataset training (ST) adapts a model to an individual dataset. In-domain multi-dataset training (MT-INN) jointly trains a model on multiple datasets belonging to the same task domain, whereas cross-domain multi-dataset training (MT-INT) combines datasets from diverse EEG decoding tasks. For cross-domain training, we investigate two variants: MT-INT(+c), which incorporates clinical EEG datasets, and MT-INT(-c), which exclusively uses datasets collected from healthy participants. These configurations allow us to distinguish beneficial transfer from adverse cross-task interference and to determine whether broader supervision helps or harms short-window decoding.

**Prediction granularity.** Each short-window model is evaluated at two prediction granularities. Short-window (SW) evaluation assesses decoder performance using a single local interval, whereas majority-vote (SW-Vote) evaluation assesses performance after aggregating local predictions within the same parent segment. This distinction is essential because the two evaluation settings address different scientific questions. SW measures whether a local segment contains enough signal for

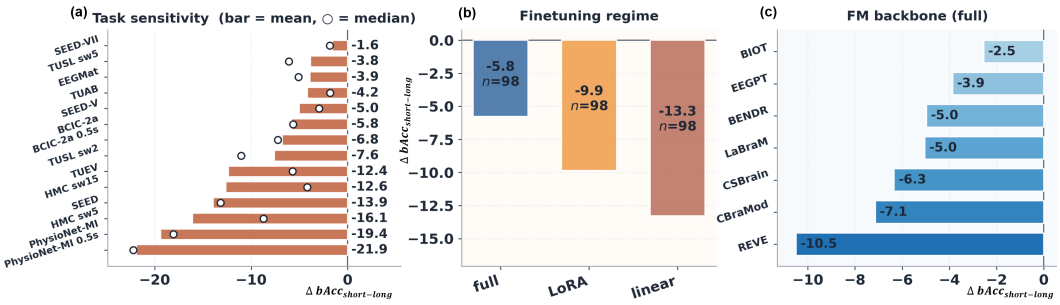


Figure 2: Short-window performance changes across tasks, adaptation strategies, and EEG foundation models. (a) Dataset-wise comparison under shortened input windows. (b) Effects of full fine-tuning, LoRA, and linear probing on performance retention. (c) Model-wise sensitivity to reduced temporal context under the same training protocol.

classification, whereas SW-Vote measures how much information can be recovered after temporally integrating multiple local decisions.

**Fine-tuning strategies.** Full fine-tuning updates the backbone and prediction head to measure downstream adaptability. Linear probing freezes the backbone to assess the intrinsic usefulness of pretrained representations without task-specific updates. LoRA introduces low-rank updates that preserve most of the original parameters while allowing efficient task adaptation. EEGNet, EEG Conformer, and ADFCNN are evaluated in their full-training settings, following the reference protocol. This setup allows us to compare whether the observed short-window trend reflects the pretrained representation itself or the degree to which it is adapted to a target task.

**Evaluation metrics.** Following common practice in EEG decoding, we adopt three standard evaluation metrics: balanced accuracy (bAcc), Cohen's kappa ($\kappa$), and macro-averaged F1-score (Macro-F1). To further quantify the impact of input duration and the effectiveness of short-window voting, we introduce two complementary metrics: *BAcc performance ratio* and *BAcc performance change*. The bAcc performance ratio (R$bAcc$) is calculated by $\mathrm{R}bAcc_{a/b} = \frac{bAcc_a}{bAcc_\mathrm{b}}$, $a, b \in \{\mathrm{short}, \mathrm{vote}, \mathrm{long}\}$. A value of 1 indicates equivalent performance to the long-window baseline, while values below or above 1 represent degradation or improvement, respectively. This normalized metric enables comparisons across datasets with different baseline accuracies. The bAcc performance change ($\Delta bAcc$), defined by $\Delta bAcc_{a-b} = bAcc_a - bAcc_\mathrm{b}$,, measures the absolute difference between short-window (or voting-based) decoding and the long-window baseline.

## 4 BENCHMARKING ANALYSIS

In this section, we present findings addressing our key research questions.

### 4.1 RQ1: How much performance is lost when short windows are available?

To systematically quantify the performance degradation caused by shortening EEG input windows, we investigate three key factors: (1) task and window duration, (2) fine-tuning strategy, and (3) foundation model backbone. **First**, Table 2 reports absolute short-window performance, showing that meaningful decoding remains possible with limited temporal context. For example, REVE achieves 41.146% bAcc on BCIC-IV-2a using only 1 s of EEG, while CSBrain achieves 75.422% on TUSL using 2-s windows. **Second**, regarding task sensitivity, Fig. 2(a) illustrates that shortening the input window leads to performance degradation across all evaluated datasets, although the magnitude of the decline varies substantially across datasets and window durations. **Third**, we investigate whether different adaptation strategies can mitigate the information loss caused by shorter windows. As illustrated in Fig. 2(b), full fine-tuning yields the smallest average bAcc reduction, followed by LoRA and linear probing, indicating that frozen pretrained representations may be insufficient to maintain long-window performance when the available temporal context is substantially reduced. In contrast,

Table 2: Short-window bAcc / $\kappa$ / F1 on four datasets (single-task, full finetuning). Best results in each column (within dataset) are in **bold**; second-best are underlined.

| | BCIC-IV-2a | | | | | | TUSL | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\text{Short}_1$ (1 s) | | | $\text{Short}_2$ (0.5 s) | | | $\text{Short}_1$ (2 s) | | | $\text{Short}_2$ (5 s) | | |
| Model | bAcc | $\kappa$ | F1 | bAcc | $\kappa$ | F1 | bAcc | $\kappa$ | F1 | bAcc | $\kappa$ | F1 |
| BENDR | 26.003 | 0.013 | 0.224 | 25.977 | 0.013 | 0.196 | 49.467 | 0.194 | 0.383 | 47.833 | 0.207 | 0.404 |
| BIOT | 28.021 | 0.040 | 0.203 | 26.454 | 0.019 | 0.211 | 59.644 | 0.333 | 0.548 | 63.361 | 0.397 | 0.595 |
| CBraMod | 36.341 | 0.151 | 0.332 | 33.019 | 0.107 | 0.295 | 63.156 | 0.410 | 0.610 | 62.000 | 0.362 | 0.563 |
| CSBrain | 39.514 | 0.194 | 0.379 | 39.062 | 0.188 | 0.378 | **75.422** | **0.605** | **0.740** | **78.806** | **0.628** | **0.745** |
| EEGPT | 38.416 | 0.179 | 0.356 | 35.286 | 0.137 | 0.326 | 66.333 | 0.421 | 0.592 | 73.889 | 0.533 | 0.668 |
| LaBraM | 31.819 | 0.091 | 0.286 | 33.225 | 0.110 | 0.291 | 61.289 | 0.381 | 0.590 | 66.667 | 0.456 | 0.616 |
| REVE | **41.146** | **0.215** | **0.397** | 40.137 | 0.202 | **0.387** | 60.522 | 0.339 | 0.508 | 71.833 | 0.511 | 0.671 |
| EEGNet | 37.500 | 0.167 | 0.332 | **40.299** | **0.204** | 0.378 | 63.044 | 0.390 | 0.576 | 60.361 | 0.349 | 0.536 |
| Conformer | 34.796 | 0.131 | 0.297 | 30.859 | 0.078 | 0.217 | 53.867 | 0.259 | 0.447 | 52.556 | 0.272 | 0.459 |

| | HMC | | | | | | Sleep-EDF | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\text{Short}_1$ (5 s) | | | $\text{Short}_2$ (15 s) | | | $\text{Short}_1$ (5 s) | | | $\text{Short}_2$ (15 s) | | |
| Model | bAcc | $\kappa$ | F1 | bAcc | $\kappa$ | F1 | bAcc | $\kappa$ | F1 | bAcc | $\kappa$ | F1 |
| BENDR | 61.847 | 0.544 | 0.631 | 68.658 | **0.633** | **0.715** | **64.940** | **0.648** | **0.733** | 69.910 | **0.695** | **0.775** |
| BIOT | 63.885 | 0.559 | 0.659 | **69.162** | 0.630 | **0.715** | 61.508 | 0.614 | 0.710 | 66.833 | 0.662 | 0.748 |
| CBraMod | 62.404 | 0.547 | 0.645 | 66.934 | 0.607 | 0.692 | 63.209 | 0.632 | 0.724 | **70.176** | **0.695** | 0.772 |
| CSBrain | 63.270 | 0.559 | 0.649 | 67.074 | 0.612 | 0.696 | 62.773 | 0.629 | 0.719 | 67.236 | 0.674 | 0.756 |
| EEGPT | **64.689** | **0.575** | **0.665** | 69.047 | 0.630 | 0.704 | 62.854 | 0.629 | 0.721 | 68.903 | 0.679 | 0.764 |
| LaBraM | 62.218 | 0.543 | 0.633 | 68.188 | 0.623 | 0.704 | 61.778 | 0.614 | 0.714 | 69.429 | 0.686 | 0.763 |
| REVE | 63.452 | 0.557 | 0.656 | 68.866 | 0.624 | 0.711 | 62.235 | 0.621 | 0.717 | 67.861 | 0.677 | 0.761 |
| EEGNet | 60.081 | 0.526 | 0.615 | 65.365 | 0.593 | 0.676 | 55.182 | 0.562 | 0.668 | 61.386 | 0.614 | 0.709 |
| Conformer | 61.355 | 0.520 | 0.612 | 66.140 | 0.582 | 0.668 | 56.439 | 0.553 | 0.648 | 61.269 | 0.616 | 0.701 |

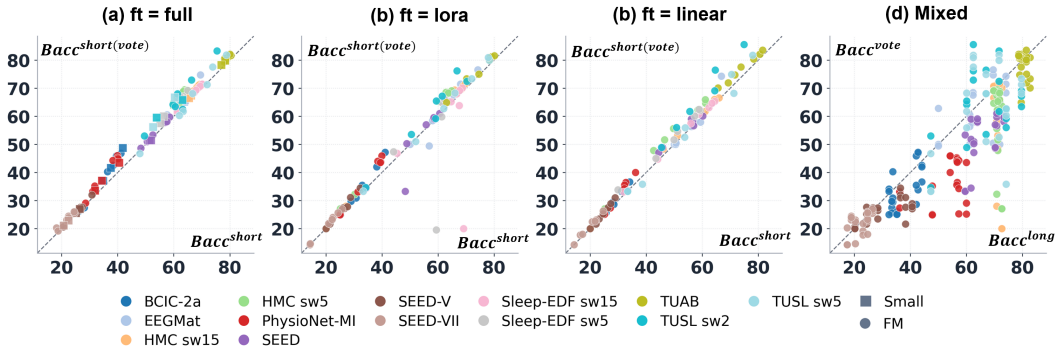


Figure 3: Effect of majority voting on short-window EEG decoding across adaptation strategies. (a–c) Comparison of individual short-window predictions and majority-vote predictions under full fine-tuning, LoRA, and linear probing. (d) Voting performance relative to the long-window baseline across EEG foundation models.

updating the entire model allows more extensive adaptation to short-window signals. **Finally**, we examine whether different EEG foundation models exhibit similar robustness to window shortening under full fine-tuning. Figure 2(c) reveals considerable variation across pretrained backbones. BIOT and EEGPT exhibit relatively small average bAcc reductions, followed by BENDR and LaBraM. A smaller performance reduction does not necessarily imply better absolute short-window decoding accuracy. As shown in Table 2, REVE achieves the highest bAcc with 1-s windows on BCIC-IV-2a despite exhibiting a relatively large average degradation across the evaluated configurations.

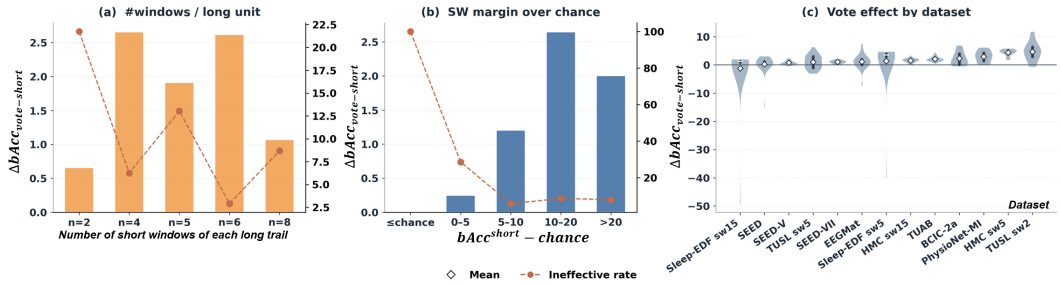


Figure 4: Influence of voting aggregation on short-window decoding. (a) Performance changes as a function of the number of short windows aggregated within the same parent segment. (b) Relationship between voting effectiveness and the margin by which individual short-window accuracy exceeds chance. (c) Dataset-wise distribution of voting outcomes.

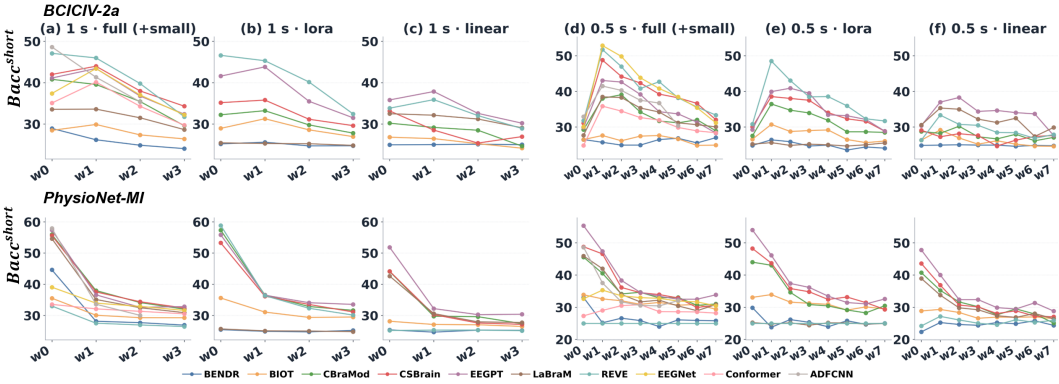


Figure 5: Temporal variation in motor imagery decoding under short-window evaluation. Fine-grained analysis helps identify informative EEG intervals for earlier decisions.

### 4.2 RQ2: Can majority voting recover the lost evidence?

We systematically evaluate majority voting from three perspectives: (1) performance gains across adaptation strategies and datasets; (2) the effect of the number of short windows; and (3) the relationship between voting effectiveness and the discriminative ability of individual short-window predictions. **First**, Fig. 3(a–c) compares short-window decoding before and after majority voting under full fine-tuning, LoRA, and linear probing. Most configurations lie close to or above the diagonal, indicating that voting generally preserves or improves short-window accuracy across adaptation strategies. However, Fig. 3(d) and Fig. 4(c) reveal substantial dataset-level variation: TUSL with 2-s windows and HMC with 5-s windows show relatively large gains, while BCIC-IV-2a and PhysioNet-MI also consistently benefit from voting. **Second**, Fig. 4(a) shows that using four or six short windows yields an average bAcc improvement of approximately 2.6 percentage points, with a lower proportion of ineffective voting outcomes. In contrast, two-window configurations provide limited gains and a higher proportion of ineffective outcomes. **Third**, Fig. 4(b) shows how voting effectiveness varies with the margin by which individual short-window performance exceeds the task-specific chance level. When short-window accuracy is near or below chance, voting provides limited improvement and is frequently ineffective. Performance gains emerge once individual short-window performance exceeds chance, but do not increase monotonically as this margin grows.

### 4.3 RQ3: Does the temporal position affect EEG decoding performances?

To identify the temporal intervals in which discriminative information is concentrated in motor imagery EEG, we investigate the effects of both **window length** and **temporal position after the cue**. As shown in Fig. 5, we compare 1-s and 0.5-s windows on BCIC-IV-2a and PhysioNet-MI under

Table 3: Performance gains $\Delta bAcc_{*-SW}$ under different multi-Task training configurations compared with per-dataset short window training(SW), where $*$ ∈MT-INT,MT-INT(-c), MT-INT(+c)

| | | **BCIC-IV-2a** | | | | | | **HMC** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MT-INN | | MT-INT(-c) | | MT-INT(+c) | | MT-INN | | MT-INT(-c) | | MT-INT(+c) | |
| Model | FT | SW | Vote | SW | Vote | SW | Vote | SW | Vote | SW | Vote | SW | Vote |
| BENDR | lora | -0.11 | -0.62 | -0.11 | -0.48 | -0.06 | -0.10 | -5.17 | -7.16 | -5.03 | -7.16 | -2.40 | -2.82 |
| | linear | 0.09 | -0.12 | -0.08 | -0.50 | -0.02 | -0.12 | -5.17 | -7.16 | -5.03 | -7.16 | -2.40 | -2.82 |
| BIOT | lora | 0.26 | -1.55 | -0.11 | -1.28 | -0.55 | -1.40 | 0.45 | 0.49 | -0.39 | 0.26 | -0.23 | 0.62 |
| | linear | -0.22 | -1.33 | 0.13 | 0.09 | -0.27 | -1.16 | 1.21 | 2.43 | -1.20 | -0.73 | -0.96 | -0.18 |
| CBraMod | lora | 0.70 | -0.37 | -3.75 | -3.90 | -3.82 | -4.07 | 0.78 | 1.66 | -2.49 | -2.87 | -2.00 | -1.97 |
| | linear | -0.09 | -0.41 | -1.96 | -4.75 | -2.42 | -4.66 | -0.03 | -0.76 | 0.22 | 0.80 | -0.33 | -0.36 |
| CSBrain | lora | **2.92** | **0.89** | -2.78 | -3.95 | -4.01 | -5.71 | 0.78 | 0.73 | 0.53 | 0.45 | 0.05 | 0.48 |
| | linear | -0.19 | -0.24 | -1.14 | -1.40 | -1.25 | -2.70 | 0.78 | -0.18 | 1.29 | 2.12 | 1.48 | **2.53** |
| EEGPT | lora | 0.04 | -2.36 | -6.16 | -9.30 | -4.32 | -6.38 | 0.82 | 0.95 | -0.26 | -0.17 | -1.32 | -0.62 |
| | linear | -0.24 | -2.58 | -0.61 | 0.09 | -0.61 | -0.64 | **1.62** | **3.77** | -1.50 | -1.38 | -0.89 | -0.01 |
| LaBraM | lora | 0.10 | -0.24 | **2.59** | **1.35** | **1.10** | **1.46** | -4.43 | -8.04 | **1.56** | 1.70 | **2.80** | 1.57 |
| | linear | -0.45 | -3.82 | -3.43 | -5.61 | -3.81 | -6.08 | 0.28 | -16.52 | -1.46 | -1.79 | -1.11 | -1.38 |
| REVE | lora | 0.92 | -0.09 | -7.65 | -12.38 | -6.69 | -10.94 | 1.12 | 0.82 | 1.25 | **2.42** | -1.35 | 1.42 |
| | linear | -0.25 | -4.08 | -3.38 | -4.25 | -4.09 | -5.79 | -0.70 | -0.99 | -0.51 | -0.57 | -0.82 | -0.97 |

three adaptation strategies: full fine-tuning, LoRA, and linear probing. **First**, early post-cue windows generally provide the most discriminative information. Across the two motor imagery datasets, the highest decoding performance is generally achieved within the first few windows after the cue. **Second**, shorter windows provide finer temporal localization, but their decoding performance is more sensitive to temporal position. Comparing the 1-s and 0.5-s settings reveals an important trade-off. A 1-s window integrates EEG activity over a longer period and yields smoother performance curves across adjacent positions. In contrast, 0.5-s windows provide finer temporal resolution and expose more pronounced local fluctuations. **Third**, temporal sensitivity is consistent across adaptation strategies, but its magnitude differs across models. This temporal pattern is broadly preserved under full fine-tuning, LoRA, and linear probing, which primarily reflects the temporal characteristics of the motor imagery signals rather than a particular adaptation strategy. Nevertheless, high-performing models generally exhibit more pronounced temporal peaks under full fine-tuning and LoRA, whereas linear probing often produces flatter curves and lower peak performance.

### 4.4 RQ4: Does multi-task learning help or hurt short-window decoding?

We investigate the effectiveness of multi-dataset training for short-window EEG decoding by comparing in-domain training (MT-INN) with two cross-domain settings (MT-INT(-c) and MT-INT(+c)). As shown in Table 3, the benefits of multi-dataset training vary substantially across datasets, foundation models, and adaptation strategies. In-domain training generally provides more consistent improvements, whereas cross-domain training can either enhance performance or introduce negative transfer. We further analyze how the inclusion of clinical data and the use of majority voting affect these transfer patterns. **First**, the effectiveness of multi-dataset training is highly task- and model-dependent. In-domain training improves performance for several model-task combinations, while cross-domain supervision may degrade performance even within the same dataset. Conversely, some models benefit more from cross-domain training, indicating that transfer effectiveness depends on the compatibility between source datasets and learned representations. **Second**, incorporating clinical EEG data produces heterogeneous effects. Clinical data can improve transfer performance for some model-dataset pairs while reducing gains for others. Moreover, adaptation strategies strongly influence cross-domain transfer: the same training configuration may improve performance under LoRA-based adaptation but reduce it under linear probing. These results suggest that cross-domain generalization is jointly determined by dataset composition, model architecture, and adaptation strategy. **Third**, the ability of majority voting to offset performance losses varies across multi-dataset training configurations. While voting can improve some short-window predictions, it cannot reliably offset performance losses caused by negative cross-domain transfer.

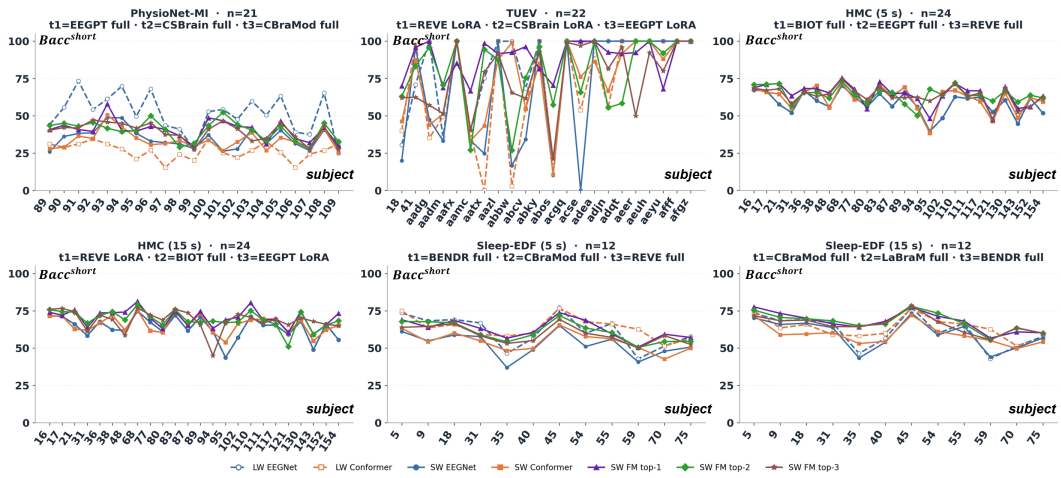


Figure 6: Subject-level comparison of long-window baselines and short-window EEG foundation models. Short-window models can achieve competitive decoding performance and improve robustness for some users with weak baseline performance.

### 4.5 RQ5: Can short-window decoding improve performance across users?

We conduct a subject-level analysis to examine whether short-window EEG foundation models (EEG-FMs) can improve decoding for users with weak long-window baselines. As shown in Fig. 6, we compare the subject-level balanced accuracy of conventional long-window models (EEGNet and EEG Conformer) with three best-performing short-window EEG-FM configurations on each dataset. **First**, short-window EEG-FMs can substantially improve decoding for some subjects with weak long-window baselines, although the benefits are task-dependent. On TUEV, conventional models exhibit large inter-subject variability, with accuracies ranging from near zero to almost 100%, whereas the best short-window EEG-FM achieves more stable performance across most subjects, including those with poor conventional baselines. Similar improvements are observed on HMC and Sleep-EDF. However, the gains are not universal: on PhysioNet-MI, short-window EEG-FMs improve decoding for several subjects with low baseline performance but may underperform long-window EEGNet for subjects with already strong baselines. **Second**, subject-level results indicate that short-window EEG-FMs can alleviate severe performance degradation in individual cases, particularly on TUEV, HMC, and Sleep-EDF. Nevertheless, substantial inter-subject variability remains on some datasets, such as PhysioNet-MI, suggesting that short-window decoding alone does not guarantee uniformly robust performance across users.

## 5 Conclusion

We introduced **ShortEEG-Bench**, a unified benchmark for evaluating EEG foundation models with reduced temporal context. The benchmark covers 11 datasets, five task families, seven pretrained models, and three conventional baselines. Our results show that EEG-FMs remain effective under short-window settings, although performance varies across datasets, architectures, and adaptation strategies. Full fine-tuning provides the most robust adaptation, while majority voting often improves on individual short-window predictions but does not consistently recover long-window performance. Temporal and subject-level analyses further reveal variations hidden by aggregate metrics, including position-dependent decoding changes and improvements for some users with weak baselines. Multi-dataset training provides selective benefits, with in-domain training generally yielding more reliable gains than cross-domain training. These findings establish input duration, temporal position, and subject variability as important dimensions for EEG-FM evaluation. Future work should explore temporally precise annotations and streaming-oriented adaptive decoding, jointly considering accuracy, latency, and reliability. ShortEEG-Bench provides a foundation for temporally resolved evaluation of EEG decoding systems.

### AI USE STATEMENT

In this work, generative AI tools were used exclusively for manuscript language editing. They were not used to design algorithms, generate experimental results, or create figures. All AI-assisted textual revisions were carefully reviewed. The authors take full responsibility for the accuracy, integrity, and final content of this work.

### REPRODUCIBILITY STATEMENT

We will open-source ShortEEG-Bench on GitHub, providing the code, model checkpoints, and detailed documentation needed to reproduce our experiments. The release includes:

- **Short-window data processing:** Preprocessing and segmentation code for constructing short-window datasets across 11 public EEG datasets, including standardized data splits, window configurations, and parent-segment mappings for majority voting.
- **Training frameworks and checkpoints:** Training code and model checkpoints for all evaluated models, covering three adaptation strategies (full fine-tuning, linear probing, and LoRA) and three multi-task training configurations (ST, MT-INN, and MT-INT), including cross-domain variants with and without clinical datasets.
- **Evaluation protocols:** Unified evaluation code for short-window (SW) prediction, short-window majority voting (SW-Vote), and long-window comparisons, together with scripts for reproducing the reported metrics and experimental results.
- **Experimental results:** Detailed window-level and subject-level results across evaluated models and datasets, enabling reproduction of performance comparisons, temporal-position analyses, majority-voting evaluations, and subject-level analyses.

## REFERENCES

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

Kai Keng Ang, Zheng Yang Chin, Chuanchu Wang, Cuntai Guan, and Haihong Zhang. Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in neuroscience*, 6:39, 2012.

Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pp. 81–84. IEEE, 2013.

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. volume 38, pp. 22541–22577, 2026.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

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, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 16(2):969–985, 2024a.

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

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Tiantian Feng, and Shrikanth Narayanan. A multi-dimensional framework for evaluating generalization in eeg foundation models. 2026.

Konstantinos Kontras, Trui Osselaer, Stylianos G Mouslech, Angeliki-Ilektra Karaiskou, Guido Gagliardi, Thomas Strypsteen, Mohammad Hossein Badiei, Anku Rani, Maarten Vanmarcke, Miguel Bhagubai, et al. Neuroatlas: Benchmarking foundation models for clinical eeg and brain-computer interfaces. *arXiv preprint arXiv:2605.14698*, 2026.

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

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

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. Eeg-fm-compass: Progress, benchmarking, and future directions for eeg foundation models. *National Science Review*, pp. nwag466, 2026.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, (2):715–729, 2021.

Sebas Lopez, G Suarez, D Jungreis, I Obeid, and Joseph Picone. Automated identification of abnormal adult eegs. In *2015 IEEE signal processing in medicine and biology symposium (SPMB)*, pp. 1–5. IEEE, 2015.

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

Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4):441–446, 2000.

Gerwin Schalk. Eeg motor movement/imagery dataset, 2009.

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

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

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

Wei Tao, Ze Wang, Chi Man Wong, Ziyu Jia, Chang Li, Xun Chen, CL Philip Chen, and Feng Wan. Adfcnn: Attention-based dual-scale fusion convolutional neural network for motor imagery brain–computer interface. *IEEE transactions on neural systems and rehabilitation engineering*, 32:154–165, 2023.

Eva von Weltin, Tameem Ahsan, Vinit Shah, Dawer Jamshed, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. Electroencephalographic slowing: A primary source of error in automatic seizure detection. In *2017 IEEE signal processing in medicine and biology symposium (SPMB)*, pp. 1–5. IEEE, 2017.

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. volume 37, pp. 39249–39280, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

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 *International conference on learning representations*, 2026.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. volume 36, pp. 78240–78260, 2023.

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.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. *Advances in Neural Information Processing Systems*, 38: 87150–87195, 2026.

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