BEYOND ACCURACY: ROBUSTNESS, INTERPRETABILITY AND EXPRESSIVENESS OF EEG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG foundation models (EEG-FMs) have been evaluated predominantly on clean, in-distribution accuracy, demonstrating modest gains over supervised baselines and weak frozen representations. This study examines whether these conclusions hold beyond clean accuracy by evaluating six EEG-FMs and a supervised baseline across ten datasets along three layers of analysis: *(i)* Robustness: we apply test-time perturbations including additive noise, random and region-based channel dropout and region-specific noise injection. Our analyses show that no single model dominates all failure modes. The most noise-robust model is among the most fragile under channel dropout and much of the dropout fragility disappears when channels are removed rather than zero-padded. *(ii)* Interpretability: using attribution methods in EEG-FMs, we show that models broadly concentrate relevance on task-appropriate brain regions consistent with known neurophysiology. *(iii)* Expressiveness: we demonstrate that the poor head-only performance previously attributed to low-quality pre-trained representations is largely explained by the pooling strategy and that EEG-FMs possess sufficient representational capacity when their token-level embeddings are preserved. Furthermore, with block-wise probing and attention analysis we show that late blocks are repurposed during fine-tuning, while early blocks already hold task-related information. Our results show that conclusions about EEG-FMs depend on evaluation choices and we recommend that future evaluation of EEG-FMs should report robustness per perturbation type, produce attribution maps and examine multiple pooling strategies.

1 INTRODUCTION

Brain-Computer Interfaces (BCIs) translate neural activity into control signals for communication, rehabilitation and clinical monitoring (McFarland & Wolpaw, 2011). Electroencephalography (EEG) is the most practical neuroimaging modality for BCIs: it is non-invasive, portable, inexpensive and offers millisecond temporal resolution (Nicolas-Alonso & Gomez-Gil, 2012). However, EEG signals have a low signal-to-noise ratio due to muscular and environmental artefacts and vary across subjects, sessions and recording devices (Saha & Baumert, 2020).

To address these challenges, various methods and techniques have been deployed over the years for EEG decoding. Traditional BCI systems relied on hand-crafted features and machine learning algorithms to perform EEG classification (Pfurtscheller & Neuper, 2001; McFarland et al., 2006; Bashashati et al., 2007; Handy, 2009; Rao, 2013; Nam et al., 2018). The advent of deep learning reduced the reliance on manual feature selection (LeCun et al., 2015; Schirrmeister et al., 2017; Lawhern et al., 2018; Santamaría-Vázquez et al., 2020; Barmpas et al., 2023a; Song et al., 2023), with neural networks learning spatial and temporal representations end-to-end from raw or lightly preprocessed EEG data. Although these deep models achieve competitive performance across diverse BCI paradigms, they require large labelled datasets (a persistent bottleneck in EEG research) and often lose performance when applied across cohorts or recording conditions (Lotte et al., 2018). Inspired by the success of foundation models in natural language processing and computer vision (Brown et al., 2020; Touvron et al., 2023), researchers have begun pre-training large models on EEG data with the goal of learning general-purpose representations. These EEG foundation models

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.

(EEG-FMs) pre-train on large unlabelled corpora via self-supervision, aiming to generalise across users, tasks and conditions with minimal labelled fine-tuning data.

Although several EEG-FMs (e.g., Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025; Xiao et al., 2025) have been proposed that claim competitive downstream task performance, multiple independent studies converge on the same finding: *EEG-FM downstream task performance gains over supervised baselines are modest.* Under subject-independent cross-validation, Lee et al. (2025a) found that EEG-FMs achieve marginal advantage over supervised baselines such as EEGNet, a finding later corroborated through systematic fine-tuning (Lee et al., 2025b) and few-shot calibration (Širca et al., 2026). Another large-scale study by Liu et al. (2026), covering twelve models, reached a similar conclusion: *specialist architectures trained from scratch sometimes outperformed EEG-FMs. Furthermore, linear probing of pre-trained models underperformed full fine-tuning across all models*.

These benchmarks, thorough as they are, share a common limitation: they evaluate EEG-FMs on clean, in-distribution data only. In practice, EEG recordings are inherently noisy due to electrode impedance drifts, gel dry-out, subject movements, and channel failures, leading to substantial variability across sessions and users (Saha & Baumert, 2020). Consequently, models that perform well on curated laboratory data may significantly degrade under these conditions, limiting their value in clinical or real-world BCI settings (Zhou et al., 2023; Kuruppu et al., 2026). Several works (Barmpas et al., 2024; 2023b) formalise this concern from a causal perspective, showing that data distribution shifts, arising from inter-subject variability but also sensor noise or electrode displacement, are among the main factors limiting generalisation in BCI systems. Yet, such conditions remain largely unexamined in current EEG-FM benchmarks. Among other concerns, Kuruppu et al. (2026) highlights three gaps relevant to our study: *(i)* robustness of EEG-FMs under simulated signal corruption has not been demonstrated, *(ii)* interpretability remains unaddressed despite being necessary in high-risk domains such as healthcare and *(iii)* head-only probing of pre-trained EEG-FMs consistently underperforms fine-tuning across all reviewed models, raising questions about the quality of pre-trained representations.

The goal of this work is to test whether the conclusions currently drawn about EEG-FMs hold once evaluation extends beyond clean accuracy. We adopted the benchmarking protocol of Lee et al. (2025a) and extended the model set with state-of-the-art EEG-FMs: REVE (El Ouahidi et al., 2025), BrainOmni (Xiao et al., 2025), NeuroRVQ (Barmpas et al., 2026), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025) and BIOT (Yang et al., 2023), and EEGNet (Lawhern et al., 2018) as the baseline. We also added six benchmark datasets to the original set, totalling ten datasets spanning motor imagery, motor execution, event-related potentials, working memory, sleep staging, eyes-open versus eyes-closed classification and seizure detection. We conducted three layers of analysis to investigate the gaps highlighted by Kuruppu et al. (2026): *(i)* **Robustness:** using synthetic perturbations, we examined the performance of fine-tuned EEG-FMs to isolate distinct failure modes. *(ii)* **Interpretability:** using attribution maps as a plausibility test, we examined whether relevance falls on task-appropriate regions consistent with known neurophysiology. *(iii)* **Expressiveness:** using pooling ablations on frozen and fine-tuned backbones, we examined whether weak head-only performance reflects poor representation quality or the pooling strategy, and used block-wise probing on two best-performing models to examine where task-related information resides. Our contributions can be summarised as follows:

1. **Robustness is perturbation-specific**. No single model is most robust across all tested perturbations. For example, the most noise-robust model is the most fragile under zero-padded channel dropout while all models (that accept variable input channels) are less fragile when channels are removed rather than zero-padded, pointing to the handling of zero-valued inputs rather than loss of task information.
2. **Attribution maps pass a spatial plausibility test**. Attribution maps are broadly consistent with expected topographies, with model-dependent differences. This is necessary but not sufficient evidence of trustworthy decoding.
3. **Poor head-only performance is substantially a pooling artefact**. Replacing mean pooling with token flattening improves head-only performance for all tested models. A parameter-matched control shows the gain is not explained by head size. Therefore, frozen EEG-FM representations retain more information than previous weak linear probing results suggested, and block-wise probing shows that early blocks already carry task information while fine-tuning mainly reshapes the later blocks.

Based on this analysis, we recommend that future evaluation of EEG-FMs should report robustness per perturbation type (stating the exact testing configurations, such as zero-padded channels or removed), produce attribution maps for spatial plausibility evaluation and examine multiple pooling strategies before drawing conclusions about the representation quality of EEG-FMs.

2 Related Work

While clean accuracy has been extensively benchmarked for EEG-FMs, robustness, interpretability and expressiveness remain underexplored. This section reviews related work along these axes.

2.1 Robustness and Expressiveness of EEG Foundation Models

EEG signals are inherently noisy due to electrode artefacts and channel failures (Saha & Baumert, 2020). While additive noise and channel dropout have been used as augmentation techniques during training (Cheng et al., 2020), to our knowledge no study has systematically evaluated existing EEG-FMs under test-time signal corruption. Separately, linear probing consistently underperforms full fine-tuning across all reviewed EEG-FMs, typically interpreted as evidence of low-quality representations (Liu et al., 2026; Kuruppu et al., 2026; Xiong et al., 2026). However, this interpretation conflates representation quality with the pooling strategy used to aggregate tokens before the probe. The effect of pooling on frozen representations has been examined in other domains (Rauch et al., 2026) but not in EEG-FMs, where classifier heads have only been compared under full fine-tuning (Xiong et al., 2026).

2.2 Interpretability of EEG Foundation Models

Interpreting EEG-FMs requires faithful attribution methods. We review the main approaches used in EEG classification (Zhou et al., 2023) and EEG-FMs (Ma & Wang, 2026; Xiong et al., 2026; Meyer zu Bexten et al., 2026), and the rules needed to extend them to transformer backbones. Gradient$\times$Input (G$\times$I) (Shrikumar et al., 2017; Ancona et al., 2018) computes attribution by element-wise multiplying each input feature with its gradient of the output. Gradient-weighted Class Activation Mapping (GradCAM) (Selvaraju et al., 2017) produces a coarse attribution map by computing a weighted gradient feature map to the last convolution layer in a CNN model, highlighting which spatial regions most influence the prediction. Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) decomposes a model's output into per-input relevance scores by propagating relevance backward through every layer under conservation rules. Both LRP and G$\times$I pass standard sanity checks on EEG (Ravindran & Contreras-Vidal, 2023). GradCAM works primarily on CNNs and standard LRP rules are not directly applicable to the bilinear query-key interaction inside self-attention modules. For transformers, attention rollout (Abnar & Zuidema, 2020; Chefer et al., 2021) aggregates attention across layers but does not provide a faithful decomposition of the prediction and suffers from limited resolution. AttnLRP (Achtibat et al., 2024) resolves this by extending LRP with propagation rules for the operations present in self-attention, preserving the conservation property across the transformer. Beyond attribution methods, sparse autoencoders (SAEs) have been used to extract clinically grounded concepts from EEG-FM model embeddings (Lehn-Schiøler et al., 2026).

3 Methods

In this work, we benchmarked six EEG foundation models and one supervised baseline across ten datasets. Each model was tested on clean and perturbed data to assess performance and robustness, and with various attributions examined if learned representations align with expected brain regions.

Models. The six EEG-FM models had a backbone parameter count spanning four orders of magnitude larger than the baseline model: BIOT (Yang et al., 2023) (3.19M), CBraMod (Wang et al., 2025) (4.88M), LaBraM (Jiang et al., 2024) (5.82M), NeuroRVQ (Barmpas et al., 2026) (5.87M), BrainOmni (Xiao et al., 2025) (37.76M, of which a 5.05M VQ tokeniser stayed frozen during fine-tuning) and REVE (El Ouahidi et al., 2025) (69.19M). EEGNet (Lawhern et al., 2018) (3.4K parameters) served as the supervised baseline trained from scratch on each task. We evaluated every foundation model under two regimes: head-only probing on a frozen backbone and full fine-tuning

of all parameters. The six foundation models were initialised from publicly available pre-trained checkpoints. Full architectural details, head layer specifications, and parameter breakdowns are provided in Appendix A. All models were trained following the protocol described in Appendix B, along with compute costs.

Datasets and Tasks. We evaluated the models on ten datasets: a 4-class executed movement dataset (Schirrmeister et al., 2017) (*Movement-HG*), two tasks from the Korean University OpenBMI (Lee et al., 2019) covering 2-class motor imagery (*Motor-Imagery-KU*) and visual P300 event-related potentials (*ERP*), three tasks from the PhysioNet EEG Motor dataset (Schalk et al., 2004; Goldberger et al., 2000) covering 4-class motor imagery (*Motor-Imagery-PN*), 4-class motor execution (*Movement-PN*), and eyes open versus closed (*Eyes*), a digit span working memory task (Pavlov et al., 2022) (*Memory*), a 6-class sleep staging dataset (Kemp et al., 2000) (*Sleep-EDF*), a 5-class clinical sleep staging dataset (Alvarez-Estevez & Rijsman, 2022) (*Sleep-HMC*), and a seizure detection task (Detti, 2020) (*Seizure*). All recordings are common-average referenced and bandpass filtered following each model's preprocessing. Dataset and preprocessing details are listed in Appendix C.

Robustness to Perturbations. All perturbations were applied at test time only to models trained on clean data, so any performance drop reflects the fragility of learned representations. We tested three settings (full details and parameters are given in Appendix D): *(i) Additive noise.* Independent white and pink ($1/f$) noise was added to every channel at signal-to-noise ratios $\{10, 5, 0, -3, -5\}$ dB, with an additional -15 dB level for white noise. *(ii) Random channel dropout.* Channels were independently selected for zero-padding with per-channel probabilities $p \in \{0.1, 0.3, 0.5\}$. For the four models that accept a variable channel set (LaBraM, NeuroRVQ, REVE and BrainOmni), we also removed the dropped channels instead of zeroing them. *(iii) Region-based perturbations.* For each task, we defined a primary electrode region expected to carry the task signal and a control region with no expected contribution (assignments in Appendix D). We tested both zero-padding and noise injection (with 5 dB and -3 dB) for each region.

Interpretability Analyses. To understand why models are robust or fragile and whether their learned features align with known neurophysiology, we adopted *AttnLRP* as our primary attribution method (as motivated in Section 2.2) and used G$\times$I as an alternative for BrainOmni and NeuroRVQ where AttnLRP is numerically unstable (detailed in Appendix I). We supplement AttnLRP with raw *attention maps* and *linear probes* at each transformer block to locate where task-relevant information emerges, focusing on REVE and NeuroRVQ as the two best performing models in Section 4.1. Full implementation details and model-specific adaptations are described in Appendix I. We show attribution maps for five benchmarks: Movement, Motor-Imagery, ERP, Eyes and Seizure.

4 Results

4.1 Clean Performance

Figure 1 reports balanced accuracy per benchmark for each model under full fine-tuning and head-only adaptation (full tables and figures are in Appendix F while Appendix G has a training-from-scratch analysis). Under full fine-tuning, the top six models cluster within 6.0 percentage points (0.647–0.707) on average balanced accuracy (bacc). Per-benchmark winners differ across tasks. These results confirm the small-gains of EEG-FMs over smaller models finding of other studies. NeuroRVQ leads on average (0.707) despite having an order of magnitude fewer parameters than the largest models tested. EEGNet (0.647) matches or exceeds several foundation models on some tasks. Memory only peaks at 0.603 (CBraMod), compared with chance accuracy of 0.500. The large standard deviations across folds on Seizure (0.085–0.155) highlight the difficulty of generalizing across patients in clinical seizure data. BIOT fails to learn on almost every task and is excluded from subsequent analyses. Head-only adaptation underperforms full fine-tuning for every foundation model, with average performance spanning 0.432–0.633 (REVE performing best), all below our deep learning baseline EEGNet (0.647). The model heads differ in size and pooling (Appendix A), and we examine the effect of pooling strategy in Section 4.5.

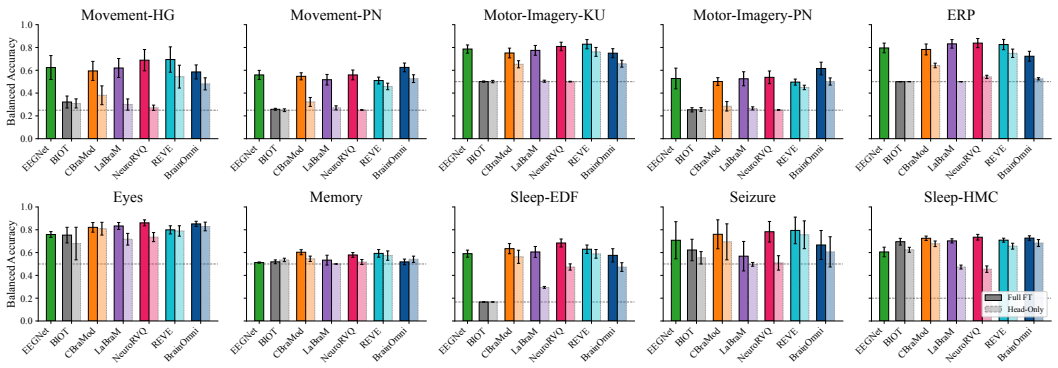


Figure 1: Classification balanced accuracy of full fine-tuned and head-only fine-tuned EEG-FMs and a deep learning baseline across 10 downstream tasks: Movement-HG, Motor-Imagery-KU, Movement-PN, Motor-Imagery-PN, ERP, Eyes, Memory, Sleep-EDF, Seizure and Sleep-HMC. Bars show mean balanced accuracy over ten subject-independent cross-validation folds (error bars show one standard deviation). The horizontal lines mark chance accuracy, $1/n_{\text{classes}}$.

4.2 Robustness Analysis

Figure 2 shows mean balanced accuracy over eight datasets. To use the same dataset set across all four perturbation conditions, we exclude the two small-montage sleep datasets (Sleep-EDF: two bipolar derivations; Sleep-HMC: four derivations) from all averages because their limited channel counts make random channel dropout unsuitable. Dataset-specific perturbation results are in Appendix H.

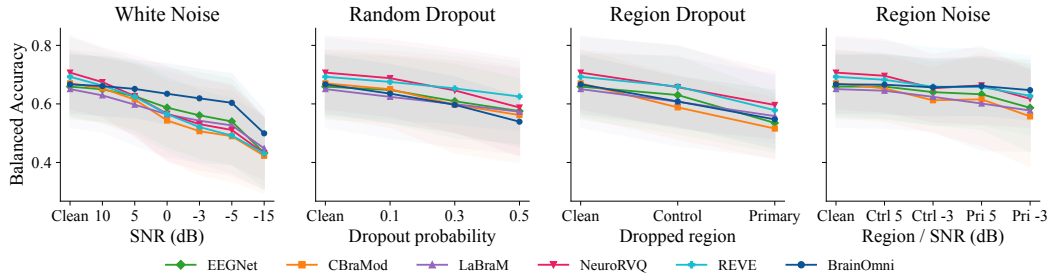


Figure 2: Balanced accuracy ± 1 standard deviation under full fine-tuning, averaged across eight datasets (excluding Sleep-EDF and Sleep-HMC): (a) additive white noise, (b) random channel dropout, (c) region dropout and (d) region noise injection. Per-task figures are in Appendix H.

White Noise. Degradation patterns appear to be task-dependent. At moderate noise (5 to 10 dB), models maintain separation, indicating a **model-limited** regime where architectural differences matter. At extreme noise (≤ -3 dB), most models converge toward a task-specific floor, defining a **signal-limited** regime in which the corruption dominates regardless of architecture: The Eyes and Motor-Imagery tasks remain above chance even at -5 dB, while the rest degrade more sharply. BrainOmni is the most noise-robust model: at -5 dB it loses only 6.3 percentage points compared to 11.9–20.1 for all other models. Pink noise produces similar patterns (see full results Appendix H).

Random Channel Dropout. At p=0.5, average losses range from -6.8 percentage points (REVE) to -12.8 (BrainOmni). Among the eight included datasets, the largest drop is -35.1 percentage points on Movement-HG and smaller on others. BrainOmni, the most noise-robust model, has the largest average loss under zero-padded dropout, showing that robustness to one perturbation type does not transfer to another. When dropped channels are removed entirely rather than zero-padded (Appendix H.2.2), the most affected models recover substantially (from -12.8 to -2.4 percentage points for BrainOmni and from -11.9 to -4.8 for NeuroRVQ), indicating that their fragility stems from how they handle zero-valued inputs rather than from the loss of task-related information itself.

Region Dropout. Zero-padding the primary (task-related) region hurts every model on average. LaBraM, NeuroRVQ and REVE have the smallest losses, from -9.2 to -11.4 percentage points. Removing the dropped region instead of zero-padding it reduces the average loss for all four models that accept variable channel sets, with BrainOmni showing the largest gain from -12.0 to -3.8 percentage points (Appendix H.3). Dropping control regions preserves most performance on average, with the notable exception of Movement-PN and Motor-Imagery-PN where most models degrade more under control dropout. We attribute this to eye-movement artefact contamination in these datasets (Lee et al., 2025a) and therefore exclude them from all interpretability analyses.

Region Noise. BrainOmni has the smallest average loss under primary-region noise at -3 dB (-2.0 percentage points), while CBraMod has the largest loss (-11.3 percentage points). Every model loses more on average under primary than control noise at this level. At $+5$ dB, losses remain visible under primary-region noise but are small under control-region noise (-1.7 to 0.0 percentage points).

4.3 Attribution Maps

Following the methods and techniques described in Section 3 (implementation details in Appendix I), class-averaged attribution topographic maps were extracted for each EEG-FM across five benchmarks. The maps in Figure 3 often emphasise task-appropriate regions: central and centroparietal electrodes for motor tasks (Movement-HG, Motor-Imagery-KU), occipital sites for ERP and occipital and anterior electrodes for Eyes. This alignment varies by task and model. This provides evidence of a connection to expected neurophysiology, one of the concerns raised by Kuruppu et al. (2026).

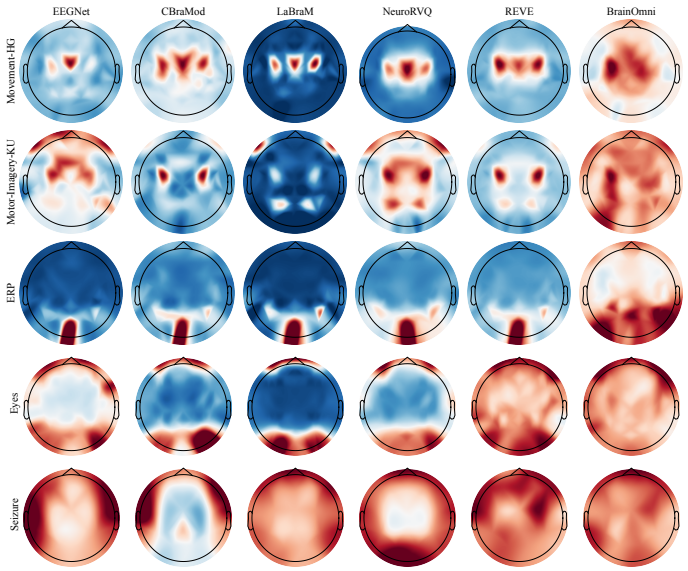


Figure 3: Class-averaged attribution topographic maps. Columns (Models): EEGNet, CBraMod, LaBraM, NeuroRVQ, REVE and BrainOmni. Rows (Benchmarks): Movement-HG, Motor-Imagery-KU, ERP, Eyes and Seizure.

Localisation. Models differ in how sharply they localise relevance. LaBraM, REVE, CBraMod, and NeuroRVQ produce tightly focused peaks over the expected cortical areas. BrainOmni is the most diffuse, spreading relevance across the scalp rather than concentrating on brain regions.

Model-Specific Patterns. On Eyes task, diverging from the occipital focus of other models, REVE and BrainOmni present diffused behaviour. On Motor-Imagery-KU, NeuroRVQ, LaBraM and EEGNet show parietal relevance alongside the expected central peak. These three models also exhibit frontal activation consistent with potential eye movement artefacts, whereas REVE and CBraMod do not. On Seizure, CBraMod concentrates relevance over temporal electrodes, while the other models produce diffused maps.

4.4 BLOCK-WISE PROBING

To understand where task information emerges, we train linear probes on intermediate representations extracted after each fine-tuned and pre-trained transformer block (details and additional analyses in Appendix I.5). For computational reasons, we focus on REVE and NeuroRVQ as the two highest-performing models.

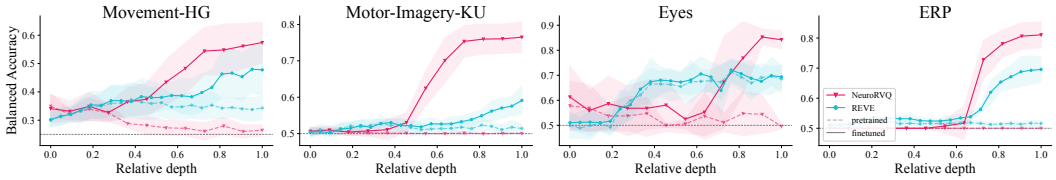
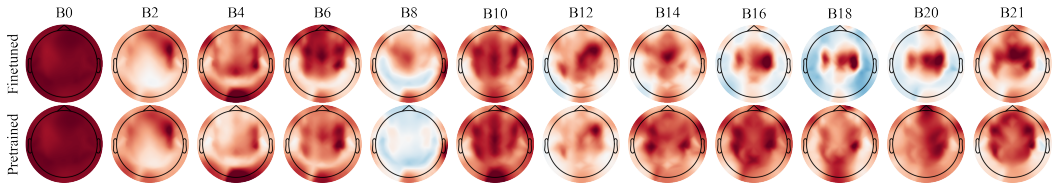


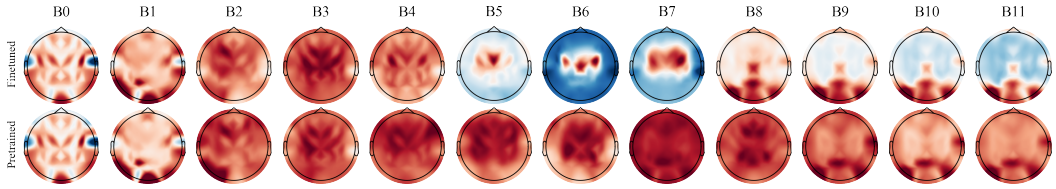
Figure 4: Mean-pooled linear probing accuracy ($\pm$ 1 standard deviation over ten folds). Tasks from left to right: Movement-HG, Motor-Imagery-KU, Eyes and ERP. Relative depth runs from the output of the first block (0) to the output of the last (1). Dashed lines: pre-trained, solid lines: fine-tuned.

Figure 4 shows mean-pooled probing accuracy by relative depth (0 = first block output, 1 = final block output) for pre-trained and fine-tuned checkpoints across four benchmarks. For both models, pre-trained and fine-tuned curves remain close in early blocks and tend to diverge in middle to late blocks. This suggests that task-relevant, linearly decodable signal is similar in early blocks before and after fine-tuning, while later blocks gain task-discriminative structure. The exact depth and size of this separation vary by task, with little separation for REVE on Eyes. Complementary analyses (where block-wise flatten probing is performed) show that task information is already present earlier in the pre-trained model (Appendices I.5.1 and I.6).

Attention maps show a similar transition point for both models: pre-trained and fine-tuned patterns overlap in early blocks and diverge at the same depth where probing accuracy separates (Figure 5). Under fine-tuning on the Movement-HG task, REVE's later blocks sharpen onto task-relevant regions, whereas NeuroRVQ's only do so in middle blocks while later blocks shift attention occipitally. The full set of tasks is given in Appendix I.4.1.



(a) REVE



(b) NeuroRVQ

Figure 5: Per-block attention topographic maps in the Movement-HG task. Top row of each panel: fine-tuned model. Bottom row: pre-trained model. Blocks (B) ordered from input to output.

4.5 HEAD POOLING STRATEGY

Poor head-only performance of some models can reflect representation quality or pooling strategy. To further examine this, we compare mean pooling, attention pooling, token flattening and a mean-tile control for LaBraM, NeuroRVQ, REVE, BrainOmni and CBraMod. Mean-tile repeats the pooled vector to match the flattened input size and head parameter count, but does not preserve individual

tokens. We use linear classifiers for this ablation and evaluate both head-only adaptation and full fine-tuning on four tasks (Appendix J).

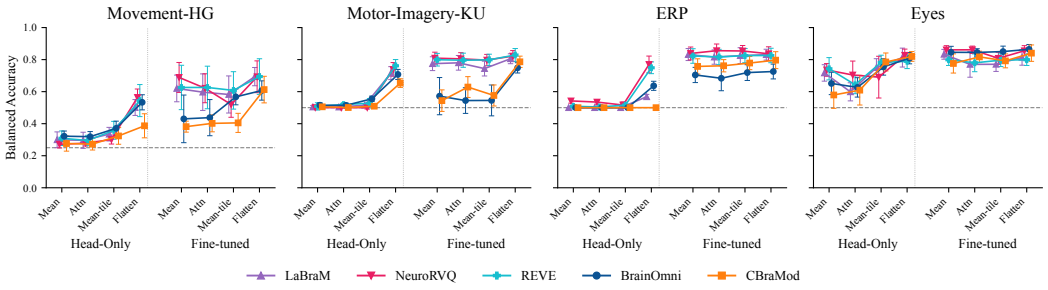


Figure 6: Balanced Accuracy $\pm$ 1 standard deviation of mean, mean-tile, attention and flatten pooling for LaBraM, NeuroRVQ, REVE, BrainOmni and CBraMod under head-only adaptation and full fine-tuning. Mean-tile matches the flatten head's parameter count without preserving tokens.

Under head-only adaptation, flatten pooling gives the highest average accuracy for every model (Figure 6 and Table 20). Mean-tile improves several models but does not recover the flattening gain, showing that matching the head's parameter count alone is insufficient. Attention pooling does not lead to substantial improvements over mean pooling. Flatten pooling largely closes the head-only gap with EEGNet for the strongest models: NeuroRVQ reaches 0.725 and REVE 0.711, compared with EEGNet's 0.741 on the same four tasks. These results show that poor head-only performance reflects selected pooling strategy rather than weak pre-trained representations. Under full fine-tuning, the effect of pooling is smaller for most models. The mean-to-flatten gain is negligible for NeuroRVQ and smaller for REVE and LaBraM, but remains present for BrainOmni and CBraMod. During full fine-tuning, the model backbone can be adapted along with the classification head, thereby reducing the impact of suboptimal pooling strategies.

5 Discussion

Prior benchmarking works (Lee et al., 2025a;b; Liu et al., 2026) have examined the performance of EEG-FMs on clean in-distribution data, reporting modest gains over supervised baselines and weak frozen representations. In this work, we examined *(i)* whether these conclusions change under synthetic perturbations, *(ii)* whether attributions are spatially plausible and *(iii)* whether weak head-only performance reflects weak representational quality of EEG-FMs. Across six EEG-FMs and ten datasets, we find that several conclusions depend on evaluation choices.

Robustness. Robustness is a property of perturbation type and model, and EEG-FMs do not consistently outperform EEGNet. At extreme noise (≤ -3 dB), models start to converge towards a signal-limited floor that varies by task, while at moderate noise and under channel dropout model differences remain decisive. Under additive noise, BrainOmni leads by a wide margin compared to other EEG-FMs and EEGNet. However, no single model dominates all perturbation types: BrainOmni is the most fragile under zero-padded random dropout, where REVE leads (potentially due to its masked-channel pre-training strategy that other models do not follow). Furthermore, our analysis shows that this channel dropout fragility depends largely on the evaluation protocol. When channels are removed rather than zero-padded, the four variable-channel models (NeuroRVQ, LaBraM, BrainOmni and REVE) have smaller average losses (Appendix H.2.2). Dropout robustness therefore cannot be compared across studies unless the handling of missing channels is reported. Inference-time quantisation is not necessary for BrainOmni's noise-robustness advantage, as shown in Appendix H.4. Two plausible sources remain: *(a)* compression of input channels into latent variables before the transformer and *(b)* multimodal EEG and MEG pre-training. Cross-modal transfer reported by Gokce et al. (2026) motivates testing the latter, but does not establish its role in robustness.

Interpretability. Attribution maps generally place relevance on task-appropriate brain regions (central electrodes for motor tasks, occipital for visual ERP and occipital/frontal for eyes open/closed) with several exceptions (e.g. REVE and BrainOmni produce diffuse maps on Eyes). Region dropout

supports this by removing electrodes rather than measuring sensitivity: removing primary regions caused larger average losses than removing control regions. Together, these results partially address the concern of Kuruppu et al. (2026) that EEG-FMs must connect to known neurophysiology to be considered trustworthy. However, spatial plausibility is necessary but not sufficient for trustworthy decoding (Jacovi & Goldberg, 2020). On Motor-Imagery-KU, for example, NeuroRVQ, LaBraM and EEGNet show frontal relevance consistent with eye-movement artefacts alongside the expected central peak, meaning a broadly plausible map can co-exist with artefacts. Differences in attribution patterns between models remain, meaning that architectural or pre-training decisions influence what features models rely on. For example, BrainOmni's attribution maps are generally diffuse, which may reflect that its transformer operates over abstract tokens rather than physical electrodes and/or its inference-time VQ module introduces a non-differentiable step that disrupts relevance propagation. Establishing which features actually drive model predictions will require feature-level methods, such as sparse autoencoders that separate features within model activations (Lehn-Schiøler et al., 2026).

Expressiveness. The poor performance of pre-trained EEG-FMs, previously interpreted as evidence of low-quality representations (Lee et al., 2025b; Kuruppu et al., 2026; Xiong et al., 2026), is substantially a pooling strategy artefact. Replacing mean pooling with flatten pooling improves head-only accuracy for all five models tested and narrows the gap with EEGNet to 1.6–3.0 percentage points for the strongest models, suggesting that mean pooling discards task-discriminative information present in the token-level embeddings. The effect is smaller under full fine-tuning, as the backbone can adapt to the pooling strategy. Frozen EEG-FM representations therefore retain more task information than previous mean-pooled probing results suggested. Linear probing across transformer depth shows that fine-tuning reshapes the later blocks, where linear decodability of the downstream task grows. Complementary analyses on pre-trained representations of the best models motivate the hypothesis that fine-tuning repurposes these blocks from the pre-training objective, a pattern previously observed in pre-trained NLP transformers (Voita et al., 2019; Merchant et al., 2020). We recommend that future evaluations of pre-trained EEG representations test multiple pooling strategies, to avoid conflating pooling design with representation quality.

6 Limitations

Our evaluation has several limitations. All models are trained on clean data only, leaving open whether noise augmentation or adversarial training would close the observed robustness gaps. No dropout protocol ranks all six models fairly: zero-padding penalises variable-channel models, while true removal excludes fixed-channel baselines. Small-montage datasets were also excluded from robustness experiments. More broadly, our synthetic perturbations do not capture the full scope of signal corruption seen in deployment, such as eye-movement and muscle artefacts or electrode shift and rotation. The roles of multimodal pre-training and latent compression in BrainOmni's noise robustness cannot be disentangled without matched pre-training ablations. Our attribution analyses do not examine incorrect or low confidence trials to show what goes wrong when decoding fails and we leave this for future work. Lastly, attribution maps reveal sensitivity, while region perturbations test the effect of changing inputs, and neither establishes which model features drive a prediction.

7 Conclusion

In this work, we attempted to go beyond clean accuracy to evaluate the robustness, interpretability and expressiveness of six recent EEG-FMs across ten datasets. We found that several conclusions currently drawn about EEG-FMs depend on evaluation choices. Robustness depends on both perturbation type and model. No single model is most robust across all perturbations. Attribution maps generally place relevance on task-appropriate regions, which is a necessary but not sufficient condition for trustworthy decoding. Finally, the weak head-only performance previously taken as evidence of poor representations is substantially a pooling artefact. We therefore recommend that future evaluations of EEG-FMs report robustness per perturbation type (stating the exact testing configurations), produce attribution maps for spatial plausibility evaluation and examine multiple pooling strategies before drawing conclusions about representation quality.

8 AI USE DISCLOSURE

AI was used to assist in improving writing style, grammar and readability of this paper. All AI-assisted text was reviewed by the authors, and we take full ownership for the content of this work.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 135–168. PMLR, 2024. doi: 10.48550/arXiv.2402.05602.

Diego Alvarez-Estevez and Roselyne Rijsman. Haaglanden Medisch Centrum sleep staging database. *PhysioNet*, 2022. doi: 10.13026/t79q-fr32.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. doi: 10.48550/arXiv.1711.06104.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140.

Konstantinos Barmpas, Yannis Panagakis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. BrainWave-Scattering Net: a lightweight network for EEG-based motor imagery recognition. *Journal of Neural Engineering*, 20(5):056014, 2023a. ISSN 1741-2552. doi: 10.1088/1741-2552/acf78a.

Konstantinos Barmpas, Yannis Panagakis, Stylianos Bakas, Dimitrios A. Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Improving generalization of CNN-based motor-imagery EEG decoders via dynamic convolutions. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 1997–2005, 2023b. doi: 10.1109/TNSRE.2023.3265304.

Konstantinos Barmpas, Yannis Panagakis, Georgios Zoumpourlis, Dimitrios A. Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. A causal perspective on brainwave modeling for brain–computer interfaces. *Journal of Neural Engineering*, 21(3):036001, 2024. doi: 10.1088/1741-2552/ad3eb5.

Konstantinos Barmpas, Na Lee, Dimitrios Chalatsis, William Raftery, Yannis Panagakis, Dimitrios A. Adamos, Nikolaos Laskaris, Alexandros Koliousis, Dario Farina, and Stefanos Zafeiriou. NeuroRVQ: Multi-scale biosignal tokenization for generative foundation models. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2510.13068.

Robert J. Barry, Adam R. Clarke, Stuart J. Johnstone, Christopher A. Magee, and Jacqueline A. Rushby. EEG differences between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology*, 118(12):2765–2773, 2007. ISSN 1388-2457. doi: 10.1016/j.clinph.2007.07.028.

Ali Bashashati, Mehrdad Fatourechi, Rabab K Ward, and Gary E Birch. A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural Engineering*, 4(2):R32–R57, 2007. doi: 10.1088/1741-2560/4/2/r03.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. doi: 10.48550/arXiv.2005.14165.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 387–396, 2021. doi: 10.1109/ICCV48922.2021.00045.

Joseph Y. Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *CoRR*, abs/2007.04871, 2020. doi: 10.48550/arXiv.2007.04871.

Paolo Detti. Siena Scalp EEG Database. *PhysioNet*, 2020. doi: 10.13026/5d4a-j060.

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2510.21585.

Jacob Gildenblat and contributors. PyTorch library for CAM methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

Abdulkadir Gokce, Yingtian Tang, and Martin Schrimpf. Multimodal scaling laws for task & data-optimized models of visual cortex. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=OQ6jQHJPTT`.

Ary L. Goldberger, Luis A. N. 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. doi: 10.1161/01.CIR.101.23.e215.

Todd C. Handy. *Brain Signal Analysis: Advances in Neuroelectric and Neuromagnetic Methods*. MIT Press, Cambridge, MA, 2009.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386.

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. doi: 10.48550/arXiv.2405.18765.

B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. 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. doi: 10.1109/10.867928.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. doi: 10.48550/arXiv.1412.6980.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, and Yogatheesan Varatharajah. EEG foundation models: a critical review of current progress and future directions. *Journal of Neural Engineering*, 23(2):021001, 2026. doi: 10.1088/1741-2552/ae4455.

V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. 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.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *GigaScience*, 8(5):giz002, 2019. doi: 10.1093/gigascience/giz002.

Na Lee, Stylianos Bakas, Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Assessing the capabilities of large brainwave foundation models. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2025a. doi: 10.1109/MLSP62443.2025.11204282.

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 *Forty-second International Conference on Machine Learning*, 2025b. doi: 10.48550/arXiv.2507.01196.

William Lehn-Schiøler, Magnus Ruud Kjær, Rahul Thapa, Magnus Guldberg Pedersen, Anton Mosquera Storgaard, Nick Williams, Radu Gatej, Tue Lehn-Schiøler, Andreas Brink-Kjær, Sadasivan Puthusserypady, Sándor Beniczky, James Zou, and Lars Kai Hansen. Mechanistic interpretability of EEG foundation models via sparse autoencoders. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2605.13930.

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. doi: 10.1093/nsr/nwag466.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. doi: 10.48550/arXiv.1711.05101.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Fabrice Yger. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. doi: 10.1088/1741-2552/aab2f2.

Hansong Ma and Junxiao Wang. EEG-Xplain: Decoding neural black-boxes of EEG foundation models. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2609.15687.

Dennis J. McFarland and Jonathan R. Wolpaw. Brain–computer interfaces for communication and control. *Communications of the ACM*, 54(5):60–66, 2011. doi: 10.1145/1941487.1941506.

D.J. McFarland, C.W. Anderson, K.-R. Muller, A. Schlogl, and D.J. Krusienski. BCI meeting 2005-workshop on BCI signal processing: feature extraction and translation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):135–138, 2006. doi: 10.1109/TNSRE.2006.875637.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 33–44, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.4.

Justus Meyer zu Bexten, Nico Scherf, Bogdan Franczyk, and Simon M. Hofmann. From Clever Hans to scientific discovery: Interpreting EEG foundational transformers with LRP. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2605.11885.

Chang S. Nam, Anton Nijholt, and Fabien Lotte. *Brain–Computer Interfaces Handbook: Technological and Theoretical Advances*. CRC Press, 2018.

Christa Neuper, Michael Wörtz, and Gert Pfurtscheller. ERD/ERS patterns reflecting sensorimotor activation and deactivation. In Christa Neuper and Wolfgang Klimesch (eds.), *Event-Related Dynamics of Brain Oscillations*, volume 159 of *Progress in Brain Research*, pp. 211–222. Elsevier, 2006. doi: 10.1016/S0079-6123(06)59014-4.

L. F. Nicolas-Alonso and J. Gomez-Gil. Brain–computer interfaces, a review. *Sensors*, 12(2):1211–1279, 2012. doi: 10.3390/s120201211.

Yuri G. Pavlov, Dauren Kasanov, Alexandra I. Kosachenko, Alexander I. Kotyusov, and Niko A. Busch. Pupillometry and electroencephalography in the digit span task. *Scientific Data*, 9(1):325, 2022. doi: 10.1038/s41597-022-01414-2.

G. Pfurtscheller and C. Neuper. Motor imagery and direct brain–computer communication. *Proceedings of the IEEE*, 89(7):1123–1134, 2001. doi: 10.1109/5.939829.

Rajesh P. N. Rao. *Brain-Computer Interfacing: An Introduction*. Cambridge University Press, 2013.

Lukas Rauch, René Heinrich, Houtan Ghaffari, Lukas Miklautz, Ilyass Moummad, Bernhard Sick, and Christoph Scholz. Unmute the patch tokens: Rethinking probing in multi-label audio classification. In *The Fourteenth International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2509.24901.

A. Sujatha Ravindran and Jose Contreras-Vidal. An empirical comparison of deep learning explainability approaches for EEG using simulated ground truth. *Scientific Reports*, 13:17709, 2023. doi: 10.1038/s41598-023-43871-8.

S. Saha and M. Baumert. Intra- and inter-subject variability in EEG-based sensorimotor brain–computer interface: A review. *Frontiers in Computational Neuroscience*, 13, 2020. doi: 10.3389/fncom.2019.00087.

Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Fernando Vaquerizo-Villar, and Roberto Hornero. EEG-Inception: A novel deep convolutional neural network for assistive ERP-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2773–2782, 2020. doi: 10.1109/TNSRE.2020.3048106.

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: a general-purpose brain–computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004. doi: 10.1109/TBME.2004.827072.

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. ISSN 1065-9471. doi: 10.1002/hbm.23730.

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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, pp. 3145–3153, 2017. doi: 10.48550/arXiv.1704.02685.

Urban Širca, Lovro Brulec, and Maryam Alimardani. Parameter-efficient fine-tuning of EEG foundation models for plug-and-play motor imagery BCIs. In *2026 14th International Conference on Brain-Computer Interface (BCI)*, pp. 1–7, 2026. doi: 10.1109/BCI69045.2026.11435102.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise, 2017. Presented at the ICML 2017 Workshop on Visualization for Deep Learning.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2302.13971.

Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4395–4405, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1448.

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. doi: 10.48550/arXiv.2412.07236.

Esther Werth, Peter Achermann, and Alexander A. Borbély. Brain topography of the human sleep EEG: antero-posterior shifts of spectral power. *NeuroReport*, 8(1):123–127, 1996. doi: 10.1097/00001756-199612200-00025.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2505.18185.

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 *Forty-third International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2508.17742.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2305.10351.

Xinliang Zhou, Chenyu Liu, Jinan Zhou, Zhongruo Wang, Liming Zhai, Ziyu Jia, Cuntai Guan, and Yang Liu. Interpretable and robust AI in EEG systems: A survey. *CoRR*, abs/2304.10755, 2023. doi: 10.48550/arXiv.2304.10755.

A Model Architectures

Our model set spans the more recent EEG-FM releases (NeuroRVQ, REVE and BrainOmni) and the established baselines from prior EEG-FM benchmarks (BIOT, LaBraM and CBraMod). The following EEG-FMs differ in many aspects, including positional encoding, pre-training objective and tokenisation, and form the model set used in our analyses. BIOT (Yang et al., 2023) was one of the first foundation models for bio-signals. It tokenises each channel into fixed-length segments and uses contrastive pre-training on clinical EEG and ECG recordings. LaBraM (Jiang et al., 2024), inspired by the vision transformer, segments EEG into channel patches and trains a vector-quantised (VQ) neural tokeniser to build a codebook of neural codes. A transformer-based model is then pre-trained to predict masked tokens from this codebook. CBraMod (Wang et al., 2025) separates spatial and temporal attention into two parallel streams using a criss-cross transformer and utilises asymmetric convolutional positional encoding to adapt to arbitrary electrode layouts. NeuroRVQ (Barmpas et al., 2026) extends the VQ approach of LaBraM to multiple frequency scales through four parallel temporal branches, each with its own residual VQ (RVQ) codebook. Combined with a phase-and-amplitude-aware training loss, the NeuroRVQ tokeniser achieves high-fidelity signal reconstruction across all frequency bands. REVE (El Ouahidi et al., 2025) represents the largest EEG pre-training effort to date, training on over 60,000 hours of data and 25,000 subjects. It introduces a 4D positional encoding scheme that generates embeddings from the 3D electrode coordinates and the temporal patch index, removing the need for fixed lookup tables. BrainOmni (Xiao et al., 2025) is the first model to pre-train on both EEG and MEG data. It encodes sensor properties such as 3D position, orientation, and type, and compresses channels into a fixed set of latent source variables on which the transformer operates, retaining its VQ tokeniser at inference time.

Table 1 summarises the architecture of each model, including spatial encoding technique, pooling strategy after the last backbone block and the number of trainable parameters in head-only training mode. Parameter counts are reported for the Eyes task (2 classes) (Goldberger et al., 2000) but head sizes vary slightly across tasks due to input-dependent dimensions: *(i)* BIOT applies an ELU activation followed by a single linear layer, *(ii)* LaBraM applies dropout followed by a single linear layer, *(iii)* NeuroRVQ inserts a LayerNorm before a single linear layer, *(iv)* REVE flattens its tokens then applies RMSNorm, dropout and a single linear layer, *(v)* BrainOmni uses a two-layer MLP (consisting of dropout, linear layer, SELU, linear layer) and *(vi)* CBraMod stacks three linear layers separated by ELU activations and dropout. For EEGNet, we treat the entire model as backbone.

Table 1: Model architectural details.

Model	Pooling	Backbone Parameters	Head Parameters	%
EEGNet	—	3.4K	—	—
BIOT	Mean	3.19M	257	<0.1
CBraMod	Flatten	4.88M	41.12M	89.4
LaBraM	Mean	5.82M	201	<0.1
NeuroRVQ	Mean	5.87M	2.4K	<0.1
BrainOmni*†	Mean + Flatten	37.76M	6.29M	14.3
REVE	Flatten	69.19M	393.2K	0.6

*Mean pools over temporal tokens per channel; flatten concatenates across latent source variables.
†BrainOmni keeps its 5.05M VQ tokeniser frozen during full fine-tuning.

B Training Protocol

All models are trained on clean data only (datasets described in Appendix C), using ten-fold leave-many-subjects-out cross-validation (LOMSO-CV). We train for a maximum of 20 epochs with early stopping with patience of 5 epochs. Each model uses the hyperparameters recommended by its original authors. All models are optimised with AdamW Loshchilov & Hutter (2019) except BIOT, which uses Adam Kingma & Ba (2015). LaBraM and CBraMod use cross-entropy with label smoothing of 0.1 for multiclass tasks and binary cross-entropy for binary tasks. BrainOmni uses cross-entropy with label smoothing of 0.1. Additional model-specific settings include: LaBraM uses drop-path 0.1, NeuroRVQ uses layer-wise learning rate (LR) decay 0.975, BIOT uses attention

dropout 0.2, REVE and BrainOmni cosine schedulers use $\eta_{\min} = \text{LR} \times 0.1$ and LaBraM and CBraMod use $\eta_{\min} = 10^{-6}$. Per-model learning rates and schedules are listed in Table 2.

All experiments run on a single NVIDIA H100 GPU. Batch size is set per model and per benchmark to the largest value that fits in memory. We report balanced accuracy as mean $\pm$ standard deviation across the ten LOMSO folds. The Average column in result tables averages the per-task means across tasks rather than across folds.

Table 2: Per-model learning rate, weight decay, and schedule.

Model	Learning Rate	Weight decay	Scheduler	Warm-up
EEGNet	1×10^{-3}	0	Cosine	—
LaBraM	5×10^{-4}	5×10^{-2}	Cosine	4 epochs
NeuroRVQ	5×10^{-4}	1×10^{-2}	Linear Warm-up + Decay	4 epochs
CBraMod	5×10^{-4}	5×10^{-2}	Cosine	—
BrainOmni	1×10^{-4}	1×10^{-5}	Linear Warm-up + Cosine	10% steps
REVE	1×10^{-4}	1×10^{-5}	Linear Warm-up + Cosine	10% steps
BIOT	1×10^{-3}	1×10^{-5}	None	—

B.1 Computational Cost

Table 3 reports training/inference GPU hours and peak memory use on a single NVIDIA H100. The numbers cover every training and evaluation run in this paper, across both adaptation regimes and all perturbation conditions, on all ten datasets. The total is 875.9 GPU-hours: 838.5 for training and 37.4 for inference. Compared to the EEGNet baseline, the foundation models are far more expensive to train and evaluate.

Table 3: Computational cost of the complete 10-dataset workload on a single NVIDIA H100. GPU-hours are summed; peak memory is a maximum. Rows are ordered by total cost.

Model	Train (GPU-h)	Inference (GPU-h)	Total (GPU-h)	Peak train (GB)	Peak inf. (GB)
EEGNet	4.87	0.50	5.37	1.8	1.3
BIOT	9.44	2.16	11.60	89.1	20.0
LaBraM	20.55	2.11	22.66	98.9	8.7
CBraMod	37.71	1.30	39.01	77.6	5.3
REVE	77.03	4.08	81.12	78.2	3.5
NeuroRVQ	100.77	5.61	106.38	99.1	4.1
BrainOmni	588.09	21.66	609.75	97.3	30.5

C Datasets and Preprocessing

Table 4 summarises the ten benchmark tasks. All datasets[1] are as distributed and were not included in the pre-training corpus of any foundation model we evaluate [2].

Trial extraction (length, baseline correction and rejection criteria) follows the original publications and (Lee et al., 2025a); we refer readers to those works for full details. We apply a Common Average Reference (CAR) to all recordings before any perturbation or model-specific filtering. Two preprocessing pipelines are used: *(i)* The default pipeline (EEGNet (Lawhern et al., 2018), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), BIOT (Yang et al., 2023), NeuroRVQ (Barmpas et al., 2026), REVE (El Ouahidi et al., 2025)) resamples to 200 Hz, applies a fourth-order Butterworth bandpass at 0.5–45 Hz, and notch filters at 50 and 60 Hz. *(ii)* BrainOmni (Xiao et al., 2025) uses its own pipeline: resampling to 256 Hz, bandpass at 0.1–96 Hz, and the same notch filters as in *(i)*.

[1] In PhysioNet ME and MI, six participants (subjects 88, 89, 92, 100, 104, and 106) were excluded due to differences in either the sampling frequency or duration of the performed tasks.

[2] With the exception of NeuroRVQ whose pre-training corpus includes PhysioNet Movement and Motor-Imagery tasks

Table 4: Benchmark datasets details. Trial length refers to the extracted window duration.

Dataset	Name	Paradigm	Classes	Channels	Trial length	Subjects
High-Gamma	Movement-HG	Motor execution	4	78	4 s	14
OpenBMI-MI	Motor-Imagery-KU	Motor imagery	2	62	4 s	54
OpenBMI-ERP	ERP	Visual P300	2	62	1 s	54
PhysioNet-MI	Motor-Imagery-PN	Motor imagery	4	64	4 s	103
PhysioNet-ME	Movement-PN	Motor execution	4	64	4 s	103
PhysioNet Eyes	Eyes	Eyes open/closed	2	64	4 s	103
Pavlov Memory	Memory	Working memory	2	63	4 s	65
Sleep EDF	Sleep-EDF	Sleep staging	6	2	30 s	78
Siena	Seizure	Seizure detection	2	19	4 s	14
HMC	Sleep-HMC	Sleep staging	5	4	30 s	151

†Sleep-EDF channels are bipolar derivations (Fpz-Cz and Pz-Oz).
‡Sleep-HMC channels are mastoid-referenced derivations (F4-M1, C4-M1, O2-M1 and C3-M2).

D Robustness Perturbations

We apply all perturbations at test time only. Each model is trained once on clean data and the same fold checkpoint is evaluated under every perturbation condition. This mimics the deployment case where data quality drops at inference but retraining is not possible. We test four families of perturbation: additive white and pink noise, random channel dropout, region-based channel dropout, and region-specific noise injection. Noise samples and dropout masks are shared across all models, so every model sees the same perturbed inputs at every level.

D.1 Additive Noise

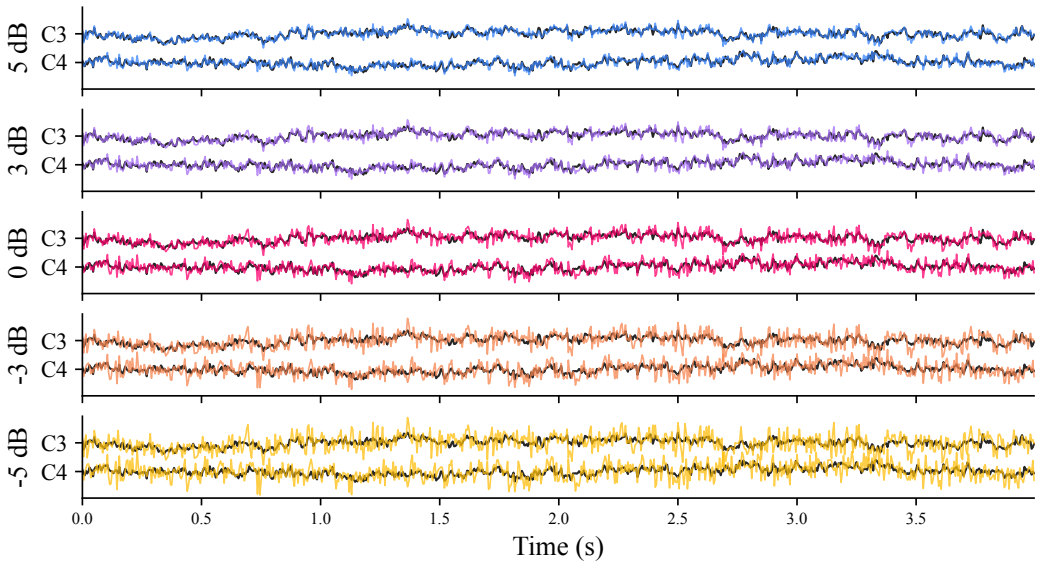


Figure 7: White noise examples of two channels at five example SNR levels applied at test-time to a single EEG trial.

White noise is sampled i.i.d. from $\mathcal{N}(0, 1)$ independently per channel and time point (illustrated in Figure 7). **Pink noise** $(1/f)$ is generated by applying the fast Fourier transform (FFT) to white noise along the time axis, scaling the magnitude spectrum by $1/\sqrt{f}$, and applying the inverse FFT. Both noise types are filtered through the same bandpass and notch pipeline used for the clean data of each model, ensuring that the perturbation falls within the recording bandwidth. The baseline recordings already contain physiological and instrumental noise, so the reported SNR values are relative to the

provided data rather than absolute. However, since all models are evaluated on the same data, the resulting comparisons reflect relative differences in robustness.

Pink noise is intended to approximate the spectral profile of biological EEG noise, but our implementation generates noise independently across channels. Real physiological noise exhibits spatial correlations due to volume conduction through the skull and scalp. This simplification means our pink noise perturbation might underestimate the disruptive effect of spatially correlated biological noise. For each trial, we compute the signal power as the mean squared amplitude after centring:

$$P_s = \frac{1}{C \cdot T} \sum_{c,t} (X_{c,t} - \bar{X}_c)^2 \tag{1}$$

where C is the number of channels and T the number of time points. The target noise power is $P_n = P_s / 10^{(\mathrm{SNR_{dB}}/10)}$. The raw noise is normalised to zero mean and unit variance per channel, then scaled by $\sqrt{P_n}$:

$$X_{\text{noisy}} = X_{\text{clean}} + \sqrt{P_n} \cdot \hat{n} \tag{2}$$

SNR levels tested: 10, 5, 0, −3, −5, −15 (white noise only) dB. At 10 dB the signal is clearly dominant; at 0 dB signal and noise have equal power; at −5 dB the noise is approximately three times stronger than the signal.

D.2 Random Channel Dropout

Table 5: Channels zero-padded at each per-channel dropout probability, as recorded by the corruption runs. Each benchmark draws its own mask; benchmarks sharing one are listed together. One mask is drawn per dataset and probability and shared across trials and folds. Sleep-EDF, Sleep-HMC are omitted.

Dataset	p	n	Dropped channels
Movement-HG (78 ch)	0.10	9	FPZ, F7, FC6, T7, CP1, CPZ, P6, FT8, PO7
	0.30	22	FP2, FPZ, F7, FC6, T7, CZ, M2, CP1, CP2, AF7, C6, CPZ, P1, P6, PO3, FT7, FT8, TP8, PO7, FT9, TPP9H, PO10
	0.50	38	FP2, FPZ, F7, FC6, T7, CZ, M2, CP5, CP1, CP2, PZ, O2, AF7, AF8, F5, F6, FCZ, C5, C1, C6, CPZ, P1, P6, PO5, PO3, PO4, FT7, FT8, TP7, TP8, PO7, FT9, TPP9H, PO9, PO10, AFZ, FTT9H, FTT10H
Motor-Imagery-KU (62 ch)	0.10	8	AF7, AF3, F8, FT9, FT10, PO9, O1, FTT10H
	0.30	18	FP2, AF7, AF3, F8, FT9, FC3, FC4, FT10, T7, CP5, P4, PO9, PO4, O1, O2, TPP9H, FTT10H, TPP10H
	0.50	30	FP2, AF7, AF3, F8, FT9, FC3, FC4, FC6, FT10, T7, CZ, TP7, CP5, CPZ, CP2, TP8, P7, P1, PZ, P4, PO9, PO4, O1, OZ, O2, FTT9H, TPP9H, FTT10H, TPP8H, TPP10H
ERP (62 ch)	0.10	8	F7, F3, T7, CZ, CP6, TPP9H, TPP10H, AF8
	0.30	18	FP2, F7, F3, T7, CZ, T8, CP1, CP6, TP10, FC3, TTP7H, TPP9H, TPP8H, TPP10H, F10, AF4, AF8, PO4
	0.50	30	FP2, F7, F3, T7, CZ, T8, CP1, CP2, CP6, TP10, P4, PO10, FC3, C1, C2, CPZ, P1, POZ, FT9, TTP7H, TPP9H, TPP8H, TPP10H, F9, F10, AF7, AF4, AF8, PO3, PO4
Movement-PN Motor-Imagery-PN Eyes (64 ch)	0.10	9	FC1, FCZ, C2, C6, CP6, P3, P6, PO8, O2
	0.30	19	FC3, FC1, FCZ, C2, C6, CP3, CP2, CP6, FP1, F1, P7, P3, P2, P6, PO7, PO4, PO8, OZ, O2
	0.50	31	FC3, FC1, FCZ, C2, C6, CP3, CP2, CP4, CP6, FP1, AF3, F3, F1, F4, F6, FT8, T8, T10, TP7, P7, P3, P2, P6, P8, PO7, PO3, PO4, PO8, O1, OZ, O2
Memory (63 ch)	0.10	9	F3, F7, CP1, P3, TP10, PO4, CP4, F6, F2
	0.30	19	FZ, F3, F7, CP1, P3, O1, P4, TP10, CP6, AF7, PO3, PO4, P2, CP4, C6, FT8, F6, AF4, F2
	0.50	31	FZ, F3, F7, CP1, P3, O1, P4, P8, TP10, CP6, T8, FP2, AF7, F1, F5, C1, TP7, P1, P5, PO3, PO4, P2, CP4, TP8, C6, C2, FT8, F6, AF8, AF4, F2
Seizure (19 ch)	0.10	4	C3, P3, FP2, C4
	0.30	7	F3, C3, P3, FP2, C4, O2, T6
	0.50	7	F3, C3, P3, FP2, C4, O2, T6

Each channel is independently selected for zero-padding with probability $p \in \{0.10, 0.30, 0.50\}$ (Bernoulli dropout). Here, p is a per-channel dropout probability, not a fixed fraction of channels. The resulting dataset-level mask is applied to all trials and folds. Zero-padding preserves the original channel dimension. We omit Sleep-EDF and Sleep-HMC from all random dropout figures and tables because of their limited channel counts.

In addition, we also test *true channel dropout*, where dropped channels are removed from the input rather than zero-padded, for the variable-channel models: REVE, NeuroRVQ, LaBraM, and BrainOmni. We define a model as variable-channel if its forward pass conditions on channel identity

(via learned channel embeddings or 3D electrode coordinates) and accepts a subset of the training channels at inference without retraining or padding. These results are reported in Appendix H.2.2.

Table 5 lists the exact channels dropped at each dropout probability for the eight included datasets. The three PhysioNet benchmarks (Eyes, Motor-Imagery-PN, Movement-PN) share the same 64-channel montage and therefore the same dropout mask.

D.3 Region-based Channel Dropout

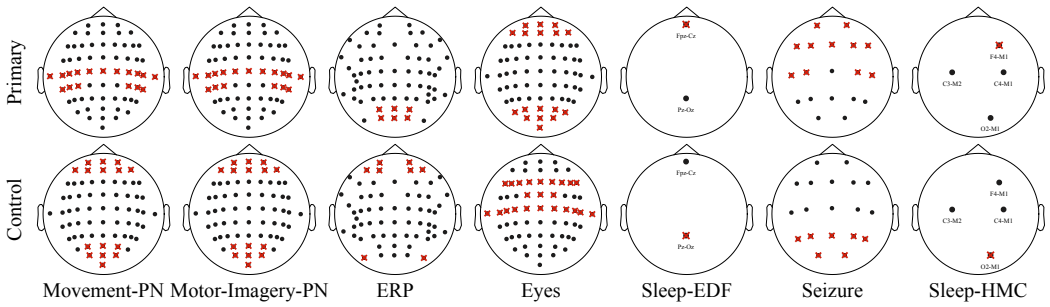


Figure 8: Primary and control channels for region-based dropout, shown in red for seven tasks. Movement-HG and Movement-PN share region assignments, as do Motor-Imagery-KU and Motor-Imagery-PN.

We group channels into anatomical regions defined by the international 10–20 system (extended to 10–5 positions where applicable). For each task we assign two region sets: a primary set covering brain regions known to generate the task-relevant signal, and a control set used for comparison outside the primary region (Figure 8). Dropping a region zeroes out all channels in that set for all models, preserving the original channel dimension. For the models that accept variable-length channel sets (REVE, NeuroRVQ, LaBraM, and BrainOmni), we also test *true removal of the dropped region* and report these results in Appendix H.3. Table 6 lists the region assignments per task.

Each primary assignment maps the scalp regions that carry the dominant neural source for the task. For motor execution and motor imagery, mu and beta event-related de-synchronisation localises over the contralateral sensorimotor cortex, so central and centroparietal sites form the primary set in both tasks (Pfurtscheller & Neuper, 2001; Neuper et al., 2006). For ERP, Lee et al. (2019) report posterior-occipital sites as carrying the target P300 response, so we mark the posterior middle group as primary, capturing the occipital component of the P300, and treat anterior and lateral posterior sites as control. Posterior alpha rhythms dominate the eyes-closed condition, and the transition to eyes-open produces band-specific topographic changes over both anterior sites (delta and beta) and posterior sites (theta and beta) (Barry et al., 2007). The eyes-open condition also adds oculomotor activity (blinks and saccades) that loads onto the anterior electrodes, so we mark both region groups as primary. For both sleep datasets, the frontal predominance of slow waves motivates using the frontal derivation as primary and the posterior derivation as control (Werth et al., 1996). However, Sleep-EDF provides only two bipolar derivations (Kemp et al., 2000), and Sleep-HMC only four, so a single region is defined by a single derivation. We therefore exclude both sleep datasets from the region-based averages, but report their per-dataset results for completeness. For Seizure, Siena's temporal and frontal seizure localisations motivate our primary region (Detti, 2020), although this shared assignment is coarse, and control channels may also carry seizure information.

We define 18 region groups using the international 10–20 system extended to 10–5 positions:

1. **Anterior**: Fp1, Fpz, Fp2, AF3, AF4, AF7, AF8, AFz.
2. **Frontal** (Left/Middle/Right): F7, F5, F3, F9 / F1, Fz, F2, FC1, FCz, FC2 / F4, F6, F8, F10.
3. **Frontotemporal** (Left/Right): FT7, FT9, FC5, FC3, FTT9h / FC4, FC6, FT8, FT10, FTT10h.
4. **Central** (Left/Middle/Right): T7, T9, C5, C3 / C1, Cz, C2 / C4, C6, T8, T10.

5. **Centroparietal** (Left/Middle/Right): TP7, TP9, CP5, CP3, TTP7h, TPP9h / CP1, CPz, CP2 / CP4, CP6, TP8, TP10, TPP8h, TPP10h.
6. **Parietal** (Left/Middle/Right): P7, P5, P3, P9 / P1, Pz, P2 / P4, P6, P8, P10.
7. **Posterior** (Left/Middle/Right): PO7, PO5, PO9, CB1 / PO3, POz, PO4, O1, Oz, O2, Iz / PO6, PO8, PO10, CB2.

Table 6: Region assignments for region-based perturbations. Primary regions are expected to contain task-relevant features; control regions provide a comparison.

Task	Primary	Control
Movement	Central L/M/R, Centroparietal L/R	Posterior Middle, Anterior
Motor imagery	Central L/M/R, Centroparietal L/R	Posterior Middle, Anterior
Visual P300	Posterior Middle	Anterior, Posterior L/R
Working memory	Frontal L/M/R	Posterior Middle, Central L/R
Eyes open/closed	Anterior, Posterior L/M/R	Frontal L/M/R, Central L/M/R
Sleep staging	EDF†: Fpz-Cz (frontal) HMC: F4 (frontal)	EDF†: Pz-Oz (posterior) HMC: O2 (posterior)
Seizure detection	Anterior, Frontal L/M/R, Central L/R	Parietal L/M/R, Posterior Middle

†Sleep-EDF channels are bipolar derivations.

D.4 Region Signal Substitution

This condition removes the signal in a region without setting it to zero. For each trial we replace every channel in the primary region with white noise. We filter the noise with the same settings as the dataset, normalise it to zero mean and unit variance, then scale it so its root mean square matches the median of the intact channels. The substituted channels keep a plausible amplitude and carry none of the original signal. We use this condition for attribution only, not for accuracy, because Zero padding is a poor control for Input $\times$ Gradient and AttnLRP.

D.5 Region Noise Injection

Instead of removing channels, white noise at a given SNR is injected into channels belonging to a single region while all other channels remain non-perturbed. This is a softer perturbation than region dropout, so models can still partially extract features from noisy channels. It simulates degraded electrodes, the most common clinical scenario. SNR levels tested: $5\,\text{dB}$ and $-3\,\text{dB}$, for each of primary and control regions. The same region assignments as Table 6 are used.

E Statistical Significance Testing

We compare models on 10-fold leave-many-subjects-out cross-validation (LOMSO-CV) with identical subject splits across models. Although the training sets overlap across folds, the comparison within each fold is paired because all models are evaluated on exactly the same held-out subjects. A two-sided Wilcoxon signed-rank test was therefore applied to compare models to our baseline. When we average over datasets, we pair over every dataset and fold pair. We adjust the p-values for multiple comparisons with the Holm-Bonferroni method, at significance $p = 0.05$.

F CLEAN PERFORMANCE TABLES

Tables 7 and 8 report test balanced accuracy (mean ± standard deviation across ten folds) for all models under full fine-tuning and head-only adaptation respectively. Models are ranked by average balanced accuracy. EEGNet is included in both tables as the supervised baseline.

Table 7: Clean balanced accuracy (full fine-tuned), mean ± std over 10 folds, ranked by Average. Best per column in **bold**, second best underlined. Average equally weights the 10 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Sleep-EDF, Seizure, Sleep-HMC). * on the Average marks a model significantly better than EEGNet, $^{\nabla}$ significantly worse (two-sided Wilcoxon signed-rank over the shared folds, Holm-corrected).

#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Sleep-EDF	Seizure	Sleep-HMC	Average
1	NeuroRVQ	0.689±0.089	0.559±0.041	0.809±0.035	0.538±0.053	**0.838±0.038**	**0.861±0.025**	0.579±0.020	**0.684±0.033**	0.783±0.085	**0.735±0.024**	**0.707***
2	REVE	**0.694±0.105**	0.510±0.028	**0.829±0.038**	0.496±0.025	0.826±0.043	0.799±0.034	0.593±0.032	0.629±0.035	**0.794±0.111**	0.709±0.017	0.688*
3	CBraMod	0.594±0.079	0.547±0.029	0.752±0.040	0.501±0.032	0.783±0.045	0.821±0.040	**0.603±0.021**	0.634±0.042	0.760±0.121	0.725±0.018	0.672*
4	BrainOmni	0.585±0.058	**0.625±0.036**	0.750±0.038	**0.615±0.053**	0.723±0.041	0.852±0.022	0.518±0.023	0.576±0.055	0.667±0.121	0.727±0.019	0.664
5	LaBraM	0.620±0.079	0.517±0.044	0.775±0.041	0.526±0.059	0.832±0.036	0.833±0.029	0.534±0.041	0.606±0.046	0.568±0.123	0.702±0.018	0.651
6	EEGNet	0.624±0.100	0.559±0.038	0.786±0.034	0.528±0.086	0.796±0.040	0.758±0.024	0.512±0.005	0.590±0.029	0.708±0.155	0.605±0.040	0.647
7	BIOT	0.322±0.050	0.258±0.008	0.501±0.005	0.253±0.018	0.500±0.000	0.753±0.065	0.519±0.016	0.167±0.000	0.623±0.089	0.695±0.028	0.459$^{\nabla}$

Table 8: Clean balanced accuracy (head-only), mean ± std over 10 folds, ranked by Average. Best per column in **bold**, second best underlined. EEGNet has no frozen-backbone variant - it trains from scratch, so its row repeats its supervised score as the reference the foundation models are read against. Average equally weights the 10 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Sleep-EDF, Seizure, Sleep-HMC). * on the Average marks a model significantly better than EEGNet, $^{\nabla}$ significantly worse (two-sided Wilcoxon signed-rank over the shared folds, Holm-corrected).

#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Sleep-EDF	Seizure	Sleep-HMC	Average
1	EEGNet	**0.624±0.100**	**0.559±0.038**	**0.786±0.034**	**0.528±0.086**	**0.796±0.040**	0.758±0.024	0.512±0.005	**0.590±0.029**	0.708±0.155	0.605±0.040	**0.647**
2	REVE	0.544±0.095	0.458±0.027	0.761±0.037	0.449±0.018	0.749±0.035	0.790±0.043	**0.574±0.039**	0.589±0.036	**0.756±0.116**	0.656±0.024	0.633
3	BrainOmni	0.481±0.050	0.527±0.032	0.658±0.029	0.502±0.030	0.525±0.010	**0.829±0.036**	0.541±0.026	0.473±0.036	0.606±0.126	**0.684±0.029**	0.583$^{\nabla}$
4	CBraMod	0.381±0.078	0.322±0.037	0.653±0.030	0.284±0.040	0.642±0.019	0.809±0.053	0.545±0.022	0.563±0.055	0.694±0.150	0.677±0.023	0.557$^{\nabla}$
5	NeuroRVQ	0.271±0.022	0.252±0.003	0.501±0.003	0.252±0.003	0.542±0.013	0.736±0.038	0.517±0.021	0.474±0.026	0.509±0.060	0.455±0.027	0.451$^{\nabla}$
6	BIOT	0.309±0.038	0.250±0.012	0.501±0.010	0.256±0.016	0.500±0.000	0.679±0.135	0.536±0.014	0.167±0.000	0.554±0.052	0.624±0.021	0.438$^{\nabla}$
7	LaBraM	0.300±0.047	0.270±0.017	0.504±0.008	0.266±0.013	0.500±0.000	0.717±0.049	0.500±0.000	0.295±0.008	0.497±0.016	0.473±0.016	0.432$^{\nabla}$

G PRE-TRAINING VERSUS ARCHITECTURE

Pre-training improves average performance for most models compared with training the same architecture from scratch under the same schedule (Table 9). The gain is small for CBraMod, while BIOT shows no benefit from pre-training on these tasks.

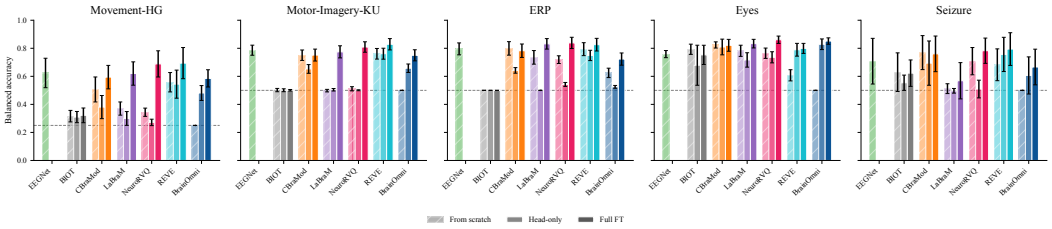


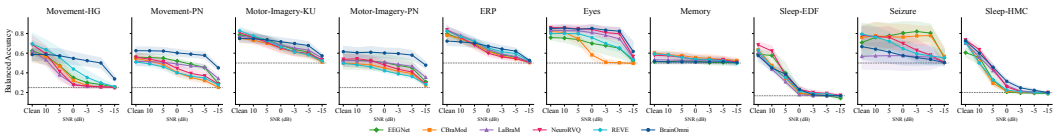
Figure 9: From-scratch training (hatched bars), pre-trained frozen backbone (light bars) and full fine-tuning (saturated bars) on Movement-HG, Motor-Imagery-KU, ERP, Eyes and Seizure. Bars show mean balanced accuracy over 10 folds; error bars show ± 1 standard deviation. The dashed horizontal line marks chance-level balanced accuracy.

Table 9: Pre-training versus architecture. Balanced accuracy for the same architecture trained from random initialisation, adapted with a frozen pre-trained backbone, and fully fine-tuned from the pre-trained checkpoint. Restricted to the 5 datasets with from-scratch runs (Movement-HG, Motor-Imagery-KU, ERP, Eyes, Seizure); Average weighs them equally. EEGNet is from-scratch. Best per column in **bold**, second best underlined.

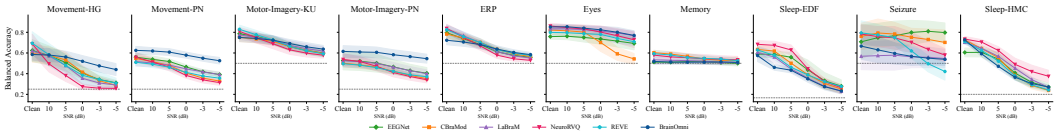
Model	Regime	Movement-HG	Motor-Imagery-KU	ERP	Eyes	Seizure	Average
NeuroRVQ	From scratch	0.344	0.511	0.719	0.765	0.708	0.609
	Head-only	0.271	0.501	0.542	0.736	0.509	0.512
	Full FT	0.689	0.809	**0.838**	**0.861**	0.783	**0.796**
REVE	From scratch	0.558	0.761	0.794	0.607	0.682	0.680
	Head-only	0.544	0.761	0.749	0.790	0.756	0.720
	Full FT	**0.694**	**0.829**	0.826	0.799	**0.794**	0.788
CBraMod	From scratch	0.506	0.751	0.799	0.824	0.769	0.730
	Head-only	0.381	0.653	0.642	0.809	0.694	0.636
	Full FT	0.594	0.752	0.783	0.821	0.760	0.742
LaBraM	From scratch	0.370	0.498	0.735	0.782	0.512	0.580
	Head-only	0.300	0.504	0.500	0.717	0.497	0.504
	Full FT	0.620	0.775	0.832	0.833	0.568	0.726
BrainOmni	From scratch	0.250	0.500	0.625	0.500	0.500	0.475
	Head-only	0.481	0.658	0.525	0.829	0.606	0.620
	Full FT	0.585	0.750	0.723	0.852	0.667	0.715
BIOT	From scratch	0.316	0.502	0.500	0.793	0.630	0.548
	Head-only	0.309	0.501	0.500	0.679	0.554	0.509
	Full FT	0.322	0.501	0.500	0.753	0.623	0.540
EEGNet	From scratch	0.624	0.786	0.796	0.758	0.708	0.735

H ROBUSTNESS RESULTS

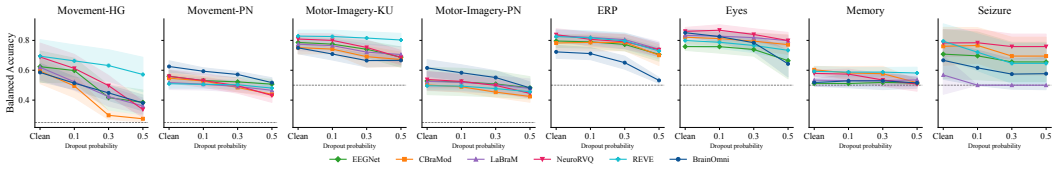
Figures 10a–10e show robustness curves under full fine-tuning. Figures 11a–11e show robustness curves under head-only adaptation.



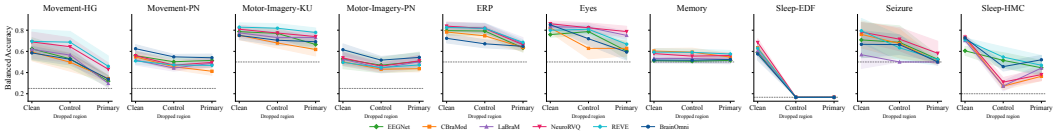
(a) Additive white noise.



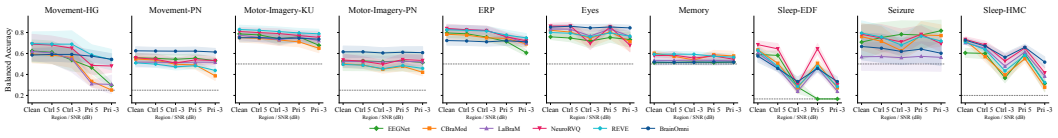
(b) Additive pink noise.



(c) Random channel dropout.

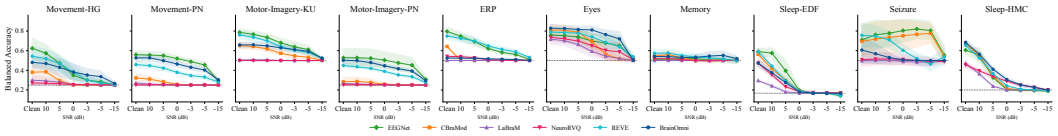


(d) Region dropout.

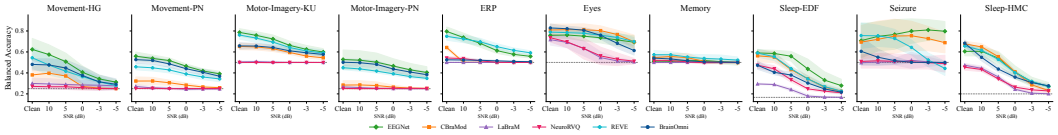


(e) Region noise injection.

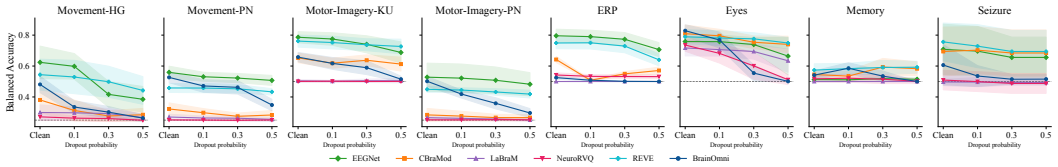
Figure 10: Robustness evaluation under five perturbation types (full fine-tuned models): (a) Additive White Noise, (b) Additive Pink Noise, (c) Random Channel Dropout, (d) Region Dropout and (e) Region Noise Injection. Each line shows mean balanced accuracy over ten folds as a function of perturbation level; shaded bands show ± 1 standard deviation across folds and dashed lines mark chance level. Noise and region panels cover ten datasets; random dropout panels cover eight.



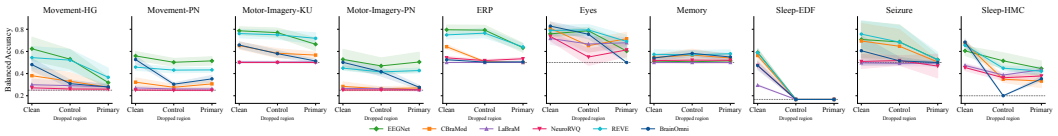
(a) Additive white noise degradation under head-only adaptation.



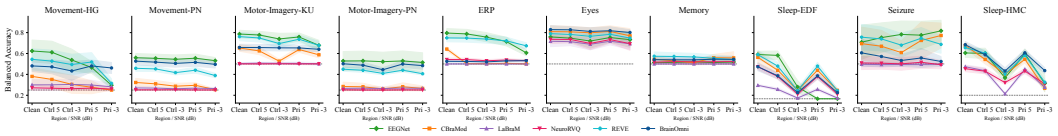
(b) Additive pink noise degradation under head-only adaptation.



(c) Random channel dropout under head-only adaptation.



(d) Region dropout under head-only adaptation.



(e) Region noise injection under head-only adaptation.

Figure 11: Robustness evaluation under five perturbation types (head-only adaptation): (a) Additive White Noise, (b) Additive Pink Noise, (c) Random Channel Dropout, (d) Region Dropout and (e) Region Noise Injection. Each line shows mean balanced accuracy over ten folds as a function of perturbation level; shaded bands show ± 1 standard deviation across folds and dashed lines mark chance level. Noise and region panels cover ten datasets; random dropout panels cover eight.

H.1 Additive Noise

Tables 10 and 11 report per-benchmark changes in percentage points from clean at -5 dB white and pink noise respectively. Models are ranked by average degradation across eight datasets, excluding Sleep-EDF and Sleep-HMC.

White-noise degradation varies across tasks and models. At -5 dB, several models approach chance on Movement-HG and ERP, while most models remain above chance on Eyes and all do on Motor-Imagery-KU. Both sleep datasets approach chance under white noise at this level, but remain above chance under pink noise. BrainOmni appears to be the most noise robust model, remaining above chance even at -15 dB for certain tasks. It also remains the most noise robust model with both sleep datasets included (-14.2 vs. -17.9 to -25.9 percentage points for all other models at -5 dB additive white noise). Several architectural choices may contribute. BrainOmni compresses input channels to 16 latent source variables before the transformer; the original paper's ablation supports this channel-compression hypothesis, with Table 8 in (Xiao et al., 2025) showing that replacing the cross-attention compression with electrode-level self-attention degrades clean balanced accuracy on every dataset they test. They attribute the gain to elimination of redundant information between adjacent electrodes. The ablation considers clean inputs without perturbation, but the same bottleneck

may underlie the noise robustness shown in our results. Pre-training on both EEG and MEG may also contribute. The quantiser ablation (Appendix H.4) shows that BrainOmni retains its noise-robustness advantage without quantisation. Latent compression and multimodal pre-training thus remain possible explanations, and separating them requires matched pre-training variants.

Table 10: Per-benchmark degradation under White -5 dB (full fine-tuning). Δ percentage points from clean, paired per fold. Ranked by Average. Best per column in **bold**, second best underlined. Average equally weights the 8 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Seizure); Sleep-EDF, Sleep-HMC excluded from the average. * on the Average marks a model significantly better than EEGNet, $^\nabla$ significantly worse (two-sided Wilcoxon signed-rank over the shared folds, Holm-corrected).

#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Sleep-EDF	Seizure	Sleep-HMC	Average
1	BrainOmni	**−8.5**	**−4.8**	**−7.2**	**−3.5**	**−10.2**	**−2.7**	**−0.5**	**−40.7**	−13.1	−50.5	**−6.3***
2	EEGNet	−34.2	−10.5	−17.6	−7.6	−23.5	−10.9	−0.7	−42.4	**+9.6**	**−41.1**	−11.9
3	LaBraM	−36.3	−6.1	−14.3	−5.9	−26.5	−8.6	−0.6	−43.9	−0.9	−49.9	−12.4
4	CBraMod	−33.9	−22.6	−16.6	−10.9	−23.7	−31.9	−6.7	−46.2	+1.9	−52.5	−18.1$^\nabla$
5	NeuroRVQ	−43.0	−19.0	−21.3	−13.0	−29.5	−6.3	−4.5	−49.3	−20.2	−53.4	−19.6$^\nabla$
6	REVE	−39.6	−16.5	−24.4	−13.4	−23.0	−14.5	−6.9	−44.5	−22.2	−50.5	−20.1$^\nabla$

Table 11: Per-benchmark degradation under Pink -5 dB (full fine-tuning). Δ percentage points from clean, paired per fold. Ranked by Average. Best per column in **bold**, second best underlined. Average equally weights the 8 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Seizure); Sleep-EDF, Sleep-HMC excluded from the average. * on the Average marks a model significantly better than EEGNet, $^\nabla$ significantly worse (two-sided Wilcoxon signed-rank over the shared folds, Holm-corrected).

#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Sleep-EDF	Seizure	Sleep-HMC	Average
1	BrainOmni	**−14.6**	**−10.0**	**−11.2**	**−7.0**	**−13.9**	−8.2	**−1.0**	−34.3	−12.8	−45.6	**−9.8***
2	EEGNet	−31.2	−16.8	−18.7	−12.4	−23.8	**−6.6**	−1.2	**−31.1**	**+8.9**	**−33.4**	−12.7
3	LaBraM	−32.9	−12.6	−15.6	−12.6	−28.1	−6.7	−2.5	−37.7	−2.2	−43.6	−14.2
4	CBraMod	−30.8	−21.9	−15.6	−15.0	−21.7	−27.9	−7.5	−38.4	−5.8	−48.7	−18.3$^\nabla$
5	REVE	−38.6	−15.4	−23.7	−12.6	−24.8	−9.0	−6.8	−35.3	−37.3	−46.7	−21.0$^\nabla$
6	NeuroRVQ	−43.1	−24.6	−23.0	−19.8	−31.0	−13.3	−4.5	−42.0	−20.2	−35.9	−22.4$^\nabla$

H.2 Random Channel Dropout

H.2.1 Zero Padded Channels

Table 12 reports per-benchmark degradation at p=0.50 random channel dropout. We omit Sleep-EDF and Sleep-HMC from all random channel dropout figures and tables. Channel dropout is model-limited (with some task-specific variation), suggesting that architectural choices, particularly how models encode or handle missing information, play a role. REVE has the smallest average loss under zero-padded random dropout. A possible contributing factor is REVE's pre-training, which drops 10% of channels on top of token masking and could plausibly give the model exposure to missing-channel scenarios.

Table 12: Per-benchmark degradation under Rand. drop p=0.50 (full fine-tuning). Δ percentage points from clean, paired per fold. Ranked by Average. Best per column in **bold**, second best underlined. Average equally weights the 8 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Seizure); Sleep-EDF, Sleep-HMC excluded from the average. * on the Average marks a model significantly better than EEGNet, $^\nabla$ significantly worse (two-sided Wilcoxon signed-rank over the shared folds, Holm-corrected).

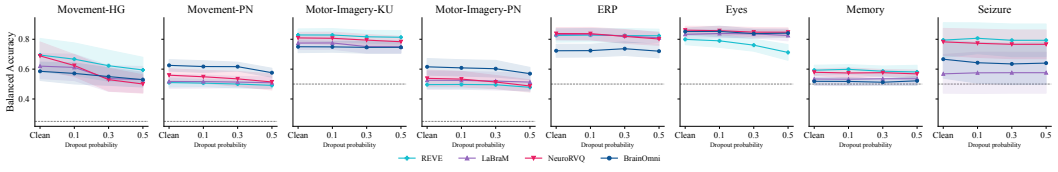
#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Seizure	Average
1	REVE	**−12.3**	**−3.0**	**−2.6**	**−4.0**	−9.7	−6.6	−1.1	−14.8	**−6.8**
2	LaBraM	−25.9	−5.1	−6.9	−4.7	−9.1	**−3.5**	**+0.2**	−6.8	−7.7
3	EEGNet	−23.9	−5.2	−9.8	−4.6	−9.0	−9.4	**+0.2**	−5.2	−8.4
4	CBraMod	−31.9	−10.8	−7.9	−7.7	**−8.2**	−5.0	−8.6	−6.5	−10.8$^\nabla$
5	NeuroRVQ	−35.1	−12.8	−12.4	−9.3	−10.1	−6.2	−7.1	**−2.4**	−11.9$^\nabla$
6	BrainOmni	−20.4	−10.8	−8.4	−13.3	−19.1	−20.8	−0.3	−8.9	−12.8$^\nabla$

H.2.2 True Random Channel Dropout

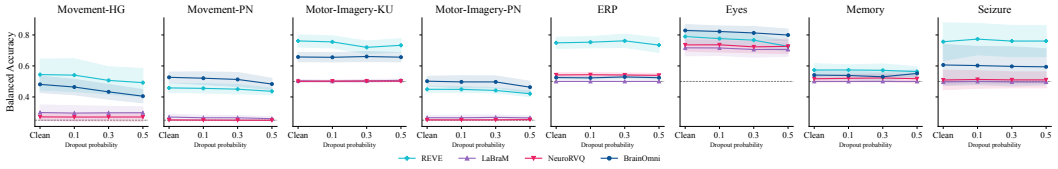
In the main analysis, dropped channels are zeroed-out for all models to ensure a fair comparison. Here, we report results for true channel dropout, where dropped channels are removed from the input rather than zero-padded. This setting is restricted to models that accept variable-length channel sets: REVE, NeuroRVQ, LaBraM, and BrainOmni. At p=0.50, average losses range from 2.0 to 4.8 percentage points across the four models. Figures 12a and 12b show per-benchmark curves under full fine-tuning and head-only adaptation respectively, and Table 13 reports per-benchmark degradation at p=0.50.

Table 13: Per-benchmark degradation under Rand. removal p=0.50 (full fine-tuning). Δ percentage points from clean, paired per fold. Ranked by Average. Best per column in **bold**, second best underlined. Average equally weights the 8 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Seizure); Sleep-EDF, Sleep-HMC excluded from the average.

#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Seizure	Average
1	LaBraM	−9.0	**−0.9**	−2.5	**−1.2**	−2.2	**−0.8**	+0.2	**+0.8**	**−2.0**
2	BrainOmni	**−5.6**	−4.9	**−0.4**	−4.6	−0.3	−1.1	**+0.3**	−2.6	−2.4
3	REVE	−10.0	−1.9	−1.5	−1.8	**−0.2**	−8.7	−1.2	−0.1	−3.2
4	NeuroRVQ	−18.8	−4.5	−2.6	−5.1	−3.7	−1.4	−1.1	−1.6	−4.8



(a) Full fine-tuning.



(b) Head-only adaptation.

Figure 12: Per-benchmark degradation under true random channel dropout for the models that accept variable-length channel sets: REVE, NeuroRVQ, LaBraM, and BrainOmni.

H.3 Region-based Perturbations

Tables 14 and 15 report per-benchmark degradation under primary and control region dropout respectively. Table 16 reports degradation under primary region noise.

Across the eight datasets in the average, LaBraM, NeuroRVQ and REVE have the smallest losses under primary-region dropout (9.2, 11.1 and 11.4 percentage points). BrainOmni and EEGNet lose 12.0 and 12.4 points, and CBraMod has the largest loss (15.4 points). On Movement-HG, all models lose 23.5–32.3 percentage points, consistent with reliance on central electrodes for motor execution. Motor-Imagery-KU follows the same pattern as Movement-HG but with smaller drops, as expected for imagery versus overt movement. ERP losses range from 7.5 to 17.6 points, consistent with posterior electrodes carrying the P300 signal. On Eyes, BrainOmni (25.8 points) and CBraMod (19.3 points) lose more than the other models (7.4–15.5 points). Sleep-EDF reaches chance when either channel is zero-padded, with losses of 40.9–51.7 percentage points. Sleep-HMC remains above chance under primary-region dropout, with balanced accuracy of 0.364–0.521. All six models lose more under control than primary dropout on Motor-Imagery-PN and three of six do on Movement-PN. This reversal is unexpected: neuroscience predicts that motor features should concentrate in central electrodes. Given that Lee et al. (2025a) identified eye-movement artefact contamination in these datasets (PhysioNet), we exclude Movement-PN and Motor-Imagery-PN from interpretability

analyses. Under primary region noise, including sleep datasets, BrainOmni remains the most robust model at -3 dB (-6.1 vs. -12.5 to -17.1 percentage points for other models).

Table 14: Per-benchmark degradation under Region drop (pri.) (full fine-tuning). Δ percentage points from clean, paired per fold. Ranked by Average. Best per column in **bold**, second best underlined. Average equally weights the 8 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Seizure); Sleep-EDF, Sleep-HMC excluded from the average. * on the Average marks a model significantly better than EEGNet, $^{\nabla}$ significantly worse (two-sided Wilcoxon signed-rank over the shared folds, Holm-corrected).

#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Sleep-EDF	Seizure	Sleep-HMC	Average
1	LaBraM	−32.3	**−3.0**	**−4.4**	−2.4	−15.7	−8.2	−0.9	−43.9	**−6.8**	−25.9	**−9.2***
2	NeuroRVQ	−25.7	−6.3	−6.9	−2.7	−17.6	**−7.4**	−1.8	−51.7	−20.1	−35.4	−11.1
3	REVE	**−23.5**	−4.3	−5.0	−2.4	−14.1	−13.1	−1.7	−46.3	−27.2	−24.0	−11.4
4	BrainOmni	−24.7	−8.5	−5.9	−7.2	**−7.5**	−25.8	**+0.5**	**−40.9**	−16.7	−20.6	−12.0
5	EEGNet	−30.5	−4.5	−12.1	**−2.3**	−16.3	−15.5	+0.2	−42.4	−18.3	**−16.0**	−12.4
6	CBraMod	−25.1	−13.5	−13.4	−6.5	−14.6	−19.3	−5.7	−46.8	−25.6	−36.1	−15.4$^{\nabla}$

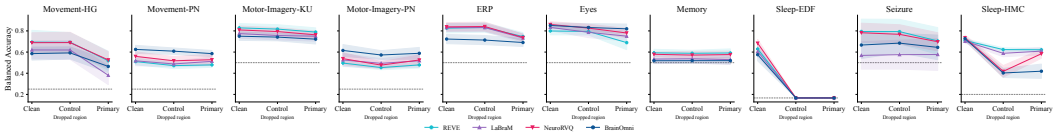
Table 15: Per-benchmark degradation under Region drop (ctrl.) (full fine-tuning). Δ percentage points from clean, paired per fold. Ranked by Average. Best per column in **bold**, second best underlined. Average equally weights the 8 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Seizure); Sleep-EDF, Sleep-HMC excluded from the average. * on the Average marks a model significantly better than EEGNet, $^{\nabla}$ significantly worse (two-sided Wilcoxon signed-rank over the shared folds, Holm-corrected).

#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Sleep-EDF	Seizure	Sleep-HMC	Average
1	EEGNet	−9.4	−5.7	−1.7	−5.8	**−0.4**	**+2.8**	−0.5	−42.4	−2.2	**−9.0**	**−2.9**
2	REVE	**−0.7**	**−3.9**	**−1.0**	**−4.9**	−0.9	+0.9	−0.2	−46.3	−16.1	−16.2	−3.3
3	LaBraM	−5.5	−7.5	−4.5	−8.3	−1.1	−0.7	**−0.1**	−43.9	−6.8	−43.0	−4.3$^{\nabla}$
4	NeuroRVQ	−4.4	−8.9	−3.6	−7.8	−2.3	−3.8	−2.4	−51.7	−6.7	−42.5	−5.0$^{\nabla}$
5	BrainOmni	−6.0	−7.7	−4.5	−9.8	−5.2	−13.2	−0.2	**−40.9**	**−0.3**	−27.1	−5.9$^{\nabla}$
6	CBraMod	−9.7	−9.2	−7.3	−7.0	−3.6	−19.3	−1.0	−46.8	−8.6	−45.4	−8.2$^{\nabla}$

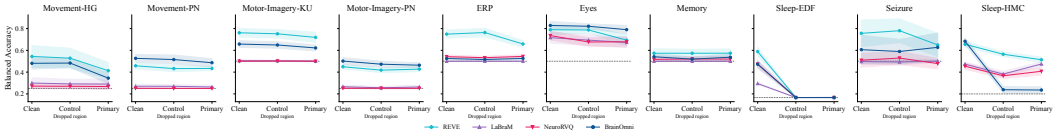
Table 16: Per-benchmark degradation under Primary Region noise -3 dB (full fine-tuning). Δ percentage points from clean, paired per fold. Ranked by Average. Best per column in **bold**, second best underlined. Average equally weights the 8 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Seizure); Sleep-EDF, Sleep-HMC excluded from the average. * on the Average marks a model significantly better than EEGNet, $^{\nabla}$ significantly worse (two-sided Wilcoxon signed-rank over the shared folds, Holm-corrected).

#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Sleep-EDF	Seizure	Sleep-HMC	Average
1	BrainOmni	**−4.1**	−1.2	**−1.4**	−0.8	**−1.2**	**−0.8**	−0.1	**−24.3**	−6.4	**−21.0**	**−2.0***
2	REVE	−15.3	−7.2	−4.1	−3.9	−7.7	−3.5	−2.9	−34.4	−7.6	−38.5	−6.5
3	EEGNet	−32.8	−2.8	−10.8	−1.4	−18.9	−2.6	**+0.6**	−42.4	**+10.9**	−29.2	−7.2
4	LaBraM	−31.5	**−0.0**	−5.3	−0.6	−13.2	−6.9	−0.4	−36.6	−0.5	−31.8	−7.3
5	NeuroRVQ	−20.8	−2.7	−5.5	**−0.4**	−11.8	−18.2	−3.4	−38.3	−8.9	−31.9	−9.0
6	CBraMod	−34.3	−16.0	−10.2	−8.1	−9.2	−10.9	−2.7	−36.3	+1.0	−44.7	−11.3$^{\nabla}$

Similarly to H.2.2, we report results for true region dropout, where dropped regions are removed from the input rather than zero-padded. This setting is restricted to models that accept variable-length channel sets: REVE, NeuroRVQ, LaBraM, and BrainOmni. Figures 13a and 13b show per-benchmark curves under full fine-tuning and head-only adaptation respectively, and Table 17 reports per-benchmark degradation under primary region dropout. As under true random dropout (Appendix H.2.2), all four variable-channel models improve on average when dropped regions are removed rather than zero-padded: average primary-region losses fall from 9.2–12.0 to 3.8–6.8 percentage points, with BrainOmni showing the largest gain.



(a) Full fine-tuning.



(b) Head-only adaptation.

Figure 13: Per-benchmark degradation under true primary region dropout for the models that accept variable-length channel sets: REVE, NeuroRVQ, LaBraM, and BrainOmni.

Table 17: Per-benchmark degradation under Region removal (primary) (full fine-tuning). Δ percentage points from clean, paired per fold. Ranked by Average. Best per column in **bold**, second best underlined. Average equally weights the 8 datasets (Movement-HG, Movement-PN, Motor-Imagery-KU, Motor-Imagery-PN, ERP, Eyes, Memory, Seizure).

#	Model	Movement-HG	Movement-PN	Motor-Imagery-KU	Motor-Imagery-PN	ERP	Eyes	Memory	Sleep-EDF	Seizure	Sleep-HMC	Average
1	BrainOmni	**−12.1**	−4.0	−2.8	−2.7	**−3.3**	**−3.2**	**+0.1**	**−40.9**	−2.2	−30.8	**−3.8**
2	LaBraM	−23.9	**−0.5**	**−2.5**	**−0.8**	−9.8	−8.3	−0.5	−43.9	**+0.9**	−9.4	−5.7
3	NeuroRVQ	−16.3	−3.1	−4.6	−1.3	−10.1	−8.0	−0.3	−51.7	−8.5	−15.0	−6.5
4	REVE	−17.6	−3.1	−4.0	−1.7	−7.9	−11.0	−0.1	−46.3	−8.9	**−8.7**	−6.8

H.4 Vector Quantisation Ablation

BrainOmni is the most robust under additive noise (Section 4.2). We tested whether its vector quantiser, which maps the compressed latents to codebook entries before the transformer, accounts for this advantage.

We skipped the quantiser at inference and passed the continuous latents straight to the transformer. Figure 14 compares both variants across the same eight datasets used in the robustness averages. At -5 dB, the average loss rises from 6.3 to 8.4 percentage points when VQ is skipped. The skipped variant still loses less than every other foundation model (12.4–20.1 percentage points) and EEGNet (11.9 percentage points). Across all ten datasets, skipping VQ raises mean clean balanced accuracy from 0.664 to 0.681. Inference-time quantisation is therefore not critical for BrainOmni's noise-robustness. Its multimodal pre-training on MEG and EEG, or the latent compression in its tokeniser remain as two plausible sources of noise robustness. Separating them needs matched pre-training runs, which lie outside our compute budget.

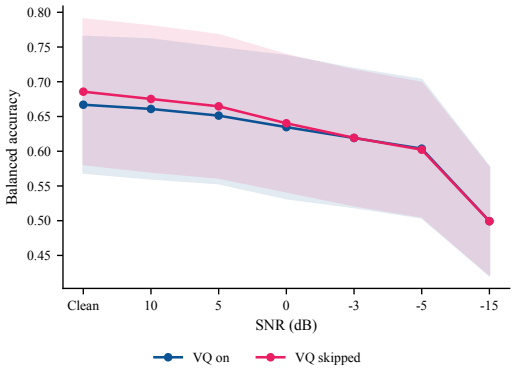


Figure 14: BrainOmni with and without its inference-time vector quantiser, averaged over the eight datasets in the perturbation set. Mean balanced accuracy under additive white noise; shading shows one standard deviation across datasets.

I Interpretability

This section provides details of the interpretability methods used in the main paper: attribution techniques (AttnLRP, Gradient $\times$ Input and GradCAM), raw attention extraction, linear probing and block truncation ablation. Each subsection covers method and results together. Table 18 summarises which methods were applied to each model.

Table 18: Interpretability methods applied to each model.

Method	LaBraM	NeuroRVQ	REVE	CBraMod	BrainOmni	EEGNet
Attribution	LRP	Grad$\times$Input	LRP	LRP	Grad$\times$Input	LRP
GradCAM	✓	✓	✓	✓	—	✓
Raw attention	—	✓	✓	—	—	—
Linear probing	—	✓	✓	—	—	—

I.1 Attention-Aware Layer-Wise Relevance Propagation

AttnLRP. We apply AttnLRP (Achtibat et al., 2024), a recent extension of LRP to transformers. LRP propagates relevance backward through each layer using conservation rules, producing attributions that sum to the model output. LRP passes label- and model-weight randomisation sensitivity checks on EEG and accurately recovers ground-truth spatial features (Ravindran & Contreras-Vidal, 2023). AttnLRP extends LRP by adding dedicated rules for the operations inside self-attention (the bilinear query-key and attention-value products, and the softmax). We use AttnLRP's standard

propagation rules as defined in (Achtibat et al., 2024): an ε-stabilised rule for linear layers, an identity rule for activations and normalisations, and an equal-split bilinear rule for the two attention matrix multiplications, with a Taylor-decomposition rule for the softmax. These are applied via the authors' own LXT library (Achtibat et al., 2024), and we implement custom auto-grad functions for operations without LXT equivalents.

Gradient $\times$ Input. For NeuroRVQ and BrainOmni, AttnLRP does not produce interpretable maps. NeuroRVQ's multi-scale convolutional encoder produces $O(0.01)$ token embeddings, so the ε-rule's $R/(z_j + \varepsilon)$ division by near-zero activations causes relevance to compound ($\sim 100\times$ per block) and destroy spatial structure across the 12 blocks. Likewise, BrainOmni's frozen tokeniser and VQ discretisation on the forward pass create a similar barrier for relevance propagation. For both models, we use Gradient $\times$ Input ($\partial y/\partial x \cdot x$), which produces clearer spatial maps where LRP fails.

We therefore use Gradient $\times$ Input (G$\times$I) as a complementary attribution method for these models. G$\times$I is among the methods that pass sanity checks and recover ground-truth features in simulated EEG (Ravindran & Contreras-Vidal, 2023). To further verify that Gradient $\times$ Input is a reasonable alternative rather than a model-specific workaround, we also compute it for the models where AttnLRP is numerically stable (LaBraM, REVE, CBraMod, EEGNet). Figure 15 shows class-averaged Gradient $\times$ Input topographic maps across five benchmarks and the remaining EEG-FMs. The spatial patterns agree with the AttnLRP maps in Figure 3 for the LRP-stable models. However, G$\times$I has known limitations such as noise sensitivity (Smilkov et al., 2017) and does not adhere to conservation rules like LRP, which we accept in exchange for numerical stability. **This supports using G$\times$I as an alternative for the two models where AttnLRP is unsuitable.**

Both methods (AttnLRP and Gradient $\times$ Input.) produce a relevance map of the same shape as the input ($C \times T$). We restrict the analysis to correctly classified test samples whose prediction confidence exceeds the 75th percentile, excluding ambiguous trials whose attributions would otherwise add noise to the class average. Each map is normalised per sample by the 99th percentile of $|R|$ and clipped to $[-1, 1]$. We average $|R|$ over time, trials, and folds. Taking the absolute value discards sign, so the maps show the magnitude of relevance rather than whether evidence supports or opposes the prediction. The resulting class-averaged topographic maps are shown in Figure 3; per-class (non-averaged) AttnLRP maps appear alongside GradCAM in Figure 17.

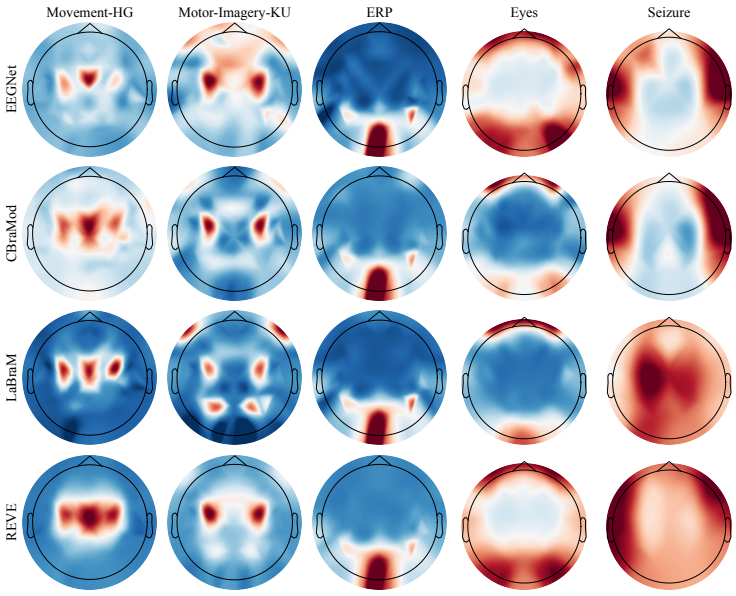
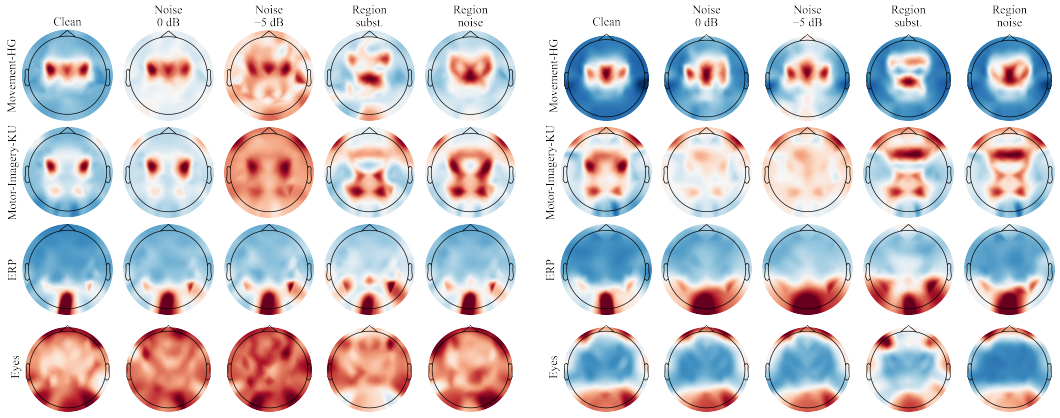


Figure 15: Class-averaged Gradient $\times$ Input topographic maps. Rows (Models): EEGNet, CBraMod, LaBraM and REVE. Columns (Benchmarks): Movement-HG, Motor-Imagery-KU, ERP, Eyes and Seizure.

I.2 Attribution under Perturbation

The robustness results (Section 4.2) show how much performance degrades, while the attribution maps (Section 4.3) show where models assign relevance on clean data. This section connects the two by examining attribution under perturbation. We compare REVE and NeuroRVQ, the two highest-performing models in full fine-tuning (Section 4.1), on Movement-HG, Motor-Imagery-KU, ERP and Eyes. Figure 16 shows averaged maps from clean inputs and correctly-classified samples under four perturbation conditions: white noise at 0 dB and -5 dB, primary-region noise at 5 dB, and primary-region substitution. Substitution replaces regional signals with filtered white noise scaled to the intact channels, avoiding potential zero-padding artefacts (Appendix D.4). Under stronger noise, Movement-HG and Motor-Imagery-KU maps become more diffuse, though Movement-HG retains a central focus and Motor-Imagery-KU gains frontal relevance. ERP retains a posterior focus, while Eyes changes little: NeuroRVQ preserves its spatial pattern and REVE remains diffuse. Substitution shifts relevance towards neighbouring motor electrodes, preserves the posterior focus on ERP and causes smaller changes on Eyes. Some spatial patterns therefore persist despite overall accuracy losses. However, we only plot correct trials in the highest confidence quartile within each condition, so the selected trials may differ and do not show what happens when predictions fail. Further analyses of attention under perturbed conditions are presented in Appendix I.4.2.

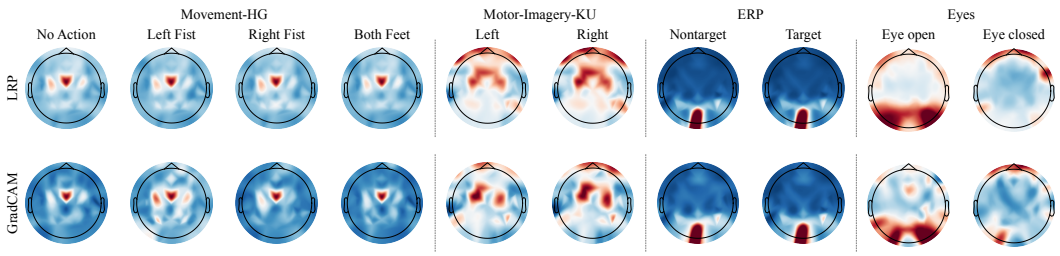


(a) REVE (b) NeuroRVQ

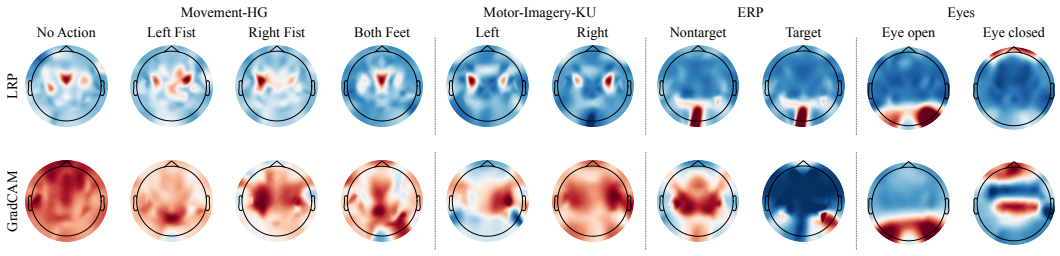
Figure 16: Class-averaged attribution maps, averaged over folds, for REVE (AttnLRP) and NeuroRVQ (Input $\times$ Gradient). Rows: Movement-HG, Motor-Imagery-KU, ERP and Eyes. Columns: clean input, white noise at 0 dB and -5 dB, matched white-noise substitution in the primary region, and primary-region white noise at 5 dB. Substitution replaces zero-padded region dropout in this analysis.

I.3 GRADCAM

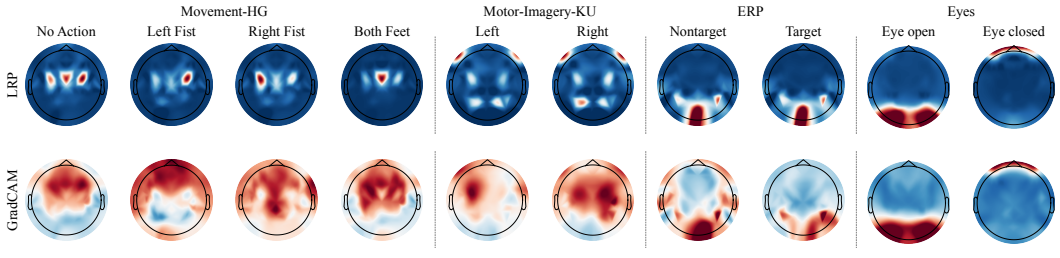
We apply GradCAM (Selvaraju et al., 2017) as a supplementary attribution method using the pytorch-grad-cam library (Gildenblat & contributors, 2021), targeting the last transformer block for transformer-based models and the final convolutional layer for EEGNet. The same sample selection, normalisation, and averaging pipeline as AttnLRP is applied (see I.1). GradCAM requires a one-to-one mapping between spatial tokens and electrodes. BrainOmni compresses channels into latent source variables, breaking this mapping, so its GradCAM maps are not spatially interpretable at the electrode level. Figure 17 shows the comparison results between GradCAM and AttLRP, with AttnLRP producing clearer, more interpretable topographic maps.



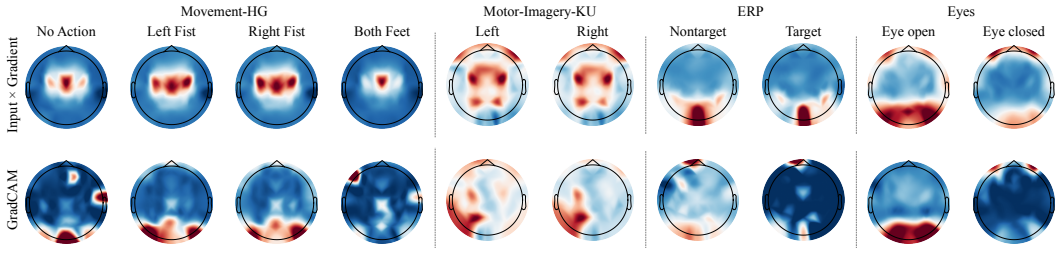
(a) EEGNet



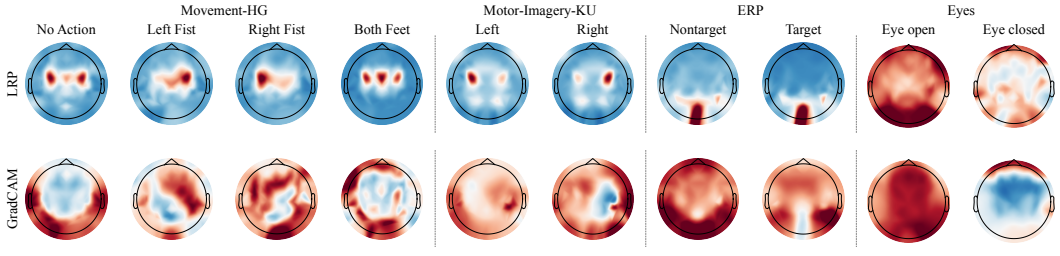
(b) CBraMod



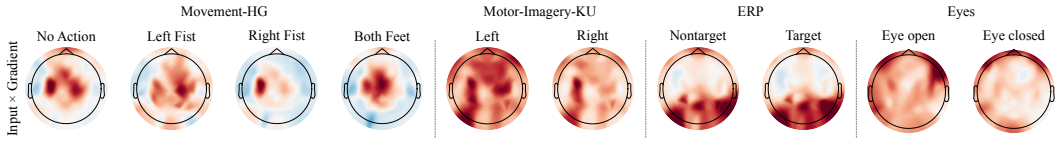
(c) LaBraM



(d) NeuroRVQ



(e) REVE



(f) BrainOmni

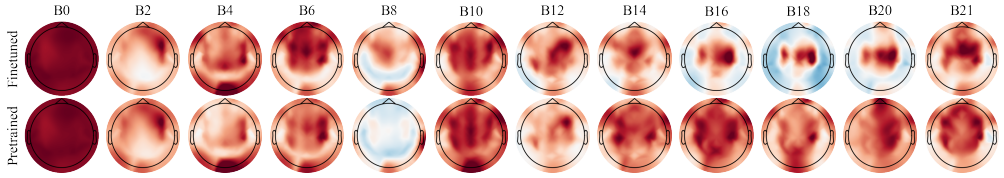
Figure 17: Per-class attribution maps (AttnLRP / GradCAM) across all benchmarks for each model. Note: BrainOmni has no direct mapping of attention to channels making it incompatible with GradCAM.

I.4 Attention Maps

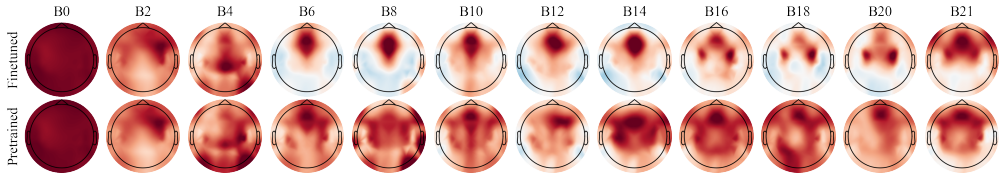
We extract raw softmax attention weights from each transformer block for REVE and NeuroRVQ, the two highest-performing architectures. Both models produce standard ($batch$, $heads$, n_{tokens}, n_{tokens}) attention matrices. Per-channel attention is computed by summing each attention matrix over the query dimension, then averaging over heads, frequencies (for NeuroRVQ), time patches, samples, and all cross-validation folds. NeuroRVQ carries a CLS token, which is excluded.

I.4.1 Per-block Attention: Pre-trained vs. Fine-tuned

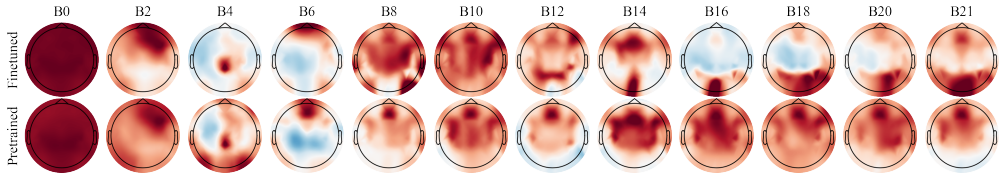
Figures 18 and 19 compare per-block attention topographic maps between fine-tuned (top row) and pre-trained (bottom row) models across four benchmarks. Each column corresponds to one transformer block, ordered from input (left) to output (right). Early blocks show near-identical attention under both training modes. For REVE, the divergence varies by task. On Movement-HG, consistent divergence appears from block 16 onward, with some earlier blocks already showing small differences. On Motor-Imagery-KU, fine-tuned attention diverges from block 6 and focuses on more task-appropriate regions than the head-only maps. On ERP, differences emerge as early as block 4, but only blocks 14–21 settle on the occipital focus. On Eyes, fine-tuned and pre-trained attention are both diffuse across all blocks with no clear divergence point, consistent with the pre-trained checkpoint already carrying task signal (Figure 4). For NeuroRVQ, the fine-tuned and pre-trained attention patterns diverge around block five on ERP and motor tasks (Movement-HG, Motor-Imagery-KU) and around block eight on Eyes. This is broadly consistent with the probing transitions in Section 4.4. On Motor-Imagery-KU, ERP and Movement-HG, middle blocks produce attention maps with clearer alignment to expected cortical regions than the final blocks.



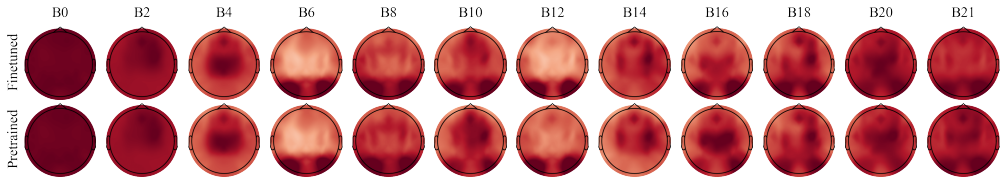
(a) Movement-HG



(b) Motor-Imagery-KU

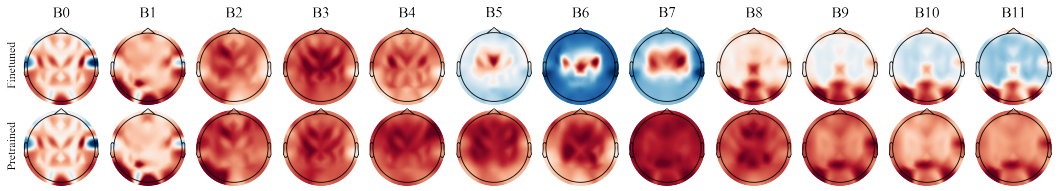


(c) ERP

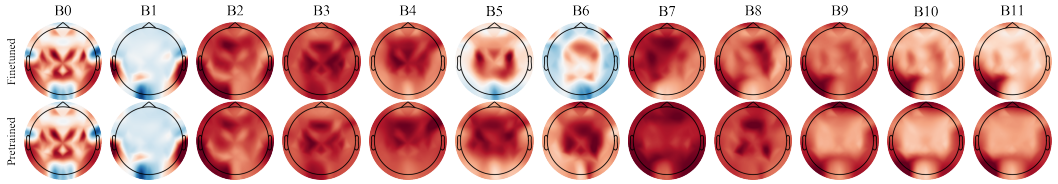


(d) Eyes

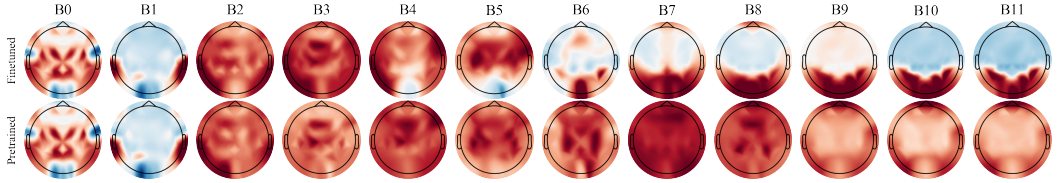
Figure 18: REVE per-block attention topographic maps across datasets. Top: fine-tuned. Bottom: pre-trained. Blocks (B) ordered from input (left) to output (right).



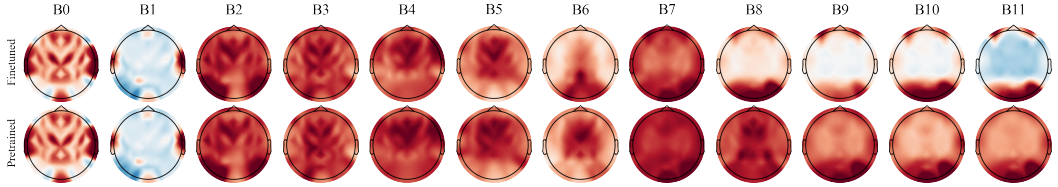
(a) Movement-HG



(b) Motor-Imagery-KU



(c) ERP

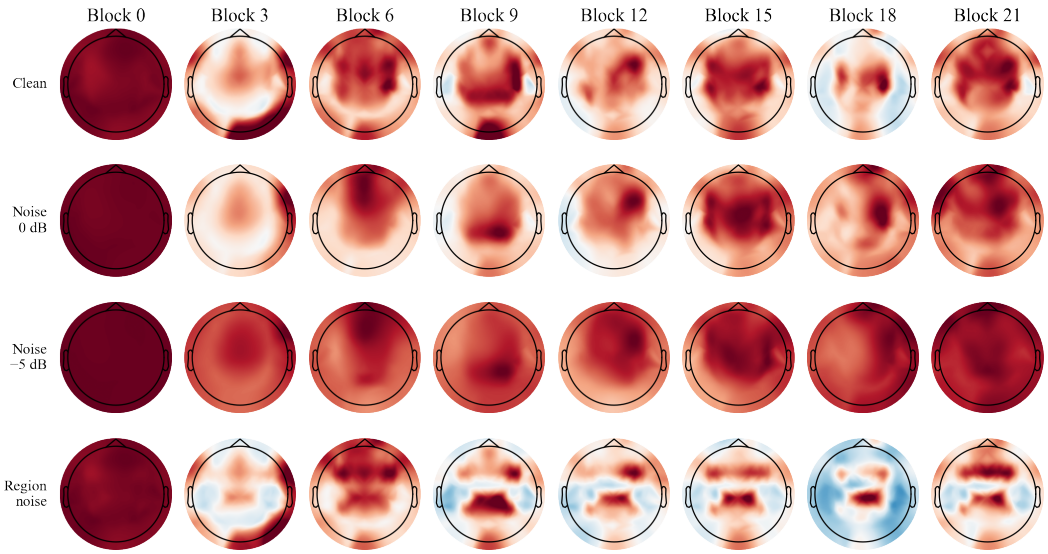


(d) Eyes

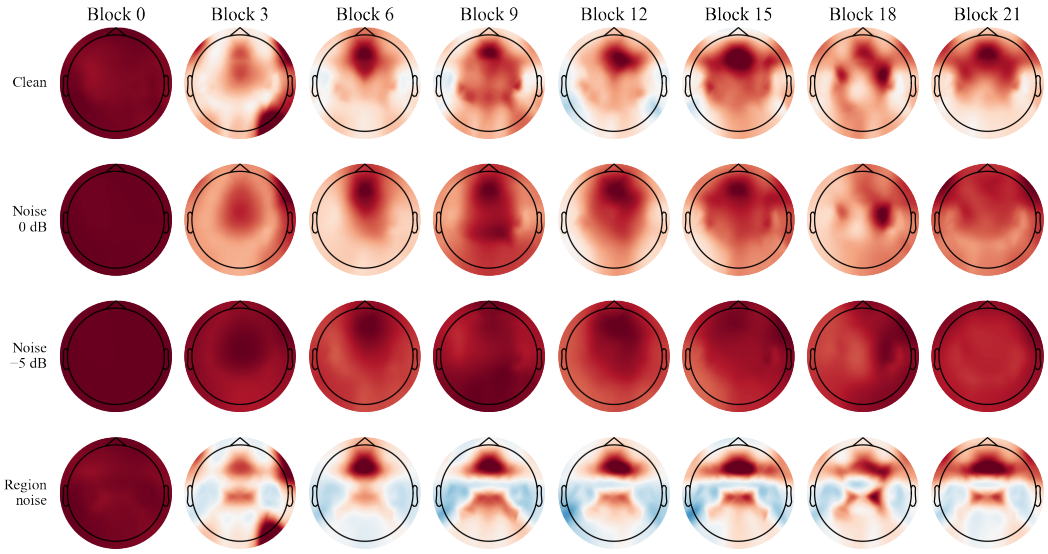
Figure 19: NeuroRVQ per-block attention topographic maps across datasets. Top: fine-tuned. Bottom: pre-trained. Blocks (B) ordered from input (left) to output (right).

I.4.2 Attention under Perturbation

Figures 20a–21b and Figures 22a–23b show per-block attention topographic maps for REVE and NeuroRVQ under the same perturbation conditions used in Section I.2, excluding noise substitution. Each column corresponds to one transformer block, ordered from input (left) to output (right), and each row to one perturbation condition. REVE's attention patterns shift progressively across blocks under noise, with attention near-uniform at -5 dB. NeuroRVQ's per-block attention is more stable under noise. The spatial patterns at 0 dB closely resemble clean attention across block six specifically. Later blocks also remain stable across most tasks. Region noise leaves both models close to the clean condition.

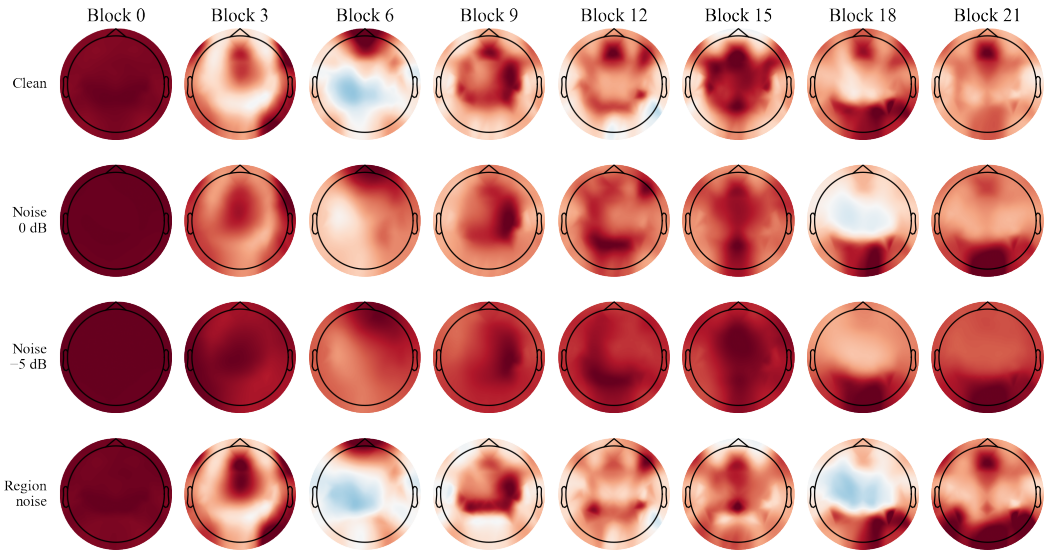


(a) Movement-HG

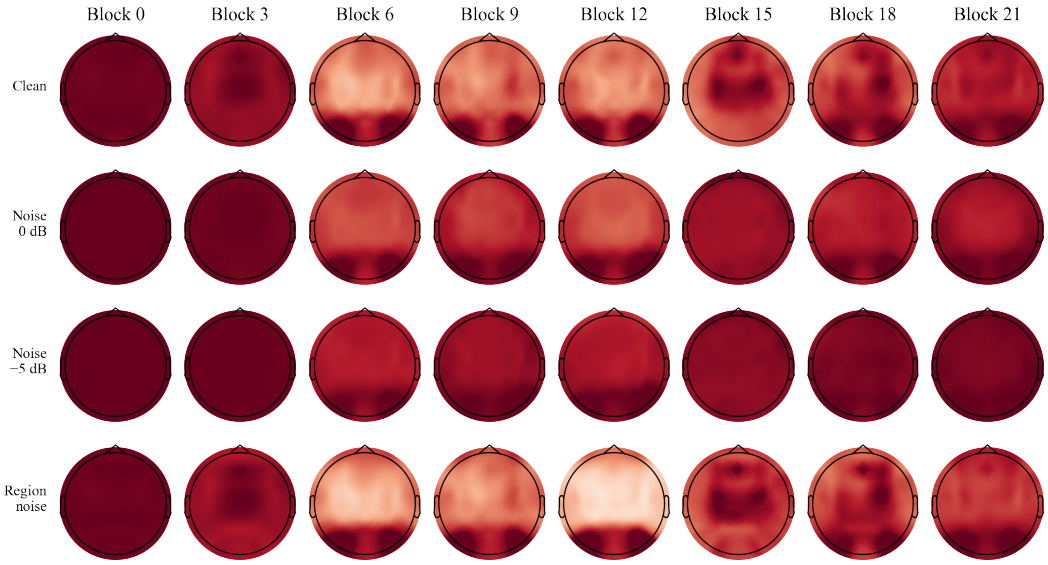


(b) Motor-Imagery-KU

Figure 20: REVE per-block attention under perturbation across datasets. Blocks ordered from input (left) to output (right) - *Plot 1/2*.

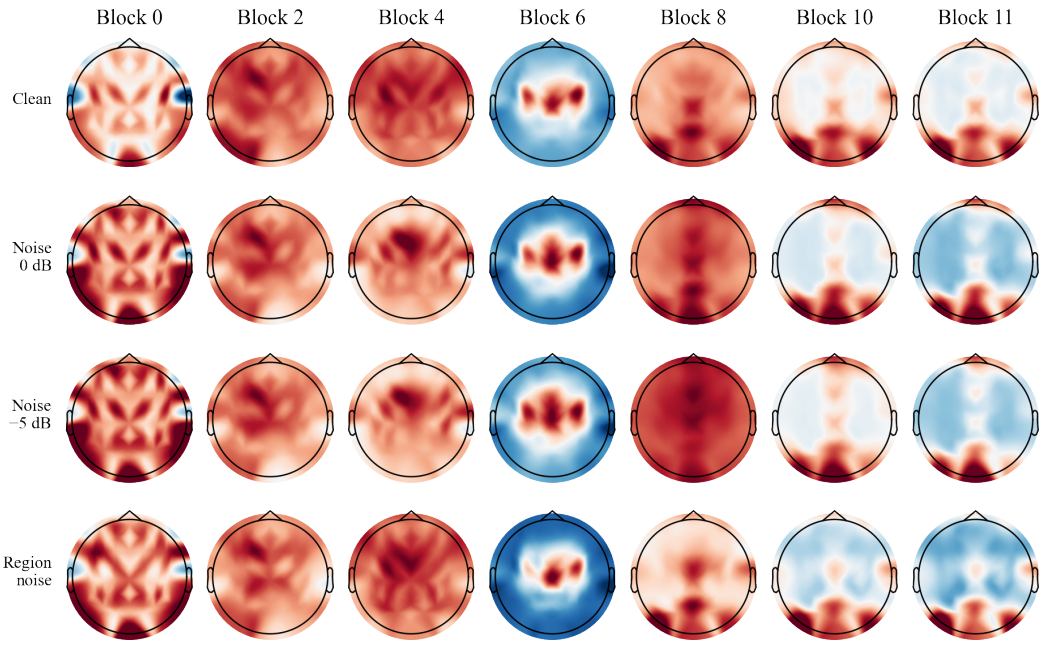


(a) ERP

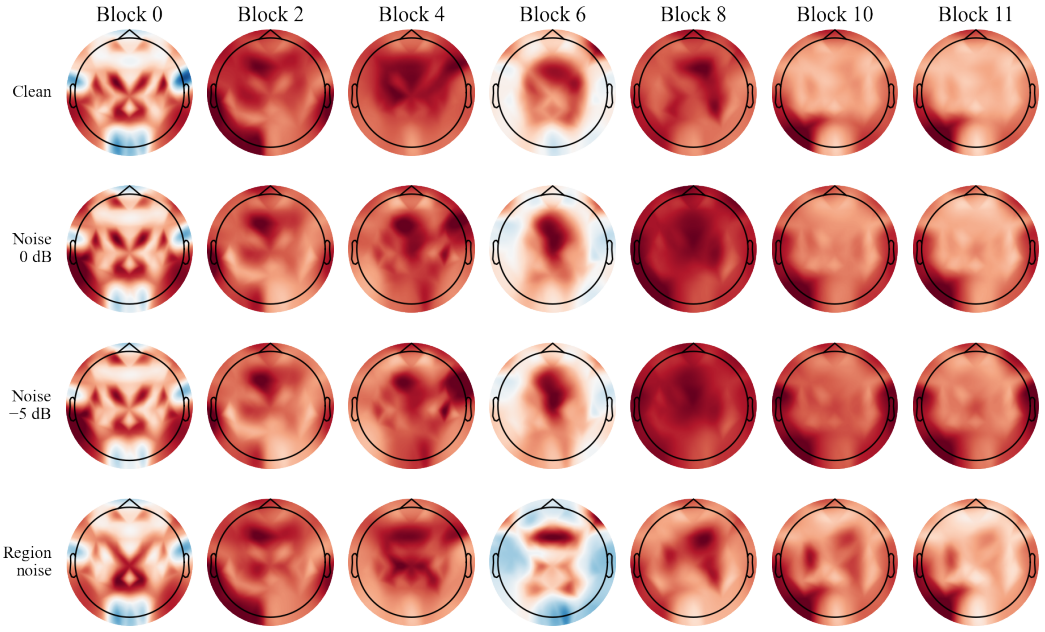


(b) Eyes

Figure 21: REVE per-block attention under perturbation across datasets. Blocks ordered from input (left) to output (right) - *Plot 2/2*.

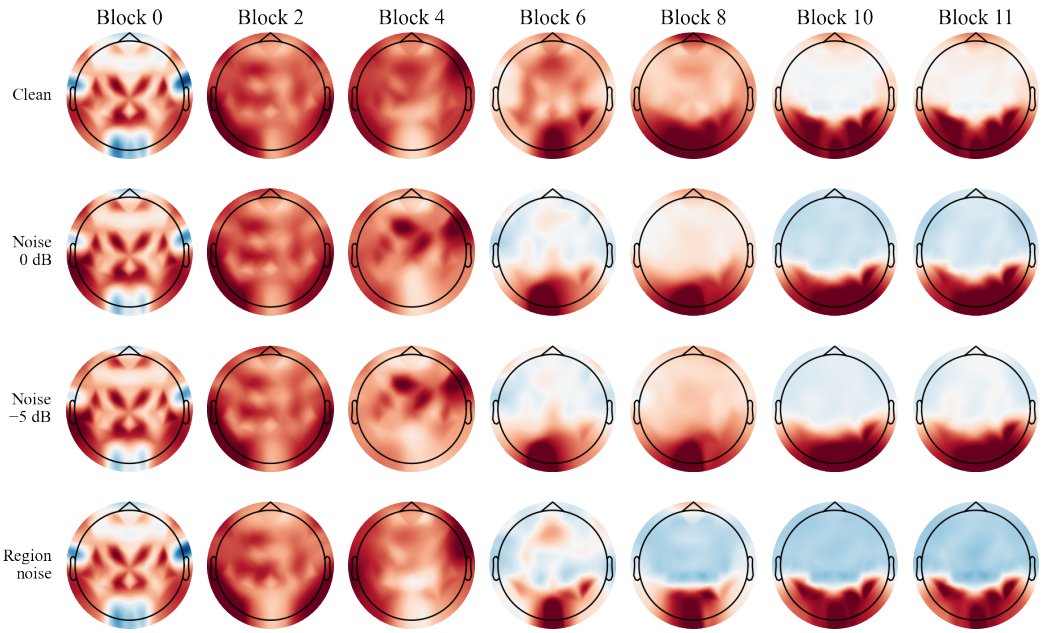


(a) Movement-HG

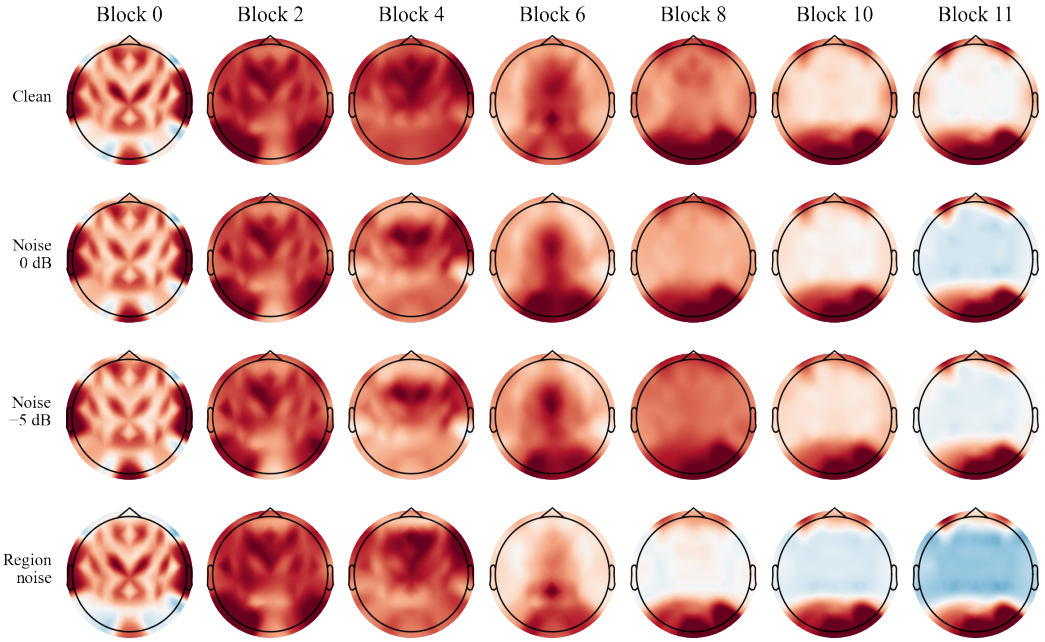


(b) Motor-Imagery-KU

Figure 22: NeuroRVQ per-block attention under perturbation across datasets. Blocks ordered from input (left) to output (right) - *Plot 1/2*.



(a) ERP



(b) Eyes

Figure 23: NeuroRVQ per-block attention under perturbation across datasets. Blocks ordered from input (left) to output (right) - *Plot 2/2*.

I.5 Block-wise Probing

To locate where task-relevant information emerges, we train a single layer probe at the output of every transformer block. A single linear layer maps from the block's representation to the number of classes. Probes are trained for up to 20 epochs with Adam, cosine annealing, and early stopping (patience 5 epochs). Probes are trained on the same ten folds as the model checkpoints, and we report the mean balanced accuracy $\pm$ std across folds for each transformer block. We index blocks from 0 to $L-1$: index 0 refers to the first block output. Relative depth is $i/(L-1)$, with $L=12$ for NeuroRVQ and $L=22$ for REVE. We use two pooling strategies: *(i)* **mean pooling**: averages all tokens to a single vector *(ii)* **flatten pooling**: concatenates all tokens into a single vector, applies RMSNorm, and fits the linear probe on the full flattened dimension.

I.5.1 Flatten Probing

Section 4.4 reports linear probing with mean pooling. Figure 24 shows the pre-trained and fine-tuned mean-pooled curves overlaid. Figure 25 shows the same analysis with flatten pooling, which preserves positional structure by concatenating all token representations before the probe.

Flatten pooling reveals task information that mean pooling misses in NeuroRVQ's pre-trained representations. Comparing each pooling method's best block, accuracy rises from 0.500 to 0.822 on ERP, from 0.509 to 0.754 on Motor-Imagery-KU, and from 0.349 to 0.627 on Movement-HG. REVE also benefits from flatten pooling. Fine-tuning reduces the mean–flatten gap for NeuroRVQ, but does not close it for REVE. At the final block, flatten pooling still leads by 18.3, 17.5 and 11.0 percentage points on Movement-HG, Motor-Imagery-KU and ERP.

NeuroRVQ's pre-trained flatten probes peak at index 2 on Movement-HG and ERP (the third block), index 3 on Motor-Imagery-KU (the fourth block), and index 0 on Eyes (the first block). Accuracy then declines towards the output. REVE's peaks occur later, at indices 17, 13, 8 and 20 on Movement-HG, Motor-Imagery-KU, ERP and Eyes. Across these four tasks, the mean relative decline from each task's peak to its final block is 23.1% for NeuroRVQ and 2.0% for REVE. Fine-tuning reduces the decline for NeuroRVQ but does not remove it on every task: its Movement-HG flatten probe peaks at 0.651 and ends at 0.588.

One possible explanation is that later pre-trained blocks specialise for the pre-training objective over downstream task features, as observed in NLP transformers (Voita et al., 2019). However, the probing results alone do not establish this mechanism. Probes for the pre-training task, subject identity and recording session could test which information persists at each depth and how fine-tuning changes it (Merchant et al., 2020; Saha & Baumert, 2020).

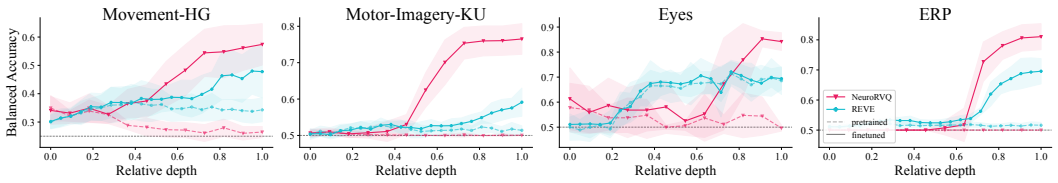


Figure 24: Linear probing balanced accuracy (mean pooling) by relative block depth. Pre-trained (dashed) and fine-tuned (solid) curves shown for NeuroRVQ and REVE.

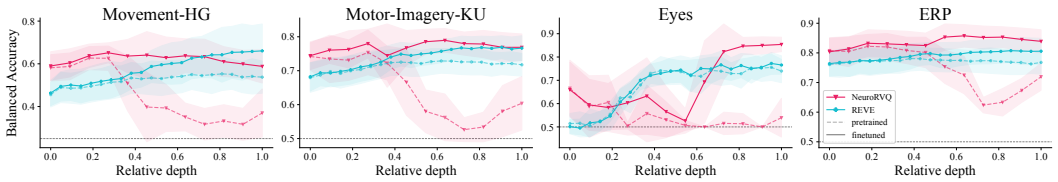


Figure 25: Linear probing balanced accuracy (flatten pooling) by relative block depth. Pre-trained (dashed) and fine-tuned (solid) curves shown for NeuroRVQ and REVE. Flatten pooling reveals pre-trained task information that mean pooling misses (Figure 4).

I.6 Depth Truncation

The probing results show that task-relevant representations are already present in early to middle blocks (Appendix I.5.1). To test whether downstream performance reflects this, we run a depth truncation ablation. We attach a flatten-pooling classification head (see Appendix J) after each of several intermediate blocks and fine-tune the resulting truncated stack. Figure 26 shows the results. NeuroRVQ approaches full-depth performance by block 6, and REVE by block 14. One hypothesis is that the later layers contribute little to these downstream tasks and instead serve the pre-training objective, though this ablation does not establish this on its own.

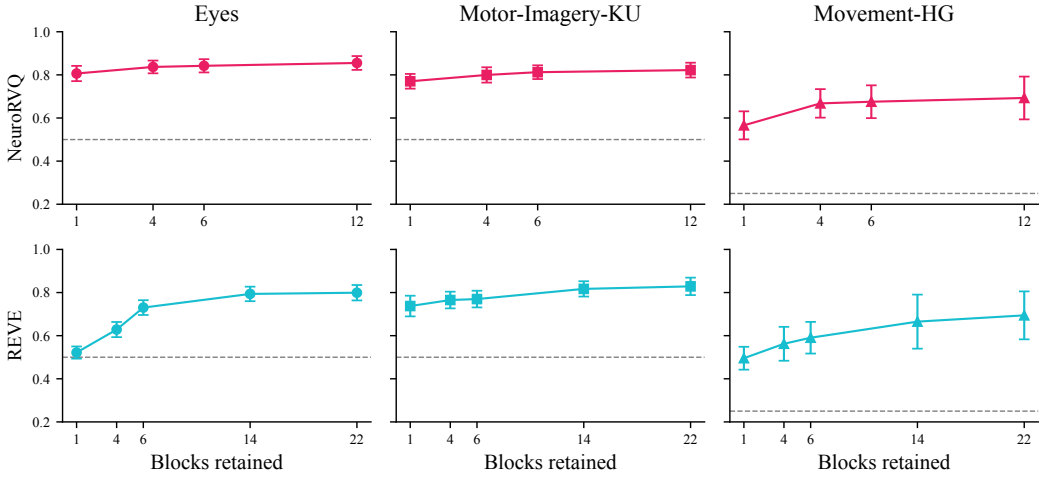


Figure 26: Depth truncation, where k counts the transformer blocks retained. For each k, we discard later blocks, attach a flatten-pooling head, and fine-tune end-to-end. Balanced accuracy $\pm$ std is shown as a function of k. NeuroRVQ approaches full-depth performance by block 6, and REVE by block 14.

J Pooling Strategy

We compare four readouts for LaBraM, NeuroRVQ, REVE, BrainOmni and CBraMod on Movement-HG, Motor-Imagery-KU, ERP and Eyes. Each readout is tested under head-only adaptation and full fine-tuning on the same ten folds.

Mean pooling averages all spatial and temporal tokens. Attention pooling uses a learned query to weigh the tokens before summing them. Flatten pooling concatenates all tokens. Mean-tile repeats the token mean to match the flattened input size and uses the same head as flatten. It therefore matches the head's parameter count without retaining individual tokens. Table 19 reports these counts. For some models, the heads used in this ablation differ from the original heads used in the rest of this study.

Table 19: Head-only trainable parameters for the four readouts on Eyes (binary, 64 channels). Flatten and mean-tile have the same head structure and parameter count.

Pooling	REVE	NeuroRVQ	LaBraM	BrainOmni	CBraMod
Mean	2,050	2,401	201	1,026	201
Attention	2,562	3,201	401	1,538	401
Mean-tile	393,218	614,401	153,601	393,218	51,201
Flatten	393,218	614,401	153,601	393,218	51,201

Table 20 reports test balanced accuracy. Under head-only adaptation, flatten pooling gives the highest average for all five models and exceeds mean pooling significantly in each case. Mean-tile improves

several models but underperforms compared to flatten, so head size alone does not explain the gain. Attention pooling also underperforms flatten. Training the attention readout during pre-training might help, but this remains a hypothesis.

Under full fine-tuning, the mean-to-flatten gain is smaller for most models. NeuroRVQ changes little (0.799 to 0.802), while REVE improves from 0.762 to 0.787 and LaBraM from 0.765 to 0.786. The effect remains larger for BrainOmni (0.638 to 0.736) and CBraMod (0.615 to 0.759).

Table 20: Pooling ablation, for head-only and full fine-tuning. Test balanced accuracy (mean over 10 folds). Within each regime, models are ordered by their best Average. The best per column is **bold**, second best *underlined*. * marks an Average significantly better than Mean pooling for that model (Wilcoxon, Holm-corrected).

Setting	Model	Pooling	Movement-HG	Motor-Imagery-KU	ERP	Eyes	Average
Head-Only	NeuroRVQ	Mean	0.271	0.501	0.542	0.736	0.512
		Mean-tile	0.301	0.500	0.516	0.691	0.502
		Attn	0.278	0.500	0.535	0.704	0.504
		Flatten	**0.564**	<u>0.739</u>	**0.770**	**0.827**	**0.725***
	REVE	Mean	0.309	0.508	0.509	0.743	0.517
		Mean-tile	0.359	0.528	0.516	0.783	0.546*
		Attn	0.295	0.520	0.504	0.643	0.491
		Flatten	<u>0.544</u>	**0.761**	<u>0.749</u>	0.790	<u>0.711</u>*
	BrainOmni	Mean	0.322	0.515	0.500	0.651	0.497
		Mean-tile	0.371	0.555	0.500	0.754	0.545*
		Attn	0.319	0.515	0.500	0.627	0.490
		Flatten	0.533	0.707	0.636	0.809	0.672*
	LaBraM	Mean	0.300	0.504	0.500	0.717	0.505
		Mean-tile	0.335	0.515	0.500	0.750	0.525*
		Attn	0.296	0.504	0.500	0.599	0.475
		Flatten	0.501	0.716	0.568	0.813	0.649*
	CBraMod	Mean	0.276	0.505	0.500	0.579	0.465
		Mean-tile	0.324	0.509	0.500	0.785	0.530*
		Attn	0.272	0.501	0.500	0.609	0.471
		Flatten	0.387	0.653	0.500	<u>0.820</u>	0.590*
Fine-tuned	NeuroRVQ	Mean	0.689	0.809	0.838	<u>0.861</u>	<u>0.799</u>
		Mean-tile	0.520	0.797	<u>0.854</u>	0.808	0.745
		Attn	0.621	0.805	**0.856**	<u>0.861</u>	0.786
		Flatten	<u>0.693</u>	<u>0.822</u>	0.836	0.855	**0.802**
	REVE	Mean	0.626	0.799	0.823	0.799	0.762
		Mean-tile	0.608	0.799	0.825	0.796	0.757
		Attn	0.626	0.796	0.817	0.780	0.755
		Flatten	**0.694**	**0.829**	0.826	0.799	0.787*
	LaBraM	Mean	0.620	0.775	0.832	0.833	0.765
		Mean-tile	0.582	0.744	0.825	0.771	0.730
		Attn	0.597	0.780	0.816	0.770	0.741
		Flatten	**0.694**	0.805	0.831	0.814	0.786*
	CBraMod	Mean	0.381	0.544	0.757	0.776	0.615
		Mean-tile	0.405	0.576	0.779	0.792	0.638*
		Attn	0.402	0.629	0.762	0.817	0.652*
		Flatten	0.613	0.786	0.797	0.840	0.759*
	BrainOmni	Mean	0.430	0.572	0.704	0.846	0.638
		Mean-tile	0.567	0.545	0.719	0.850	0.670
		Attn	0.438	0.544	0.684	0.845	0.628
		Flatten	0.607	0.749	0.726	**0.864**	0.736*